

۱۱

$$V_K^* = \max_{\pi} \mathbb{E} \left\{ \sum_{t=0}^{K-1} \gamma^t r_t \right\} \leq \max \left\{ \sum_{t=0}^{K-1} \gamma^t R_{\max} \right\} =$$

$$R_{\max} \sum_{t=0}^{k-1} \gamma^t = R_{\max} \frac{1-\gamma^k}{1-\gamma} \leq \frac{R_{\max}}{1-\gamma}$$

٢٠١  
دراستی

متاخرة Value '  $V_k^*$

Policy بھینہ اسیت۔ می خواصیں ثابت کرنے لئے  $\sqrt{V_{k+1}} \geq V_k^*$  باید رکھا کریں گے اور  
برقرار رکھو۔ Policy  $I_k$  کا کام کام درستہ ہی کریں، خواصیں (اسیت) کے

$$\sqrt{J_K^R} = \sqrt{J_K}$$

$$V_{k+1}^{\pi} = V_k^{\pi} + \gamma^k r_k$$

از مکاری طریق  $\leq$  Reward درستیعه خواهد داشتند

$$\sqrt{\gamma_{k+1}}^* = \sqrt{\gamma_k}^* = \sqrt{\gamma_k}^* + \gamma^k r_k \geq \sqrt{\gamma_k}^* = \sqrt{\gamma_k}^* \longrightarrow$$

کاهشی تحویل بعد از حلقه !  $\frac{R_{max}}{1-\gamma}$  نیز باند است  $\leftarrow$  بحواب

will converge with  معاذل

$$V_{k+1}^*(s) = \max_{\alpha} \sum_{s'} P(s'|s, \alpha) [R(s, \alpha) + \gamma V_k^*(s)]$$

$$\lim V_k^*(s) = \lim_{\alpha} \max_{s'} \sum p(s' | s, \alpha) [R(s_{2\alpha}) + \gamma V_{k-1}^*(s')]$$

$$= \max_a \sum_{s'} p(s'|s, a) \left( R(s, a) + \gamma \lim_{k \rightarrow \infty} V_{k-1}^*(s') \right)$$

$$\Rightarrow \text{Bellman Equation} \quad V^*(s) = \max_{\alpha} \sum_{s'} P(s'|s, \alpha) (R(s, \alpha) + \gamma V^*(s'))$$

$$\lim_{K \rightarrow \infty} V_{K-1}^* = \lim_{K \rightarrow \infty} V_K^* = V^* \quad \text{لما زادت}\ K \text{،}\ V_K^* \text{ ينضم إلى}\ V^*$$

: General Reward ٢.١

م. حواصي حسنه reward حاصل من سالء  
Value iteration  $\Rightarrow$  م. حواصي حسنه reward  
كانع ميكنه ،

$\hat{V}_K^*$   $\rightarrow$  مقدار return بعد K عام بافرض  
، reward متغير بـ  $r_0$

$$\hat{V}_K^* = V_K^* + r_0 + \gamma r_0 + \dots + \gamma^{K-1} r_0 = V_K^* + r_0 \left( \frac{1 - \gamma^K}{1 - \gamma} \right)$$

حران ماهر دلخور انتقام - تغير ترموم  $\rightarrow$  Policy

رادرس  $\Leftarrow$  وقتى بعض حسنه  $\hat{V}_K^*$  مالو جمع  $(V_K^*)$  بالذ

حالا دلخور .  $\hat{\pi}_K = \pi_K \Leftarrow$

$$\hat{V}_K^* = \max_{\alpha} \sum_{s'} P(s'|s, \alpha) [R(s, \alpha) + r_0 + \gamma \hat{V}_{K-1}^*(s')]$$

$$= \max_{\alpha} \sum_{s'} P(s'|s, \alpha) [R(s, \alpha) + r_0 + \gamma \hat{V}_{K-1}^*(s') + \gamma (r_0 + \dots)]$$

$$= \underbrace{\max_{\alpha} \sum_{s'} P(s'|s, \alpha) [R(s, \alpha) + \gamma V_{K-1}^*(s')]}_{= V_K^*(s)} + \left( r_0 + r_0 \left( \frac{\gamma (1 - \gamma^{K-1})}{1 - \gamma} \right) \right)$$

$$= V_K^*(s) + r_0 \left( \frac{1 - \gamma + \gamma - \gamma^K}{1 - \gamma} \right) = V_K^*(s) + r_0 \left( \frac{1 - \gamma^K}{1 - \gamma} \right)$$

،  $V_K^*(s) = 0 : K \rightarrow \infty$   $\Leftarrow$

$$\hat{V}_K^*(s) = V_K^*(s) + r_0 \left( \frac{1 - \gamma^K}{1 - \gamma} \right)$$

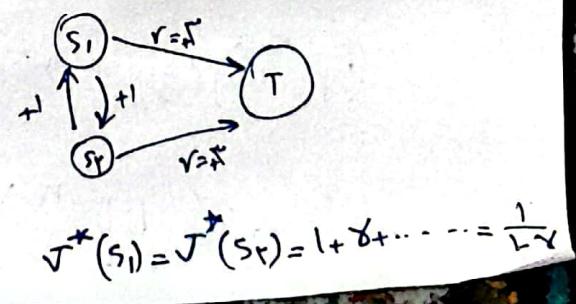
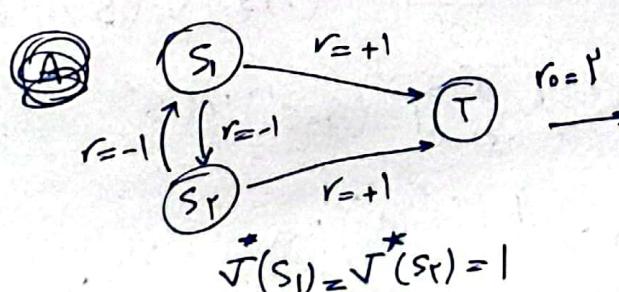
حال خواصی داشته که

$$\lim_{K \rightarrow \infty} \hat{V}_K^*(s) = \lim_{K \rightarrow \infty} \left( V_K^*(s) + r_0 \left( \frac{1 - \gamma^K}{1 - \gamma} \right) \right) =$$

$$\left( \lim_{K \rightarrow \infty} V_K^*(s) \right) + \left( \frac{r_0}{1 - \gamma} \lim_{K \rightarrow \infty} (1 - \gamma^K) \right) = V^*(s) + \frac{r_0}{1 - \gamma}$$

بعنوان  $V_K^*$  به صفتار بجای  $V^*$  converge به  $V^*$  باشد  
Policy ایسی مجاز است که  $V_K^*$  به این  $V^*$  converge کند  
 $\lim_{K \rightarrow \infty} \frac{r_0}{1 - \gamma}$  به این  $V^*$  باشد فعلاً  $V^*$  قدرتی ایسی مجاز است  
که  $V_K^*$  به این  $V^*$  converge کند و  $r_0$  این  $V^*$  را terminal state نهاده کنند  
و  $V_K^*$  به این  $V^*$  converge کند و  $r_0$  این  $V^*$  را non-terminal state نهاده کنند

و  $\frac{r_0}{1 - \gamma}$  به این  $V^*$  converge کند و  $r_0$  این  $V^*$  را non-terminal state نهاده کنند  
Policy  $\pi$  به این  $V^*$  converge کند و  $r_0$  این  $V^*$  را non-terminal state نهاده کنند  
و  $V_K^*$  به این  $V^*$  converge کند و  $r_0$  این  $V^*$  را terminal state نهاده کنند  
و  $V_K^*$  به این  $V^*$  converge کند و  $r_0$  این  $V^*$  را terminal state نهاده کنند



نهایی Policy یعنی چگونه از مکانی کجا حرکت کنیم تا به Terminal States برسیم

1 Policy Turn ۳.۱

برای policy iteration

$$\pi_{k+1}(s) = \underset{a \in A}{\operatorname{argmax}} Q^{\pi_k}(s, a)$$

$$Q^{\pi_k}(s, \pi_{k+1}(s)) = \max_{a \in A} Q^{\pi_k}(s, a) \geq Q^{\pi_k}(s, \pi_k(s)) \\ = V^{\pi_k}(s)$$

برای bellman operator

$$T^{\pi_{k+1}} V^{\pi_k} \geq V^{\pi_k}$$

برای این را اثبات کنیم

$\forall s, V(s) \geq u(s)$  باشد،  $V$  value func داشته باشد و  $u$  خواصی داشته باشد

$$(T^\pi V)(s) = R^\pi(s) + \gamma \sum_{s'} P^\pi(s, s') V(s')$$

$$(T^\pi u)(s) = R^\pi(s) + \gamma \sum_{s'} P^\pi(s, s') u(s')$$

$$\rightarrow (T^\pi V)(s) \geq (T^\pi u)(s) \quad \forall s \in S$$

$$(T^\pi V)(s) \geq (T^{\pi_k} V)(s), \quad \forall s$$

،  $\pi_{k+1}$  داریم  $V(s) \geq V_k(s)$  ،  $\forall s$  (جایگزینی  $\pi_k$  با  $\pi_{k+1}$ )

$$\pi_{k+1}(s) = \underset{a \in A}{\operatorname{argmax}} Q^{\pi_k}(s, a)$$

$$Q^{\pi_k}(s, \pi_{k+1}(s)) = \max_a Q^{\pi_k}(s, a) \geq Q^{\pi_k}(s, \pi_k(s)) = V^{\pi_k}$$

حالاً باید این قدر را در نظر اعمال می کنیم،

$$(T^{\pi_{k+1}} V^{\pi_k})(s) = Q^{\pi_k}(s, \pi_{k+1}(s)) \geq V^{\pi_k}(s) \quad \forall s \in S$$

،  $T^{\pi_{k+1}}$  را خواهیم داشت  $T^{\pi_{k+1}}$  بار اعمال کنیم  $\sqrt{n}$  حالاً

$$V^{\pi_{k+1}} = \lim_{n \rightarrow \infty} (T^{\pi_{k+1}})^n V^{\pi_k} \geq (T^{\pi_{k+1}})^{n-1} V^{\pi_k} \geq \dots \geq V^{\pi_k}$$

،  $T^{\pi_{k+1}}$  باید نقطه ثابتی فلکی اسے،  $T^{\pi_k}$  جونه

، عبارت بالا ایجاد می شود.

،  $T^{\pi_{k+1}}$  بعین نیاز نداشته اند موقع خواصی داشته باشند.

$$\max_a Q^{\pi_k}(s, a) > V^{\pi_k}_{k+1}(s)$$

$$\left( \overbrace{V^{\pi_K}}^{\text{فقط}} + \overbrace{\tau_{K+1}}^{\text{لهم}} \right) (s) \geq V^{\pi_K}(s)$$

and Strict Global Optimality only ~~exists~~

وہ مدرسی ادارے Policy, Curriculum

کانٹھی کمپ  $\leftarrow$  گلکنڈ

$$\max_{\pi} \mathbb{E}_{\pi} [S^T] = \sqrt{\pi_{K+1}(S)} = \sqrt{\pi_K(S)} \Rightarrow \text{all cards are equal}$$

باختلاف  $\alpha$   $Q^{\pi_K}(s, \alpha) = V^{\pi_K}(s)$   $\forall s$   $\in S$

finite in  $|A|$ ,  $|S|$  called finite state MDP  $\sqrt{\text{نحو}} \rightarrow \checkmark$

فستان ، حافظه ایمنی داد و سرمه هنگام در پیش قتل

~~strict policy~~ is going to give strict rules ~~improve~~  
strict policy improvement

لہے بیتے بعضیہ سڑکی ٹھوکا رکھیں۔ ٹھوکا رکھنے کا ایک طریقہ Policy Iteration ہے جو دو تا

~~Confidence~~ ~~mapping~~ ~~Policy~~ ~~suboptimal~~

جوہ تعداد finite ( $|A|^{|\Sigma|}$ ) میں policy بدلنے کی

$\pi_{k+1} = \pi_k$   ~~$\pi_k \neq \pi_k$~~  ← جواب ممکن باشد to Policy iteration

~~فوجہ کسی نے اسی دارالحکم (جی ان ٹی نیشنل برائی) میں بھینی بھنیں برقیں نہیں تے~~

$$\max_{\alpha} \sum_{s'} P(s' | s, \alpha) [R(s, \alpha) + \gamma V^{\pi_K}(s')] >$$

$$\sum_{s'} p^{\pi_k}(s' | s, \pi_k(s)) [R(s, \pi_k(s)) + \gamma V^{\pi_k}(s')] = V^{\pi_k}(s)$$

~~soft commitment mapping~~ in the sense of Policy  $\pi_k \leftarrow$   
 $\pi_{k+1}$  whenever  $\pi_k \neq \pi_{k+1}$

جواب ایجاد کننده می باشد  $V^{\pi_K}$

Convergence! Value iter, Policy iter  $\xrightarrow{\text{interleave}} \pi^*$

با وجود آن نتایج قبلی هر دو به جواب معادله بینیکی پس از Convergence می‌رسد و برای این

$\gamma$ -contradiction  $\hookrightarrow$  Bellman optimal operator  $\Rightarrow$  داینامیک مسیرهای ممکن

mapping  
policy/value iter  $\rightarrow$  to the fixed point ~~of g~~  $\in$  bestis  
drifts  $\rightarrow$  value function

لکھاں میں چینیں Policy کاں جمعیت اپنے مفہوم کی  
جسکی دلیل ہے ۱۱-۰۸-۲۰۱۷

تعدادی انتخابی action بین سه محتوا را در میان معاکر می‌گیرد.

بھینی بھن را مانزیر کر لے۔

$$V^{\pi_k}(s) = \sum_{s'} P(s'|s, a) [R(s, a) + \gamma V^{\pi_k}(s')] \quad \forall s \in S$$

لهم  $O(|S|^r)$  ،  $\leftarrow$  مجموع حالات  $|S|$  ، معنی  $|S|$  حالات  
 $\leftarrow$  one policy iter  $\leftarrow$  one value iter  $\leftarrow$  policy improvement

$$\text{ts: } \pi_{k+1}(s) = \arg \max_{a \in A} \sum_{s'} P(s'|s, a) [R(s, a) + \gamma V^{\pi_k}(s')]$$

$|S| \leftarrow S$  حالات  $|A|$  اجراءات  $|A|$   $\leftarrow$  one policy iter  $\leftarrow$  one value iter

$O(|S|^r + |S|^r |A|)$  ، one policy iter  $\leftarrow$  one value iter

$$\text{ts: } V_{k+1}^*(s) = \max_a \sum_{s'} P(s'|s, a) [R(s, a) + \gamma V_{k+1}^*(s')]$$

$\max_a$  ، one value iter  $\leftarrow$  one value iter

$O(|S|^r |A|) \leftarrow$  one value iter

one value iter  $\leftarrow$  one value iter

one value iter  $\leftarrow$  one value iter

1. ~~finite MDP~~  $\rightarrow \gamma = 1$  - وعیت داریم که

Value iter تعریف

$$V_k^*(s) = \max_{a \in A} \sum_{s'} P(s'|s, a) [R(s, a) + \gamma V_{k+1}^*(s')]$$

Loop  $\approx$  ~~دایرکت~~  $\rightarrow k \rightarrow \infty$  وعیت با ریتارد

متبتلی خواهد شد تا  $V$  باشند.

Policy iter داریم که

$$V = R^\pi + P^\pi V \rightarrow (I - P^\pi)V = R^\pi$$

برای داشتن جواب  $V$  باید  $I - P^\pi$  وارون نباشد.

$I - P^\pi$  معکار  $P^\pi$  است  $P \leftarrow P^{-1}$  از قابلیت  $\circlearrowleft$

معکار فرآیند صفر طارد درستیغ (برای ماتریس دروغ نباشد).

converges to Policy iter  $\leftarrow$   $V$   $\leftarrow$   $\underbrace{\dots}_{\text{نهایت}}.$

نحوی  $V$  باشد terminal

و همچنان  $P^\pi$  ماتریس  $\pi$  و دستیغ  $\pi$  باشد.

بعد محضی ماتریس  $P^\pi$  از  $\pi$  که حالت خواهد بود.

و  $\pi$  diverge باشد terminal نباشد.

loop با ریتارد متبتلی باشد  $\pi$  MDP = ~~state~~

هر دو راهنمایی خواهد میگشت.

$$\begin{aligned}
 \|B^\pi V - B^\pi V'\| &= \max_s \left( \left| \left[ r(s, \pi(s)) + \gamma \sum_{s'} P(s'|s, \pi(s)) V(s') \right] \right. \right. \\
 &\quad \left. \left. - \left[ r(s, \pi(s)) + \gamma \sum_{s'} P(s'|s, \pi(s)) V'(s') \right] \right| \right) \\
 &\leq \gamma \max_s \left\{ \sum_{s'} P(s'|s, \pi(s)) |V(s') - V'(s')| \right\} \leq \\
 &\quad \gamma \max_s \sum_{s'} P(s'|s, \pi(s)) \|V - V'\| = \gamma \|V - V'\| \\
 \Rightarrow \|B^\pi V - B^\pi V'\| &\leq \gamma \|V - V'\|
 \end{aligned}$$

٢. فرض كثيرون  $V, V', V''$  fixed point في موقع دارين

$$\|V - V'\| = \|B^\pi V - B^\pi V'\| \leq \gamma \|V - V'\|$$

جع  $\gamma \leq 1$  بركلد اس  $\Leftrightarrow$  فعلا تتحقق حالتي

دارين اينه  $V = V'$  باشه

لذلك  $V \leq V'$  دارين

$$\forall s : V(s) \leq V'(s)$$

$$B^\pi V(s) \leq B^\pi V'(s)$$

$$\begin{aligned}
 B^\pi V(s) &= r(s, \pi(s)) + \gamma \sum_{s'} P(s'|s, \pi(s)) V(s') \leq r(s, \pi(s)) + \\
 &\quad \gamma \sum_{s'} P(s'|s, \pi(s)) V'(s') = B^\pi V'(s) \rightarrow \text{لذلك}
 \end{aligned}$$

دقتی کی میں  $BV - V$  برابر صفر خواهد بود.

$\|BV^* - V^*\|$  میں  $V = V^*$  باسے جاتا ہے، نتیجے میں سودہ اور صفر خواهد بود.

۵۰ خواص سلسل دھرمیہ

$$\|V - V^\pi\| \leq \frac{\|V - BV^\pi\|}{1-\gamma}, \|V - V^*\| \leq \frac{\|V - BV\|}{1-\gamma}$$

می دانیں کہ بارزای  $\bullet$  بالمسوی درج کیا گیا ہے  
برابر ① از شایع بعضی

$$\|BV - BV^\pi\| \leq \gamma \|V - V^\pi\| \rightarrow$$

~~$$\|V - BV\| + \gamma \|BV - BV^\pi\| \leq \|V - BV\| + \gamma \|V - V^\pi\|$$~~

$$\|V - BV\| + \|BV - BV^\pi\| \leq \|V - BV\| + \gamma \|V - V^\pi\|$$

$$\underbrace{\|V - BV + BV - BV^\pi\|}_{\|V - V^\pi\|} \leq \|V - BV\| + \|BV - BV^\pi\|$$

$$\Rightarrow \|V - V^\pi\| \leq \underbrace{\|BV - BV^\pi\|}_{\|V - V^\pi\|} \|V - BV\| + \gamma \|V - V^\pi\|$$

$$\Rightarrow (1-\gamma) \|V - V^\pi\| \leq \|V - BV\| \rightarrow$$

$$\|V - V^\pi\| \leq \frac{\|V - BV\|}{1-\gamma}$$

برای بخش دوم داریم

$$V^\pi = V^*, \quad B^\pi V = BV \implies \text{تساوی خواهد شد نتیجه می‌شود}$$

می‌خواهیم  $\epsilon$  را پیدا کنیم

$$V^\pi(s) \geq V^*(s) - \frac{\epsilon}{1-\gamma}, \quad \epsilon = \|BV - V\|$$

$$B^\pi V(s) = r(s, \pi(s)) + \gamma \sum_{s'} P(s'|s, \pi(s)) V(s')$$

$\pi(s)$  گزینه  $\arg \max$  است

$$B^\pi V(s) = r(s, \pi(s)) + \gamma \sum_{s'} P(s'|s, \pi(s)) V(s') =$$

$$\max_a r(s, a) + \gamma \sum_{s'} P(s'|s, a) V(s') = BV(s)$$

$$\approx \frac{\epsilon}{1-\gamma} = \underbrace{\frac{\|BV - V\|}{1-\gamma}}_{\geq \|V^* - V\|} \geq V^*(s) - V(s)$$

$$\frac{\epsilon}{1-\gamma} = \underbrace{\frac{\|B^\pi V - V\|}{1-\gamma}}_{\geq \|V - V^\pi\|} \geq V(s) - V^\pi(s)$$

از بخش اول می‌توان اینجا

⇒ مجموع می‌کسرد عبارت بالا

$$\frac{\epsilon}{1-\gamma} \geq \underbrace{V^*(s) - V(s)}_{\geq 0} + V(s) - V^\pi(s)$$

$V^*(s) - V^\pi(s) \geq 0$

$$V^\pi(s) \geq V^*(s) - \frac{\epsilon}{1-\gamma} \quad \text{وَهُوَ lower bound} \approx V$$

مقدمة (زاید) حفظ و اثبات  $\boxed{V^\pi \geq V^*}$

رسی داده را در بین حالت بحول می‌دهیم،

$$\frac{\epsilon}{1-\gamma} \text{ میان } V^\pi, V^* \text{ را درین حالت بازی } \approx -1$$

برقراره  $V^\pi \geq V^*$  بمحضه.

$$V^\pi(s) \geq V^*(s) - \frac{\epsilon}{1-\gamma} \quad \text{درستگیره} \quad 9$$

محضی برای هر  $s$   $\pi$  Policy را ساخته می‌کنیم

$$V \geq V^* \geq V^\pi \quad \text{و طبقاً} \quad V^* \geq V^\pi \quad \text{بنسبت}$$

$$\frac{\epsilon}{1-\gamma} = \frac{\|B^\pi V - V\|}{1-\gamma} \geq \|V - V^*\| \geq V(s) - V^*(s)$$

$$, V(s) - V^*(s) \geq V^*(s) - V^*(s)$$

$$\Rightarrow \frac{\epsilon}{1-\gamma} \geq V^*(s) - V^*(s) \rightarrow$$

$$V^\pi(s) \geq V^*(s) - \frac{\epsilon}{1-\gamma}$$

نیک خواص سکل دھمکی ۱۰ -  $V^* \leq V$ ,  $BV \leq V$  ~~بلا رہم~~  
جتنی تعلیم.

$$V \leq V' \rightarrow BV \leq BV' \quad \text{دھمکی بلا رہم}$$

$$\forall V \geq BV \geq B^2V \geq \dots \geq B^nV \quad \text{جیسے داریم} \quad \text{مختصر}$$

$$\Rightarrow B^{n+1}V \leq B^{n+1}V \rightarrow \cancel{B(B^nV) \leq B^nV} \quad B(B^nV) \leq B^nV \leq V$$

$$\lim_{n \rightarrow \infty} B^nV \leq V \Rightarrow V^* \leq V$$

$\underbrace{\dots}_{[1]}$ ,  $\text{fixed point}$  / موقع سرگزاس / موقع  
Point  $V^*$  Point

$$V \geq V^* \quad (\text{راہیں}), \quad V^*(s) \geq V^*(s) - \frac{\gamma \varepsilon}{1-\gamma}, \quad \text{نیک خواص سکل دھمکی} \quad ||$$

$$, \quad V^*(s) \geq V^*(s) - \frac{\gamma \varepsilon}{1-\gamma}$$

~~میری~~,  $B^\pi V^\pi = V^\pi$ ,  $B^\pi V = BV \rightarrow$

$$\|B^\pi V - V^\pi\| = \|B^\pi V - B^\pi V^\pi\| \leq \gamma \|V - V^\pi\| \leq \frac{\gamma \varepsilon}{1-\gamma}$$

$$\text{میری داریم}, \quad \frac{\|BV - \tilde{V}\|}{1-\gamma} = \frac{\|BV' - V'\|}{1-\gamma} \geq \|V^* - V'\| = \|V^* - BV\| = \|V^* - B^\pi V\|$$

$$\frac{\|BV' - V'\|}{1-\gamma} \leq \frac{\gamma \|V' - V\|}{1-\gamma} \leq \frac{\gamma \varepsilon}{1-\gamma} \quad \text{میری داریم}$$

برای ←

$$\|V^* - V^\pi\| = \|V^* - BV + BV - V^\pi\| \leq \|V^* - BV\| + \|BV - V^\pi\|$$

$$\leq \frac{\gamma\varepsilon}{1-\gamma} + \frac{\gamma\varepsilon}{1-\gamma} = \frac{2\gamma\varepsilon}{1-\gamma} \longrightarrow$$

$$\frac{V^*(s) - V^\pi(s)}{V^\pi(s)} \leq \|V^* - V^\pi\| \rightarrow \leq \frac{2\gamma\varepsilon}{1-\gamma} \longrightarrow V^\pi(s) \geq V^*(s) - \frac{2\gamma\varepsilon}{1-\gamma}$$

برای این پس بینه کاری را

$$\frac{\gamma\varepsilon}{1-\gamma} \geq \|BV - V^\pi\| = \|BV - V^*\| \geq BV(s) - V^\pi(s)$$

$$\leftarrow BV \geq BV^* = V^* \iff V^* \leq V \quad \text{برای بقیه}$$

$$\frac{\gamma\varepsilon}{1-\gamma} \geq BV(s) - V^\pi(s) \geq BV^*(s) - V^\pi(s) = V^*(s) - V^\pi(s) \quad \checkmark$$