09/05

$0<\gamma<1$

$A_t = A^\pi(s)$

$v = \max_\pi \mathbb{E} \sum_{t=0}^{\infty} \gamma^t R(S_t, A_t)$.

↳ $Q_{k+1}(s,a) = R(s,a) + \gamma \mathbb{E} V_k(s')$

$s' \sim P(s,a) \rightarrow s' = f(s,a,w)$

$\forall s,a$

$V_{k+1}(s) = \max_a Q_{k+1}(s,a) \quad \forall s$

$A^*(s) \in \arg\max_a Q(s,a)$.

⊛ $V = \max_{a_1 \cdots a_5} \min_{b_1 \cdots b_5} \sum R(s_t, a_t, b_t)$.

$S_{t+1} = f(s_t, a_t, b_t)$.

reward: $\eta = \begin{cases} +1 & \text{if caught} \\ 0 \end{cases}$

EVADER: $\min \mathbb{E} \sum_{t=1}^{\infty} \eta_t \gamma^t$

PURSUER: $\max \mathbb{E} \sum_{t=1}^{\infty} \eta_t \gamma^t$

k: iteration (counter). $k=0$, $V_0^{pursuer}(s)=0$. $V_0^{evader}(s)=0$.

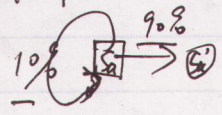
in turn

EVADER: $Q_{k+1}^{evader}(s,a) = 0.9\eta(s') + 0.1\eta(s)$.

$\gamma^a = 90\%$ $\quad + \gamma \mathbb{E} V^{pursuer}(s')$.

$= [0.9 V_k^{pursuer}(s') + 0.1 V_k^{pursuer}(s)]$

the speed

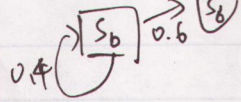


$V_{k+1}^{evader}(s_a) = \min_a Q_{k+1}(s_a, a)$. $\quad \forall s,a$.

PURSUER: $Q_{k+1}^{PURSUER}(s_b, b) = 0.6\eta(s''_b) + 0.4\eta(s'_b)$

$+ \gamma \mathbb{E} V^{EVADER}(s''_b)$.

$= [0.6 V_{k+1}^{evader}(s''_b) + 0.4 V_{k+1}^{evader}(s'_b)]$.



$V_{k+1}^{PURSUER}(s_b) = \max_b Q_{k+1}(s'_b, b)$

$\forall s,b$

$s = (x^{evader}, y^{evader}, x^{pursuer}, y^{pursuer})$

V: table of 15^4. $(15 \times 15)^2$.

09/07. Markov Chains.

$X_{t+1} = f(X_t, W_{t+1})$. f

transaction prob $\pi_i \geq 0$, $\pi e = 1$.

$\mathbb{P}(X_{t+1} | X_1, \cdots X_t) = \mathbb{P}(X_{t+1} | X_t)$.

Given $(\pi, P_0, P_1, \cdots P_t, \cdots)$

$\pi_i = prob(X_0 = i)$

$prob(X_1 = j) = \sum_{i=1}^{N} prob(X_1 = j, X_0 = i) = \sum prob(X_0 = i) \cdot prob(X_1 = j | X_0 = i)$

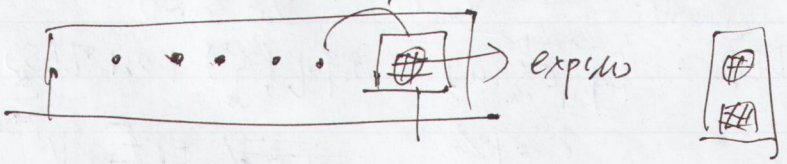
$= \sum_{i=1}^{\infty} \pi_i P_{0ij}$

$\Rightarrow P_1 = \pi P_0$

Row Row matrix

$P_2 = \pi P_0 P_1$.

pizza. one server. exp(λ) two servers



differentiate 微分

*Mathematica

~~:ESC~~

ESC alpha ESC

:2:

expo distribution is memoryless:

$P(T > t + \delta t | T > \delta t) = P(T > t)$

info that $T > \delta t$ (we have waiting) is irrelevant to predict how long to wait

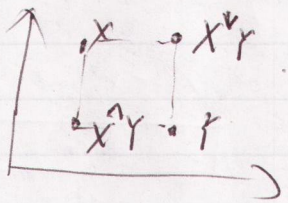
09/19

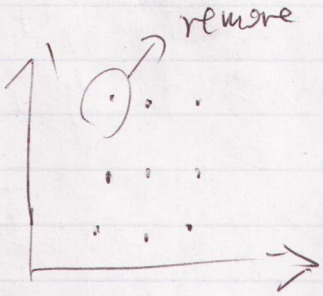
Lattice

$(\mathbb{R}^2, \leq)$

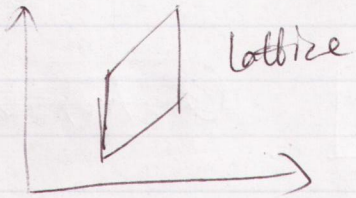
meet $x \wedge y$ $x \cap y$

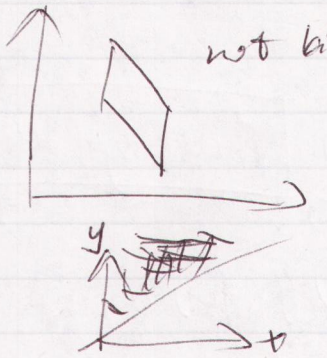
joint $x \vee y$ $x \cup y$





$\Rightarrow$ still Lattice.





$(x, y) \in \mathbb{R}^2$, $x \leq y$.

all format : $x_i - y_i \leq b_i$.

Eg. $L: (x, y, z)$: $x+y+z \leq 1$ $\quad x, y, z \geq 0$.

$(1\ 0\ 0) \in L$

$(0\ 1\ 0) \in L$

joint $(1\ 1\ 0) \notin L$.

$u = x$

$v = x+y$

$w = x+y+z$

$\Rightarrow \begin{cases} x = u \geq 0 \\ y = v - u \geq 0 \\ w = w - v \geq 0 \\ w \leq 1. \end{cases}$

easy to decide if lattice when 2 variables

s : inventory.

$s \geq 0$

a: sell $\quad 0 \leq a \leq s$.

full

$f(x, y)$ — x: product, y: product, planes

$x^+ \geq x^-, \quad y^+ \geq y^-$.

increasing difference ⇔

$$f(x^+, y^-) - f(x^-, y^-) \leq f(x^+, y^+) - f(x^-, y^+)$$

$$f(x^+, y^+) - f(x^+, y^-) \geq f(x^-, y^+) - f(x^-, y^-)$$

$f(x, y) = x \cdot y$ → Hessian Matrix $\quad \frac{\partial f}{\partial x_i \partial x_j} \geq 0 \quad \forall i \neq j$

≥ 0

$\begin{cases} f(x) \\ g(x) \end{cases}$ increasing difference

~~② $h(x) = \theta_1 f(x) + \theta_2 g(x)$,~~ ~~$\theta_1, \theta_2 \geq 0$~~

~~is increa~~

① $g(f(x))$ is increasing difference.

---

supermodular

$x, y \in$ Lattice

$$f(x \wedge y) + f(x \vee y) \geq f(x) + f(y)$$

proof: $f(x^+, y^+) + f(x^-, y^-) \geq f(x^+, y^-) + f(x^-, y^+)$

$f_1(x, y)$, $f_2(x, y)$ are supermodular.

then $f(x, y) = \theta_1 f_1(x, y) + \theta_2 f_2(x, y)$, $\theta_1, \theta_2 \geq 0$

is supermodular.

in $\mathbb{R}^n$, supermodular ⇔ increasing difference.

$f(u, v) = u^T v$

$u \in \mathbb{R}^n$, $v \in \mathbb{R}^n$

But supermodular ALWAYS ⇒ increasing difference

$f(x) = g(a^T x)$ $a \geq 0$, g convex. $f(x)$ is supermodular.

proof: $\frac{\partial f}{\partial x_i} = g'(a^T x) a_i$

$\frac{\partial^2 f}{\partial x_i \partial x_j} = \underbrace{g''(a^T x)}_{\geq 0} \underbrace{a_i a_j}_{\geq 0} \geq 0.$

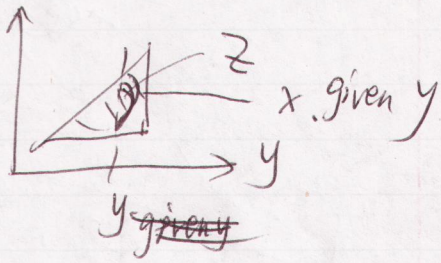
---

f is convex if $H \geq 0$

on variable = $f(x)$ must be supermodular.

$f(x,y)$ ~~super~~ supermodular. on $z = \{x, y\}$. z is feasible set

$g(y) = \sup_{x \in S_y z} f(x,y)$.



assumption $(z, \leq)$ lattice

$\Rightarrow S_z(y)$ lattice

- $g(y)$ supermodularity

$x^* = \arg\max_{x \in S_y z} f(x,y)$ (assume unique)

* sensitivity to parameters in some models

$y^+ \geq y^- \Rightarrow (x^*)^+ \geq (x^*)^-$ because $f(x,y)$ is lattice

through change parameters.

cannot $(x^*)^+ < (x^*)^-$

$$\pi(y) = \arg\max_{x \in S_y} f(x, y)$$

monotonicity y

Isotonicity y

Sep 26

complementary CDF

$$E(X) = \int_0^\infty \underline{\text{prof}(X \geq x)\, dx} = \int_0^\infty 1 - F(x)\, dx$$

$$\bar{F}_X^c(x)\, dx$$

$$= \int_0^\infty x f(x)\, dx$$

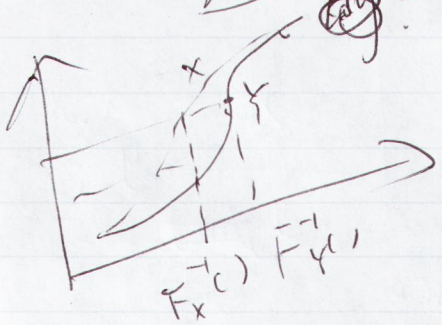
$$= \int_0^\infty 1\, dx - \int_0^\infty F(x)\, dx$$

$$= x\Big|_0^\infty - \left(xF(x)\Big|_0^\infty - \int_0^\infty x\, dF(x) \right)$$

$$= x\Big|_0^\infty - xF(x)\Big|_0^\infty + \int_0^\infty x f(x)\, dx$$

$$= \infty - 0 - \infty + 0$$

$$= \int_0^\infty x f(x)\, dx$$



$X = X^+ - X^-$, ~~$E(X) = E(X^+) + E(X^-)$~~

$$E[X^+] = \int_{-\infty}^{\infty} x^+ f(x)\, dx = \int_0^\infty x f(x)\, dx$$

$$E[X^-] = \int_{-\infty}^{\infty} x^- f(x)\, dx = \int_{-\infty}^{0} (-x) f(x)\, dx$$

$$X^+ = \begin{cases} X, & x \geq 0 \\ 0, & x \leq 0 \end{cases}, \quad X^- = \begin{cases} -X, & x \leq 0 \\ 0, & x \geq 0 \end{cases}$$

$$E[X] = E[X^+] - E[X^-]$$

$X \leq_{st} Y \Rightarrow \mathbb{E}[X] \leq \mathbb{E}[Y]$

$X \leq_{st} Y \Rightarrow h(x) \leq_{st} h(y)$. h : increasing function.

$\Rightarrow \mathbb{E}[h(x)] \leq \mathbb{E}[h(y)]$

$s_1 \leq s_2$.

$X'_{\alpha = \text{current state}}$ $X'_{s_1} \leq_{st} X'_{s_2}$. $\mathbb{E}_{x'|s_1} V(x') \leq \mathbb{E}_{x'|s_2} V(x')$.

Stochastic : First Order Dominance :

Isotone

$X \leq_{st} Y$. $F_X^c(t) = 1 - Prob(X \leq t) = Prob(X \geq t)$

$\hookrightarrow F_X^c(t) \leq F_Y^c(t)$.

"stochastic greater or less then".

For $\{X_\alpha, \alpha\}$, X_α is stochastic increasing in α.

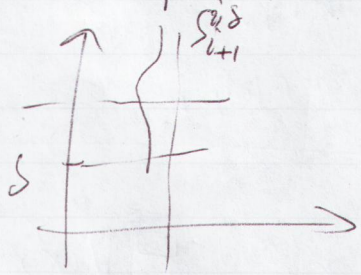
$X_{\alpha_1} \leq X_{\alpha_2}$, $\forall \alpha_1 \leq \alpha_2$.

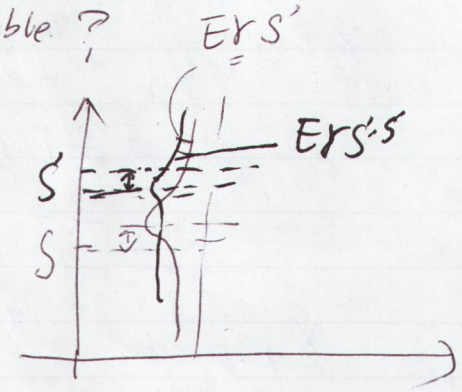
- Exp ~~random~~ value one sto decreasing

ISE 416.

~~Sep 28~~. Oct 3.

Assumptions reasonable ?





$$S_{t+1} - S_t = \kappa_t (\mu_t - S_t)\Delta t + \sigma(t, S_t) \cdot \sqrt{\Delta t}\, \varepsilon \quad \varepsilon \sim N(0,1)$$

Continuous → discrete : $dS_t \to S_{t+1} - S_t$

$dt \to \Delta t$

$dB_t \to \sqrt{\Delta t}\, N(0,1)$.

ISE 418

DP: Infinite horizon: $\max \sum_{t=0}^{\infty} \gamma^t R_t$. $R_t = R_t(S_t, A_t)$

$= R_t(S_t, A_t^{\pi}(S_t))$.

$V(s) = \max_a Q(s,a)$.

$Q(s,a) = R(s,a) + \gamma \mathbb{E} V(s') \quad \forall s$.

$V(s) = \max_a [R(s,a) + \gamma \mathop{\mathbb{E}}_{s'|s,a} V(s')]$

Algorithm: value iteration: (k是 iteration 次数)

set V_0 ~~arbitrarily~~. arbitrarily.

$V_{k+1}(s) = \max_a [R(s,a) + \gamma \mathop{\mathbb{E}}_{s'|s,a} V_k(s')]$

hope: $V_{k+1} \to V^{\infty}$.

$\mathbb{E}$ value of $\tilde{\pi}$:

$V^{\tilde{\pi}}(s_0) = V(\tilde{\pi})|_{s_0} = \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t R_t(s_t, A_{(s_t)}^{\tilde{\pi}}) \mid s_0\right]$.

$V^{\tilde{\pi}}(s) = R(s, A^{\tilde{\pi}}(s)) + \gamma \mathop{\mathbb{E}}_{s'|s,a=\tilde{\pi}(s)} V^{\tilde{\pi}}(s')$.

$V_{k+1}^{\tilde{\pi}}(s) = R(s, A^{\tilde{\pi}}(s)) + \gamma \mathop{\mathbb{E}}_{s'|s,a=A^{\tilde{\pi}}(s)} V_k^{\tilde{\pi}}(s')$.

$\boxed{V_{k+1} = TV_k}$ T: Bellman operation.

value iteration algorithm.

$V_{k+1} = T^{\tilde{\pi}} V_k$. $T^{\tilde{\pi}}$ times V_k.

($T^{\tilde{\pi}}$ is no just a number).

operator ↑ ↑ function

want → $V = TV$. fixed-point solution.

$V_{k+1} = T^{(k+1)} V_0$.

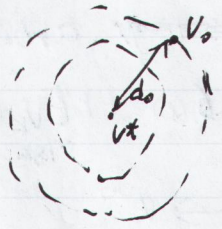
function ↓ ← supernorm

$\|V\|_{sup} = \sup_s (|V(s)|)$.

$\|V_2 - V_1\| = \sup_s |V_2(s) - V_1(s)|$.

$\|V_{k+1} - V_k\| \to 0$.

$d_0 = \|V^* - V_0\|$



$\|V_{k+1} - V^*\| < \|V_k - V^*\|$

$V_{k+1} = TV_k$.

T : contractive.

T : non expansive. $\|V_{k+1} - V^*\| \le \|V_k - V^*\|$ may not be converged.

proof $\|TV_k - TV^*\| \le \gamma \|V_k - V^*\|$

$\Downarrow$

$\|V_{k+1} - V^*\| \le \gamma \|V_k - V^*\|$.

$\le \gamma^{k+1} \|V_0 - V^*\|$

$d := \|J - J'\|$.

$J - d_1 \le J' \le J + d_1$.

By monotonicity of T, $TJ - \gamma d_1 \le TJ' \le TJ + \gamma d_2$

$-\gamma d_1 \le TJ' - TJ \le \gamma d_1$

$\|TJ' - TJ\| \le \gamma d = \gamma \|J - J'\|$.

ISE 416. 1/10.

$TJ^* = J^*$.

$\|TJ - TJ'\| \le \gamma \|J - J'\|$.

$TJ - J \equiv$ Bellman Residual.

converge. threshold.

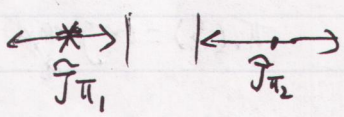
$\|J^* - J\| \le \frac{1}{1-\gamma} \|TJ - J\|$. used as ~~threshold~~

$J^* = TJ^*$ — can be computed.

$\|J^* - TJ\| = \|TJ^* - TJ\| \le \gamma \|J^* - J\| \le \frac{\gamma}{1-\gamma} \|TJ - J\|$. } tighter diff.

$\tilde{\pi} \to J_{\tilde{\pi}}$? $T^{\tilde{\pi}} J_{\tilde{\pi}} - J_{\tilde{\pi}}$. $\delta = \|T^{\tilde{\pi}} J_{\tilde{\pi}} - J_{\tilde{\pi}}\|$

$\|J_{\tilde{\pi}} - J^*_{\tilde{\pi}}\| \le \frac{1}{1-\gamma} \|T^{\tilde{\pi}} J_{\tilde{\pi}} - J_{\tilde{\pi}}\|$.



$\tilde{J} \to \cdot \; \pi_{\tilde{J}} \to J_{\pi_{\tilde{J}}} \to J_{\pi_{\tilde{J}}}$.

↳ $= \sup_{\pi} \mathbb{E} \sum_{t=0}^{\infty} \gamma^t R(S, \pi(S))$. ↳ can't again.

$J = \pi J$. can't solve this exact value.

ISE 416. 10/12.

Policy Iteration:

* Exact solve the equation as written.

never worse than value iteration

* Approximate

not necessarily true.

π_0

Iteration k. 1. Policy evaluation : $J_{\pi_k} = T_{\pi_k} J_{\pi_k}$. (1 way → "Bellman" method)

2. Policy Improvement : $s: \max_a R(s,a) + \gamma \mathbb{E} \sum R(s')$ (a instead $\pi_{k+1}(s)$)

$J' \geq J$

$\Rightarrow TJ' \geq TJ$ monotonicity

$T_{\pi_{k+1}} J_{\pi_k} = T J_{\pi_k}$.

$(T J_{\pi_k} \geq T_{\pi_k} J_{\pi_k})$.

$\pi_1(s) = \arg\max_a \eta(s,a) + \gamma \mathbb{E}_{s'|s,a} J_{\pi_0}(s')$

$J_{\pi_0}(s) = \mathbb{E} \sum_{t=0}^{\infty} \gamma^t R(s, \pi_0(s)) = \eta(s, \pi_0(s)) + \gamma \mathbb{E}_{s'|s,a} J_{\pi_0}(s')$

$s=1 \ldots s=N$

$V_{\pi_0} = \eta_{\pi_0} + \gamma P_{\pi_0} V_{\pi_0}$

P_{π_0} = row of $P^{\pi_0(i)}$

$(I - \gamma P_{\pi_0}) X = \eta_{\pi_0}$

$\Rightarrow X = (I - \gamma P_{\pi_0})^{-1} \eta_{\pi_0}$

3. $\pi_{k+1} \equiv \pi_k$. $\Rightarrow$ $T_{\pi_k} J_{\pi_k} = T_{\pi_{k+1}} J_{\pi_k} = J_{\pi_k}$.

3.8 Approximate Policy Iteration

π_k —(real value)→ $\|J_k - J_{\pi_k}\| \le \delta$.

because of (inverse of big matrix) $\Rightarrow \|T_{\pi_{k+1}} J_k - T J_k\| \le \varepsilon$.

↳ approximate value of π_k.

$$J_k = (1-\lambda)\hat{J}_{\pi_k} + \lambda \tilde{J}_{\pi_{k-1}}$$

10/18 Value iter : find V: $V = TV$. $\Rightarrow T_{k+1} = TV_k$. until convergence.

Policy iter : find π : evaluate : $V^\pi = T^\pi V^\pi$.

$V^\pi_{k+1} = T^\pi V_k^\pi$ $\quad a_t = A^\pi(s_t)$: improve : find better $a^{(t+1)}$.

Linear Optimization. · finit MDP, you get exact V.
however, you solve only small problem.

· Easy to add constaint.

finit MDP. on s, a.

find policy : $\pi = \begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_n \end{bmatrix}$ $\quad P^\pi_{sj} = \text{Prob}(S_{t+1} = j \mid S_t = s)$ $\quad A_t = A^\pi(s_t) =$ $A^\pi(s)$.

$V^\pi = \mathbb{E} \sum_{t=0}^{\infty} \gamma^t R(s_t, A^\pi(s_t))$

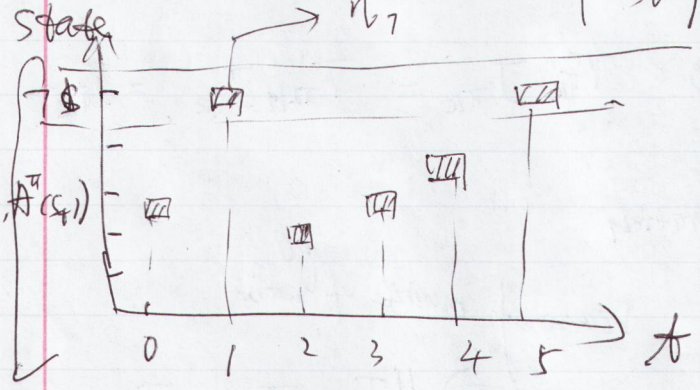
$V^\pi = (I - \gamma P^\pi_{\pi_0})^{-1} r_{\pi_0}$.

$r^\pi_s = \mathbb{E} R(s, A^\pi(s), s_{t+1})$.

$= \sum_{j=1}^{\infty} P^\pi_{sj} \cdot R(s, A^\pi(s), j)$.

$\leq \sum_{t=0}^{\infty} \gamma^t = \frac{1}{1-\gamma} = (1-\gamma)^{-1}$

if staying s, $\bar{r}_7 (1 + \gamma + \gamma^2 + \cdots)$



$\sum_{t=0}^{\infty} \gamma^t \bar{R}(s_t, \pi(s_t))$

$\mathbb{E}$ ⇒ discounted expected count of being in state 7)

$$V_s^\pi = \sum_{j=1}^{m} \bar{r}_j^\pi \cdot \mathbb{E}\left\{\begin{array}{l}\text{discounted} \\ \text{count of} \\ \text{being at} \\ \text{state } j\end{array}\right\}$$

given starting state s and policy π

(sum = 1)

$\frac{1}{(1-\gamma)}$ [discounted fraction of time on a state j]

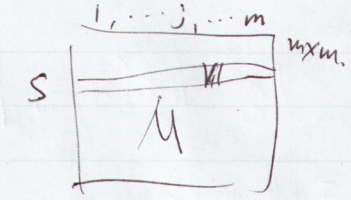
$V^\pi(s)$ = row s of $(I - \gamma P^\pi)^{-1} \bar{r}^\pi = \frac{M}{1-\gamma} \bar{r}^\pi$

$$M = \frac{(I - \gamma P^\pi)^{-1}}{\left(\frac{1}{1-\gamma}\right)} = (1-\gamma)(I - \gamma P^\pi)^{-1}$$

normalize

rows of M = discount fraction of time in state 1, 2, ⋯ m.

M_{sj} 1 2 ⋯ j ⋯ m



M : OCCUPATION measure.

like probability but not exactly real.

$I + \gamma P + \gamma^2 P^2 + (\gamma P)^3 + \cdots + \infty = (I - \gamma P)^{-1}$.

10/31.

$$Q_{k+1}(s,a) = \mathbb{E}\, r(s,a) + \gamma \sum_{s' \mid s,a} V_k(s')$$

$$= r(s,a) + \gamma \sum_{s' \mid s,a} \max_{a' \in A(s')} Q_k(s',a').$$

Temporal-Difference Learning.

$$Q_{k+1}(s,a) = [1-\alpha(s,a)]\, Q_k(s,a).$$

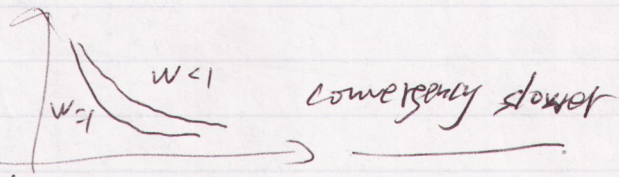
$$+ \alpha(s,a)\,[r(s,a) + \gamma \max_{a'} Q_k(s',a')]$$

(s' : s_{t+1})

$\alpha(s,a) \sim N(s,a)$: # times visiting (s,a).

$\alpha(s,a) = \left[\frac{1}{N(s,a)}\right]^w$. polynomial step size.

$$\begin{cases} \sum_k \alpha_k^2 < \infty. \\ \sum_k \alpha_k \geq \infty. \end{cases} \qquad w \in (0.5, 1].$$



$(s_0, a_0) \xrightarrow{\text{code}} (s_1, a_1) \xrightarrow{\text{code}} (s_2, s_2)$.

a_1 a_2

visit all state-action pair.

∞ often.

with probability $1-\varepsilon$: $a_k = \arg\max Q_k(s_k, a_k)$

ε : $a_k \in A(s_k)$ randomly

0.8

If the step size is too big

convergency at a bad point.

check shaking, window (max − min) ≤ Threshold

Asynchronous Q-learning.

Double Q-learning

Speedy Q-learning.

11/2. feature.

$V(s) = \varphi_1(s)\cdot\theta_1 + \cdots + \varphi_m(s)\cdot\theta_m$.

finit MDP, 5 states.

$\left.\begin{aligned}\varphi_1(s) &= 1(s=1)\\ &\vdots\\ \varphi_5(s) &= 1(s=5)\end{aligned}\right\}\ m \ll |S|$.

basic function $= \{\varphi_k\}_{1\le k\le m}$. $\simeq \varphi_k$.

$f(x) \simeq \sum_{k=0}^{k} w_k \sin(kx) + w_k^{(2)}\cos(kx)$.

~~$k=0$ $w_0^{(2)}$ =~~

$V_{\theta}(s) = \varphi_1(s)\cdot\theta_1 + \cdots + \varphi_m(s)\cdot\theta_m$.

$\{\underbrace{s^i}_{x}, \underbrace{V(s^i)}_{y}\}_{i=1\cdots N} \Rightarrow \theta:\ \min \frac{1}{2}\|\varepsilon\|_2 = \frac{1}{2}\|V_\theta^{target} - \Phi\theta\|_2$

$\varepsilon_i = V(s_i) - V_{\theta}(s_i)$.

$$\Phi = \begin{bmatrix} \varphi_1(s_1) & \cdots & \varphi_m(s_1) \\ \varphi_1(s_N) & \cdots & \varphi_m(s_N) \end{bmatrix}$$

$\theta = (\Phi^T\Phi)^{-1}(\Phi^T V^{target})$.

$\min \|\ \|_w$. $\|V\|_w = \sqrt{V^T W V}$.

W: $\dim N$

$\mathrm{diag}[w_1 w_2 \cdots w_N]$ can be determined by the occupation ~~percentage~~ prob.

11/7.

$V^\pi = T^\pi V^\pi$

$V^\pi(x) = \mathbb{E}_{y|x,\pi(x)}(r + \gamma V^\pi(y))$

$V^\pi(x) = \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t r(x_t, \pi(x_t, x_{t+1})) \mid x_0 = x\right]$

$\|T^\pi V - T^\pi V'\|_{sup} \le \gamma \|V - V'\|$

$V^\pi = T^\pi V^\pi$ $\quad V^\pi(s) = \Phi(s)\theta = \phi_1(s)\theta_1 + \dots + \phi_m(s)\theta_m$

$\min_\theta \|\Phi\theta - T^\pi \Phi\theta\|_2 \Leftarrow \min_\theta \|V^\pi - T^\pi V^\pi\|_{sup}$

$\bar{r} + \gamma P^\pi \Phi\theta$.

$V(s_1) \sim r(s_1) + \gamma P^\pi_{s_1} \Phi\theta$.

$\min_\theta \|\underbrace{(I - \gamma P^\pi)\Phi}_{\Psi}\theta - \bar{r}\|_2$ estimate P^π.

↳ QUADRATIC Bellman minimization

$(\Psi^T\Psi)^{-1}(\Psi^T\bar{r})$

$\Rightarrow \|\Phi\theta - \Pi_\Phi(T^\pi\Phi\theta)\|$

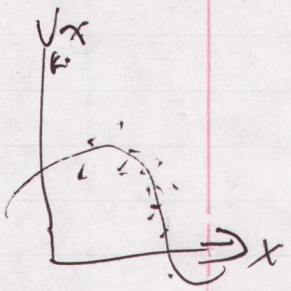
↳ Bellman Residual minimization.

11/9.

V_k : approximation $\Leftrightarrow \theta_k$.

$V_k(x) = \varphi_1(x)\theta_{k1} + \cdots + \varphi_m(x)\theta_{km}$.

Sample $x_i \sim \mu$, $i = 1, \cdots, N$.



$x_1 \to \tilde{V}_{k+1}(x_1) = \max_a \mathbb{E}_{y|x,a}[r(x_1, a, y) + \gamma V_k(y)]$

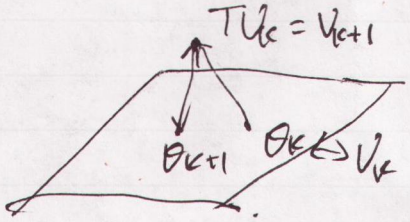
$\tilde{V}^M_{k+1}(x_1) = \max_a \frac{1}{M} \sum_{j=1}^{M} [r(x_1, a, y_{1j}) + \gamma V_k(y_{1j})]$

$x_N \to \tilde{V}^M_{k+1}(x_N) = \max_a \frac{1}{M} \sum_{j=1}^{M} [r(x_N, a, y_{Nj}) + \gamma V_k(y_{Nj})]$.

$[(x_1, \tilde{V}^M_{k+1}(x_1)), \cdots, (x_N, \tilde{V}^M_{k+1}(x_N))]$.

$$\min_{\theta_{k+1}} \| \tilde{V}^M_{k+1} - \Phi\theta \| = \left\| \begin{matrix} \tilde{V}_{k+1}(x_1) - \sum_{l=1}^{m} \varphi_l(x_1)\theta_l \\ \vdots \\ \tilde{V}_{k+1}(x_N) - \sum_{l=1}^{m} \varphi_l(x_N)\theta_l \end{matrix} \right\|$$

$+ \|\theta\| \le A$



$\sup_{V \in \mathcal{F}} \inf_{f \in \mathcal{F}} (\|f - TV\|)$.

$\downarrow$ θ_{k+1}

$\downarrow$

$V_s = \max_a r_{s,a} + \gamma P^a_{s,\cdot} V$

$= \max_a r_{s,a} + \gamma \sum_j P^a_{s,j} V_j$

$\begin{cases} V_s \geq r_{s,a} + \gamma \sum_j P^a_{s,j} \cdot V_j \quad \forall a, s \end{cases}$

~~n x n~~ n x m constraints

objective: $\min_{s} \sum_s w_s V_s \quad w_s > 0$

dual value $\neq 0$, the optimal a at state s.

Stochastic policy

$y_{sa} = \text{Prob}(a_t = a \mid s_0 = s)$

for each state.

sum of dual value is

$M(s)$.

$\max_x \min_{R' \in \mathcal{C}_b} \sum_{s,a} R'_{s,a} x_{s,a}$

s.t. $\sum_a x_{s,a} - \gamma \sum_i \sum_a P^a_{is} x_{ia} = \bar{w}_s \quad \forall s.$

$x_{s,a} \geq 0.$

Robust MDP.

$\mathcal{C}_b = \mathcal{E}(\bar{R}, \Sigma).$

$(y - \bar{R})^T \Sigma^{-1} (y - \bar{R}) \leq \alpha^2.$

$\max_\pi \min_{R'} \mathbb{E}^\pi \sum_{t \geq 0}^{\infty} \gamma^t R'(s,a).$

$\min_{y \in \mathcal{C}_b} x^T y = \min x^T y$

s.t. $(y - \bar{R})^T \Sigma^{-1} (y - \bar{R}) \leq \alpha^2.$

$\|\Sigma^{-1/2} (y - \bar{R})\|_2 \leq \alpha.$

$\max \Rightarrow -\alpha \sqrt{x^T \Sigma x} + \bar{R}^T x$

$\max_x \sum_{s,a} \bar{R}_{s,a} x_{s,a} - \alpha \sqrt{\sum_{s,a} \sum_{s',a'} x_{sa} \Sigma_{sa,s'a'} x_{s'a'}}$

s.t. $\sum_a x_{s,a} - \gamma \sum_i \sum_a P^a_{is} x_{ia} = \bar{w}_s$

$x_{s,a} \geq 0.$

10/26

COPE

$$\begin{cases} s_{t+1} = f(s_t, a_t) \\ r_{t+1} = r(s_t, a_t). \end{cases}$$

$(s_0, a_0, s_{t+1}, a_{t+1})$

MODEL - BASED

MDP : R, P. m: # of sampling

$$P^a_{ij} = \frac{1}{m} \sum_{m=1}^{m} \mathbb{1}\{s'=j \mid s,a\} = \frac{N_{saj}}{m} \xrightarrow[m\to\infty]{} P(j \mid s,a)$$

$$\forall j \quad \bar{R}_{sa} = \frac{1}{m} \sum_{m=1}^{m} r(s,a) \xrightarrow[m\to\infty]{} \bar{R}(s,a)$$

$\rightarrow \hat{\pi}$ optimal for $\hat{P}_m, \hat{R}_m$

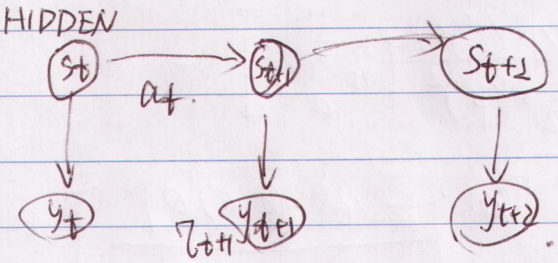
$\rightarrow V^{\hat{\pi}}$

11/21.

POMDP : Partially observable MDP.

MDP: $x_t, a_t \to x_{t+1}, a_{t+1}$.

POMDP:



$P(s_t = i \mid y_0, y_1, \cdots y_t)$.

$P(s_{t+1} = i \mid y_0, y_1, \cdots y_{t+1}) = f(P(s_t = i \mid y_0 \cdots y_t), y_{t+1})$.

P^a, R^a

$Y^a \quad Y^a_s = Prob(y_t = k \mid S_t = s, A_{t-1} = a)$.

$\max_{\pi \in \Pi} \mathbb{E} \sum_{t=0}^{\infty} R_t$.

$\max_{\theta \in H} \mathbb{E}^{\pi_\theta} \sum_{t=0}^{\infty} R_t \to \max_{\theta \in H} \int^{\pi_\theta}(s_0) = f(\theta)$

Gittis policy

$$\begin{bmatrix} \phi_1(s,a) \\ \phi_2(s,a) \\ \vdots \\ \phi_m(s,a) \end{bmatrix}$$

$Q(s,a) \simeq \phi(s,a)^T \theta \in \mathbb{R}$

$Prob(a|s) = \pi(s,a) = \dfrac{\exp(\phi(s,a)^T \theta)}{\sum_{a'} \exp(\phi(s,a)^T \theta)}$.

$[\pi(s,1), \cdots \pi(s,m)] \quad a \sim \pi$.

$$\nabla_\theta \pi(s,a) = \nabla_\theta \left[\frac{1}{\sum_{a'} \exp} \cdot \exp(\phi(s,a)^T \theta) \right]$$

$$= \left[\nabla_\theta \frac{1}{\sum_{a'} \exp} \right] \exp(\phi(s,a)^T\theta) + \frac{1}{\sum_{a'} \exp} \nabla_\theta \exp(\phi(s,a)^T\theta)$$

$$= \frac{-\exp(\phi(s,a)^T\theta)}{(\sum_{a'} \exp)^2} \left[\nabla_\theta \sum_{a'} \exp \right] + \cdots$$

$$= -\pi(s,a) \cdot \left[\frac{\sum_{a'} \nabla_\theta \exp(\phi(s,a')^T\theta)}{\sum_{a''} \exp(s,a'')} \right] + \frac{1}{\sum_{a'} \exp} \left[\phi(s,a) \exp(\phi(s,a)^T\theta) \right]$$

$$= -\pi(s,a) \left[\frac{\sum_{a'} \phi(s,a') \exp(\phi(s,a')^T\theta)}{\sum_{a''} \exp(\phi(s,a'')^T\theta)} \right] + \phi(s,a) \cdot \pi(s,a)$$

$$= \pi(s,a) \left[\phi(s,a) - \sum_{a'} \phi(s,a') \pi(s,a') \right]$$

$$\nabla_\theta \log \pi(s,a) = \frac{1}{\pi(s,a)} \nabla_\theta (\pi(s,a)) = \left[\phi(s,a) - \sum_{a'} \phi(s,a') \pi(s,a') \right]$$

$\nabla_\theta \mathbb{E}_{p_\theta} f(x)$ $\quad x \sim p_\theta$.

$$\nabla_\theta \int f(x) p(x|\theta) \, dx$$

$$= \int \nabla_\theta [f(x) p(x|\theta)] \, dx$$

$$= \int f(x) \nabla_\theta (p(x|\theta)) \, dx$$

$$= \int f(x) p(x|\theta) \nabla_\theta \log p(x|\theta) \, dx$$

$$= \mathbb{E}_{p_\theta} [f(x) \nabla_\theta \log p(x|\theta)]$$

SCORE METHOD.

$$\sim \frac{1}{N} \sum_{m=1}^{N} [f(x_m) \nabla_\theta \log p(x_m|\theta)]$$

policy search

fisher matrix (information)

Homework 3

Ruxin Xu.

Question 1. Show that a decreasing convex function of an increasing submodular function is supermodular. For simplicity, you can assume the functions are twice differentiable.

Answer.

Set: $g(x)$ is a decreasing convex function.
$f(x)$ is an increasing submodular function.

Because $g(x)$ is decreasing, $f(x)$ is increasing.
Thus $g(f(x))$ is decreasing, $g'(f(x)) \le 0$.

Because $f(x)$ is submodular,
Thus $\frac{\partial^2 f(x)}{\partial x_i \partial x_j} \le 0$, $\forall i \neq j$.

Because $f(x)$ is increasing, thus $f'(x) \ge 0$

Because $g(x)$ is convex, Thus $g''(f(x)) \ge 0$.

$\frac{\partial f(x)}{\partial x_i}$ // $\frac{\partial f(x)}{\partial x_j}$

Thus: $$\frac{\partial^2 g(f(x))}{\partial x_i \partial x_j} = \frac{\partial f'(x) \cdot g'(f(x))}{\partial x_j} = \underbrace{\frac{\partial f'(x)}{\partial x_i \partial x_j}}_{\le 0} \cdot \underbrace{g'(f(x))}_{\le 0} + \underbrace{f'(x)}_{\ge 0} \cdot \underbrace{f'(x)}_{\ge 0} \cdot \underbrace{g''(f(x))}_{\ge 0}$$

$$\ge 0 \quad \forall i \neq j.$$

Thus, it's a supermodular.

Question 2. A family of random variables $\{X_s\}$ parameterized by s is said to be stochastically convex in s if their complementary cumulative distribution function $F_s^c(t) = Prob(X_s > t)$ is convex in the parameter s for each fixed t. Show that the expectation function $h(s) = \mathbb{E}[X_s]$ is convex in s.

Answer.

$$h(s) = \mathbb{E}[X_s] = \int_0^\infty F_s^c(x)\,dx$$

$$\frac{\partial^2 h(s)}{\partial s^2} = \frac{\partial^2 \int_0^\infty F_s^c(x)\,dx}{\partial s^2} = \frac{\partial \int_0^\infty \frac{\partial F_s^c(x)}{\partial s}\,dx}{\partial s}$$

$$= \int_0^\infty \frac{\partial^2 F_s^c(x)}{\partial s^2}\,dx$$

Because $F_s^c(x)$ is convex, thus $\frac{\partial^2 F_s^c(x)}{\partial s^2} \ge 0$.

Thus, $\frac{\partial^2 h(s)}{\partial s^2} \ge 0$, $h(s) = \mathbb{E}[X_s]$ is convex in s.

Project 1

$$\min \ \frac{1}{2} w^T w + C \sum_{i=1}^{N} [\tau_m \xi_s^+ + (1-\tau_m)\xi_s^-]$$

$$\text{s.t.} \quad \left.\begin{aligned} y - Xw &\le \varepsilon + \xi_s^+ \\ -y + Xw &\le \varepsilon + \xi_s^- \end{aligned}\right\} \Rightarrow \begin{aligned} Xw - y + \varepsilon + \xi_s^+ &\ge 0 \\ -Xw + y + \varepsilon + \xi_s^- &\ge 0 \end{aligned}$$

$$\xi_s^+ \ge 0 \ , \ \xi_s^- \ge 0$$

Lagrangian Duality:

$$\mathcal{L}(w, \xi_s^+, \xi_s^-, \alpha^+, \alpha^-, \beta^+, \beta^-) = \frac{1}{2} w^T w + C \sum_{i=1}^{N} [\tau_m \xi_s^+ + (1-\tau_m)\xi_s^-] - \alpha^{+T}(Xw - y + \varepsilon + \xi_s^+) - \alpha^{-T}(-Xw + y + \varepsilon + \xi_s^-) - \beta^{+T}\xi_s^+ - \beta^{-T}\xi_s^-$$

$$\frac{\partial \mathcal{L}}{\partial w} = w - X^T\alpha^+ + X^T\alpha^- \overset{\text{set}}{=} 0 \Rightarrow w = X^T\alpha^+ - X^T\alpha^-$$

$$\frac{\partial \mathcal{L}}{\partial \xi_s^+} = C \cdot \tau_m \cdot I - \alpha^+ - \beta^+ \ , \ \frac{\partial \mathcal{L}}{\partial \xi_s^-} = C \cdot (1-\tau_m) - \alpha^- - \beta^-$$

$$\mathcal{L} = \frac{1}{2}(X^T\alpha^+ - X^T\alpha^-)^T(X^T\alpha^+ - X^T\alpha^-) + C\sum_{i=1}^{N}[\tau_m\xi_s^+ + (1-\tau_m)\xi_s^-] - \alpha^{+T}[X(X^T\alpha^+ - X^T\alpha^-)] - \alpha^{-T}[-X(X^T\alpha^+ - X^T\alpha^-)] - \alpha^{+T}(-y+\varepsilon+\xi_s^+) - \alpha^{-T}(y+\varepsilon+\xi_s^-) - \beta^{+T}\xi_s^+ - \beta^{-T}\xi_s^-$$

$$= \frac{1}{2}(X^T\alpha^+ - X^T\alpha^-)^T(X^T\alpha^+ - X^T\alpha^-) + C\sum_{i=1}^{N}[\tau_m\xi_s^+ + (1-\tau_m)\xi_s^-] - (\alpha^{+T}X - \alpha^{-T}X)(X^T\alpha^+ - X^T\alpha^-) - \alpha^{+T}(-y+\varepsilon+\xi_s^+) - \alpha^{-T}(y+\varepsilon+\xi_s^-) - \beta^{+}\xi_s^+ - \beta^{-T}\xi_s^-$$

$$= -\frac{1}{2}(X^T\alpha^+ - X^T\alpha^-)^T(X^T\alpha^+ - X^T\alpha^-) + C\sum_{i=1}^{N}[\tau_m\xi_s^+ + (1-\tau_m)\xi_s^-] - (X^T\alpha^+ - X^T\alpha^-)^T(X^T\alpha^+ - X^T\alpha^-) - \alpha^{+T}(-y+\varepsilon+\xi_s^+) - \alpha^{-T}(y+\varepsilon+\xi_s^-) - \beta^{+}\xi_s^+ - \beta^{-T}\xi_s^-$$

$$= \boxed{-\frac{1}{2}(X^T\alpha^+ - X^T\alpha^-)^T(X^T\alpha^+ - X^T\alpha^-)} + \underbrace{C\cdot\tau_m\cdot I^T\cdot\xi_s^+}_{\text{cancels}} + C\cdot(1-\tau_m)\cdot I^T\cdot\xi_s^- + \alpha^{+T}y - \alpha^{+T}\varepsilon - \alpha^{+T}\xi_s^+ - \alpha^{-T}y - \alpha^{-T}\varepsilon - \alpha^{-T}\xi_s^- - \beta^{+T}\xi_s^+ - \beta^{-T}\xi_s^-$$

$$= -\frac{1}{2}(X^T\alpha^+ - X^T\alpha^-)^T(X^T\alpha^+ - X^T\alpha^-) + \underbrace{[C\cdot\tau_m\cdot I^T - \alpha^{+T} - \beta^{+T}]}_{=0}\xi_s^+ + \underbrace{[C\cdot(1-\tau_m)\cdot I^T - \alpha^{-T} - \beta^{-T}]}_{=0}\xi_s^- + \alpha^{+T}(y-\varepsilon) - \alpha^{-T}(y+\varepsilon)$$

$$= -\frac{1}{2}\alpha^{+T}XX^T\alpha^+ - \frac{1}{2}\alpha^{-T}XX^T\alpha^- + \underbrace{\alpha^{+T}XX^T\alpha^-}_{\frac{1}{2}\alpha^{+T}XX^T\alpha^- + \frac{1}{2}\alpha^{-T}XX^T\alpha^+} + \alpha^{+T}(y-\varepsilon) - \alpha^{-T}(y+\varepsilon)$$

$\frac{1}{2}\alpha^{+T}K\alpha^{-} + \frac{1}{2}\alpha^{-T}K\alpha^{+}$ $\quad K = XX^T$

$$\max \; \alpha^{+T}XX^T\alpha^{-} - \tfrac{1}{2}\alpha^{+T}XX^T\alpha^{+} - \tfrac{1}{2}\alpha^{-T}XX^T\alpha^{-} + \alpha^{+T}(y-\varepsilon) - \alpha^{-T}(y+\varepsilon)$$

s.t. $\alpha_i^{+} \in [0, C\tau_m], \forall i$

$\alpha_i^{-} \in [0, C(1-\tau_m)], \forall i$

$$\frac{\partial L}{\partial \alpha^{+}} = XX^T\alpha^{-} - XX^T\alpha^{+} + y - \varepsilon.$$

$$\frac{\partial L}{\partial \alpha^{-}} = XX^T\alpha^{+} - XX^T\alpha^{-} - y - \varepsilon.$$

$$\frac{\partial L}{\partial \alpha^{+}} = K\alpha^{-} - K\alpha^{+} + y - \varepsilon$$

$\frac{1}{2}K\alpha^{-} + \frac{1}{2}K^T\alpha^{-}$

$$\frac{\partial L}{\partial \alpha^{-}} = K^T\alpha^{+} - K\alpha^{-} - y - \varepsilon.$$

$\frac{1}{2}K\alpha^{+} + \frac{1}{2}K^T\alpha^{+}$

CUR →