Robot Learning

Sparse Markov Decision Processes

Prof. Songhwai Oh
ECE, SNU

Sparse Markov Decision Processes (MDPs)

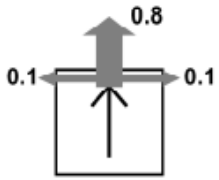
Outline

- Introduction
  - Markov Decision Processes
  - Stochastic Policies: Softmax Distribution
- Sparse Markov Decision Processes
  - Causal Sparse Tsallis Entropy
  - Sparse Bellman Equation
  - Sparse Value Iteration
- Theoretical Results
  - Optimality of Sparse Value Iteration
  - Performance Error Bounds
- Experimental Results

Markov Decision Processes (MDPs)

Definition

- A set of states $s \in S$
  - Each cell in the grid map
- A set of actions $a \in A$
  - Up, Down, Right, Left, Stay
- Transition probability
  - $P(s'|s,a)$



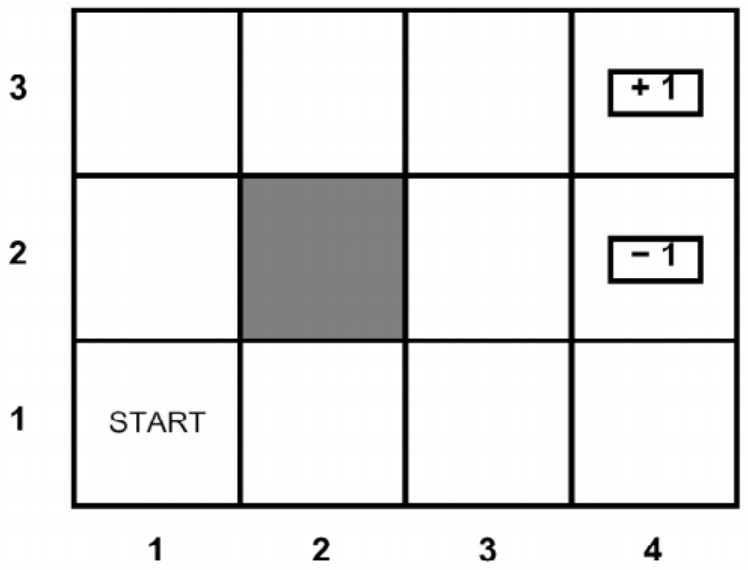
- Reward function $r(s,a)$
- Initial state distribution d
  - $P(s_0 = START) = 1$
- Terminal states (Optional)

Markov Decision Process

S: states
A: actions
T: transition model
r: reward function

Goal

π: policy $(S \rightarrow A)$

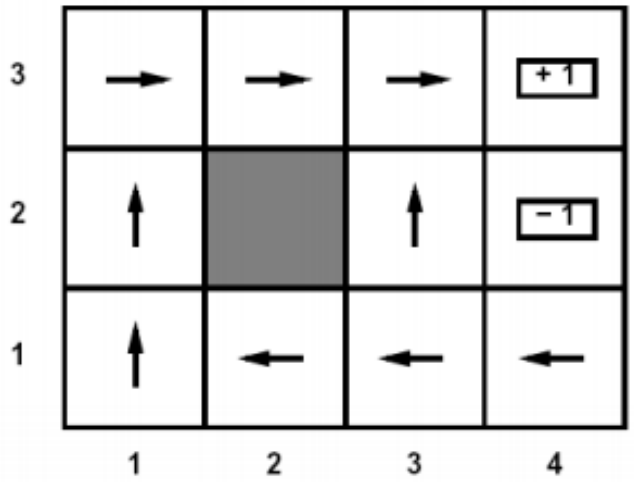


Source: AIMA

MDPs

- Goal of MDPs: find an optimal policy $\pi^*: S \rightarrow A$
  - A policy π gives an action a for each state s
  - An optimal policy maximizes the expected utility
  - Markov decision problem

$$\max_{\pi} \ \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t)\right]$$



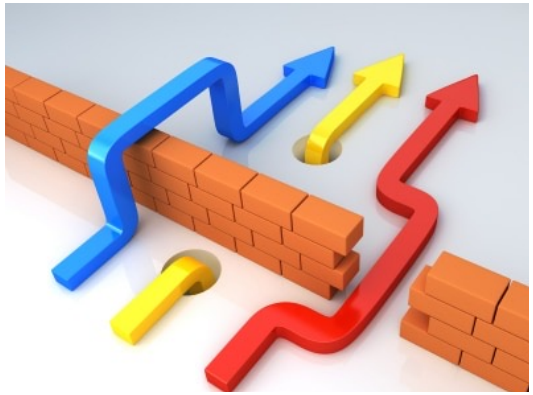
Optimal Policy

- Bellman Equation
  - Optimal condition of MDPs
  - $Q(s,a) = r(s,a) + \gamma \sum_{s'} V(s') P(s'|s,a)$
  - $V(s) = \max_a Q(s,a)$
  - $\pi(s) = \operatorname{argmax}_a Q(s,a)$

Optimal Policy

- Optimal policy of an MDP: $\pi(s) = \underset{a}{\operatorname{argmax}}\, Q(s, a)$
  - Deterministic function
  - An agent selects the exact same action at the same state
  - It can cause drawbacks in presence of multiple optimal actions

- Knowing multiple optimal action choices can be useful for real world problems
  - Go and chess
  - Driving a car and avoiding obstacles
  - Robustness against:
    - Unexpected or dynamic events
    - Modeling and estimation errors



Stochastic Policies

- A stochastic policy can provide multiple action choices
- **Softmax** distribution: widely used stochastic policy function
- Probability is exponentially proportional to the state action value Q(s,a) $\pi(a|s) = \frac{\exp Q(s,a)}{\sum_{a'} \exp Q(s,a')}$
- Softmax policy function:

Softmax Policy

Soft MDPs

- The softmax distribution is the optimal solution of a soft MDP problem:

$$\max_{\pi} \ \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t)\right] + \alpha H(\pi)$$

$$\text{subject to} \quad \forall\, s, a \qquad \sum_{a} \pi(a|s) = 1, \qquad \pi(a|s) \geq 0$$

  - Causal entropy regularization : $H(\pi) = \mathbb{E}[-\sum_{t=0}^{\infty} \gamma^t \log(\pi(a_t|s_t))]$
  - It finds an optimal policy distribution $\pi(a|s)$ (not a deterministic function)
  - $H(\pi)$ gives extra rewards to a multi-modal distribution

Softmax Policy

Entropy

- Measure of uncertainty in a distribution
- For a random variable X with distribution P, the entropy of X is defined as

$$H(X) = -\sum_{i=1}^{n} P(X = x_i) \log P(X = x_i) = \mathbb{E}(-\log P(X))$$

- Entropy of a coin flip (Bernoulli distribution)
  - p_h: probability of getting a head
  - $H(X_{coin}) = -p_h \log p_h - (1 - p_h) \log(1 - p_h)$
  - $H(X_{coin})$ has the largest value when $p_h = 0.5$ (when it is the most uncertain)
- Some facts about entropy:
  - For a distribution defined on an interval $[a, b]$, the uniform distribution has the maximum entropy.
  - For a Gaussian distribution, the larger the variance, the larger the entropy.
  - The Gaussian distribution is the maximum entropy distribution among all distributions with the same variance.

Maximum Entropy Probability Distribution

Maximum entropy principle states that the most appropriate distribution to model a given set of data is the one with the highest entropy among all those that satisfy the constraints of our prior knowledge.

Given a random variable X with $X \sim P(\cdot)$, the maximum entropy probability problem is:

$$\max \quad H(X)$$

$$\text{such that} \quad \mathbb{E}\big(f_k(X)\big) = a_k \qquad k = 1, \dots, n$$
$$\sum_x P(x) = 1$$
$$\mathrm{P}(x) \geq 0 \qquad \forall x$$

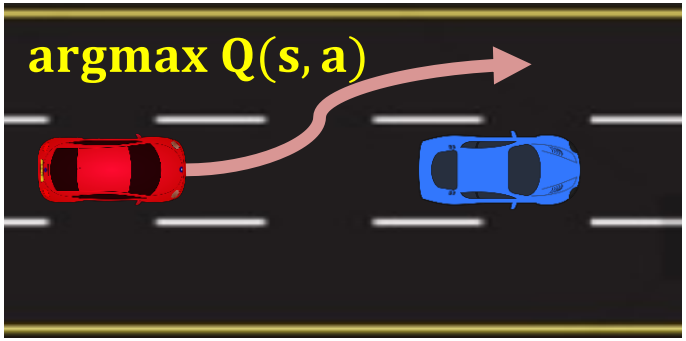
The solution is

$$P(x) = \frac{1}{Z} \exp\left(\sum_k \alpha_k f_k(x)\right),$$

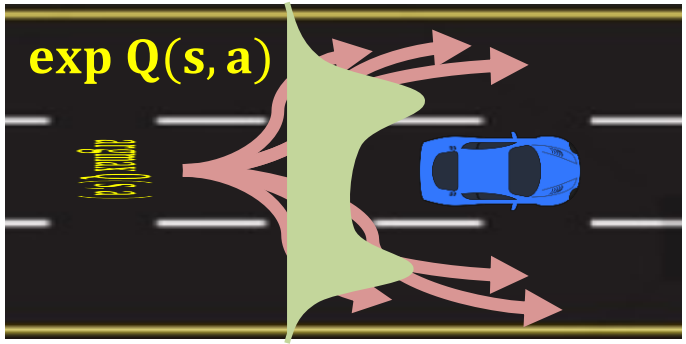
where $Z = \sum_x \exp\left(\sum_k \alpha_k f_k(x)\right)$ (partition function)

Soft Bellman Equation

- Optimal condition of soft MDPs
  - $Q(s,a) = R(s,a) + \gamma \sum_{s'} V(s')P(s'|s,a)$
  - $V(s) = \alpha \log\left(\sum_{a'} \exp\left(\frac{Q(s,a')}{\alpha}\right)\right)$
  - $\pi(a|s) = \frac{\exp\left(\frac{Q(s,a)}{\alpha}\right)}{\sum_{a'} \exp\left(\frac{Q(s,a')}{\alpha}\right)}$
- Softmax policy can represent multiple optimal actions



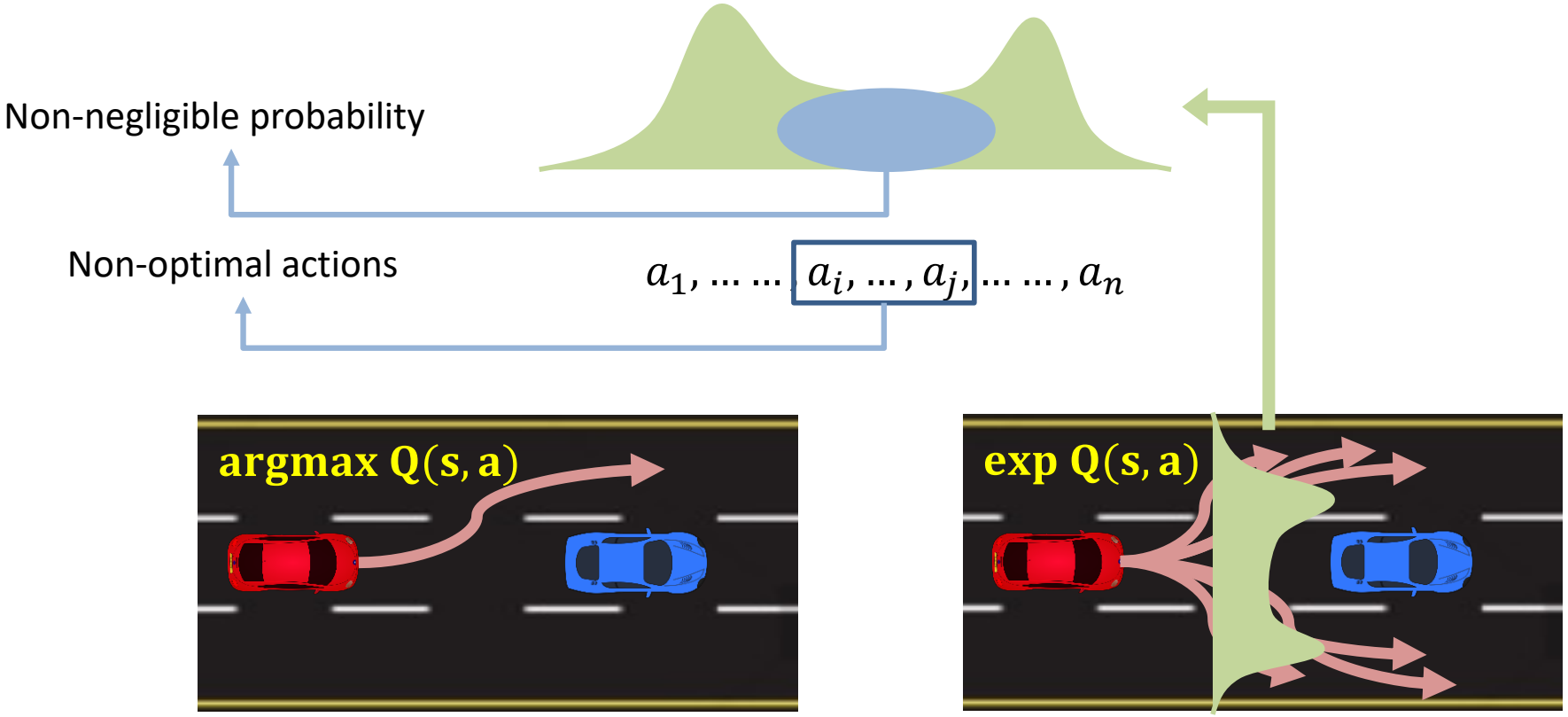
Original MDP (Deterministic)



Soft MDP (Stochastic)

Drawback of Softmax Distribution

- Softmax policy assigns non-negligible probability mass to non-optimal actions even if state-action values of these actions are dismissible
- This behavior causes a performance drop (Theorem 5)

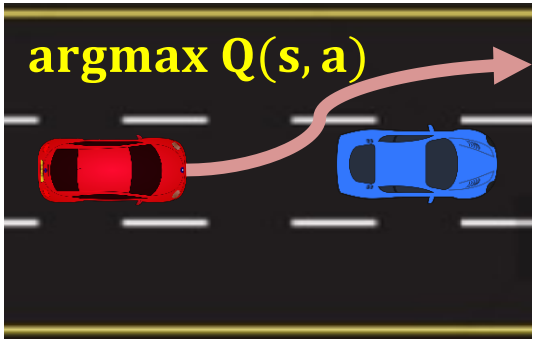


Original MDP (Deterministic)

Soft MDP (Stochastic)

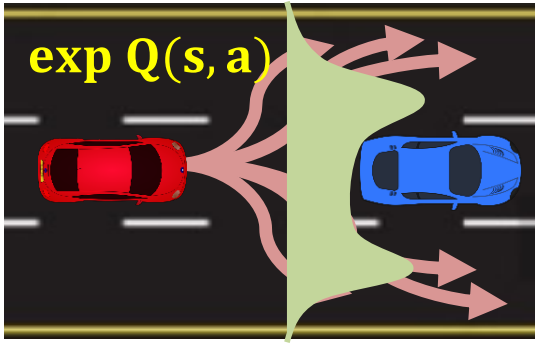
Sparse Policy Distribution

- Drawback of softmax distribution: performance drop due to the non-optimal actions
- Sparse MDPs
  - Optimal policy is a sparse and multi-modal distribution



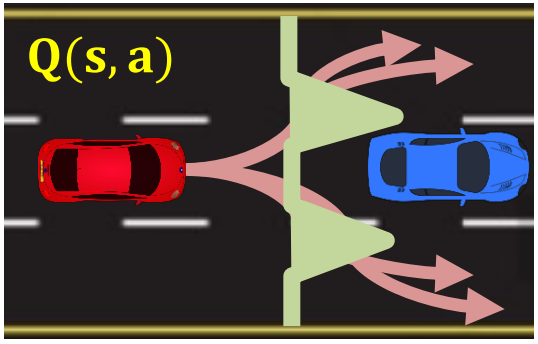
Original MDP (Deterministic)

$$\pi(s) = \text{argmax } \mathrm{Q(s, a)}$$



Soft MDP (Stochastic)

$$\pi(a|s) = \frac{\exp \mathrm{Q(s, a)}}{Z}$$



Sparse MDP (Stochastic)

$$\pi(a|s) = \max(\mathrm{Q}(s, a) - \tau, 0)$$

Sparse Markov Decision Processes

- Sparse MDP Problem:

$$\max_{\pi} \ \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t)\right] + \alpha W(\pi)$$

$$\text{subject to} \quad \forall\, s, a \qquad \sum_{a} \pi(a|s) = 1, \qquad \pi(a|s) \geq 0$$

- Causal Sparse Tsallis Entropy Regularization:

$$W(\pi) = \mathbb{E}\left[\sum_{t=0}^{\infty} \frac{\gamma^t}{2}\left(1 - \pi(a_t|s_t)\right)\right]$$

- $W(\pi)$ gives extra rewards to a multi-modal policy distribution, but weaker than $H(\pi)$

Sparsemax Policy

Tsallis Entropy

- Tsallis entropy:
  - $S_{q,k}(p) = \frac{k}{q-1}\left(1 - \sum p_i^q\right)$
  - Generalization of the standard Boltzmann–Gibbs entropy
  - Since 2000, Tsallis entropy is widely used in the field of physics, information theory, and social science
  - Tsallis entropy (nonadditive) has been used to describe complex phenomena that cannot be explained by the Boltzmann–Gibbs entropy (additive)

- Tsallis entropy has been successfully applied to explain:
  - The fluctuation of the magnetic field in the solar wind
  - The velocity distributions in dissipative dusty plasma
  - Thermostatistics of overdamped motion of interacting particles
  - Heavy tail distributions are derived from a maximum Tsallis entropy problem

Constantino Tsallis, "**Possible Generalization of Boltzmann-Gibbs statistics**," *Journal of statistical physics* 52.1 (1988): 479-487.

Sparse MDPs

- Tsallis entropy: $S_{q,k}(p) = \frac{k}{q-1}\left(1 - \sum p_i^q\right)$
  - q is called entropic-index
  - k is a positive real constant
- Special cases:
  - Boltzmann–Gibbs entropy: $S_{1,1}(p) = \lim_{q\to 1} \frac{\sum\left(1-p_i^{q-1}\right)p_i}{q-1} = \sum -p_i \log p_i$
  - Sparse Tsallis entropy: $S_{2,\frac{1}{2}}(p) = \frac{1}{2}\sum p_i(1-p_i)$
- Causal Sparse Tsallis Entropy Regularization:

$$W(\pi) = \mathbb{E}\left[\sum_{t=0}^{\infty} \frac{\gamma^t}{2}\left(1 - \pi(a_t|s_t)\right)\right]$$

  - $W(\pi)$ is an extension of the sparse Tsallis entropy to the causally conditioned random variables

Theorem 2. $\mathbb{E}\left[\sum_{t=0}^{\infty} \frac{\gamma^t}{2}\left(1 - \pi(a_t|s_t)\right)\right] = \sum_s \rho_\pi(s) \frac{1}{2} \sum_a \pi(a|s)(1-\pi(a|s))$

$\boldsymbol{S_{2,\frac{1}{2}}(\pi(\cdot\,|s))}$

Sparse Bellman Equation

Theorem 1. If a policy distribution π is the optimal solution of a sparse MDP, then π and the corresponding sparse value function V_π^{sp} necessarily satisfy following equations for all state and action pairs:

$$Q_\pi^{sp}(s,a) = r(s,a) + \sum_{s'} V_\pi^{sp}(s')T(s'|s,a)$$

$$V_\pi^{sp}(s) = \alpha\left(\frac{1}{2}\sum_{a'\in S(s)}\left(\frac{Q_\pi^{sp}(s,a')}{\alpha}\right)^2 - \tau\left(\frac{Q_\pi^{sp}(s,\cdot)}{\alpha}\right)^2 + \frac{1}{2}\right)$$

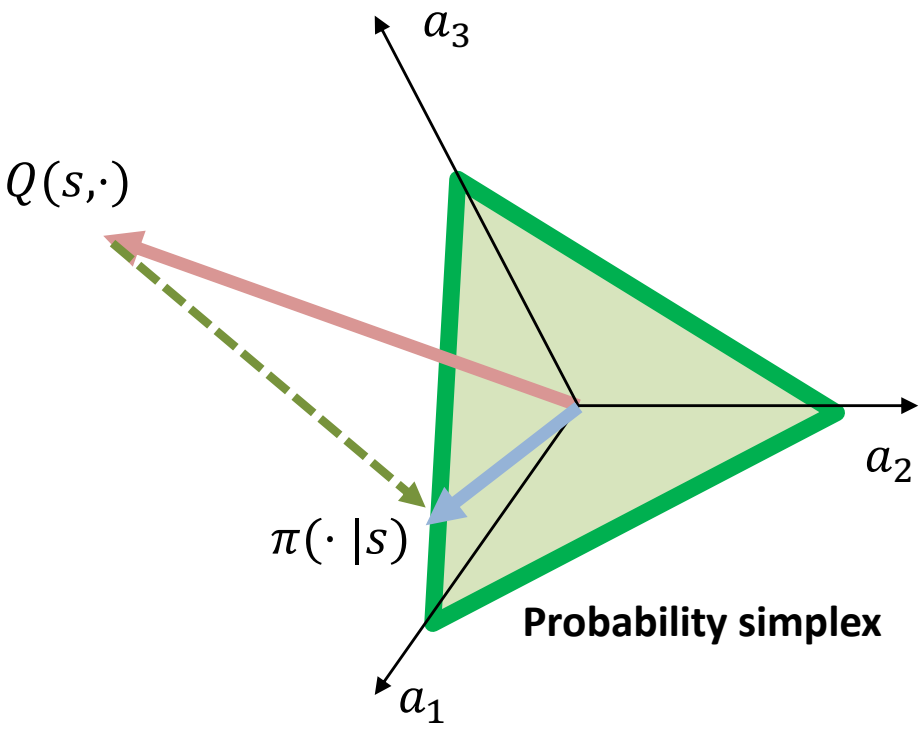
$$\pi(a|s) = \max\left(\frac{Q_\pi^{sp}(s,a)}{\alpha} - \tau\left(\frac{Q_\pi^{sp}(s,\cdot)}{\alpha}\right), 0\right)$$

where $\tau\left(\frac{Q_\pi^{sp}(s,\cdot)}{\alpha}\right) = \frac{\sum_{a'\in S(s)}\frac{Q_\pi^{sp}(s,\cdot)}{\alpha} - 1}{K}$ and $S(s)$ is a set of actions satisfying $1 + i\frac{Q_\pi^{sp}(s,a_{(i)})}{\alpha} > \sum_{j=1}^{i}\frac{Q_\pi^{sp}(s,a_{(j)})}{\alpha}$ with $a_{(i)}$ indicating the action with the i-th largest $Q_\pi^{sp}(s,a_{(i)})$ and $K = |S(s)|$.

- The optimal condition of sparse MDPs
- Theorem 1 can be proven by Karush–Kuhn–Tucker (KKT) conditions

Interpretation of Sparse Bellman Equation

- The projection of an action value to the probability simplex
- A sparse optimal policy is the same as the solution of the probability simplex projection



Probability Simplex Projection

$$\min_{\pi(\cdot|S)} \left|\left|\pi(\cdot\,|s) - Q(s,\cdot)\right|\right|_2^2$$

Optimal Projection (Sparsemax Distribution)

$$\pi(\cdot\,|s) = \max\left(\frac{Q(s,a)}{\alpha} - \tau\left(\frac{Q(s,\cdot)}{\alpha}\right), 0\right)$$

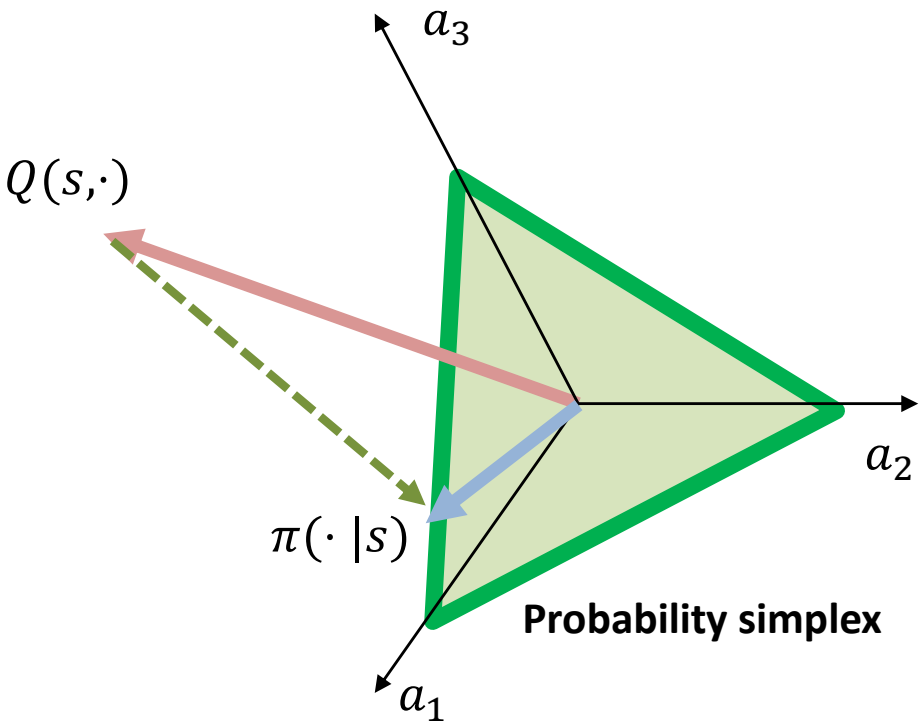
Sparsemax Operation

$$V(\mathrm{s}) = \alpha\ \mathrm{spmax}\left(\frac{Q(s,\cdot\,)}{\alpha}\right)$$

$$:= \alpha\left(\frac{1}{2}\sum_{a'\in S(s)}\left(\frac{Q(s,a')}{\alpha}\right)^2 - \tau\left(\frac{Q(s,\cdot)}{\alpha}\right)^2 + \frac{1}{2}\right)$$

Supporting Set of Sparse Optimal Policy

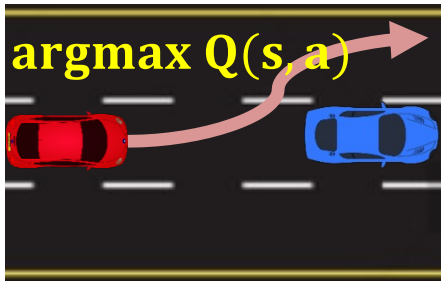
- Supporting set $S(s)$ of a sparse MDP is a set of actions with nonzero probabilities
- The cardinality of $S(s)$ can be controlled by regularization coefficient α

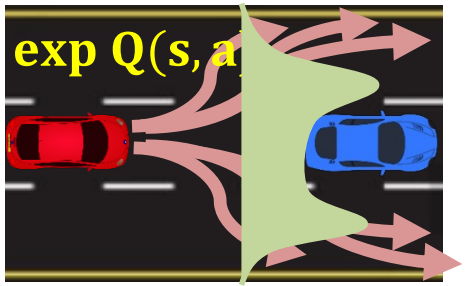


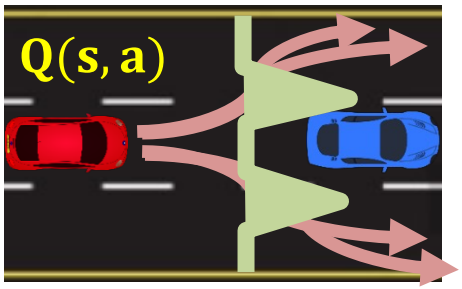
Supporting Set Condition

$$\alpha + iQ_{\pi}^{sp}\left(s, a_{(i)}\right) > \sum_{j=1}^{i} Q_{\pi}^{sp}\left(s, a_{(j)}\right)$$

MDP vs. Soft MDP vs. Sparse MDP







Problem	Original MDP (Deterministic)	Soft MDP (Stochastic)	Sparse MDP (Stochastic)
Objective function	$\mathbb{E}\left[\sum_{t=0}^{\infty}\gamma^t r(s_t,a_t)\right]$	$\mathbb{E}\left[\sum_{t=0}^{\infty}\gamma^t r(s_t,a_t)\right]+\alpha H(\pi)$	$\mathbb{E}\left[\sum_{t=0}^{\infty}\gamma^t r(s_t,a_t)\right]+\alpha W(\pi)$
$Q(s,a)$	$r(s,a)+\gamma\sum_{s'}V(s')P(s'\|s,a)$		
$V(s)$	$\max_a Q(s,a)$	$\alpha\log\left(\sum_{a'}\exp\left(\frac{Q(s,a')}{\alpha}\right)\right)$	$\alpha\left(\frac{1}{2}\sum_{a'\in S(s)}\left(\frac{Q(s,a')}{\alpha}\right)^2-\tau^2+\frac{1}{2}\right)$
$\pi(a\|s)$	$\operatorname{argmax}_a Q(s,a)$	$\frac{1}{Z}\exp\left(\frac{Q(s,a)}{\alpha}\right)$	$\max\left(\frac{Q(s,a)}{\alpha}-\tau,0\right)$

Sparse Value Iteration

- Sparse Bellman operation
  - $\mathrm{U}(x) : R^{|S|} \rightarrow R^{|S|}; U(x) = \alpha \operatorname{spmax}\left(\frac{q(s,\cdot)}{\alpha}\right)$
  - $v(s) = \alpha \operatorname{spmax}\left(\frac{q(s,\cdot)}{\alpha}\right), \quad q(s,a) = r(s,a) + \sum_{s'} x(s')T(s'|s,a)$
- **Sparse value iteration algorithm**
  - Random initial $v_0 \in R^{|S|}$
  - Repeatedly apply sparse Bellman operator

$$v_{i+1} = U(v_i)$$

  - After finding an optimal value, the policy distribution can be computed by sparse Bellman equation
- Optimality

Theorem 3. Sparse value iteration converges to the optimal value of sparse MDPs

  - Sparse Bellman operator is monotone and discounting → contraction

Value Iterations: MDP, Soft MDP, Sparse MDP

Algorithm	Value Iteration	Soft Value Iteration	Sparse Value Iteration
Bellman Operation	$U(x) = \max q(s,\cdot)$	$U(x) = \alpha \log\left(\sum_{a'} \exp\left(\frac{q(s,\cdot)}{\alpha}\right)\right)$	$U(x) = \alpha \,\mathrm{spmax}\left(\frac{q(s,\cdot)}{\alpha}\right)$
	$q(s,a) = r(s,a) + \sum_{s'} x(s')T(s'\|s,a)$		
Method	$v_{i+1} = U(v_i)$		

- Performance of Sparse and Soft Policy Distributions
  - Policy-regularized MDP always find a worse solution than the original MDP due to the regularization
    - $\mathbb{E}_{\pi^{soft}}[r(s,a)] \leq \mathbb{E}_{\pi^*}[r(s,a)]$
    - $\mathbb{E}_{\pi^{sp}}[r(s,a)] \leq \mathbb{E}_{\pi^*}[r(s,a)]$
  - However, sparse policy π^{sp} has a better lower bound than soft policy π^{soft}

Performance Error Bounds

Theorem 4. Following inequalities hold:

$$\mathbb{E}_{\pi^*}[r(s,a)] - \frac{\alpha}{1-\gamma}\frac{|\mathcal{A}|-1}{2|\mathcal{A}|} \leq \mathbb{E}_{\pi^{sp}}[r(s,a)] \leq \mathbb{E}_{\pi^*}[r(s,a)]$$

where π^* and π^{sp} are the optimal policy obtained by the original MDP and sparse MDP, respectively, and $|\mathcal{A}|$ is the number of actions.

Theorem 5. Following inequalities hold:

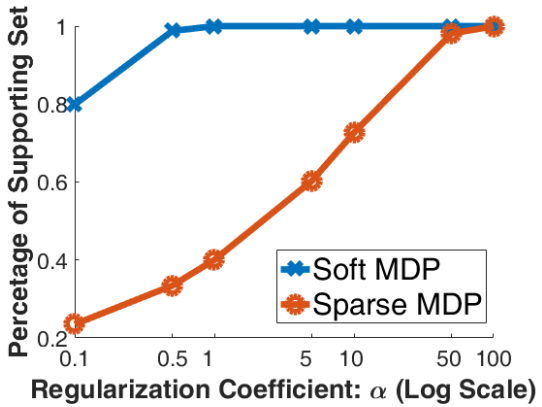
$$\mathbb{E}_{\pi^*}[r(s,a)] - \frac{\alpha}{1-\gamma}\log(|\mathcal{A}|) \leq \mathbb{E}_{\pi^{soft}}[r(s,a)] \leq \mathbb{E}_{\pi^*}[r(s,a)]$$

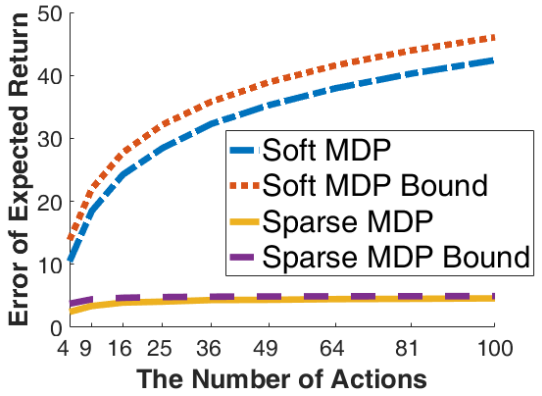
where π^* and π^{soft} are the optimal policy obtained by the original MDP and soft MDP, respectively, and $|\mathcal{A}|$ is the number of actions.

- Sparse MDP has a constant error bound while soft MDP has a logarithmic error bound

Experiment: Supporting Set, Error Bounds

- Discrete MDP problem
  - Create a transition model $P(s'|s,a)$ by discretization of unicycle dynamics
  - The reward function: $r(s) = \exp\left(\frac{|s-x_1|_2^2}{2\sigma_1^2}\right) - \exp\left(\frac{|s-x_2|_2^2}{2\sigma_2^2}\right)$
    - x_1: goal point, x_2: point to avoid
- Supporting set of optimal policy can be controlled by regularization coefficient: α
  - Counting the number of actions with positive probability while α changes
- Performance error bounds
  - Measuring performance gaps while the number of discretization level increases
  - $\left|\mathbb{E}_{\pi^{soft}}[r(s,a)] - \mathbb{E}_{\pi^*}[r(s,a)]\right|$ and $|\mathbb{E}_{\pi^{sp}}[r(s,a)] - \mathbb{E}_{\pi^*}[r(s,a)]|$





Experiment: Reinforcement Learning

- Continuous Control Reinforcement Learning
  - When applying Q-learning to a continuous action space, a finer discretization is necessary to obtain a better solution
  - As the level of discretization increases, the number of actions to be explored becomes larger

Sparse Deep Q Learning

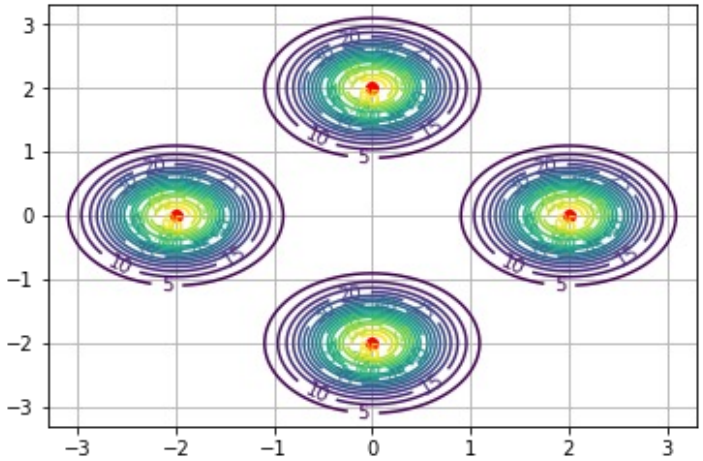
- Sparsemax distribution is used as an exploration method
- Sparse Bellman equation is used as an update rule
- Sparsemax distribution more efficiently explores the action space than other methods: softmax exploration and epsilon greedy

Algorithm:

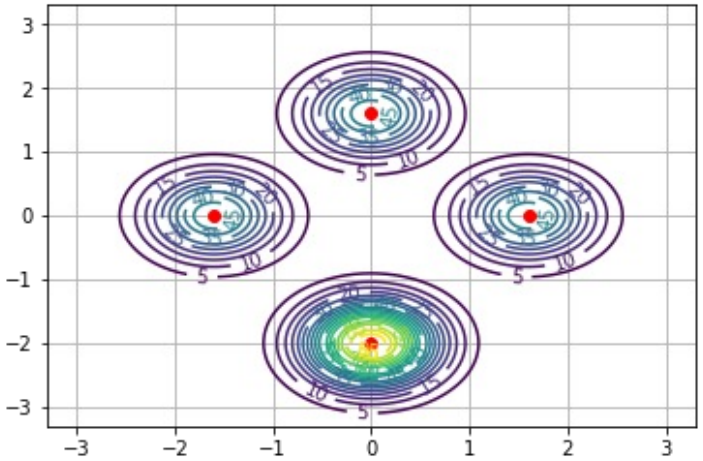
1. Initialize Q network parameters and replay memory M
2. **Exploration using sparsemax distribution $\boldsymbol{\pi^{sp}}$**
3. Add experiences to the replay memory M
4. Sample batch from M
5. Update Q network by minimizing L2 loss: $\sum_i |y_i - Q(s_i, a_i; \theta)|^2$ where $\boldsymbol{y_i = r_i + spmax(Q(s_i, \cdot; \theta))}$
6. Repeat 2~5.

Toy Example

- Multiple goal environments
  - An agent follows point mass dynamics and tries to reach one of distributed multiple modes
  - Two cases: multiple global optima case and multiple local optima case
- We verify that
  - Sparsemax exploration can successfully learn multi-modal optimal actions
  - Sparsemax exploration is helpful to escape locally optimal actions
  - It is compared with epsilon greedy and softmax exploration



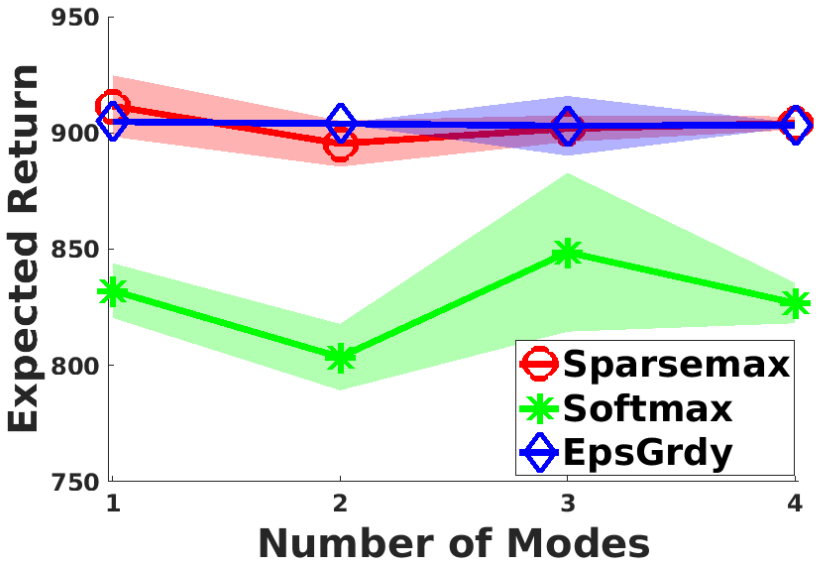
Multiple global optima case

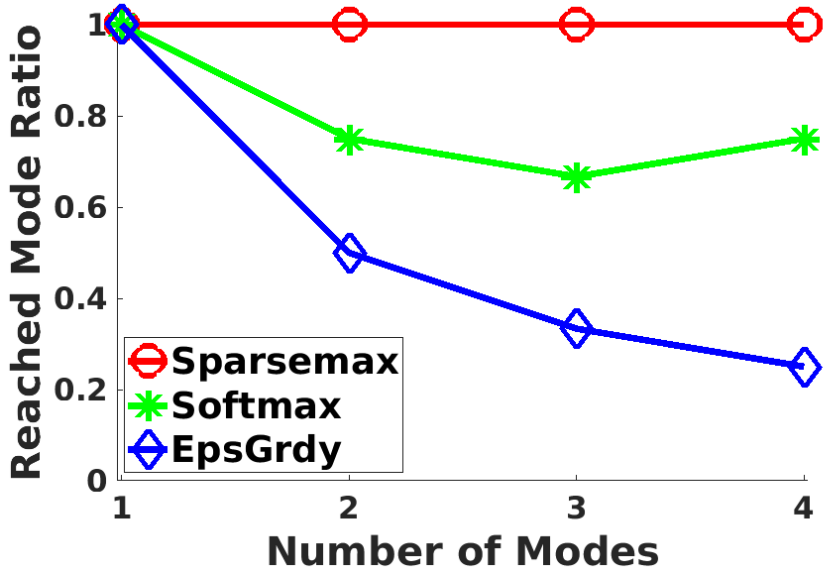


Multiple local optima case

Results: Multiple Global Optima Case

- Sparsemax exploration reaches every modes of multi-modal rewards function
- Softmax exploration shows performance drops since softmax assigns non-zero probability to non optimal actions
- Sparsemax exploration efficiently learns multi-modal optimal actions

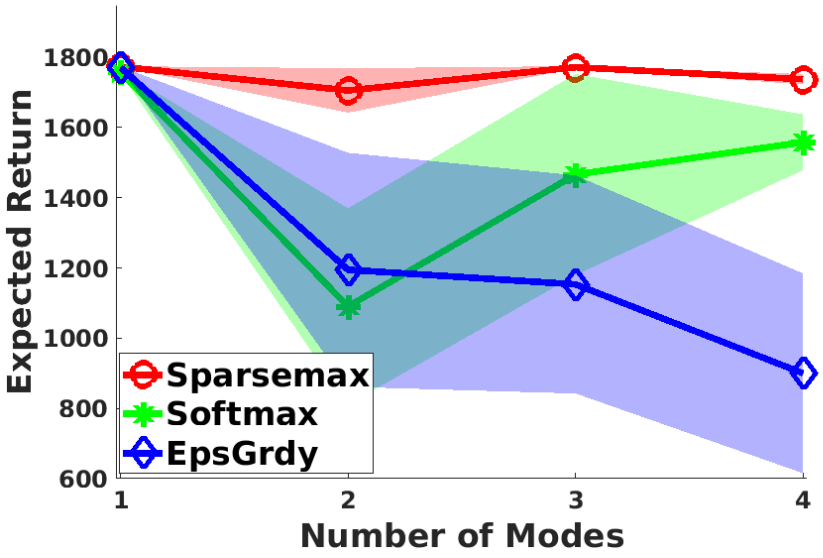


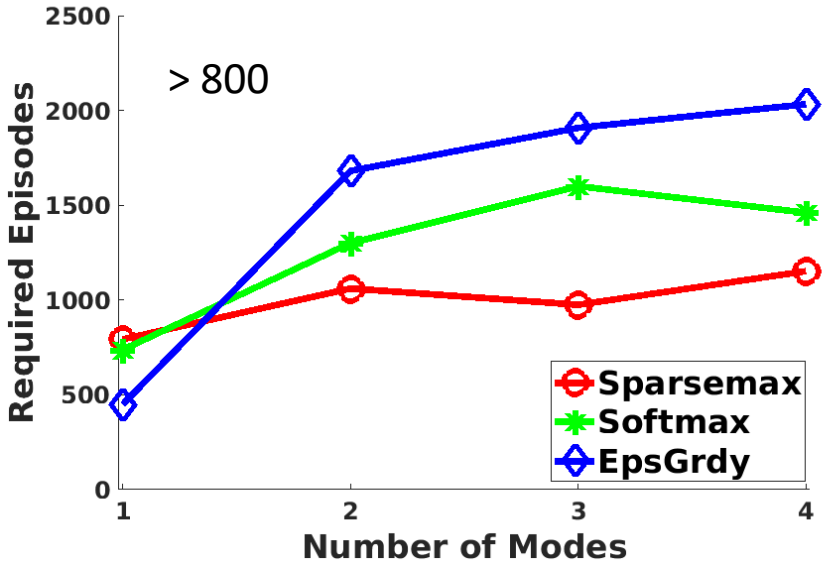


* Averages from 500 test episodes after training with 3000 episodes

Results: Multiple Local Optima Case

- Sparsemax exploration escapes local optima faster than other exploration methods (see required episodes)
- Epsilon greedy method cannot escape local optima when the number of modes increases
- Softmax exploration usually reaches the global optimum but it shows performance drop

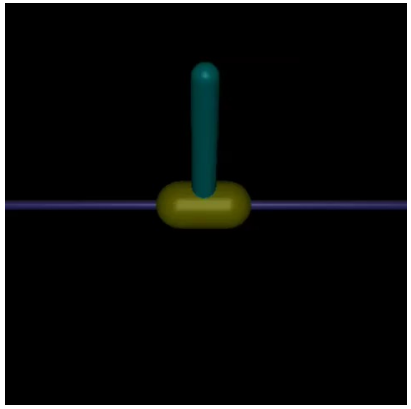




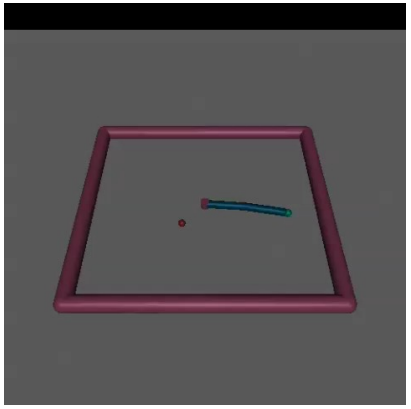
* Averages from 500 test episodes sampled from a greedy policy after training with 3000 episodes

Experiment: Reinforcement Learning

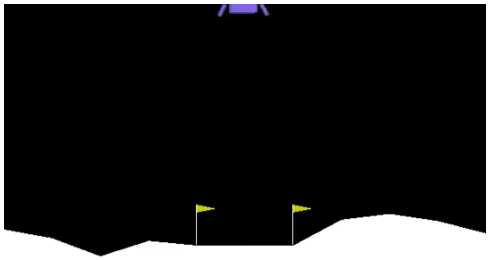
- We test 10 algorithms including our exploration methods on MuJoCo simulator
  - By combining three exploration methods and three Bellman update rules, nine methods are tested
  - Deep deterministic policy gradient (DDPG) is an actor-critic method which can handle continuous action directly
- We measure the performance of each algorithm as the number of actions increases
- Considered problems: Inverted pendulum, Reacher, Lunar Lander, Walker2D



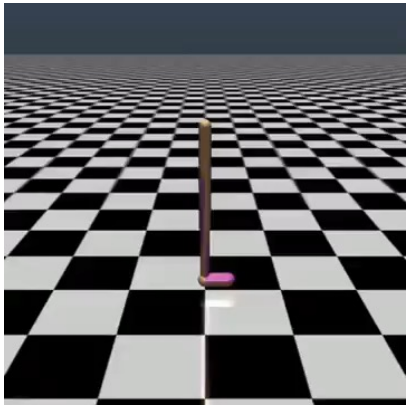
Inverted Pendulum



Reacher



Lunar Lander



Walker2D

Experiment: Reinforcement Learning

- Results

Task	$\lvert\mathcal{A}\rvert$	Sps+SpsB	Sps+SftB	Sps+B	Stf+SpsB	Stf+StfB	Stf+B	Eps+SpsB	Eps+StfB	Eps+B	DDPG
Inv. Pend.	2001^1	1000.0	1000.0	1000.0	1000.0	1000.0	1000.0	1000.0	1000.0	1000.0	1000.0
Reacher	51^2	**-4.9**	-5.5	-5.0	-5.5	-5.6	-5.5	-5.6	-5.6	-5.5	-5.9
Lun. Lan. Cont.	51^2	216.5	**223.4**	215.5	214.7	211.2	212.9	-324.7	-337.1	-349.5	216.5
Walker2D	3^6	1218.9	1189.8	**1853.6**	1580.0	1222.4	1269.5	1416.7	1244.7	690.5	1312.2

Maximum average return with consecutive 100 episodes for five different random seeds

(algorithms are named as <exploration method>+<update rule>)

Task	Threshold	Sps+SpsB	Sps+SftB	Sps+B	Stf+SpsB	Stf+StfB	Stf+B	Eps+SpsB	Eps+StfB	Eps+B	DDPG
Inv. Pend.	980	673	**573**	957	583	835	700	1693	1717	1488	1009
Reacher	-7.0	1155	1205	1256	1363	**1064**	2636	2502	2588	2298	2298
Lun. Lan. Cont.	170.0	574	**397**	480	494	753	529	-	-	-	2213
Walker2D	1000.0	1341	1447	1194	**1149**	1403	1429	1333	1440	-	1388

The number of episodes required to cross a given threshold

  - The sparsemax exploration generally outperforms the other exploration methods with respect to the maximum average return
  - The soft Bellman update rule shows worse maximum average return than other update rules while the soft Bellman update rule generally shows fast reaching speed
  - DDPG shows a slower convergence speed than sparsemax and softmax exploration since training actor and critic networks requires more episodes

Video


CPSLAB
http://cpslab.snu.ac.kr
Sparse Markov Decision Processes with Causal Sparse Tsallis Entropy Regularization for Reinforcement Learning
Kyungjae Lee, Sungjoon Choi, and Songhwai Oh
CPSLAB, ECE
Seoul National University

Conclusion

- Sparse MDPs with causal sparse Tsallis entropy regularization give a sparse solution and multi-modal distribution
- The mathematical analysis of the proposed sparse MDPs
  - Optimality condition of sparse MDPs: sparse Bellman equation
  - Optimal convergence of sparse value iteration
  - Performance error bounds of the optimal solution of sparse and soft MDPs
- Sparsemax exploration shows significantly better performance compared to epsilon-greedy, softmax exploration, and DDPG, as the number of actions increases
- The proposed sparse MDP can be an efficient alternative to problems with a large number of possible actions and even a continuous action space