An Alternative Softmax Operator for Reinforcement Learning

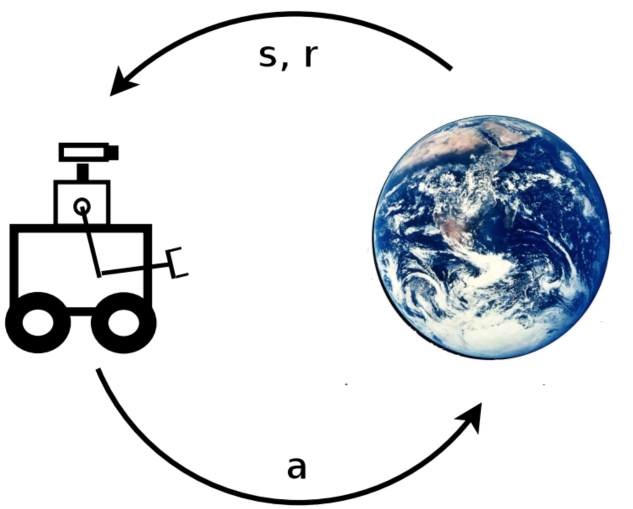
Michael L. Littman

Brown University Department of Computer Science

with Kavosh Asadi, Seungchan Kim, George Konidaris

Reinforcement Learning

Markov Decision Process (MDP): $< S, A, R, \mathcal{T}, \gamma >$



$$\max_{\pi} \mathbb{E}\left[R = \sum_{t=0}^{\infty} \gamma^t r_t\right]$$

What does agent learn?

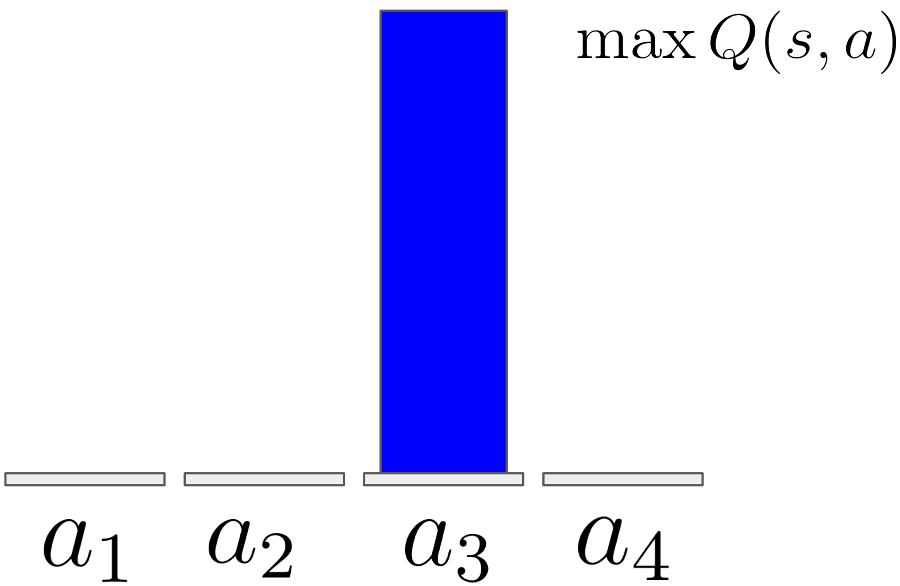
Policy $\pi : S \rightarrow A$

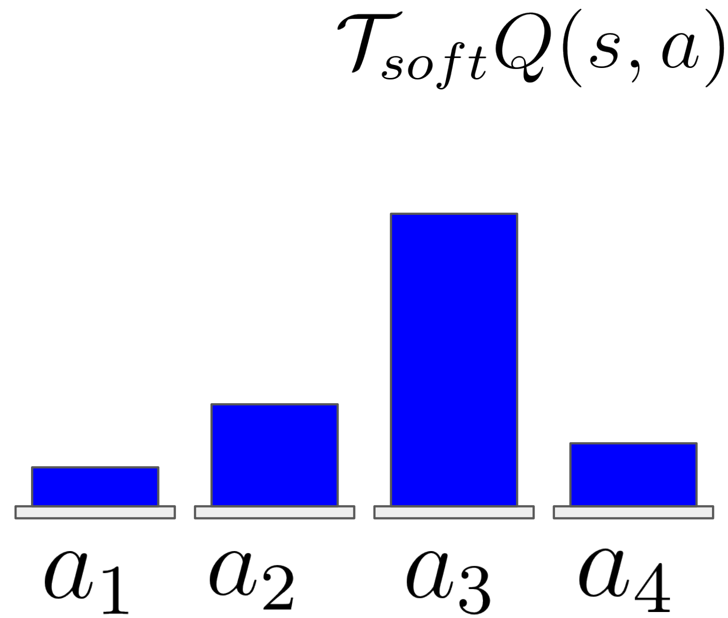
Value Function: $$V_\pi(s) = \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t r_t \middle| \pi, s_0 = s\right]$$

Action-Value Function: $$Q_\pi(s, a) = \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t r_t \middle| \pi, s_0 = s, a_0 = a\right]$$

Transition Function: $$T(s' | s, a)$$

Max vs Softmax: action selection strategy





- Greedy action-selection strategy!
- Optimal actions are chosen.
- No explorations.

- More explorations
- Suboptimal actions can be chosen.
- Less exploitations

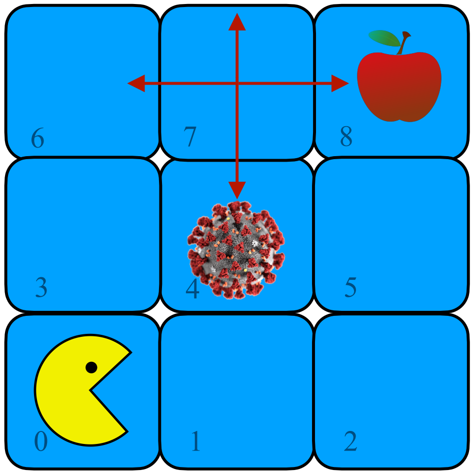
$$\bigotimes : \mathcal{R}^{|\mathcal{A}|} \rightarrow \mathcal{R}$$

$$\max_{a \in \mathcal{A}} Q(s, a)$$

$$\text{mean } Q(s, \cdot)$$

$$\epsilon \text{ mean } Q(s, \cdot) + (1 - \epsilon) \max_{a \in \mathcal{A}} Q(s, a)$$

$$\frac{\sum_{a \in \mathcal{A}} e^{\beta Q(s,a)} Q(s, a)}{\sum_{a \in \mathcal{A}} e^{\beta Q(s,a)}}$$



Generalized Value Iteration

[Littman & Szepesvári, 1996]

$$Q^*(s,a) = R(s,a) + \gamma \int_{s'} T(s'|s,a) \max_{a'} Q^*(s',a') ds'$$

$$Q(s,a) = R(s,a) + \gamma \int_{s'} T(s'|s,a) \bigotimes Q(s',\cdot) ds'$$

initialize $\widehat{Q}_0$, repeat:

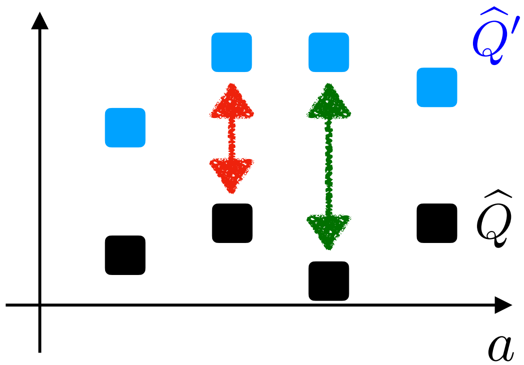
$$\widehat{Q}_{t+1}(s,a) \leftarrow R(s,a) + \gamma \int_{s'} T(s'|s,a) \bigotimes \widehat{Q}_t(s',\cdot) ds'$$

until convergence

convergence if: $$K_{\bigotimes} := \sup_{\widehat{Q},\widehat{Q}'} \frac{|\bigotimes \widehat{Q}(s,\cdot) - \bigotimes \widehat{Q}'(s,\cdot)|}{||\widehat{Q}(s,\cdot) - \widehat{Q}'(s,\cdot)||_\infty} \leq 1$$

Bellman will be a γ-contraction

[Asadi & Littman, 2017]



$\bigotimes = \max_a \widehat{Q}(s,a)$ ✓

$\bigotimes = \text{mean } \widehat{Q}(s,\cdot)$ ✓

$\bigotimes = \text{median } \widehat{Q}(s,\cdot)$ ✓

and their convex combinations

$$\text{boltz}_\beta \ \widehat{Q}(s,\cdot) = \frac{\sum_a e^{\beta \widehat{Q}(s,a)} \widehat{Q}(s,a)}{\sum_a e^{\beta \widehat{Q}(s,a)}}$$ ✗

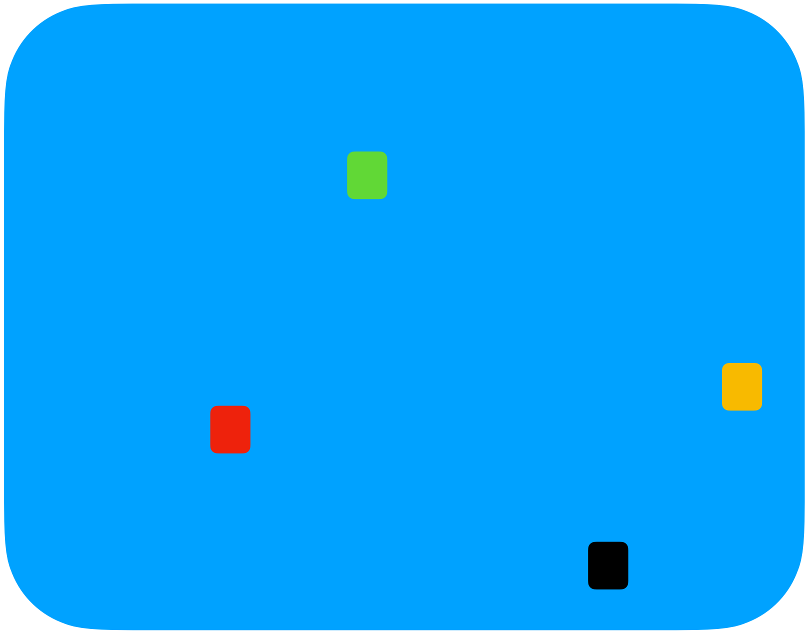
$$\sup_{\widehat{Q},\widehat{Q}'} \frac{|\bigotimes \widehat{Q}(s,\cdot) - \bigotimes \widehat{Q}'(s,\cdot)|}{||\widehat{Q}(s,\cdot) - \widehat{Q}'(s,\cdot)||_\infty} \leq 1$$

$$mm_\omega \ \widehat{Q}(s,\cdot) = \frac{\log \frac{1}{|\mathcal{A}|} \sum_a e^{\omega \widehat{Q}(s,a)}}{\omega}$$ ✓

$$mm_{\omega\to\infty}([1,2,3,4]) = 4$$
$$mm_{\omega=100}([1,2,3,4]) = 3.9861$$
$$mm_{\omega=10}([1,2,3,4]) = 3.8614$$
$$mm_{\omega=2}([1,2,3,4]) = 3.3794$$
$$mm_{\omega=1}([1,2,3,4]) = 3.0539$$
$$mm_{\omega=0}([1,2,3,4]) = 2.5$$

$$\lim_{\omega\to\infty} mm_\omega(\mathbf{x}) = \max(\mathbf{x})$$
$$\lim_{\omega\to 0} mm_\omega(\mathbf{x}) = mean(\mathbf{x})$$

Contraction Mapping



Properties

- Non-Expansion
- Differentiable
- Limits: goes to max, mean, min.
- Policy to achieve the value can be extracted.
- The maximum entropy such policy is Boltzmann (!), with some beta.

$$mm_\omega\ \widehat{Q}(s,\cdot) = \frac{\log \frac{1}{|\mathcal{A}|}\sum_a e^{\omega \widehat{Q}(s,a)}}{\omega}$$

GVI Planning Algorithm

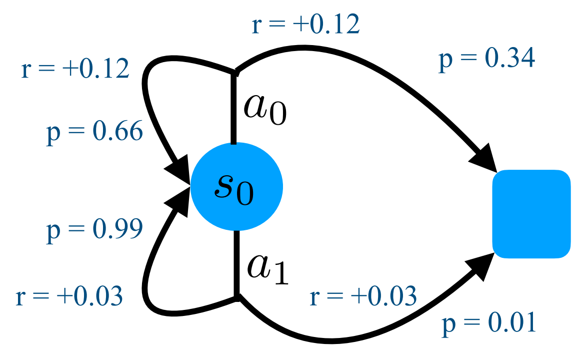
Input: initial $\hat{Q}(s,a)\ \forall s \in \mathcal{S}\ \forall a \in \mathcal{A}$ and $\delta \in \mathcal{R}^+$
repeat
  diff $\leftarrow 0$
  for each $s \in \mathcal{S}$ **do**
    for each $a \in \mathcal{A}$ **do**
      $Q_{copy} \leftarrow \hat{Q}(s,a)$
      $\hat{Q}(s,a) \leftarrow \sum_{s' \in \mathcal{S}} \mathcal{R}(s,a,s')$
        $+ \gamma \mathcal{P}(s,a,s') \bigotimes \hat{Q}(s',.)$
      diff $\leftarrow \max\left\{\text{diff}, |Q_{copy} - \hat{Q}(s,a)|\right\}$
    end for
  end for
until diff $< \delta$

An Example

[Asadi & Littman, 2017]

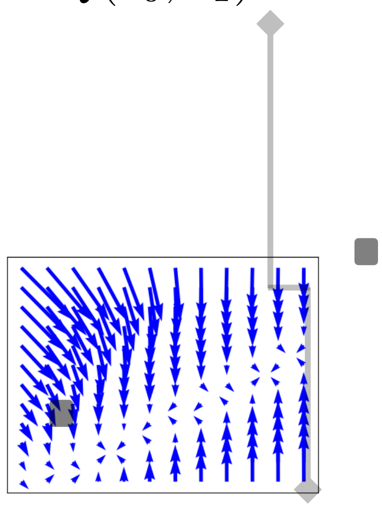
$$\widehat{Q}_{t+1}(s,a) \leftarrow R(s,a) + \gamma \textstyle\sum_{s'} T(s'|s,a) \bigotimes \widehat{Q}_t(s',\cdot)ds'$$

$$\Delta_{t+1} := \widehat{Q}_{t+1}(s_0,\cdot) - \widehat{Q}_t(s_0,\cdot)$$

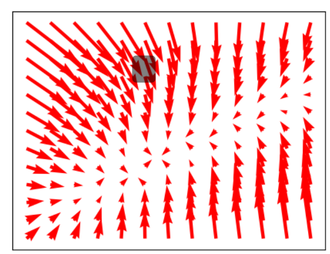


$\widehat{Q}(s_0,a_1)$

$\widehat{Q}(s_0,a_0)$

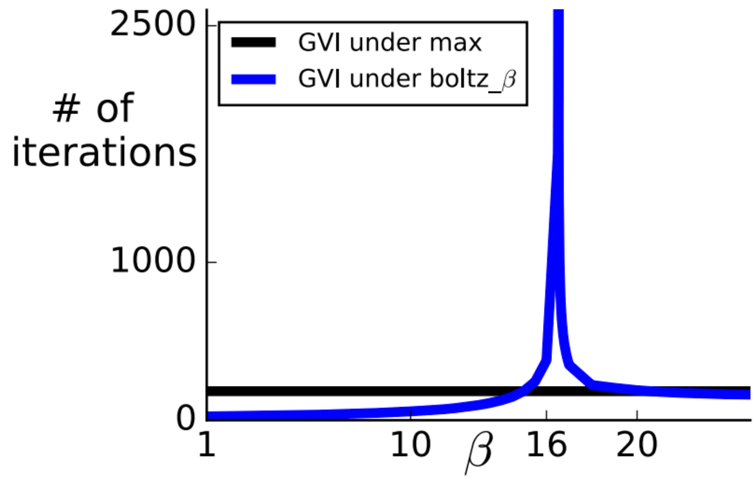


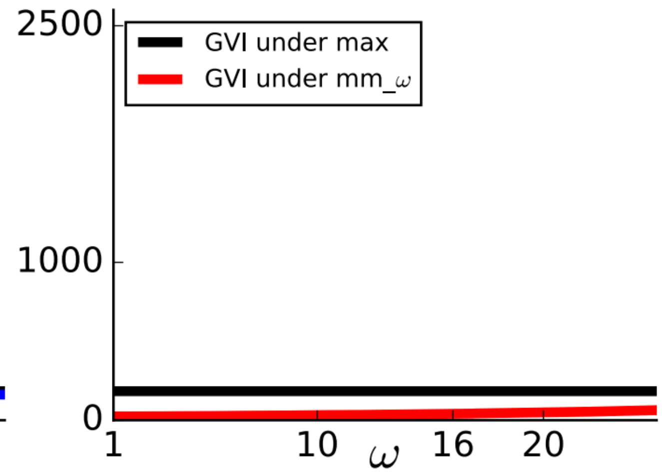
$\bigotimes = \text{boltz}_\beta$

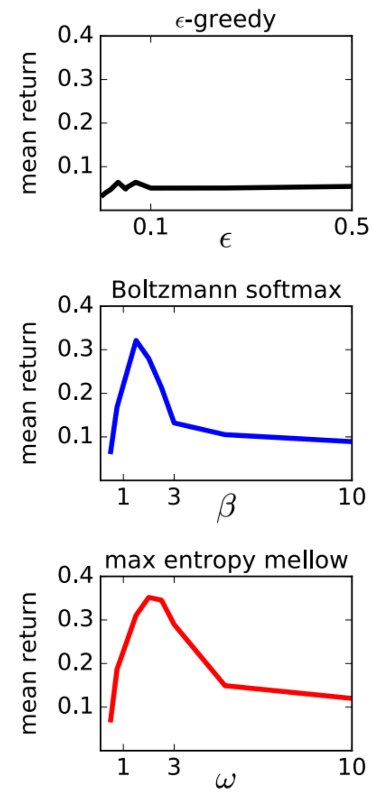


$\bigotimes = \text{mm}_\omega$

Convergence Time







Random MDPs

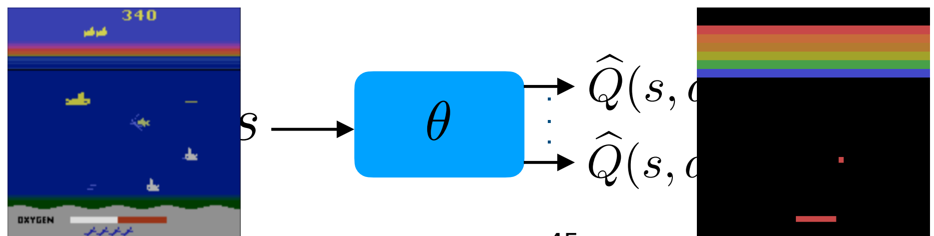
	MDPs, no terminate	MDPs, > 1 fixed points	average iterations
boltz_β	8 of 200	3 of 200	231.65
mm_ω	**0**	**0**	**201.32**

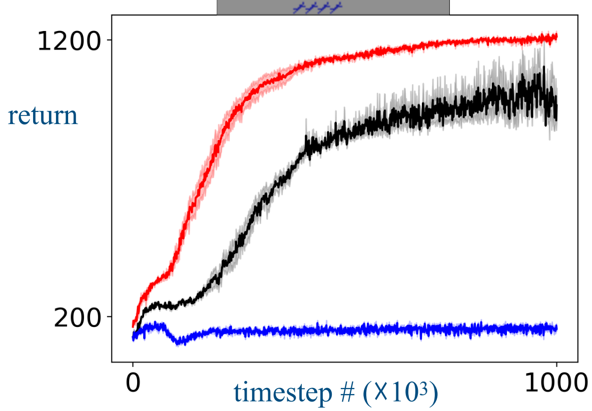
Deep Q-Learning with Mellowmax

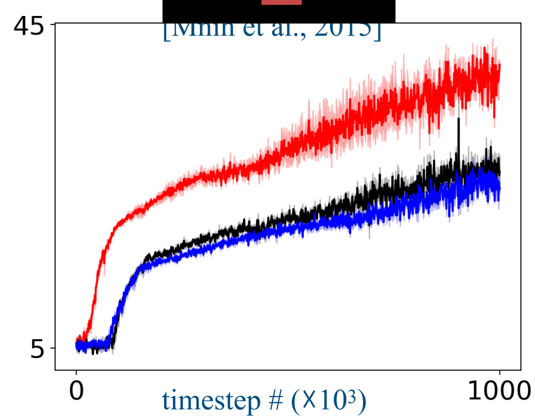
[Kim, Asadi, Konidaris & Littman, 2019]

$$\theta \leftarrow \theta + \alpha\Big(R(s,a) + \gamma \bigotimes \widehat{Q}(s',\cdot;\theta) - \widehat{Q}(s,a;\theta)\Big)\nabla_\theta \widehat{Q}(s,a;\theta)$$

$$\bigotimes = \mathrm{mm}_\omega$$





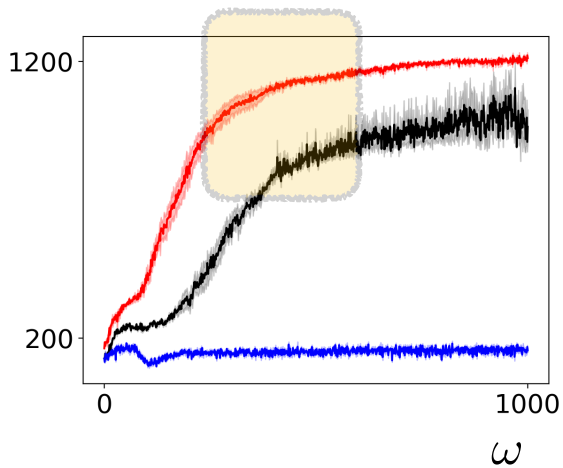


generalized DQN with mm_ω
DQN with target network
DQN no target network

A Regularization Perspective

[Geist et al., 2019]

$$Q(s,a) = R(s,a) + \gamma \int_{s'} T(s'|s,a)\text{mm}_\omega Q(s',\cdot)ds'$$



$$\ldots(\pi(a|s)) + \ln(|\mathcal{A}|)$$

$$\ldots a|s)Q(s,a) - \frac{1}{\omega}\Omega\big(\pi(\cdot|s)\big)\Big)$$

$$\ldots Q(s,a)$$

$$\ldots = mm_\omega\ Q(s,\cdot)$$

Conclusion

- Mellowmax provides an alternative to Boltzmann exploration or epsilon greedy:
  - Better convergence guarantees
  - Rich value-dependent exploration.
  - Useful smoothness behavior.