REPLICATION

AN ALTERNATIVE SOFTMAX OPERATOR FOR REINFORCEMENT LEARNING

簡右群 | 黃宇禓 | 邱俊耀 | 陳筱霓

SOFTMAX BACKGROUND

WHAT MAKES A SOFTMAX OPERATOR IDEAL

- **Approximates maximization**
  to perform reward-seeking behavior
- **Non-expansion**
  to ensure convergence to a unique fixed point
- **Differentiable**
  to work with gradient-based optimization
- **Avoids starvation of non-maximizing actions**

POPULAR SOFTMAX OPERATORS AND THEIR DRAWBACKS

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

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

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

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

HOW DOES MELLOMAX SOLVE THESE PROBLEM

MELLOWMAX SOFTMAX OPERATOR

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

PROPERTIES OF MELLOWMAX

1. NON-EXPANSION

$$\left|mm_\omega(\mathbf{X}) - mm_\omega(\mathbf{Y})\right| = \max_i \left|x_i - y_i\right|$$

2. MAXIMIZATION

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

3. DIFFERENTIABLE

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

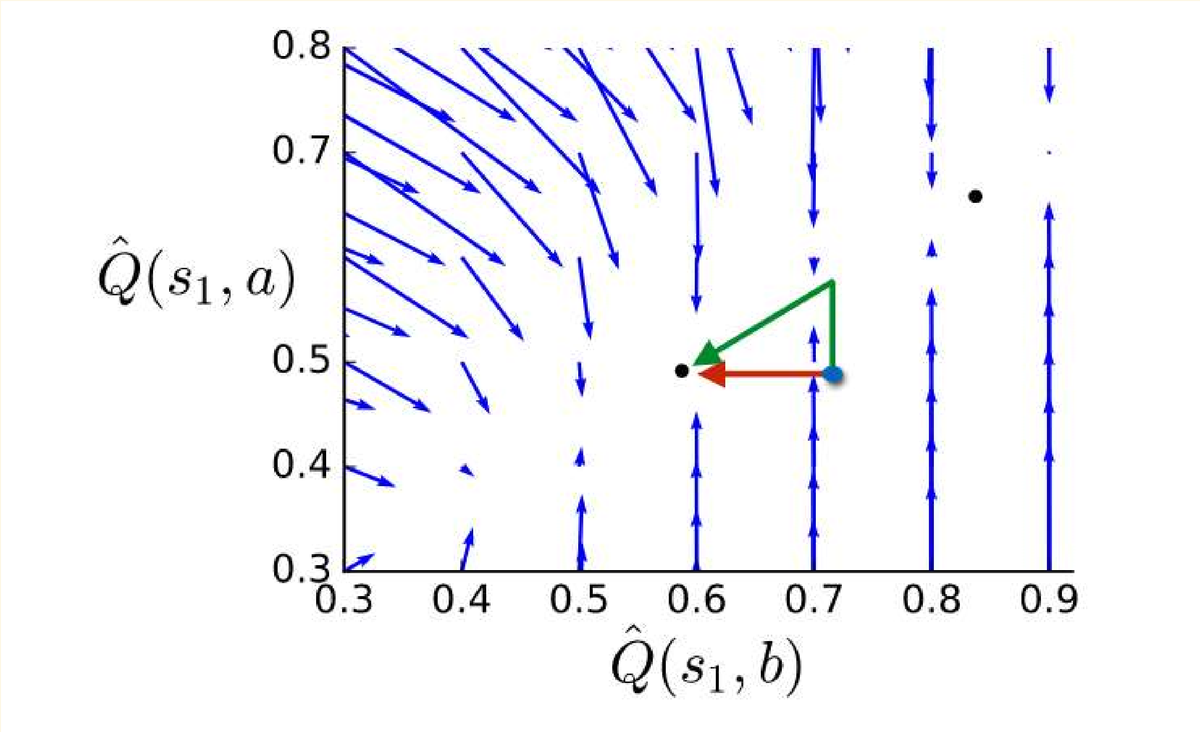
4. AVERAGING

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

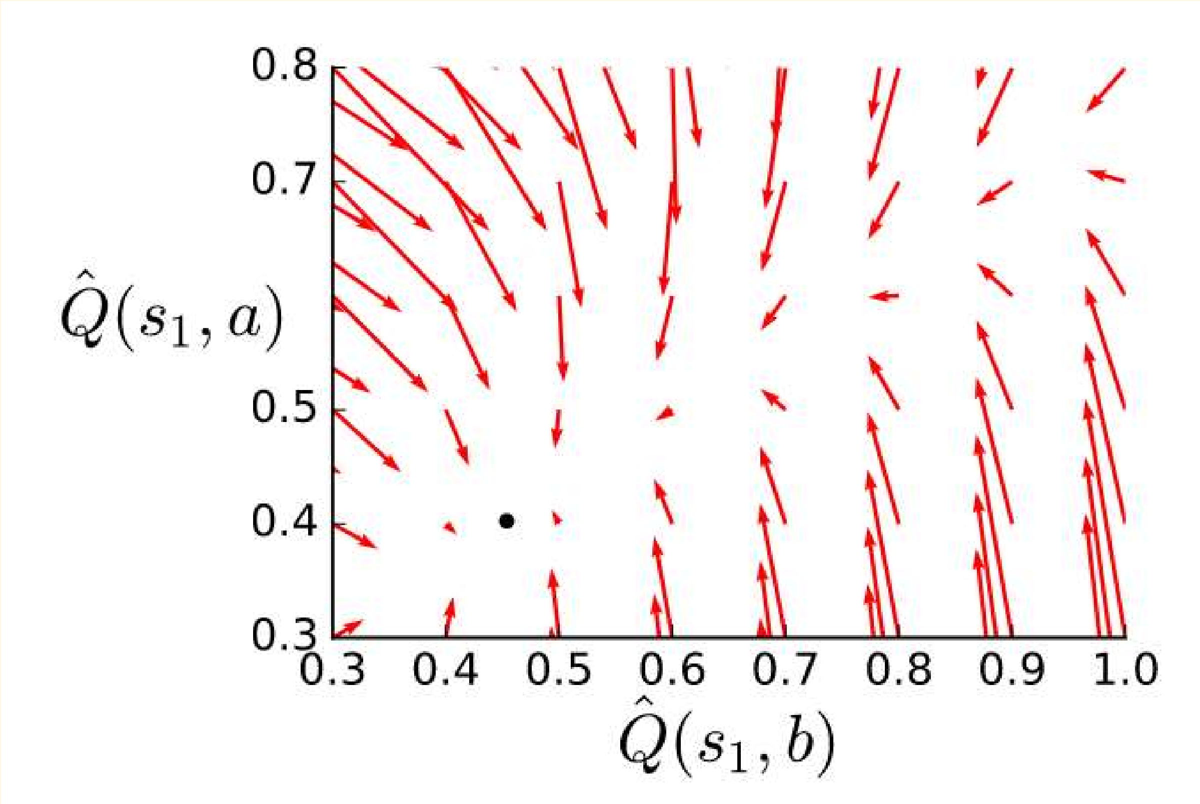
ω can't be 0

UNIQUE FIXED POINT?

GVI UNDER $boltz_\beta$ HAS MULTIPLE FIXED POINTS



GVI UNDER mm_ω



LEARNING WITH MELLOWMAX

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

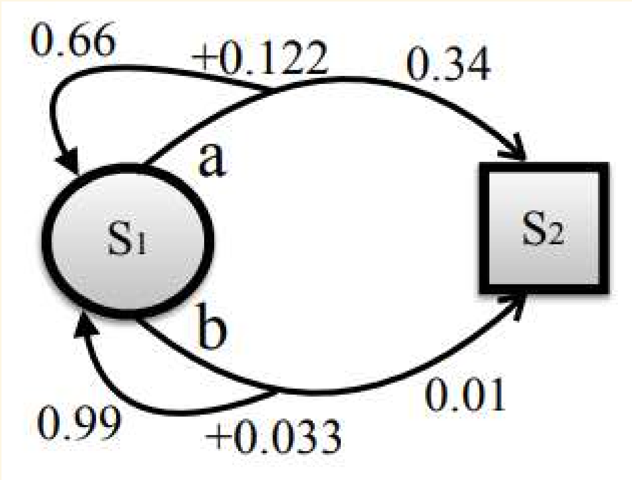
where β is the root for:

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

EXPERIMENTS AND RESULTS

ENVIRONMENT: SIMPLE MDP

- S_1 is the initial state and S_2 is the terminal state.
- The unsigned numbers denote the transition probabilities.
- The signed numbers denote the rewards.



ALGORITHM: GENERALIZED VALUE ITERATION (GVI)

Algorithm 1 GVI algorithm

Require: 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 $s \in S$ **do**
        for each $a \in A$ **do**
            $Q_{copy} \leftarrow \hat{Q}(s,a)$
            $\hat{Q}(s,a) \leftarrow \sum_{s' \in S} \mathcal{R}(s,a,s') + \gamma \mathcal{P}(s,a,s') \bigotimes \hat{Q}(s',\cdot)$
            diff $\leftarrow max\{$diff, $|Q_{copy} - \hat{Q}(s,a)\}$
        end for
    end for
until diff $< \delta$

NUMBER OF ITERATION TO CONVERGE

Simple MDP with GVI Algorithm

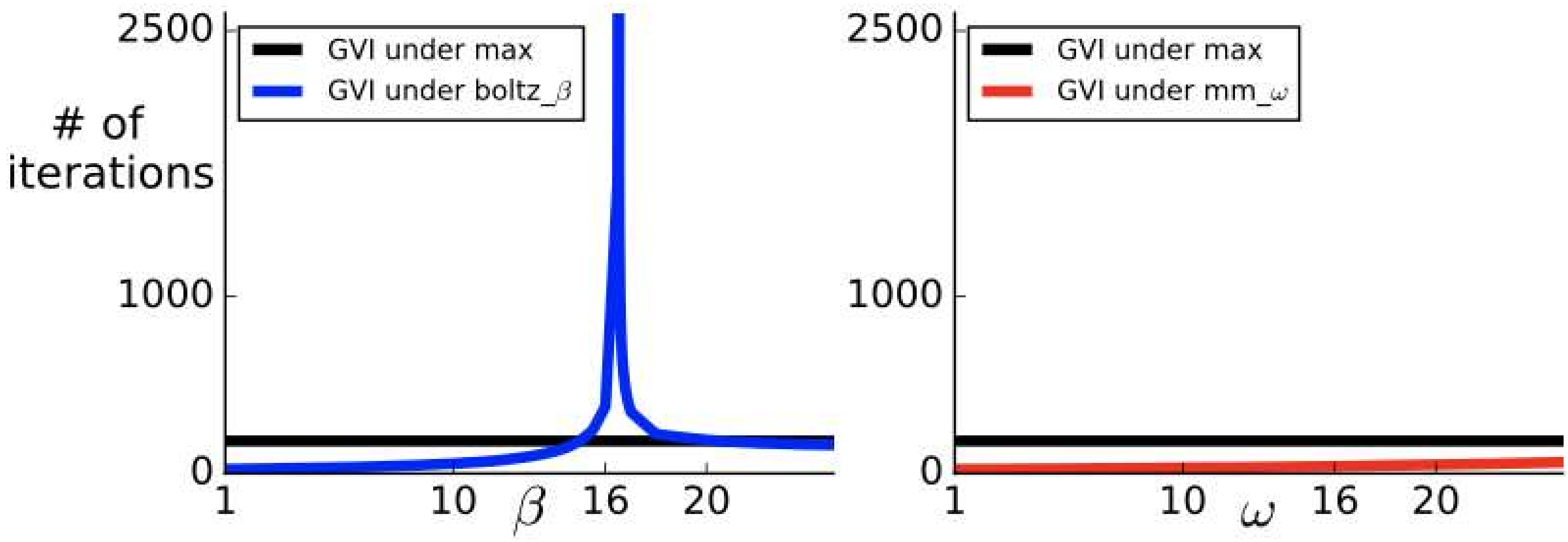
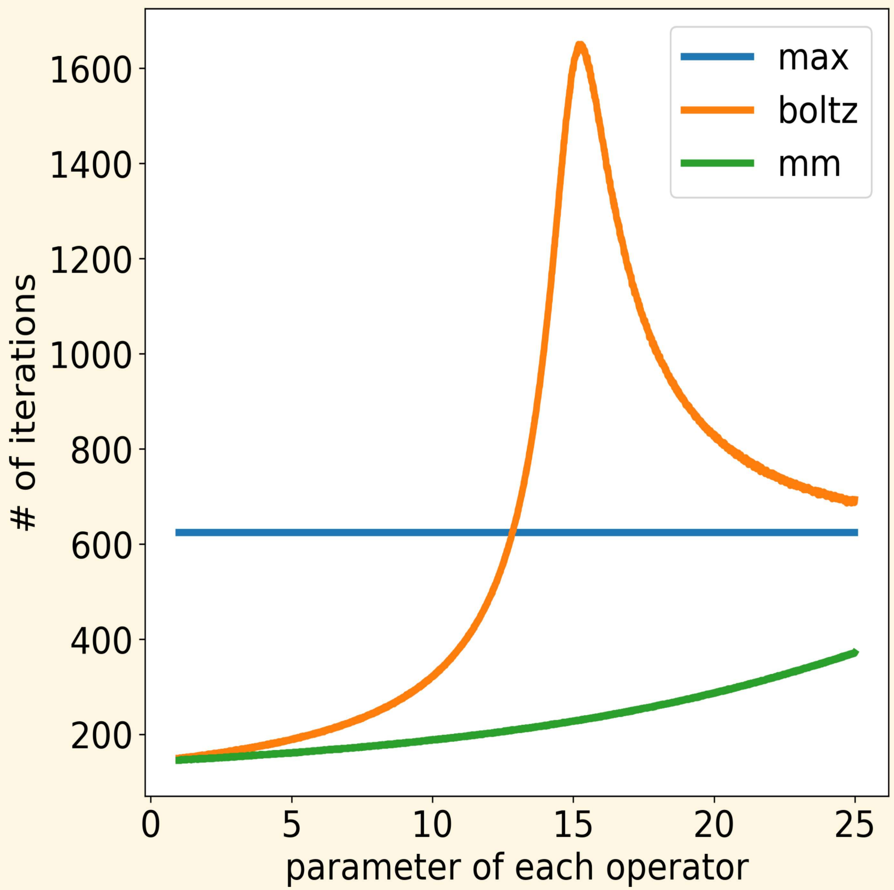


Figure 7. Number of iterations before termination of GVI on the example MDP. GVI under mm_ω outperforms the alternatives.

NUMBER OF ITERATION TO CONVERGE

Simple MDP with GVI Algorithm



ALGORITHM: SARSA

Algorithm 2 SARSA Algorithm

Require: initial $\hat{Q}(s,a)\ \ \forall s \in \mathcal{S}\ \ \forall a \in \mathcal{A}$, α, and policy $\pi(s, \hat{Q}(s,\cdot))$

for each episode **do**

Initialize s

$a \sim$ policy $\pi(s, \hat{Q}(s,\cdot))$

repeat

Take action a, observe r, s'

$a \sim$ policy $\pi(s, \hat{Q}(s',\cdot))$

$\hat{Q}(s,a) \leftarrow \hat{Q}(s,a) + \alpha[r + \gamma\hat{Q}(s',a') - \hat{Q}(s,a)]$

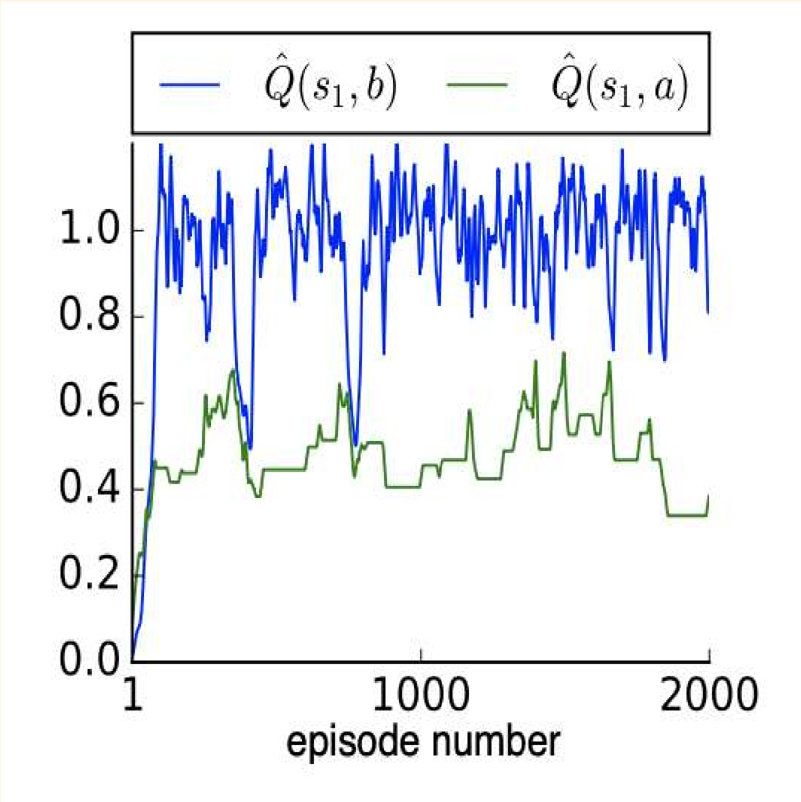
$s \leftarrow s'$, $a \leftarrow a'$

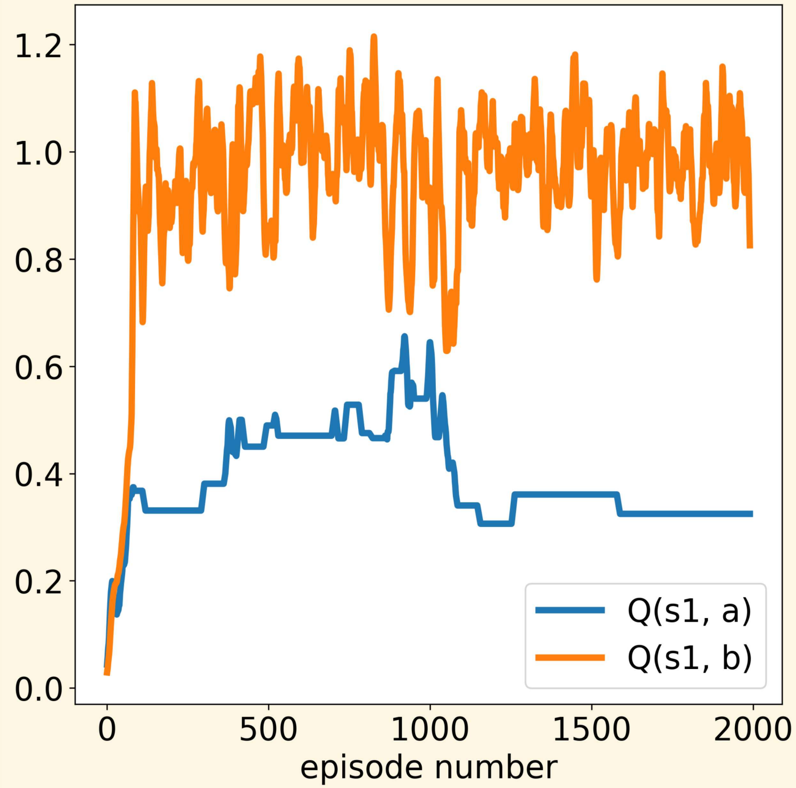
until s is terminal

end for

STABILITY OF Q-FUNCTION

Q-function is unstable when using Boltzmax

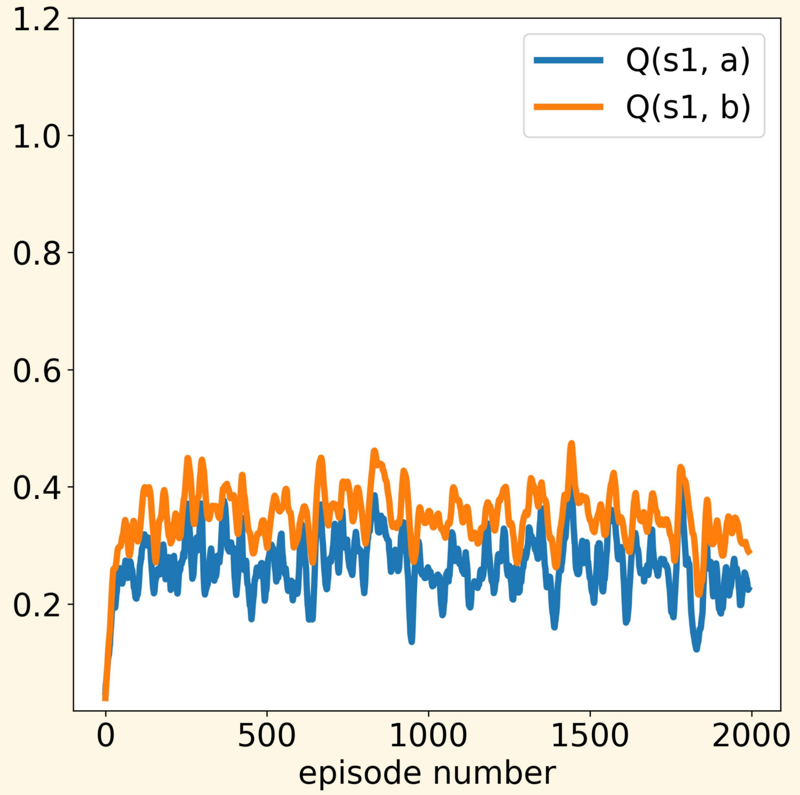




Simple MDP with SARSA Algorithm

STABILITY OF Q-FUNCTION

When using Mellowmax



Simple MDP with SARSA Algorithm

ENVIRONMENT: RANDOM MDP

- $|S| \in \{2, 3, \ldots, 10\}$
- $|A| \in \{2, 3, 4, 5\}$
- Transition probabilities $p \sim U[0, 0.01]$
- Add 2 noise to transition probabilities p
  - $\epsilon_1 \sim \mathcal{N}(1, 0.1) \times \mathcal{B}(1, 0.5)$
  - $\epsilon_2 \sim \mathcal{N}(100, 1) \times \mathcal{B}(1, 0.1)$

MULTIPLE FIX POINT

Random MDP with GVI algorithm

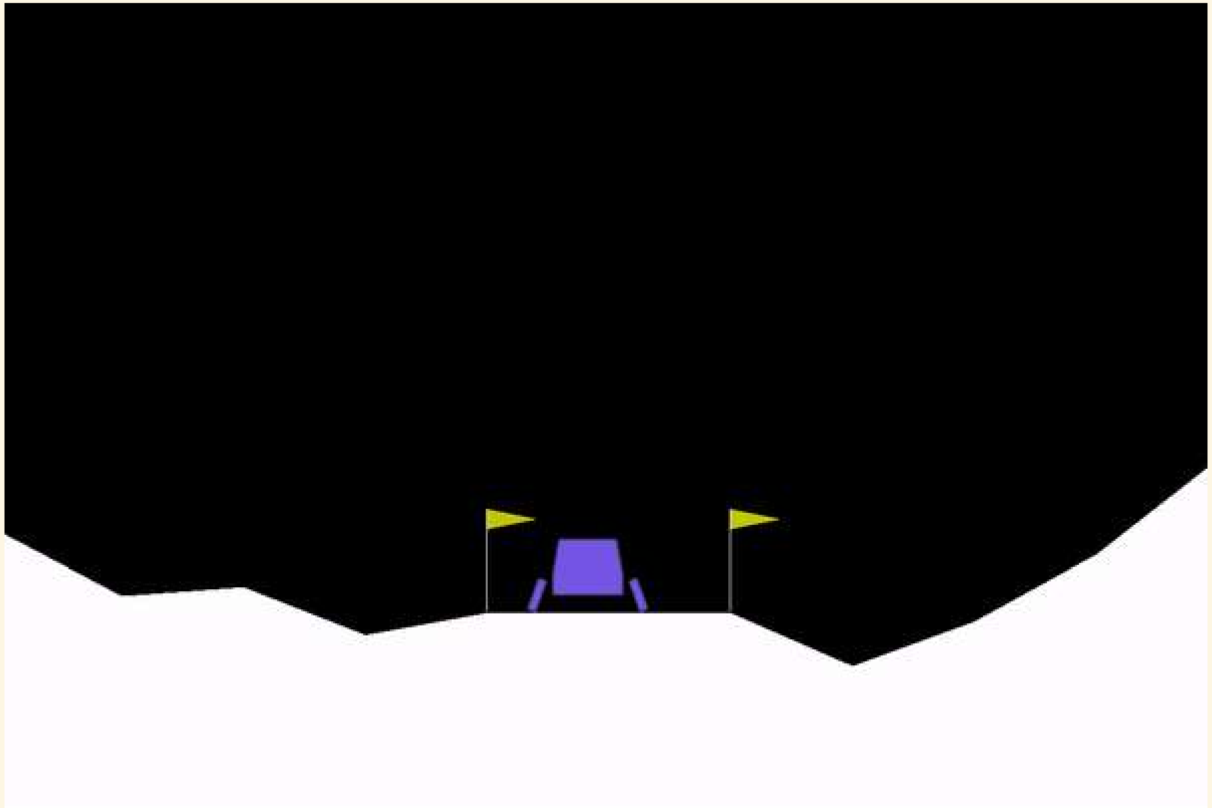
	MDPs, no terminate	MDPs, > 1 fixed points	average iterations
boltz_β	8 of 200	3 of 200	231.65
mm_ω	**0**	**0**	**201.32**

MULTIPLE FIX POINT

Random MDP with GVI algorithm

Policy	MDPs, no terminate	MDPs, >1 fixed points	average iterations
$boltz_{\beta}$	3 of 200	5 of 200	181.00
mm_{ω}	0 of 200	4 of 200	178.72

ENVIRONMENT: LUNARLANDER-V2



ALGORITHM: REINFORCE

Algorithm 3 REINFORCE Algorithm

Require: initialize θ, discount factor γ and step size α

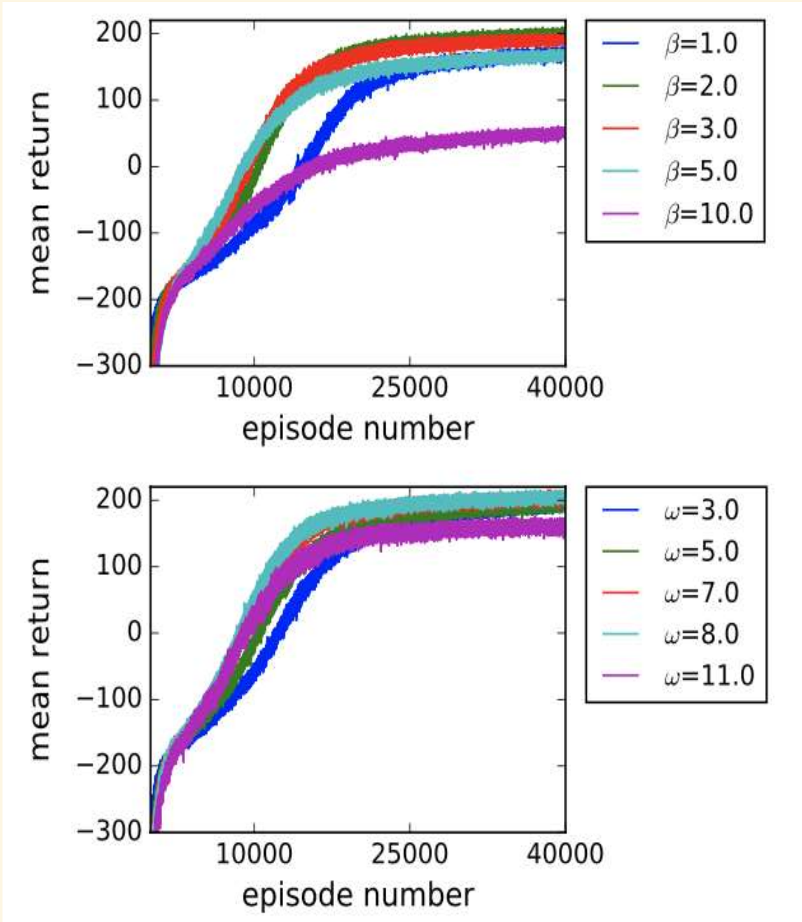
repeat

Sample trajectory $\tau \sim P_\mu^{\pi_\theta}$

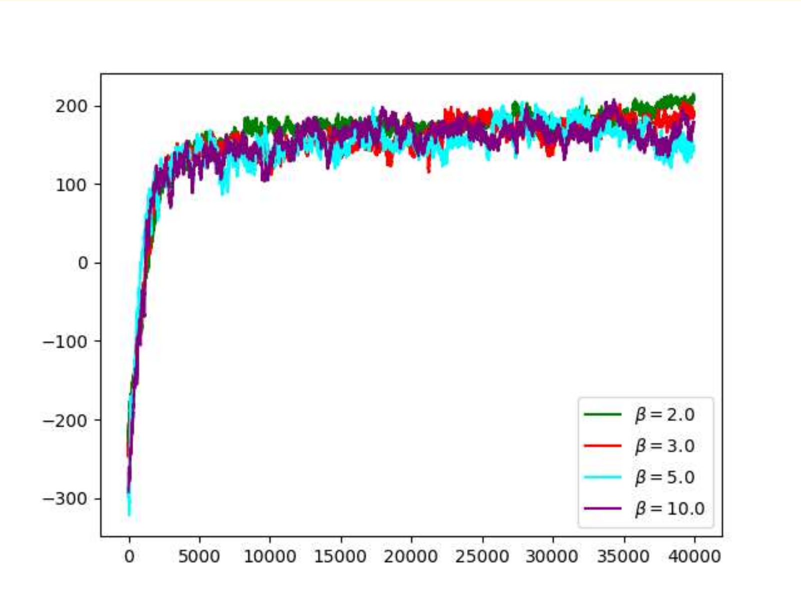
$\theta \leftarrow \theta + \alpha(\sum_{t=0}^{\infty} \gamma^t G_t(\tau) \nabla_\theta \log \pi_\theta(a_t|s_t))$

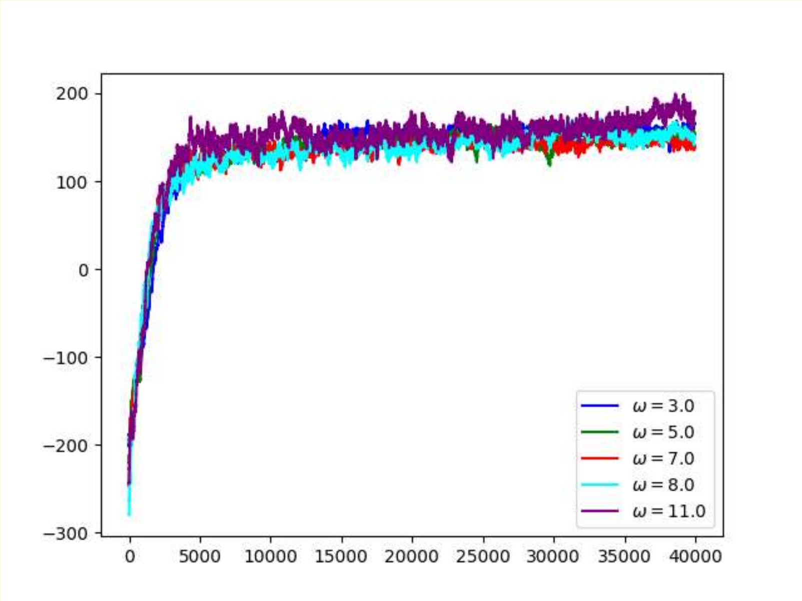
until termination

LUNARLANDER-V2 WITH REINFORCE



LUNARLANDER-V2 WITH REINFORCE





CONCLUSION

- Computationally expensive (need to solve β)
- Does not improve significantly

THANKS FOR LISTENING