Reward Shaping

Adds a small remard for each transition

based on domain knowledge to encourage

positive actions (like a heuristic)

Potential Based Reward Slaping

Potential function D(s) Basically a heuristic for

State S larger value mean state is better

 $E(X) = \sum P(X) \cdot X$

(an calculate shaped reward with I

· Using potential functions guarantees aly will

learn optimal policy

· However, no guarantee that it will conveye fouter



Policy Iteration

Evaluates the value of policies directly Improves the policy over time

2 steps during each iteration:

1) policy evaluation

2) policy Improvement

Repeat until policy converges

Policy evaluation

Find the value of each state $V^{T}(s)$ under the

current policy

```
\label{eq:localization} \begin{array}{l} \textbf{Algorithm - Policy evaluation} \\ \textbf{Input:} \ \pi \ \text{the policy for evaluation,} \ V^\pi \ \text{value function, and MDP} \ M = \langle S, s_0, A, P_a(s' \mid s), r(s, a, s') \rangle \\ \textbf{Output:} \ \text{Value function} \ V^\pi \\ \textbf{Repeat} \\ \Delta \leftarrow 0 \\ \textbf{For each} \ s \in S \\ \underbrace{V'^\pi(s) \leftarrow \sum_{s' \in S} P_{\pi(s)}(s' \mid s) \left[ r(s, a, s') + \gamma \ V^\pi(s') \right]}_{Policy \ \text{evaluation equation}} \ V \text{CS} \ \ \textbf{based only on Policy action} \\ \Delta \leftarrow \max(\Delta, |V^{r\pi}(s) - V^\pi(s)|) \\ V^\pi \leftarrow V^{r\pi} \\ \textbf{Until} \ \Delta \leq \theta \ \ \ \textbf{Until} \ \ \textbf{Cenyergence} \end{array}
```

18 clone by hand: Use a table Or solve simultaneous equations

Store Policy
T(So) T(h) T(h)
ao ay az

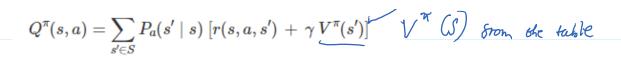
	GUV	V(S)	VCI)
mit	O	\mathcal{O}	0
./	2	~	3 Values on each line depend on
2)	-2	4 values in prev line
3	0.5	-2.7	4.3
4	0.5	-2.7	4.3 Until 2 lines are the same Cor close enough)

Once VTCs) found for all states, more to policy Improvement

Policy Improvement

- 1) Calculate Q(s,a) for all authors a in state s
- 2) set policy for s T(s) = action with highert Q Usay
- 3) Repeat for all states

$$Q^{\pi}(s,a) = \sum_{s' \in S} P_a(s' \mid s) \left[r(s,a,s') \, + \, \gamma \, \underbrace{V^{\pi}(s')}
ight]^{\pi} \, \mathcal{C}$$
 from the table



Rit Together

Algorithm – Policy Iteration

Input: MDP $M = \langle S, s_0, A, P_a(s' \mid s), r(s, a, s')
angle$

Output: Policy π

Until π does not change

Set V^π to arbitrary value function; e.g., $V^\pi(s)=0$ for all s.

Set π to arbitrary policy; e.g. $\pi(s)=a$ for all s, where $a\in A$ is an arbitrary action.

Repeat

 $\begin{aligned} & \text{Compute } V^\pi(s) \text{ for all } s \text{ using policy evaluation} \\ & \text{For each } s \in S \\ & \pi(s) \leftarrow \text{argmax}_{a \in A(s)} Q^\pi(s,a) \end{aligned}$