An Alternative Softmax for RL

Kushagra Chandak

May 31, 2021

Motivation and papers/references

- Motivated by the importance of softmax operators in RL, especially for exploration.
- Papers/References:
  1. An Alternative Softmax Operator for RL. *Asadi et al.* ICML-17
  2. DeepMellow: Removing the Need for a Target Network in Deep Q-Learning. *Kim et al.* IJCAI-19
  3. A Theory of Regularized Markov Decision Processes. *Geist et al.* ICML-19
  4. Leverage the Average: an Analysis of KL Regularization in Reinforcement Learning. *Vieillard et al.* NIPS-20
  5. Simons Institute tallk, UCB: https://simons.berkeley.edu/talks/alternative-softmax-operator-reinforcement-learning

Max vs Softmax

- $\max Q(s, a)$:
  - Greedy action-selection strategy.
  - No exploration.
- If Q values are not learned, then the greedy strategy is not a good idea.
- $\mathcal{T}_{\text{soft}} Q(s, a)$:
  - Probability distribution over action set.
  - More exploration; less exploitation. Suboptimal actions can be chosen.
- Ideal softmax operator:

  P1. Parameter settings that allow for maximisation.
  P2. Non-expansion.*
  P3. Differentiable.
  P4. Avoids starving non-maximizing actions.

Aggregation of Values

- How to aggregate values of a state given a finite action set?
- Define operator over sets of action values. $\bigotimes : \mathcal{R}^{|\mathcal{A}|} \to \mathcal{R}$.
- $\max = \max_{a \in \mathcal{A}} Q(s, a)$ (No P3 and P4)
- $\text{mean} = \text{mean}\, Q(s, .)$ (No P1)
- $\text{eps}_\epsilon = \epsilon \text{mean} Q(s, .) + (1 - \epsilon) \max_{a \in \mathcal{A}} Q(s, a)$ (No P3)
- $\text{boltz}_\beta = \frac{\sum_a e^{\beta Q(s,a)} Q(s,a)}{\sum_a e^{\beta Q(s,a)}}$ (No P2)

Generalized Value Iteration

- Generalized Bellman Equation:
  $Q(s,a) = R(s,a) + \gamma \int_{s'} T(s'|s,a) \bigotimes Q(s',.) ds'$

Generalized Value Iteration

- Generalized Bellman Equation:
  $Q(s,a) = R(s,a) + \gamma \int_{s'} T(s'|s,a) \bigotimes Q(s',.)ds'$
- Convergence if

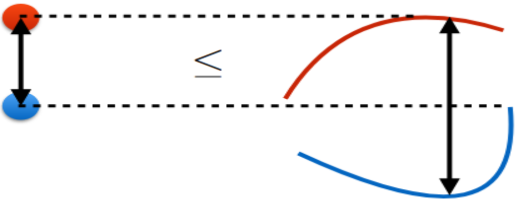
$$K_{\bigotimes} = \sup_{Q,Q'} \frac{|\bigotimes Q(s,.) - \bigotimes Q'(s,.)|}{\|Q(s,.) - Q'(s,.)\|_\infty} \leq 1$$

Generalized Value Iteration

- Generalized Bellman Equation:
  $Q(s,a) = R(s,a) + \gamma \int_{s'} T(s'|s,a) \bigotimes Q(s',.)ds'$
- Convergence if

$$K_{\bigotimes} = \sup_{Q,Q'} \frac{|\bigotimes Q(s,.) - \bigotimes Q'(s,.)|}{\|Q(s,.) - Q'(s,.)\|_\infty} \leq 1$$

- Non-expansion or the Lipschitz property. Puts upper-bound on the aggregation operator. Bellman operator is a γ-contraction.



$$\left|\max_a Q_1(s,a) - \max_a Q_2(s,a)\right| \leq \max_a \left|Q_1(s,a) - Q_2(s,a)\right|$$

Operators and non-expansion

- $\bigotimes = \max_a Q(s,a)$ ✓

Operators and non-expansion

- $\bigotimes = \max_a Q(s, a)$ ✓
- $\bigotimes = \text{mean } Q(s, .)$ ✓

Operators and non-expansion

- $\bigotimes = \max_a Q(s,a)$ $\checkmark$
- $\bigotimes = \text{mean}\, Q(s,.)$ $\checkmark$
- $\bigotimes = \text{median} Q(s,.)$ $\checkmark$

Operators and non-expansion

- $\bigotimes = \max_a Q(s, a)$ ✓
- $\bigotimes = \text{mean } Q(s, .)$ ✓
- $\bigotimes = \text{median} Q(s, .)$ ✓
- Convex combinations of the above.

Operators and non-expansion

- $\bigotimes = \max_a Q(s,a)$ ✓
- $\bigotimes = \text{mean}\ Q(s,.)$ ✓
- $\bigotimes = \text{median} Q(s,.)$ ✓
- Convex combinations of the above.
- $\text{boltz}_\beta\ Q(s,.)$ ×

Operators and non-expansion

- $\bigotimes = \max_a Q(s, a)$ ✓
- $\bigotimes = \text{mean } Q(s, .)$ ✓
- $\bigotimes = \text{median} Q(s, .)$ ✓
- Convex combinations of the above.
- $\text{boltz}_\beta\, Q(s, .)$ ×
- $\text{mm}_\omega\, Q(s, .) = \frac{\log \frac{1}{\mathcal{A}} \sum_a e^{\omega Q(s,a)}}{\omega}$ ✓

Mellowmax

- Properties of $\mathrm{mm}_\omega\, Q(s,.) = \frac{\log \frac{1}{\mathcal{A}} \sum_a e^{\omega Q(s,a)}}{\omega}$

Mellowmax

- Properties of $\mathrm{mm}_\omega\, Q(s,.) = \frac{\log \frac{1}{\mathcal{A}} \sum_a e^{\omega Q(s,a)}}{\omega}$
- $\lim_{\omega \to \infty} \mathrm{mm}_\omega(x) = \max(x)$

Mellowmax

- Properties of $\mathrm{mm}_\omega Q(s,.) = \frac{\log \frac{1}{\mathcal{A}} \sum_a e^{\omega Q(s,a)}}{\omega}$
- $\lim_{\omega \to \infty} \mathrm{mm}_\omega(x) = \max(x)$
- $\lim_{\omega \to 0} \mathrm{mm}_\omega(x) = \mathrm{mean}(x)$

Mellowmax

- Properties of $\text{mm}_\omega\, Q(s,.) = \frac{\log \frac{1}{\mathcal{A}} \sum_a e^{\omega Q(s,a)}}{\omega}$
- $\lim_{\omega \to \infty} \text{mm}_\omega(x) = \max(x)$
- $\lim_{\omega \to 0} \text{mm}_\omega(x) = \text{mean}(x)$
- $\lim_{\omega \to -\infty} \text{mm}_\omega(x) = \min(x)$

Mellowmax

- Properties of $\mathrm{mm}_\omega\, Q(s,.) = \frac{\log \frac{1}{\mathcal{A}} \sum_a e^{\omega Q(s,a)}}{\omega}$
- $\lim_{\omega\to\infty} \mathrm{mm}_\omega(x) = \max(x)$
- $\lim_{\omega\to 0} \mathrm{mm}_\omega(x) = \mathrm{mean}(x)$
- $\lim_{\omega\to -\infty} \mathrm{mm}_\omega(x) = \min(x)$
- mm is a contraction mapping, i.e., it's a non-expansion. (unlike boltz)

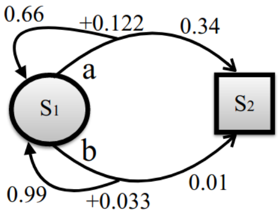
Mellowmax

- Properties of $\mathrm{mm}_\omega \, Q(s,.) = \frac{\log \frac{1}{\mathcal{A}} \sum_a e^{\omega Q(s,a)}}{\omega}$
- $\lim_{\omega \to \infty} \mathrm{mm}_\omega(x) = \max(x)$
- $\lim_{\omega \to 0} \mathrm{mm}_\omega(x) = \mathrm{mean}(x)$
- $\lim_{\omega \to -\infty} \mathrm{mm}_\omega(x) = \min(x)$
- mm is a contraction mapping, i.e., it's a non-expansion. (unlike boltz)
- Differentiable. (unlike eps_ϵ)

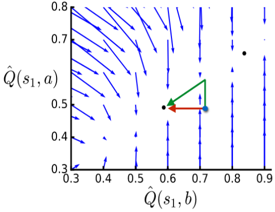
Mellowmax

- Properties of $\text{mm}_\omega\, Q(s, .) = \frac{\log \frac{1}{\mathcal{A}} \sum_a e^{\omega Q(s,a)}}{\omega}$
- $\lim_{\omega \to \infty} \text{mm}_\omega(x) = \max(x)$
- $\lim_{\omega \to 0} \text{mm}_\omega(x) = \text{mean}(x)$
- $\lim_{\omega \to -\infty} \text{mm}_\omega(x) = \min(x)$
- mm is a contraction mapping, i.e., it's a non-expansion. (unlike boltz)
- Differentiable. (unlike eps_ϵ)
- The max entropy mm policy is Boltzmann, with some β.

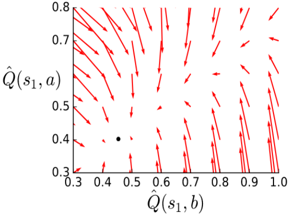
Example



0.66
+0.122
0.34
a
S1
S2
b
0.99
+0.033
0.01



0.8
0.7
0.5
0.4
0.3
0.3 0.4 0.5 0.6 0.7 0.8 0.9
$\hat{Q}(s_1, a)$
$\hat{Q}(s_1, b)$



0.8
0.7
0.5
0.4
0.3
0.3 0.4 0.5 0.6 0.7 0.8 0.9 1.0
$\hat{Q}(s_1, a)$
$\hat{Q}(s_1, b)$

More problems with boltz

- SARSA with Boltzmann softmax policy.

More problems with boltz

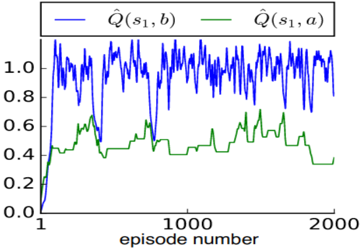
- SARSA with Boltzmann softmax policy.
- Known to converge in tabular setting with decreasing epsilon-greedy exploration.

More problems with boltz

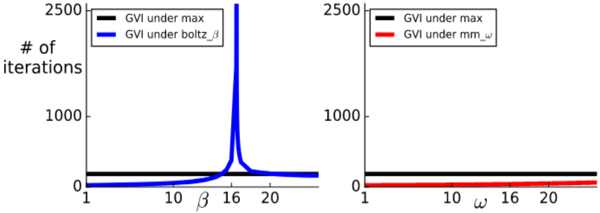
- SARSA with Boltzmann softmax policy.
- Known to converge in tabular setting with decreasing epsilon-greedy exploration.
- Also, to a region in the function approximation setting.

More problems with boltz

- SARSA with Boltzmann softmax policy.
- Known to converge in tabular setting with decreasing epsilon-greedy exploration.
- Also, to a region in the function approximation setting.
- First example to show that SARSA fails to converge in the tabular setting with Boltzmann policy: unstable value estimates.



Convergence Time



	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**

Max entropy mm policy: Boltzmann softmax

$$\pi_{\mathrm{mm}}(s) = \operatorname*{argmin}_{\pi} \sum_{a \in \mathcal{A}} \pi(a|s) \log\big(\pi(a|s)\big) \tag{2}$$

$$\text{subject to} \begin{cases} \sum_{a \in \mathcal{A}} \pi(a|s)\hat{Q}(s,a) = \mathrm{mm}_\omega(\hat{Q}(s,.)) \\ \pi(a|s) \geq 0 \\ \sum_{a \in \mathcal{A}} \pi(a|s) = 1\,. \end{cases}$$

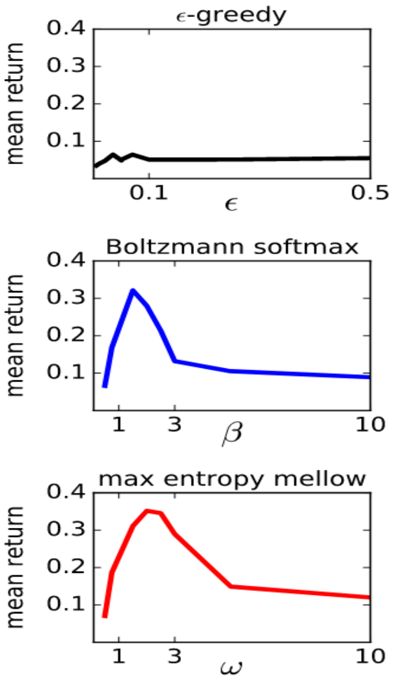
$$\pi_{mm}(a|s) = \frac{e^{\beta \hat{Q}(s,a)}}{\sum_{a \in \mathcal{A}} e^{\beta \hat{Q}(s,a)}} \quad \forall a \in \mathcal{A}\,,$$

where β is a value for which:

$$\sum_{a \in \mathcal{A}} e^{\beta\left(\hat{Q}(s,a) - \mathrm{mm}_\omega \hat{Q}(s,.)\right)} \left(\hat{Q}(s,a) - \mathrm{mm}_\omega \hat{Q}(s,.)\right) = 0$$

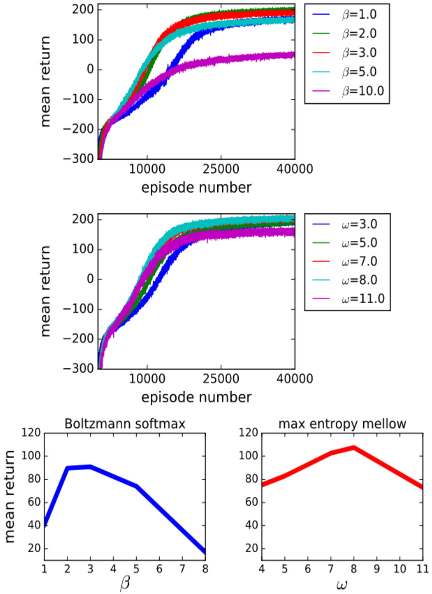
Experiment: Multi-passenger taxi

- Evaluated SARSA with epsilon-greedy, Boltzmann softmax, max entropy mm.
- Challenge: Many locally optimal policies.
- Exploration is important: need to set carefully to avoid over or under exploration.
- Epsilon-greedy performs poorly.
- Boltzmann softmax and max entropy mm achieved significantly higher avg reward.
- Conclusion: Greater stability doesn't mean less effective exploration.



Experiment: Lunar Lander

- Evaluated REINFORCE
- Max entropy mm for the last layer of the neural net policy.
- Continuous state space with 8 dimensions.
- Four discrete actions.
- Reward is +100 for landing in the designated area; -100 otherwise.
- Solving the domain is defined as maintaining mean episode return higher than200 in 100 consecutive episodes.



Deep Q-learning with mm

- Motivation: Removing target network. (towards online RL)

Deep Q-learning with mm

- Motivation: Removing target network. (towards online RL)
  - Remove delays in the update of value functions – faster learning.

Deep Q-learning with mm

- Motivation: Removing target network. (towards online RL)
  - Remove delays in the update of value functions – faster learning.
  - Better allocation of memory resources.

Deep Q-learning with mm

- Motivation: Removing target network. (towards online RL)
  - Remove delays in the update of value functions – faster learning.
  - Better allocation of memory resources.
  - Simpler algorithm.

Deep Q-learning with mm

- Motivation: Removing target network. (towards online RL)
  - Remove delays in the update of value functions – faster learning.
  - Better allocation of memory resources.
  - Simpler algorithm.
- Extra properties.

Deep Q-learning with mm

- Motivation: Removing target network. (towards online RL)
  - Remove delays in the update of value functions – faster learning.
  - Better allocation of memory resources.
  - Simpler algorithm.
- Extra properties.
  - Convexity.

Deep Q-learning with mm

- Motivation: Removing target network. (towards online RL)
  - Remove delays in the update of value functions – faster learning.
  - Better allocation of memory resources.
  - Simpler algorithm.
- Extra properties.
  - Convexity.
  - Monotonic non-decrease.

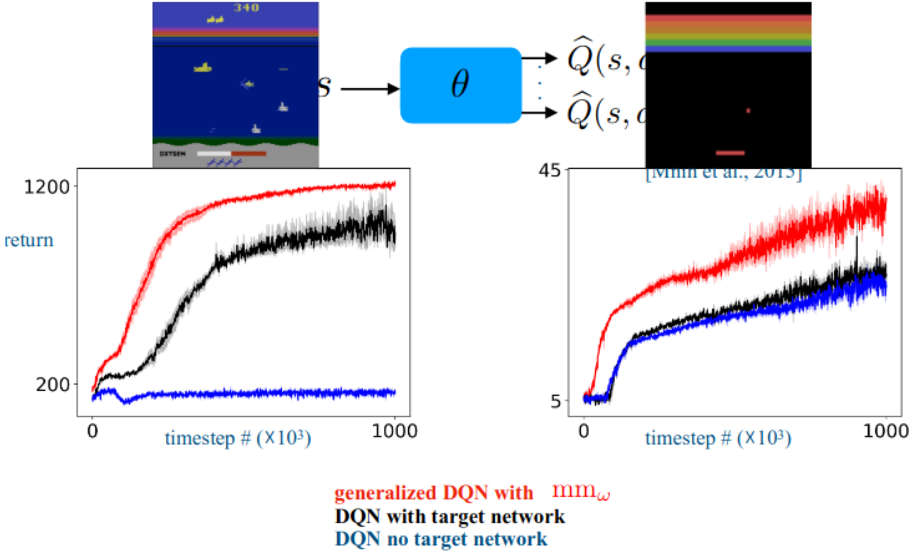
Deep Q-learning with mm

- Motivation: Removing target network. (towards online RL)
  - Remove delays in the update of value functions – faster learning.
  - Better allocation of memory resources.
  - Simpler algorithm.
- Extra properties.
  - Convexity.
  - Monotonic non-decrease.
- In DQN with mm: Alleviation of overestimation.

Deep Q-learning with mm

- Motivation: Removing target network. (towards online RL)
  - Remove delays in the update of value functions – faster learning.
  - Better allocation of memory resources.
  - Simpler algorithm.
- Extra properties.
  - Convexity.
  - Monotonic non-decrease.
- In DQN with mm: Alleviation of overestimation.
- DeepMellow: max $\rightarrow$ mm; No target network.

Mellowmax experiments



Regularization Perspective

- Regularization term: $\Omega(\pi(.|s)) = \pi(a|s) \ln \pi(a|s) + \ln |\mathcal{A}|$ $(\mathrm{KL}(\pi_s||\mathcal{U}))$

Regularization Perspective

- Regularization term: $\Omega(\pi(.|s)) = \pi(a|s) \ln \pi(a|s) + \ln |\mathcal{A}|$ $(\mathrm{KL}(\pi_s || \mathcal{U}))$
- Optimization problem:

$$\max_{\pi} \sum_{a} \pi(a|s) Q(s,a) - \frac{1}{\omega} \Omega(\pi(.|s))$$

$$= \frac{\ln \frac{1}{|\mathcal{A}|} \sum_a e^{\omega Q(s,a)}}{\omega} = \mathrm{mm}_\omega \, Q(s,.)$$

Regularization Perspective

- Regularization term: $\Omega(\pi(.|s)) = \pi(a|s)\ln\pi(a|s) + \ln|\mathcal{A}|$ $(\mathrm{KL}(\pi_s||\mathcal{U}))$
- Optimization problem:

$$\max_{\pi} \sum_{a} \pi(a|s) Q(s,a) - \frac{1}{\omega}\Omega(\pi(.|s))$$

$$= \frac{\ln \frac{1}{|\mathcal{A}|}\sum_a e^{\omega Q(s,a)}}{\omega} = \mathrm{mm}_\omega\, Q(s,.)$$

- mm is the convex-conjugate of Ω.

Regularization Perspective

- Regularization term: $\Omega(\pi(.|s)) = \pi(a|s)\ln \pi(a|s) + \ln|\mathcal{A}|$ $(\mathrm{KL}(\pi_s||\mathcal{U}))$
- Optimization problem:

$$\max_{\pi} \sum_{a} \pi(a|s)Q(s,a) - \frac{1}{\omega}\Omega(\pi(.|s))$$

$$= \frac{\ln \frac{1}{|\mathcal{A}|}\sum_a e^{\omega Q(s,a)}}{\omega} = \mathrm{mm}_\omega\, Q(s,.)$$

- mm is the convex-conjugate of Ω.
- In general, regularizing the evaluation step in MPI algorithms useful and never detrimental.

Conclusion

- An alternative to Boltzmann exploration and epsilon-greedy.

Conclusion

- An alternative to Boltzmann exploration and epsilon-greedy.
- Better convergence guarantees.

Conclusion

- An alternative to Boltzmann exploration and epsilon-greedy.
- Better convergence guarantees.
- Useful smoothness behaviour (stable algorithms).

Conclusion

- An alternative to Boltzmann exploration and epsilon-greedy.
- Better convergence guarantees.
- Useful smoothness behaviour (stable algorithms).
- Rich value-dependent exploration (like Boltzmann).

Backup

- Strongly convex function: Quadratic lower bound on the growth of the function.
- Convex-conjugate of $\Omega(\pi_s)$ (strongly convex): $\forall q_s \in \mathrm{R}^{\mathcal{A}}, \Omega^*(q_s) = \max_{\pi_s} \langle \pi_s, q_s \rangle - \Omega(\pi_s)$
- KL regularization is beneficial: Improved performance.
  - Strong performance bound: Linear dependency to the horizon and averaging of the estimation errors.

Backup: Boltzmann SARSA

Input: initial $\hat{Q}(s,a)\ \forall s \in \mathcal{S}\ \forall a \in \mathcal{A}$, α, and β
for each episode **do**
  Initialize s
  $a \sim$ Boltzmann with parameter β
  repeat
    Take action a, observe r, s'
    $a^{'} \sim$ Boltzmann with parameter β
    $\hat{Q}(s,a) \leftarrow \hat{Q}(s,a) + \alpha\Big[r + \gamma\hat{Q}(s',a') - \hat{Q}(s,a)\Big]$
    $s \leftarrow s^{'}, a \leftarrow a^{'}$
  until s is terminal
end for

Backup: GVI

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 $< \delta$