Replication: An Alternative Softmax Operator for Reinforcement Learning

組員: 彭沛鈞 0710022
黃柏維 0716084
詹凱傑 0716049
廖唯辰 0716092

Outline

- Introduction
  - Ideal softmax operator
- Boltzmann softmax
- Mellowmax softmax
  - Mellowmax's Properties
  - Mellowmax Policy
- Experiment
  - Handcrafted Simple MDP
  - Random MDPs
  - Taxi Domain
  - Lunar Lander Domain
- Conclusion

Introduction

- The issue is about decision making between the action that has highest expected reward and avoiding starving the other actions.

- In Reinforcement Learning, we often use the softmax operators for value-function optimization and softmax poicies for action selection.

Ideal softmax operator:

1. approximate maximization
2. non-expansion, convergence to a unique fixed point
3. differentiable
4. avoids starvation

Common operator:

Let $X = x_1, x_2, ..., x_n$ ***,then we define***

1. $max(X) = max_{i \in \{1,2,...,n\}}\ x_i$

2. $mean(X) = \dfrac{1}{n}\displaystyle\sum_{i=1}^{n} x_i$

3. $eps_\epsilon(X) = \epsilon\ mean(X) + (1-\epsilon)max(X)$

4. $boltz_\beta(X) = \dfrac{\sum_{i=1}^{n} x_i e^{\beta x_i}}{\sum_{i=1}^{n} e^{\beta x_i}}$

Boltzmann softmax

- $$boltz_\beta(X) = \frac{\sum_{i=1}^{n} x_i e^{\beta x_i}}{\sum_{i=1}^{n} e^{\beta x_i}}$$
- approximates max as $\boldsymbol{\beta} \rightarrow \infty$, mean as $\boldsymbol{\beta} \rightarrow 0$
- differentiable
- not a non-expansion operator

Mellowmax softmax

$$mm_\omega(X) = \frac{\log(\frac{1}{n}\sum_{i=1}^{n} e^{\omega x_i})}{\omega}$$

Mellowmax's Properties

- **Non-Expansion**

$$|mm_\omega(X) - mm_\omega(Y)| <= max_i \ |x_i - y_i|$$

- **Maximization**

$$\lim_{\omega \to \infty} mm_\omega(X) = max(X)$$

- **Derivatives**

$$\frac{\partial mm_\omega(X)}{\partial x_i} = \frac{e^{\omega x_i}}{\sum_{i=1}^{n} e^{\omega x_i}}$$

- **Averaging**

$$\lim_{\omega \to 0} mm_\omega(X) = mean(X)$$

Mellowmax Policy

Define the maximum entropy mellowmax policy of a state s as:

$$\pi_{mm}(s) = argmin_{\pi} \sum_{a \in A} \pi(a|s)\log(\pi(a|s))$$

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

The probability of taking an action under the maximum entropy mellowmax policy has the form:

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

,where $\boldsymbol{\beta}$ is a value for which:

$$\sum_{a \in A} e^{\beta(Q(s,a) - mm_{\omega} Q(s,.))}(Q(s,a) - mm_{\omega} Q(s,.)) = 0$$

Experiment

Handcrafted Simple MDP

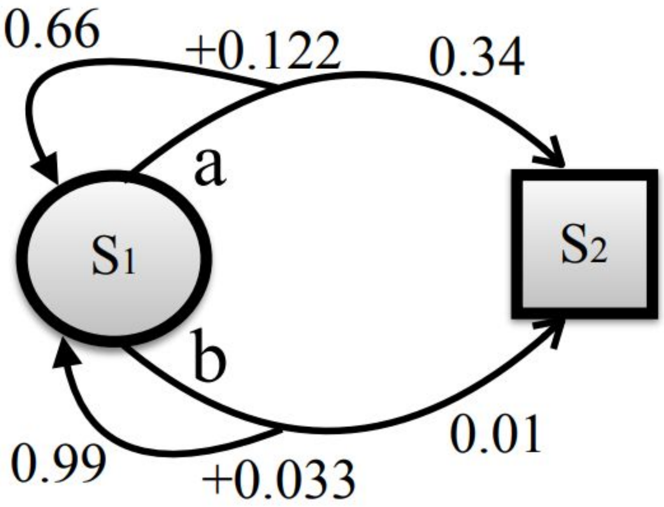


Figure 1. A simple MDP with two states, two actions, and $\gamma = 0.98$. The use of a Boltzmann softmax policy is not sound in this simple domain.

Handcrafted Simple MDP - SARSA

Paper

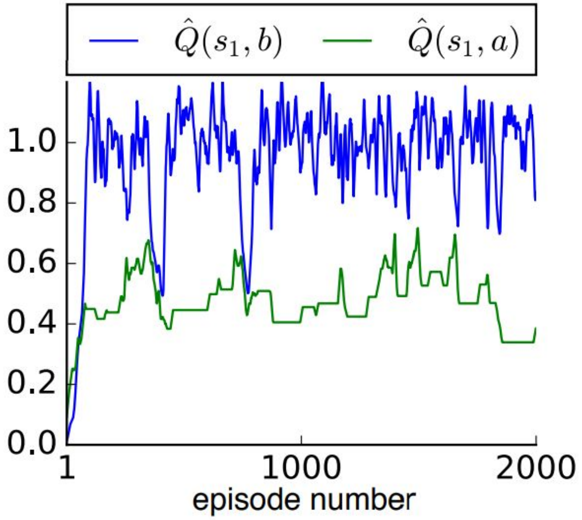
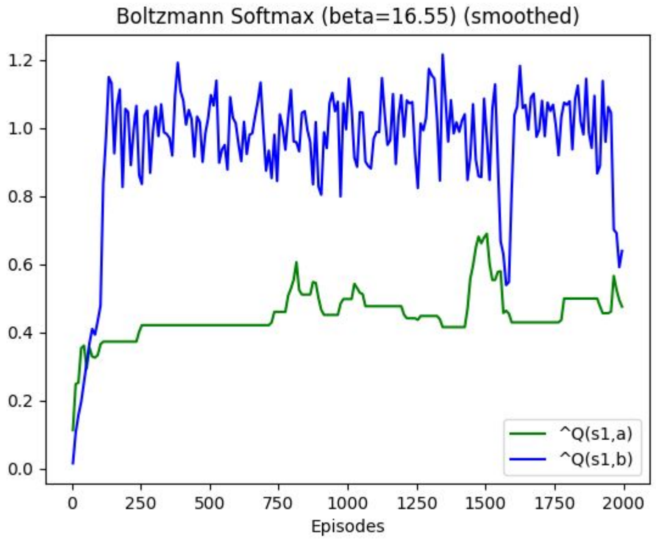


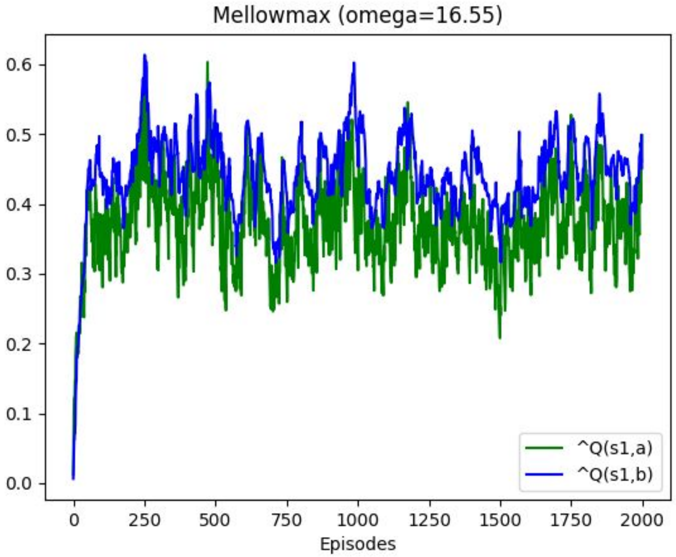
Figure 2. Values estimated by SARSA with Boltzmann softmax. The algorithm never achieves stable values.

Replication

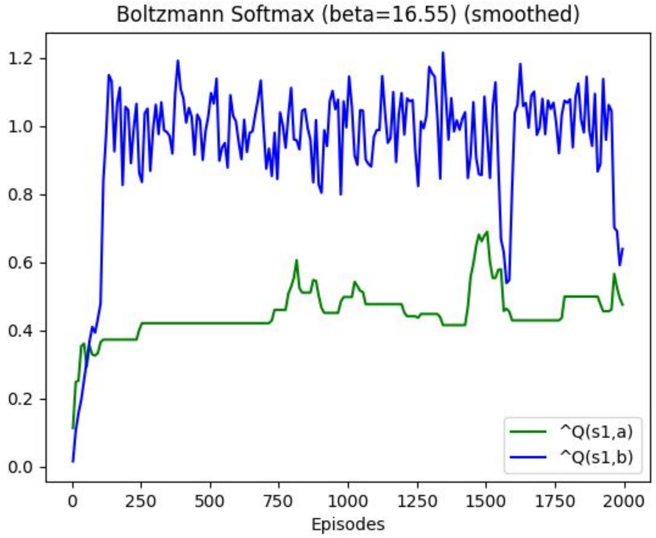


Handcrafted Simple MDP - SARSA

Replication



Replication



Handcrafted Simple MDP - Generalized Value Iteration (GVI)

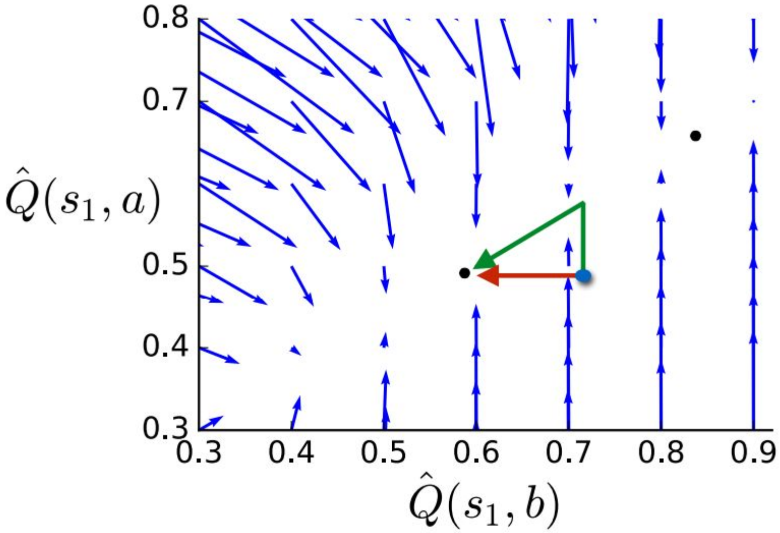
$$\mathop{\mathbb{E}}_{\pi}\left[r + \gamma \hat{Q}(s', a') \middle| s, a\right] =$$

$$\mathcal{R}(s, a) + \gamma \sum_{s' \in \mathcal{S}} \mathcal{P}(s, a, s') \underbrace{\sum_{a' \in \mathcal{A}} \pi(a'|s') \hat{Q}(s', a')}_{\mathrm{boltz}_{\beta}\left(\hat{Q}(s', \cdot)\right)} .$$

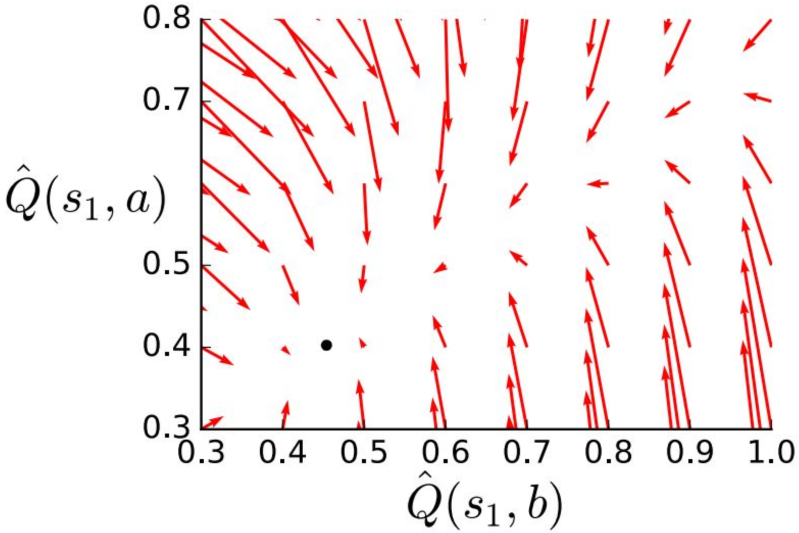
This matches the GVI update (1) when $\bigotimes = \mathrm{boltz}_{\beta}$.

Handcrafted Simple MDP - Generalized Value Iteration (GVI)

Boltzmann Softmax (paper)

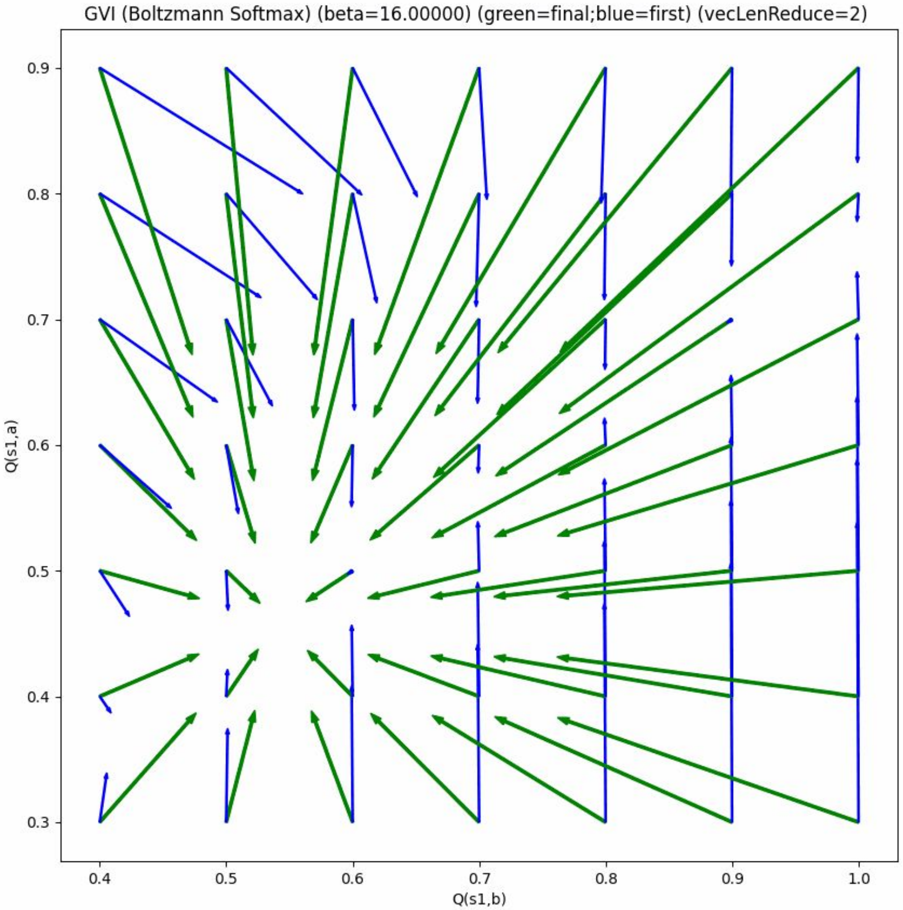


Mellowmax (paper)

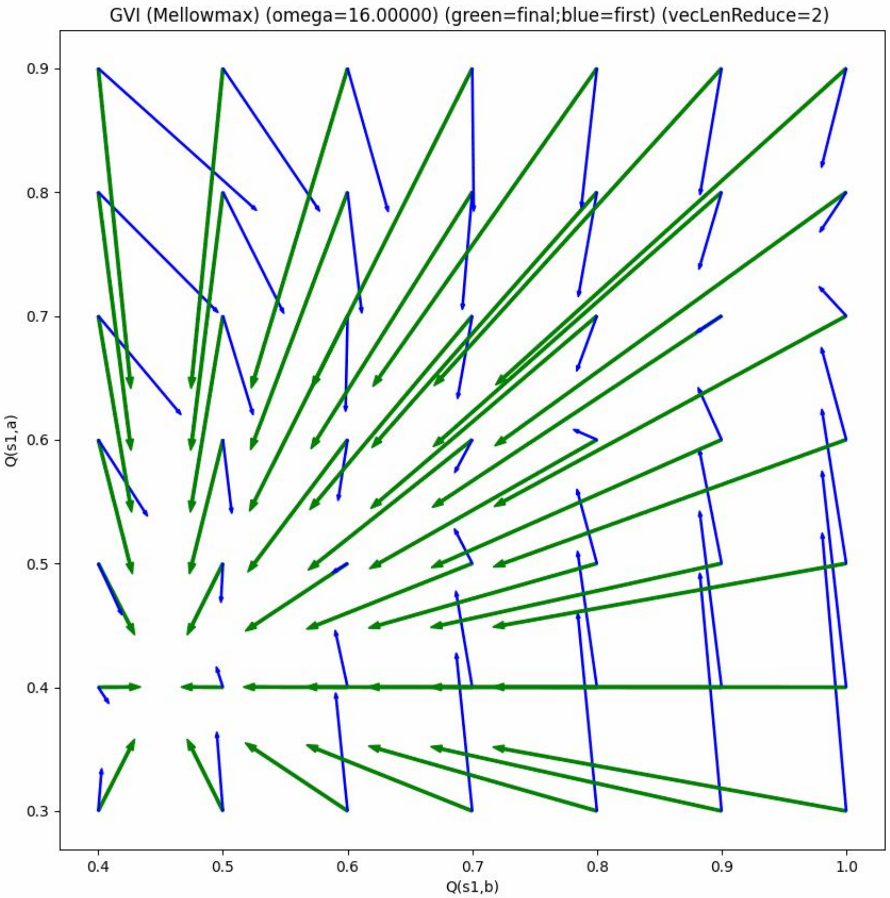


Boltzmann Softmax's has > 1 convergence points!

Boltzmann Softmax (Replication)



Mellowmax (Replication)



Handcrafted Simple MDP - Generalized Value Iteration (GVI) - Vector Fields

(paper)

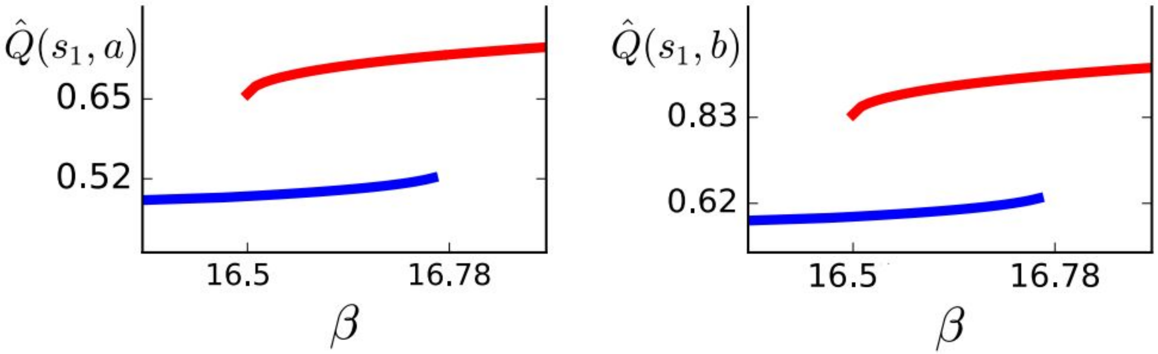
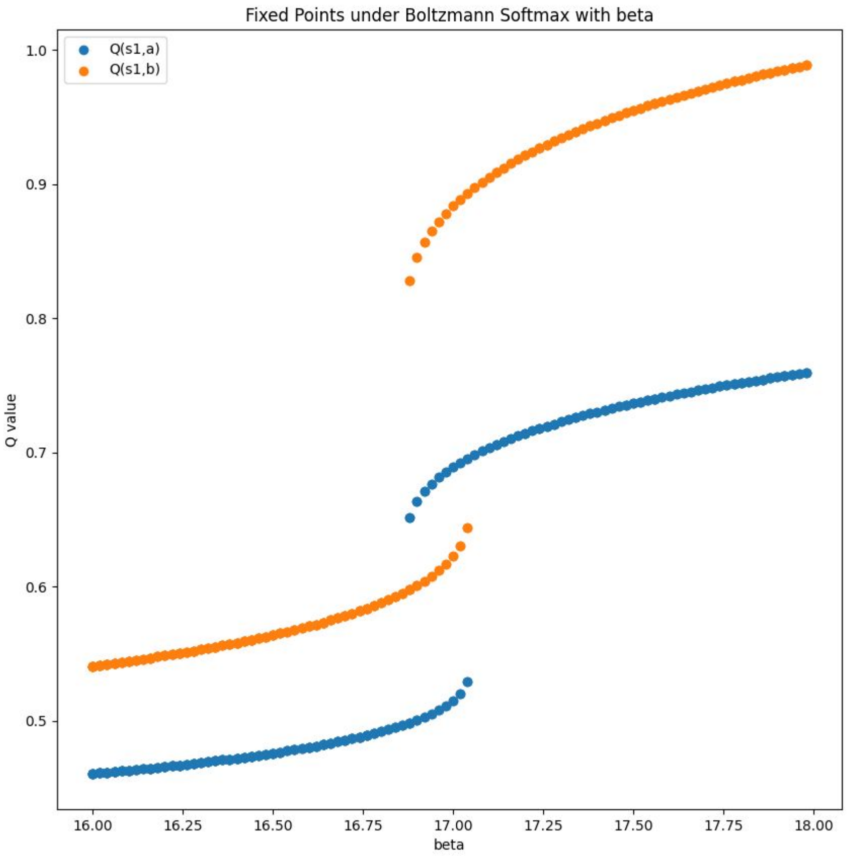
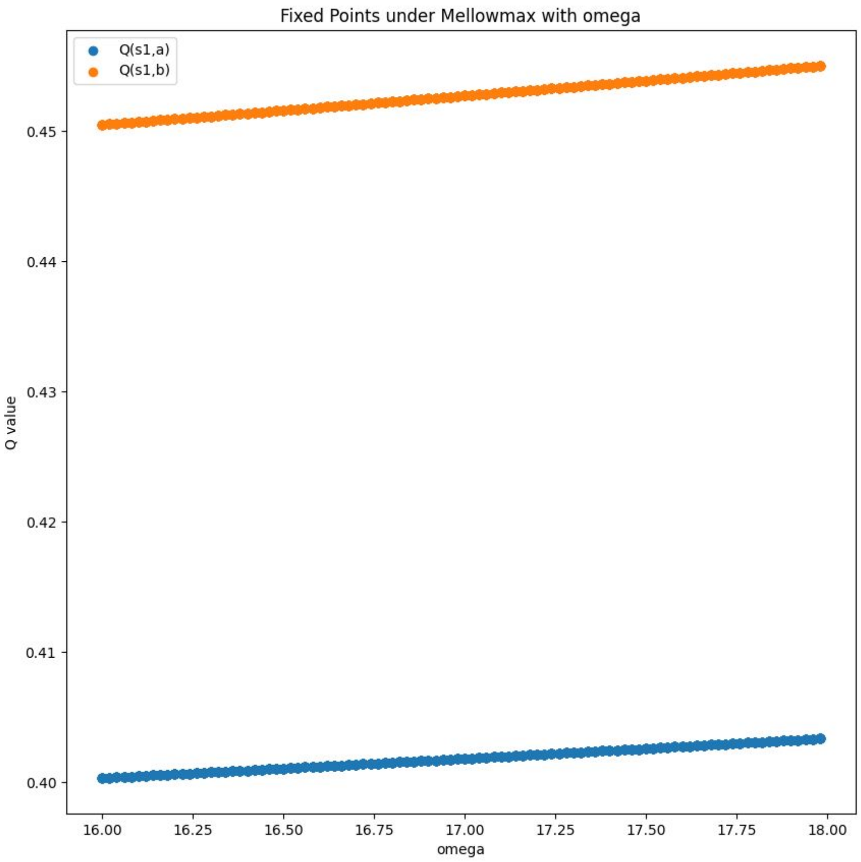


Figure 4. Fixed points of GVI under boltz$_\beta$ for varying β. Two distinct fixed points (red and blue) co-exist for a range of β.

Boltzmann Softmax(Replication)



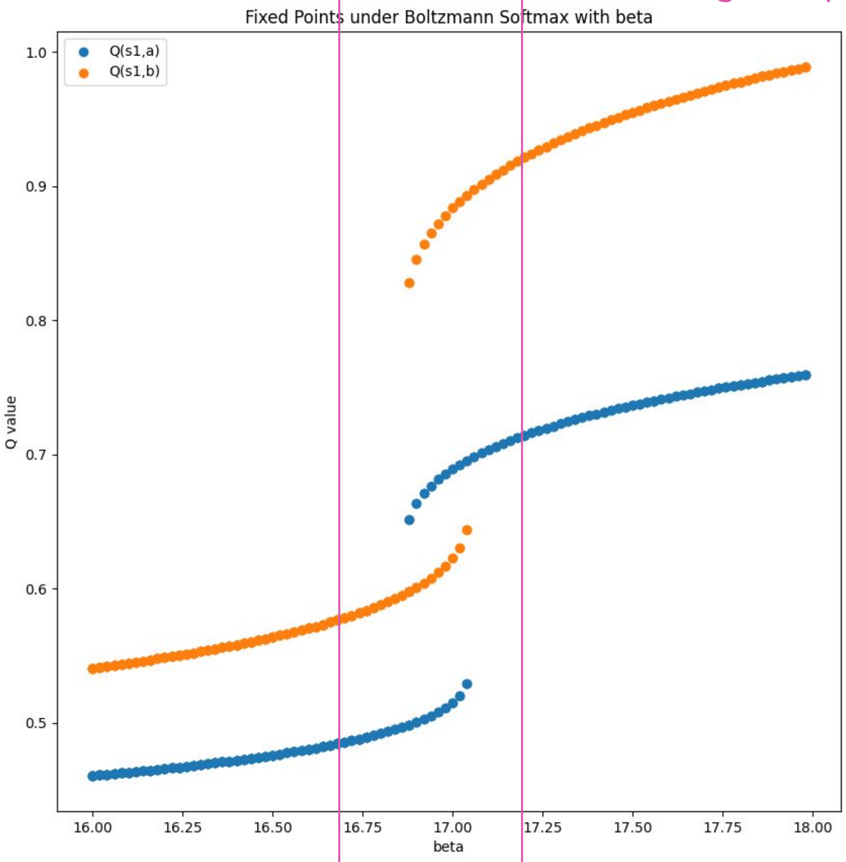
Mellowmax (Replication)

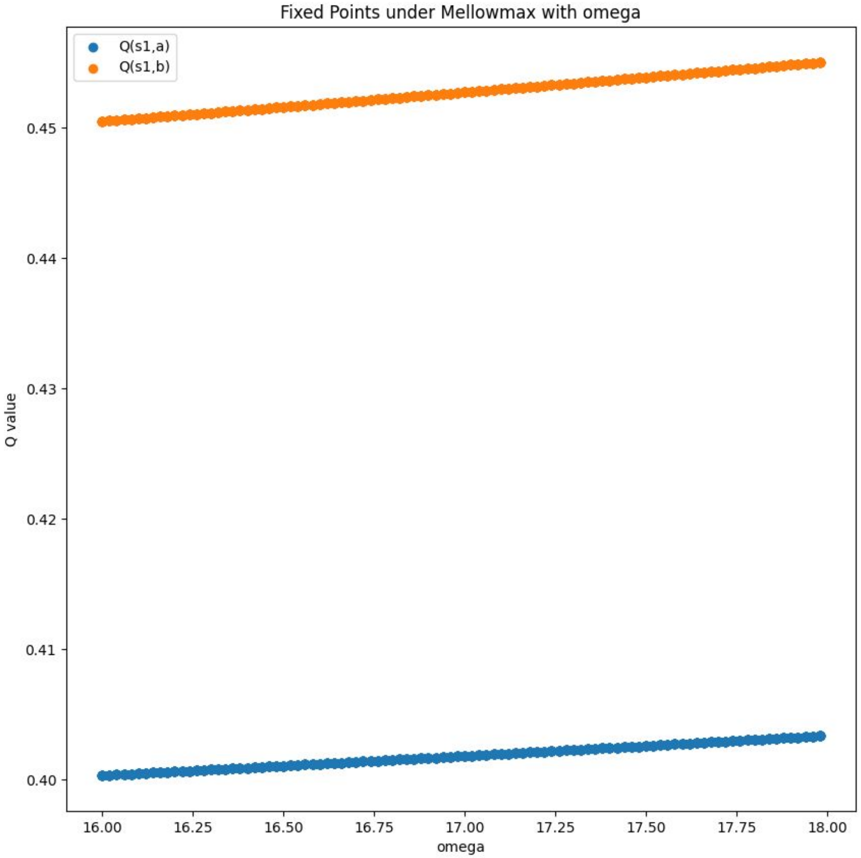


Boltzmann Softmax(Replication)

Mellowmax (Replication)

> 1 convergence points





Random MDPs

- The previous experiments are applied to a specifically handcrafted MDP.
- To be more naturally, in the following slides, more randomly constructed MDPs will be tested by GVI.

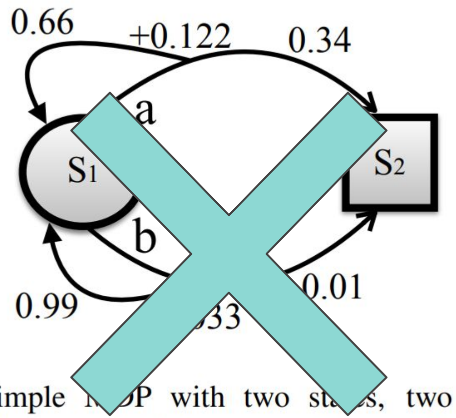


Figure 1. A simple MDP with two states, two actions, and $\gamma = 0.98$. The use of a Boltzmann softmax policy is not sound in this simple domain.

Get rid of handcrafted one.

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**

Paper's settings:

- Number of **states** sample from {2,3,...,10} uniformly at random.
- Number of **actions** sample from {2,3,...,5} uniformly at random.
- Construction of P & R:
  - For each entry, we do:
    1. Sample from [0,0.01] unifromly at random
    2. With prob 0.5: + a value sampled from N(1,0.1)
    3. With prob 0.1: + a value sampled from N(100,1)
    4. For P's entries, normalize it. For R's entries, divide values by max value and then *0.5
- Sample 200 MDPs
- Max iteration: 1000 (force to terminate if > this value)

Random MDPs

Our settings:

- Number of **states** sample from {2,3,4,5} uniformly at random.
- Number of **actions** sample from {2,3,4} uniformly at random.
- Construction of P & R:
    - For each entry, we do:
        i. Sample from [0,0.01] unifromly at random
        ii. With prob 0.5: + a value sampled from N(1,0.1)
        iii. With prob 0.1: + a value sampled from N(100,1)
        iv. For P’s entries, normalize it. For R’s entries, divide values by max value and then *0.5

Random MDPs

	Avg of all MDP's all trials		
	Avg # no terminate	Avg # > 1 fixed points	Avg iterations
Boltzmann Softmax	0.00675	0.03	1085.0973
Mellowmax	**0**	**0**	1048.794

Our settings:

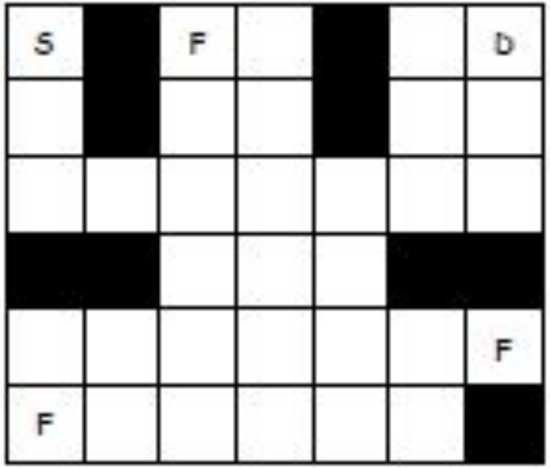
- How we decide one GVI has > 1 fixed points:
  - Do 100 trials for each MDP.
  - For each trial, each Q value is sampled from [0,30).
  - Find #unique convergece points of all 100 trials.
  - #unique>1 means >1 fixed points
- Sample 200 MDPs
- Max iteration: 2000

Random MDPs

Comparison

	Boltzmann Softmax	Mellowmax
> 1 fixed point	Sometimes	No
Number of iterations needed	More	Less

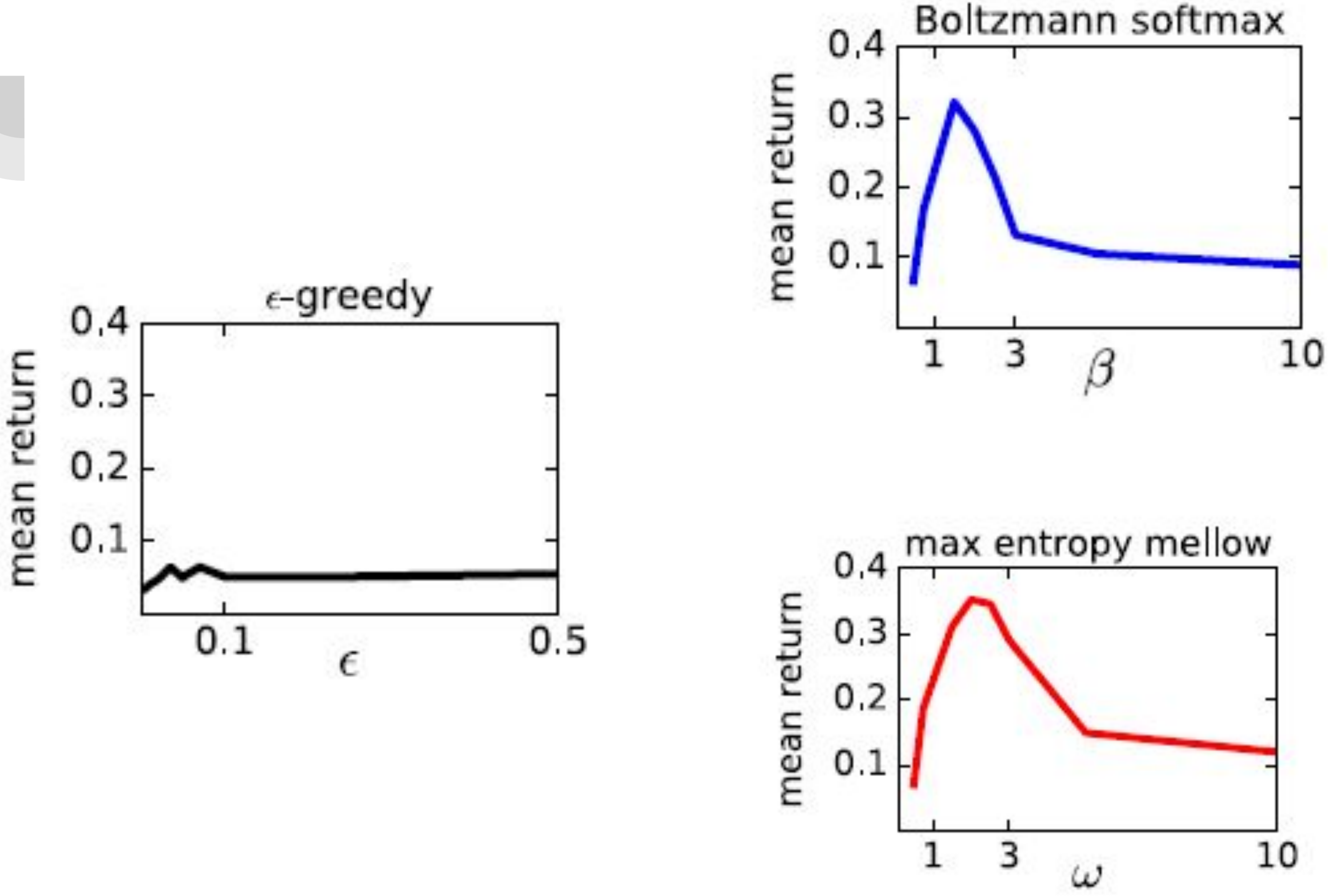
Taxi Domain



Reward +1 for delivering one passenger

Reward +3 for delivering two passenger

Reward +15 for delivering three passenger


ϵ-greedy
mean return
0.4
0.3
0.2
0.1
0.1
0.5
ϵ
Boltzmann softmax
mean return
0.4
0.3
0.2
0.1
1
3
10
β
max entropy mellow
mean return
0.4
0.3
0.2
0.1
1
3
10
ω

Lunar Lander Domain

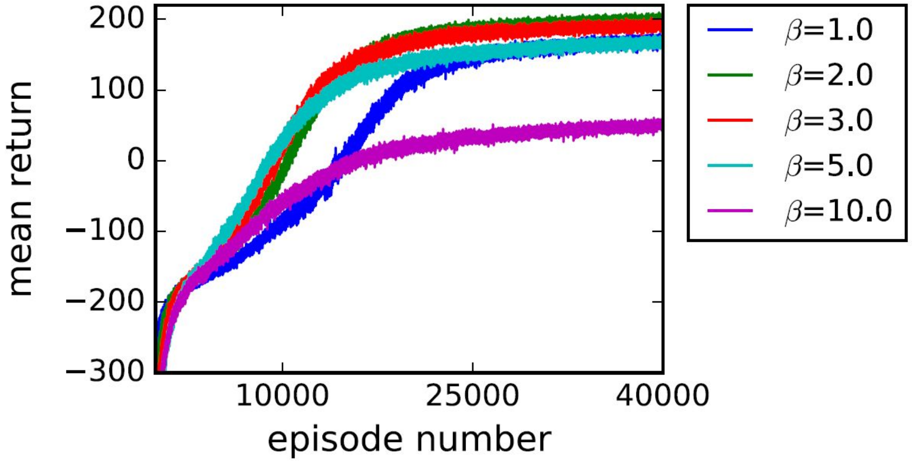
paper expiriment settings:

- Boltzmann: beta: 1, 2, 3, 5, 10
- Mellowmax: omega: 3, 5, 7, 8, 11
- learning rate: 0.005
- network: a hidden layer comprised of 16 units with RELU activation functions + a second layer with 16 units and softmax activation functions
- batch episode size: 10
- training: 40000 episodes x 400-run averages

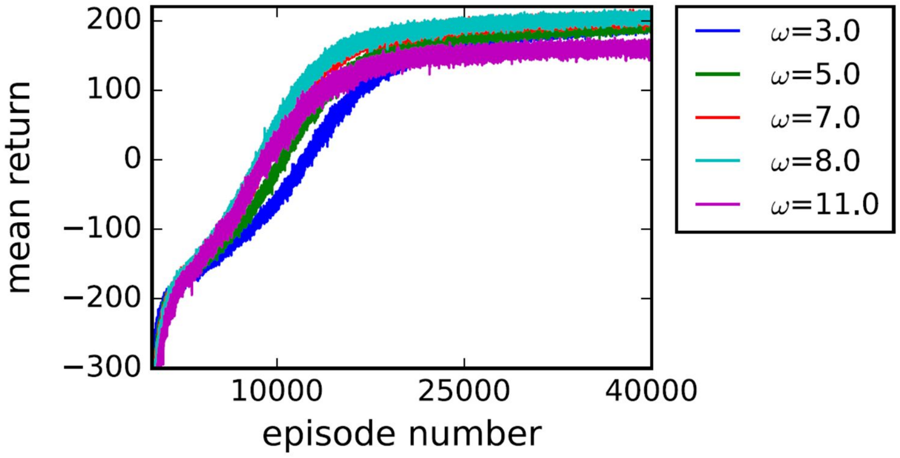
our different settings:

- training: 15000 episodes x 3-run averages

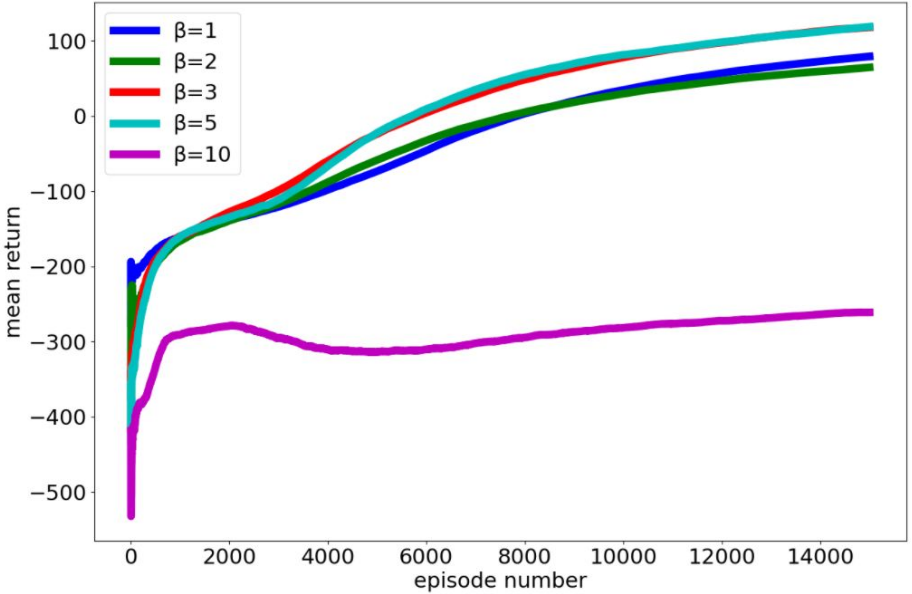
Boltzmann Softmax (Paper)



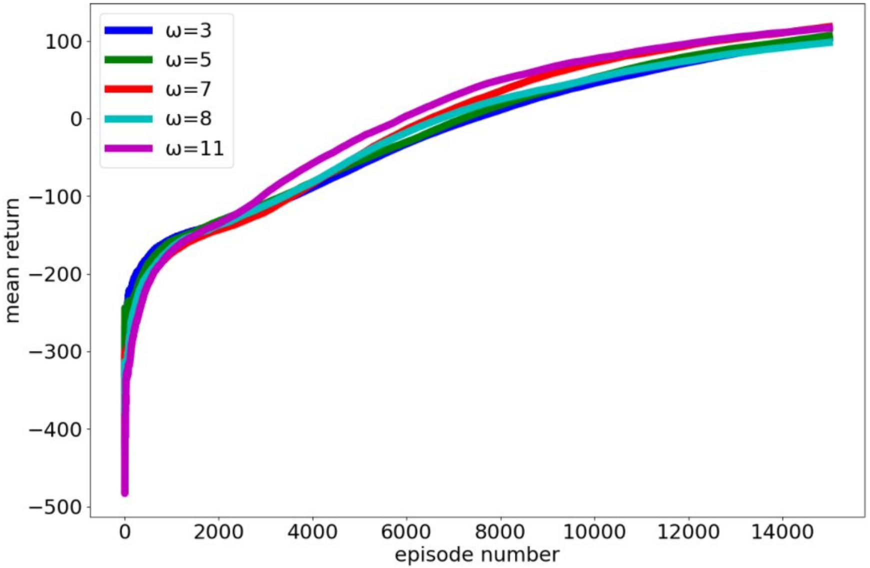
Mellowmax (Paper)



Boltzmann Softmax (Replication)

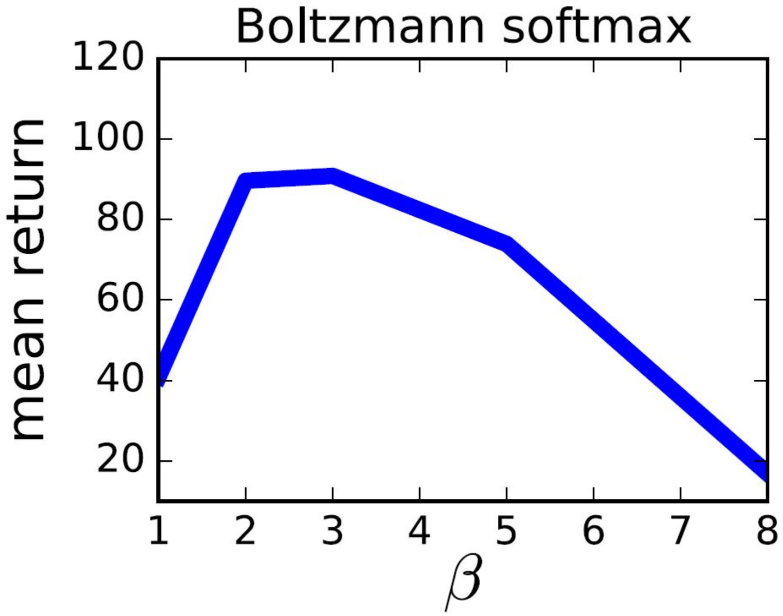


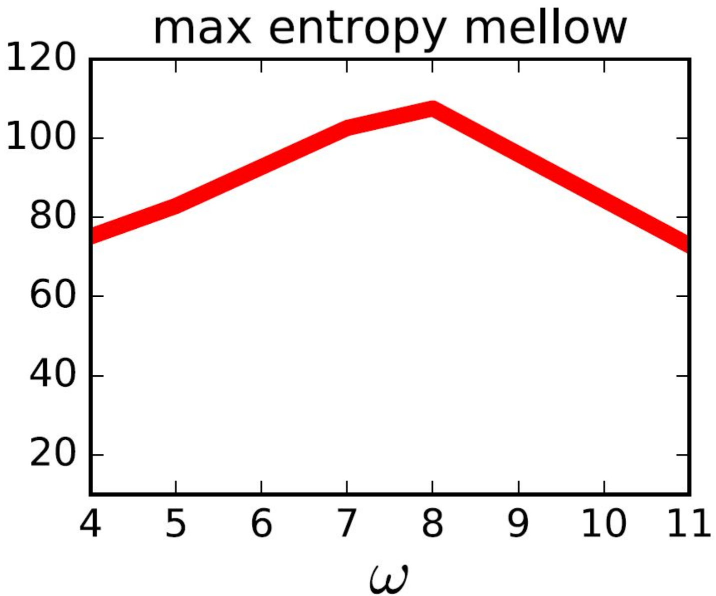
Mellowmax (Replication)



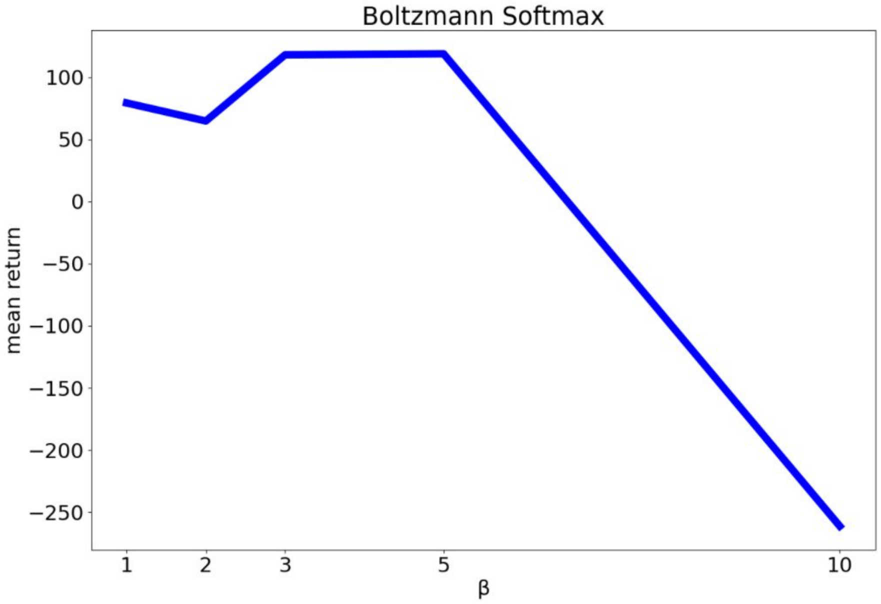
Boltzmann Softmax (Paper)

Mellowmax (Paper)

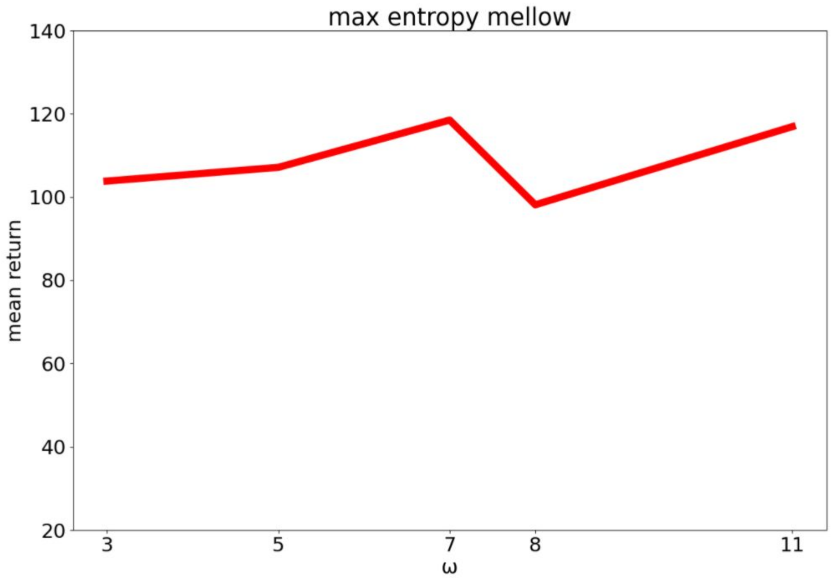




Boltzmann Softmax (Replication)



Mellowmax (Replication)



Conclusion

- Advantages of Mellowmax :
  - Non-expansion: GVI convergence guarrantee
  - More stable
- Disadvantages:
  - Needs more time to solve beta, but may get better updates
- Mellowmax operator can be an alternative to Boltzmann softmax operator