

Sparse Markov Decision Processes with Causal Sparse Tsallis Entropy Regularization for Reinforcement Learning: Supplementary Material

Kyungjae Lee, Sungjoon Choi, and Songhwai Oh

In this supplementary material, we provide proofs of lemmas and theorems in the main paper and complete experimental results. This material consists of two sections. In Section I, we first derive the sparse Bellman equations and corresponding sparse value iteration method. We also prove the optimality of sparse value iteration and its error bounds. To derive the error bounds of sparse value iteration, we prove the bounds of sparsemax operation. Finally, we show that the error bounds of sparse value iteration is tighter than that of soft value iteration. In Section II, the full experimental results are shown.

I. ANALYSIS

A. Notations and Properties

We first introduce notations and properties used in the paper. In Table I, all notations and definitions are summarized. For notational simplicity, we denote the expectation of a discounted sum, $\mathbb{E}[\sum_{t=0}^{\infty} \gamma^t f(s_t, a_t) | \pi, d, T]$, by $\mathbb{E}_{\pi}[f(s, a)]$, where $f(s, a)$ is a function of a state and an action, such as a rewards function, $\mathbf{r}(s, a)$, or an indicator function, $\mathbb{1}_{\{s'=s, a'=a\}}$. We also denote the expectation conditioned on an initial state, $\mathbb{E}[\sum_{t=0}^{\infty} \gamma^t f(s_t, a_t) | \pi, s_0 = s, T]$, by $\mathbb{E}_{\pi}[f(s, a) | s_0 = s]$. The utility, value, state visitation can be compactly expressed as below in terms of vectors and matrices:

$$\begin{aligned} J_{\pi}^{sp} &= d^{\top} G_{\pi}^{-1} \mathbf{r}_{\pi}^{sp}, & V_{\pi}^{sp} &= G_{\pi}^{-1} \mathbf{r}_{\pi}^{sp} \\ J_{\pi}^{soft} &= d^{\top} G_{\pi}^{-1} \mathbf{r}_{\pi}^{soft}, & V_{\pi}^{soft} &= G_{\pi}^{-1} \mathbf{r}_{\pi}^{soft}, \quad \rho_{\pi} = d^{\top} G_{\pi}^{-1} \end{aligned}$$

where $x^{\top}$ is the transpose of vector x , $G_{\pi} = (I - \gamma T_{\pi})$, sp indicates a sparse MDP problem which is defined as follows:

$$\begin{aligned} &\underset{\pi}{\text{maximize}} \quad \mathbb{E} \left[\sum_{t=0}^{\infty} \gamma^t \mathbf{r}(s_t, a_t) \middle| \pi, d, T \right] + \alpha W(\pi) \\ &\text{subject to} \quad \forall s \quad \sum_{a'} \pi(a'|s) = 1, \\ &\quad \forall s, a \quad \pi(a'|s) \geq 0, \end{aligned} \quad (1)$$

and $soft$ indicates a soft MDP problem which is defined as follows:

$$\begin{aligned} &\underset{\pi}{\text{maximize}} \quad \mathbb{E} \left[\sum_{t=0}^{\infty} \gamma^t \mathbf{r}(s_t, a_t) \middle| \pi, d, T \right] + \alpha H(\pi) \\ &\text{subject to} \quad \forall s \quad \sum_{a'} \pi(a'|s) = 1, \\ &\quad \forall s, a \quad \pi(a'|s) \geq 0. \end{aligned} \quad (2)$$

K. Lee, S. Choi, and S. Oh are with the Department of Electrical and Computer Engineering and ASRI, Seoul National University, Seoul 08826, Korea (e-mail: kyungjae.lee@cplslab.snu.ac.kr; sungjoon.choi@cplslab.snu.ac.kr; songhwai@snu.ac.kr).

B. Sparse Bellman Equation from Karush-Kuhn-Tucker conditions

The following theorem explains the optimality condition of the sparse MDP from Karush-Kuhn-Tucker (KKT) conditions.

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

$$\begin{aligned} Q_{\pi}^{sp}(s, a) &= \mathbf{r}(s, a) + \gamma \sum_{s'} V_{\pi}^{sp}(s') T(s'|s, a) \\ V_{\pi}^{sp}(s) &= \alpha \left[\frac{1}{2} \sum_{a \in S(s)} \left(\left(\frac{Q_{\pi}^{sp}(s, a)}{\alpha} \right)^2 - \tau \left(\frac{Q_{\pi}^{sp}(s, \cdot)}{\alpha} \right)^2 \right) + \frac{1}{2} \right] \\ \pi(a|s) &= \max \left(\frac{Q_{\pi}^{sp}(s, a)}{\alpha} - \tau \left(\frac{Q_{\pi}^{sp}(s, \cdot)}{\alpha} \right), 0 \right) \end{aligned} \quad (3)$$

where $\tau \left(\frac{Q_{\pi}^{sp}(s, \cdot)}{\alpha} \right) = \frac{\sum_{a \in S(s)} \frac{Q_{\pi}^{sp}(s, a)}{\alpha} - 1}{K_s}$ and $S(s)$ is a set of the actions $a_{(i)}$ which has i -th largest action value $\frac{Q_{\pi}^{sp}(s, a_{(i)})}{\alpha}$ and satisfies $1 + i \frac{Q_{\pi}^{sp}(s, a_{(i)})}{\alpha} > \sum_{j=0}^i \frac{Q_{\pi}^{sp}(s, a_{(j)})}{\alpha}$ and K_s is the cardinality of $S(s)$.

Proof. The KKT conditions of (1) are as follows:

$$\begin{aligned} \forall s, a \quad &\sum_{a'} \pi(a'|s) - 1 = 0, \quad -\pi(a|s) \leq 0 \quad (4) \\ \forall s, a \quad &\lambda_{sa} \geq 0 \quad (5) \\ \forall s, a \quad &\lambda_{sa} \pi(a|s) = 0 \quad (6) \\ \forall s, a \quad &\frac{\partial L(\pi, c, \lambda)}{\partial \pi(a|s)} = 0 \quad (7) \end{aligned}$$

where c and λ are Lagrangian multipliers for the equality and inequality constraints, respectively, and (4) is the feasibility of primal variables, (5) is the feasibility of dual variables, (6) is the complementary slackness and (7) is the stationarity condition. The Lagrangian function of (1) is written as follows:

$$\begin{aligned} L(\pi, c, \lambda) &= -J_{\pi}^{sp} + \sum_s c_s \left(\sum_{a'} \pi(a'|s) - 1 \right) - \sum_{s,a} \lambda_{sa} \pi(a|s) \end{aligned}$$

where the maximization of (1) is changed into the minimization problem, i.e., $\min_{\pi} -J_{\pi}^{sp}$. First, the derivative of J_{π}^{sp} can

Terms	sparse MDP	soft MDP
Utility	$J_{\pi}^{sp} \triangleq \mathbb{E}_{\pi} [\mathbf{r}(s', a') + \frac{\alpha}{2}(1 - \pi(a' s'))]$ $= \sum_s d(s) V_{\pi}^{sp}(s) = \sum_s \mathbf{r}_{\pi}^{sp}(s) \rho_{\pi}(s)$	$J_{\pi}^{soft} \triangleq \mathbb{E}_{\pi} [\mathbf{r}(s', a') - \alpha \log(\pi(a' s'))]$ $= \sum_s d(s) V_{\pi}^{soft}(s) = \sum_s \mathbf{r}_{\pi}^{soft}(s) \rho_{\pi}(s)$
Value	$V_{\pi}^{sp}(s)$ $\triangleq \mathbb{E}_{\pi} [\mathbf{r}(s', a') + \frac{\alpha}{2}(1 - \pi(a' s')) s_0 = s]$ $= \mathbf{r}_{\pi}^{sp}(s) + \gamma \sum_{s'} V_{\pi}^{sp}(s') T_{\pi}(s' s)$	$V_{\pi}^{soft}(s)$ $\triangleq \mathbb{E}_{\pi} [\mathbf{r}(s', a') - \alpha \log(\pi(a' s')) s_0 = s]$ $= \mathbf{r}_{\pi}^{soft}(s) + \gamma \sum_{s'} V_{\pi}^{soft}(s') T_{\pi}(s' s)$
Action value	$Q_{\pi}^{sp}(s, a) \triangleq$ $\mathbf{r}(s, a) + \gamma \sum_{s'} V_{\pi}^{sp}(s') T(s' s, a)$	$Q_{\pi}^{soft}(s, a) \triangleq$ $\mathbf{r}(s, a) + \gamma \sum_{s'} V_{\pi}^{soft}(s') T(s' s, a)$
Expected State Reward	$\mathbf{r}_{\pi}^{sp}(s) \triangleq$ $\sum_{a'} (\mathbf{r}(s, a') + \frac{\alpha}{2}(1 - \pi(a' s))) \pi(a' s)$	$\mathbf{r}_{\pi}^{soft}(s) \triangleq$ $\sum_{a'} (\mathbf{r}(s, a') - \alpha \log(\pi(a' s))) \pi(a' s)$
Policy Regularization	$W(\pi) \triangleq \mathbb{E}_{\pi} [\frac{1}{2}(1 - \pi(a s))]$ $= \sum_{s,a} \frac{1}{2}(1 - \pi(a s)) \pi(a s) \rho(s)$	$H(\pi) = \mathbb{E}_{\pi} [-\pi(a s) \log(\pi(a s))]$ $= \sum_{s,a} -\pi(a s) \log(\pi(a s)) \rho(s)$
Max Approximation	$\text{spmax}(z) \triangleq \frac{1}{2} \sum_{i=1}^K (z_{(i)}^2 - \tau(z)^2) + \frac{1}{2}$	$\text{logsumexp}(z) \triangleq \log \sum_i \exp(z_i)$
Value Iteration Operator	$U^{sp}(x)(s) = \alpha \text{spmax} \left(\frac{\mathbf{r}(s, \cdot) + \gamma \sum_{s'} x(s') T(s' s, \cdot)}{\alpha} \right)$	$U^{soft}(x)(s) = \alpha \text{logsumexp} \left(\frac{\mathbf{r}(s, \cdot) + \gamma \sum_{s'} x(s') T(s' s, \cdot)}{\alpha} \right)$
State Visitation	$\rho_{\pi}(s) \triangleq \mathbb{E}_{\pi} [\mathbb{1}_{\{s'=s\}}] = d(s) + \gamma \sum_{s', a'} T(s s', a') \rho_{\pi}(s', a')$	
State Action Visitation	$\rho_{\pi}(s, a) \triangleq \mathbb{E}_{\pi} [\mathbb{1}_{\{s'=s, a'=a\}}] = \pi(a s) d(s) + \gamma \sum_{s', a'} \pi(a s) T(s s', a') \rho_{\pi}(s', a')$	
Transition Probability given π	$T_{\pi}(s' s) \triangleq \sum_a T(s' s, a) \pi(a s)$	

TABLE I: Notations and Properties

be obtained by using the chain rule.

$$\begin{aligned}
\frac{\partial J_{\pi}}{\partial \pi(a|s)} &= d^{\top} G_{\pi}^{-1} \frac{\partial \mathbf{r}_{\pi}^{sp}}{\partial \pi(a|s)} + \gamma d^{\top} G_{\pi}^{-1} \frac{\partial T_{\pi}}{\partial \pi(a|s)} G_{\pi}^{-1} \mathbf{r}_{\pi}^{sp} \\
&= \rho_{\pi}^{\top} \frac{\partial \mathbf{r}_{\pi}^{sp}}{\partial \pi(a|s)} + \gamma \rho_{\pi}^{\top} \frac{\partial T_{\pi}}{\partial \pi(a|s)} V_{\pi}^{sp} \\
&= \rho_{\pi}(s) \left(\mathbf{r}(s, a) + \frac{\alpha}{2} - \alpha \pi(a|s) + \gamma \sum_{s'} V_{\pi}^{sp}(s') T(s'|s, a) \right) \\
&= \rho_{\pi}(s) \left(Q_{\pi}^{sp}(s, a) + \frac{\alpha}{2} - \alpha \pi(a|s) \right).
\end{aligned}$$

Here, the partial derivative of Lagrangian is obtained as follows:

$$\begin{aligned}
\frac{\partial L(\pi, c, \lambda)}{\partial \pi(a|s)} \\
= -\rho_{\pi}(s) (Q_{\pi}^{sp}(s, a) + \frac{\alpha}{2} - \alpha \pi(a|s)) + c_s - \lambda_{sa} = 0.
\end{aligned}$$

First, consider a positive $\pi(a|s)$ where the corresponding Lagrangian multiplier λ_{sa} is zero due to the complementary slackness. By summing $\pi(a|s)$ with respect to action a , Lagrangian multiplier c_s can be obtained as follows:

$$\begin{aligned}
0 &= -\rho_{\pi}(s) (Q_{\pi}^{sp}(s, a) + \frac{\alpha}{2} - \alpha \pi(a|s)) + c_s \\
\pi(a|s) &= \left(-\frac{c_s}{\rho_{\pi}(s)\alpha} + \frac{1}{2} + \frac{Q_{\pi}^{sp}(s, a)}{\alpha} \right) \\
\sum_{\pi(a'|s) > 0} \pi(a'|s) &= \sum_{\pi(a'|s) > 0} \left(-\frac{c_s}{\rho_{\pi}(s)\alpha} + \frac{1}{2} + \frac{Q_{\pi}^{sp}(s, a')}{\alpha} \right) = 1 \\
\therefore c_s &= \rho_{\pi}(s) \alpha \left[\frac{\sum_{\pi(a'|s) > 0} \frac{Q_{\pi}^{sp}(s, a')}{\alpha} - 1}{K} + \frac{1}{2} \right]
\end{aligned}$$

where K is the number of positive elements of $\pi(\cdot|s)$. By replacing c_s with this result, the optimal policy distribution

is induced as follows.

$$\begin{aligned}
\pi(a|s) &= \left(-\frac{c_s}{\rho_{\pi}(s)\alpha} + \frac{1}{2} + \frac{Q_{\pi}^{sp}(s, a)}{\alpha} \right) \\
&= \frac{Q_{\pi}^{sp}(s, a)}{\alpha} - \frac{\sum_{\pi(a'|s) > 0} \frac{Q_{\pi}^{sp}(s, a')}{\alpha} - 1}{K}
\end{aligned}$$

As this equation is derived under the assumption that $\pi(a|s)$ is positive. For $\pi(a|s) > 0$, following condition is necessarily fulfilled,

$$\frac{Q_{\pi}^{sp}(s, a)}{\alpha} > \frac{\sum_{\pi(a'|s) > 0} \frac{Q_{\pi}^{sp}(s, a')}{\alpha} - 1}{K}.$$

We notate this supporting set as $S(s) = \{a | 1 + K \frac{Q_{\pi}^{sp}(s, a)}{\alpha} > \sum_{\pi(a'|s) > 0} \frac{Q_{\pi}^{sp}(s, a')}{\alpha}\}$. $S(s)$ contains the actions which has larger action values than threshold

$$\tau(Q_{\pi}^{sp}(s, \cdot)) = \frac{\sum_{\pi(a'|s) > 0} \frac{Q_{\pi}^{sp}(s, a')}{\alpha} - 1}{K}.$$

By using these notations, the optimal policy distribution can be rewritten as follows:

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

By substituting $\pi(a|s)$ with this result, the following opti-

ality equation of V_π^{sp} is induced.

$$\begin{aligned}
V_\pi^{sp}(s) &= \sum_a \pi(a|s) \left(Q_\pi^{sp}(s, a) + \frac{\alpha}{2} (1 - \pi(a|s)) \right) \\
&= \sum_a \pi(a|s) \left(Q_\pi^{sp}(s, a) - \frac{\alpha}{2} \pi(a|s) \right) + \frac{\alpha}{2} \sum_a \pi(a|s) \\
&= \sum_{a \in S(s)} \pi(a|s) \\
&\quad \times \left(Q_\pi^{sp}(s, a) - \frac{\alpha}{2} \left(\frac{Q_\pi^{sp}(s, a)}{\alpha} - \tau \left(\frac{Q_\pi^{sp}(s, \cdot)}{\alpha} \right) \right) \right) + \frac{\alpha}{2} \\
&= \sum_{a \in S(s)} \pi(a|s) \frac{\alpha}{2} \left(\frac{Q_\pi^{sp}(s, a)}{\alpha} + \tau \left(\frac{Q_\pi^{sp}(s, \cdot)}{\alpha} \right) \right) + \frac{\alpha}{2} \\
&= \alpha \left[\frac{1}{2} \sum_{a \in S(s)} \left(\left(\frac{Q_\pi^{sp}(s, a)}{\alpha} \right)^2 - \tau \left(\frac{Q_\pi^{sp}(s, \cdot)}{\alpha} \right)^2 \right) + \frac{1}{2} \right]
\end{aligned}$$

To summarize, we obtain the sparse Bellman equation as follows:

$$\begin{aligned}
Q_\pi^{sp}(s, a) &= \mathbf{r}(s, a) + \gamma \sum_{s'} V_\pi^{sp}(s') T(s'|s, a) \\
V_\pi^{sp}(s) &= \alpha \left[\frac{1}{2} \sum_{a \in S(s)} \left(\left(\frac{Q_\pi^{sp}(s, a)}{\alpha} \right)^2 - \tau \left(\frac{Q_\pi^{sp}(s, \cdot)}{\alpha} \right)^2 \right) + \frac{1}{2} \right] \\
\pi(a|s) &= \max \left(\frac{Q_\pi^{sp}(s, a)}{\alpha} - \tau \left(\frac{Q_\pi^{sp}(s, \cdot)}{\alpha} \right), 0 \right).
\end{aligned}$$

□

C. Causal Sparse Tsallis Entropy

In this section, the connection between $W(\pi)$ and Tsallis entropy is explained. The Tsallis entropy is defined as follows:

$$S_{q,k}(p) = \frac{k}{q-1} \left(1 - \sum_i p_i^q \right),$$

where p is a probability mass function, q is a parameter called *entropic-index*, and k is a positive real constant.

The following theorem shows that $W(\pi)$ is equivalent to the discounted expected sum of special case of Tsallis entropy when $q = 2$ and $k = \frac{1}{2}$.

Theorem 2. The proposed policy regularization $W(\pi)$ is an extension of the Tsallis entropy with parameters $q = 2$ and $k = \frac{1}{2}$ to the version of causal entropy, i.e.,

$$W(\pi) = \mathbb{E}_\pi [S_{2, \frac{1}{2}}(\pi(\cdot|s))].$$

Proof. The proof is simply done by rewriting our regular-

ization as follows:

$$\begin{aligned}
W(\pi) &= \mathbb{E} \left[\sum_{t=0}^{\infty} \gamma^t \frac{1}{2} (1 - \pi(a_t|s_t)) \middle| \pi, d, T \right] \\
&= \sum_{s,a} \frac{1}{2} (1 - \pi(a|s)) \mathbb{E} \left[\sum_{t=0}^{\infty} \gamma^t \mathbb{1}_{\{s_t=s, a_t=a\}} \middle| \pi, d, T \right] \\
&= \sum_{s,a} \frac{1}{2} (1 - \pi(a|s)) \rho_\pi(s, a) \\
&= \sum_s \rho_\pi(s) \sum_a \frac{1}{2} (1 - \pi(a|s)) \pi(a|s) \\
&= \sum_s \rho_\pi(s) \frac{1}{2} \left(\sum_a \pi(a|s) - \sum_a \pi(a|s)^2 \right) \\
&= \sum_s \rho_\pi(s) \frac{1}{2} \left(1 - \sum_a \pi(a|s)^2 \right) \\
&= \sum_s S_{2, \frac{1}{2}}(\pi(\cdot|s)) \rho_\pi(s) = \mathbb{E}_\pi [S_{2, \frac{1}{2}}(\pi(\cdot|s))].
\end{aligned}$$

□

D. Upper and Lower Bounds for Sparsemax Operation

In this section, we prove the lower and upper bounds of $\text{spmax}(z)$ defined as

$$\text{spmax}(z) \triangleq \frac{1}{2} \sum_{i=1}^K \left(z_{(i)}^2 - \tau(z)^2 \right) + \frac{1}{2}. \quad (8)$$

Note that definitions of K and τ are the same as in Theorem 1.

The lower bound and upper bound of $\text{spmax}(z)$ is as follows,

$$\max(z) \leq \alpha \text{spmax}\left(\frac{z}{\alpha}\right) \leq \max(z) + \alpha \frac{d-1}{2d}. \quad (9)$$

Note that the proof of lower bound of (9) is provided in [1]. However, we find another interesting way to prove (9) by using the Cauchy-Schwartz inequality and the nonnegative property of a quadratic equation.

We first prove $\max(z) \leq \text{spmax}(z)$ and next prove $\text{spmax}(z) \leq \max(z) + \frac{d-1}{2d}$. For simplicity of derivation, we assume that $\alpha = 1$ but the original inequalities can be simply obtained by replacing z with $\frac{z}{\alpha}$.

Lower Bound of SparseMax Operation. For all $z \in \mathbb{R}^d$, $\max(z) \leq \text{spmax}(z)$ holds.

Proof. We prove that, for all z , $\text{spmax}(z) - z_{(1)} \geq 0$ where $z_{(1)} = \max(z)$ by definition. The proof is done by simply

rearranging the terms in (8),

$$\begin{aligned}
& \text{spxmax}(z) - z_{(1)} \\
&= \frac{1}{2} \sum_{i=1}^K (z_{(i)}^2 - \tau(z)^2) + \frac{1}{2} - z_{(1)} \\
&= \frac{1}{2} \sum_{i=1}^K z_{(i)}^2 - \frac{K}{2} \left(\frac{\sum_{i=1}^K z_{(i)} - 1}{K} \right)^2 + \frac{1}{2} - z_{(1)} \\
&= \frac{1}{2} \sum_{i=1}^K z_{(i)}^2 - \frac{1}{2K} \left(\sum_{i=1}^K z_{(i)} - 1 \right)^2 + \frac{1}{2} - z_{(1)} \\
&= \frac{K \sum_{i=1}^K z_{(i)}^2 - \left(\sum_{i=1}^K z_{(i)} - 1 \right)^2 - 2Kz_{(1)} + K}{2K} \\
&= \frac{1}{2K} \left(Kz_{(1)}^2 + K \sum_{i=2}^K z_{(i)}^2 \right. \\
&\quad \left. - \left(z_{(1)} + \sum_{i=2}^K z_{(i)} - 1 \right)^2 - 2Kz_{(1)} + K \right).
\end{aligned}$$

The quadratic term can be decomposed as follows:

$$\begin{aligned}
& \left(z_{(1)} + \sum_{i=2}^K z_{(i)} - 1 \right)^2 \\
&= z_{(1)}^2 + \left(\sum_{i=2}^K z_{(i)} \right)^2 + 1 + 2z_{(1)} \sum_{i=2}^K z_{(i)} - 2z_{(1)} - 2 \sum_{i=2}^K z_{(i)}.
\end{aligned}$$

By putting this result into the equation and rearranging them, three terms are obtained as follows:

$$\begin{aligned}
& \text{spxmax}(z) - z_{(1)} \\
&= \frac{1}{2K} \left((K-1)z_{(1)}^2 - 2z_{(1)} \left\{ \sum_{i=2}^K z_{(i)} + K - 1 \right\} \right. \\
&\quad \left. + K \sum_{i=2}^K z_{(i)}^2 + 2 \sum_{i=2}^K z_{(i)} + K - \left(\sum_{i=2}^K z_{(i)} \right)^2 \right).
\end{aligned}$$

Then, $K \sum_{i=2}^K z_{(i)}^2 + 2 \sum_{i=2}^K z_{(i)} + K$ can be replaced with $K \sum_{i=2}^K (z_{(i)} + 1)^2 - 2(K-1) \sum_{i=2}^K z_{(i)}$ and we also decompose the second term $-2z_{(1)} \left\{ \sum_{i=2}^K z_{(i)} + K - 1 \right\}$ into two parts: $-2z_{(1)} \left\{ \sum_{i=2}^K (z_{(i)} + 1) \right\}$ and $2z_{(1)}$, and rearrange the equation as follows,

$$\begin{aligned}
&= \frac{1}{2K} \left((K-1)z_{(1)}^2 - 2z_{(1)} \left\{ \sum_{i=2}^K (z_{(i)} + 1) \right\} \right. \\
&\quad \left. + K \sum_{i=2}^K (z_{(i)} + 1)^2 - 2(K-1) \sum_{i=2}^K z_{(i)} - \left(\sum_{i=2}^K z_{(i)} \right)^2 \right).
\end{aligned}$$

Again, we change $-2(K-1) \sum_{i=2}^K z_{(i)} - \left(\sum_{i=2}^K z_{(i)} \right)^2$ into $-\left(\sum_{i=2}^K (z_{(i)} + 1) \right)^2 + (K-1)^2$ by adding and subtracting $(K-1)^2$ as follow,

$$\begin{aligned}
&= \frac{1}{2K} \left((K-1)z_{(1)}^2 - 2z_{(1)} \left\{ \sum_{i=2}^K (z_{(i)} + 1) \right\} \right. \\
&\quad \left. + K \sum_{i=2}^K (z_{(i)} + 1)^2 - \left(\sum_{i=2}^K (z_{(i)} + 1) \right)^2 + (K-1)^2 \right).
\end{aligned}$$

Then, the term $(K-1)z_{(1)}^2 - 2z_{(1)} \left\{ \sum_{i=2}^K (z_{(i)} + 1) \right\}$ is reformulated as $(K-1) \left(z_{(1)} - \frac{\sum_{i=2}^K (z_{(i)} + 1)}{K-1} \right)^2 - (K-1) \left(\frac{\sum_{i=2}^K (z_{(i)} + 1)}{K-1} \right)^2$. By using this reformulation, we can obtain following equation.

$$\begin{aligned}
&= \frac{(K-1)}{2K} \left[z_{(1)} - \frac{\sum_{i=2}^K (z_{(i)} + 1)}{K-1} \right]^2 + \\
&\quad \frac{1}{2K} \left(- \frac{\left(\sum_{i=2}^K (z_{(i)} + 1) \right)^2}{K-1} + K \sum_{i=2}^K (z_{(i)} + 1)^2 - \left(\sum_{i=2}^K (z_{(i)} + 1) \right)^2 \right. \\
&\quad \left. + (K-1)^2 \right).
\end{aligned}$$

Finally, we can obtain three terms by rearranging the above equation,

$$\begin{aligned}
&= \frac{(K-1)}{2K} \left[z_{(1)} - \frac{\sum_{i=2}^K (z_{(i)} + 1)}{K-1} \right]^2 \\
&\quad + \frac{1}{2K} \left(K \sum_{i=2}^K (z_{(i)} + 1)^2 - K \frac{\left(\sum_{i=2}^K (z_{(i)} + 1) \right)^2}{K-1} \right) + \frac{(K-1)^2}{2K} \\
&= \frac{(K-1)}{2K} \left[z_{(1)} - \frac{\sum_{i=2}^K (z_{(i)} + 1)}{K-1} \right]^2 \\
&\quad + \frac{K-1}{2} \left[\sum_{i=2}^K \frac{(z_{(i)} + 1)^2}{K-1} - \left(\sum_{i=2}^K \frac{(z_{(i)} + 1)}{K-1} \right)^2 \right] + \frac{(K-1)^2}{2K}
\end{aligned}$$

where the first and third terms are quadratic and always nonnegative. The second term is also always nonnegative by the Cauchy-Schwartz inequality. The Cauchy-Schwartz inequality is written as $(\mathbf{p}^\top \mathbf{q})^2 \leq \|\mathbf{p}\|^2 \|\mathbf{q}\|^2$. Let $z_{2:K} = [z_{(2)}, \dots, z_{(K)}]^\top$, then, by setting $\mathbf{p} = z_{2:K} + \mathbf{1}$ and $\mathbf{q} = \frac{1}{K-1} \mathbf{1}$ where $\mathbf{1}$ is a $K-1$ dimensional vector of ones, it can be shown that the second term is nonnegative. Therefore, $\text{spxmax}(z) - z_{(1)}$ is always nonnegative for all z since three remaining terms are always nonnegative, completing the proof. $\square$

Now, we prove the upper bound of sparsemax operation.

Upper Bound of SparseMax Operation. For all $z \in \mathbb{R}^d$, $\text{spxmax}(z) \leq \max(z) + \frac{d-1}{2d}$ holds.

Proof. First, we decompose the summation of (8) into two

terms as follows:

$$\begin{aligned}
\text{spmax}(z) &= \frac{1}{2} \sum_{i=1}^K \left(z_{(i)}^2 - \tau(z)^2 \right) + \frac{1}{2} \\
&= \frac{1}{2} \sum_{i=1}^K (z_{(i)} - \tau(z)) (z_{(i)} + \tau(z)) + \frac{1}{2} \\
&\leq \frac{1}{2} \sum_{i=1}^K p_i^*(z) (z_{(i)} + \tau(z)) + \frac{1}{2} \\
&= \frac{1}{2} \sum_{i=1}^K p_i^*(z) z_{(i)} + \frac{\tau(z)}{2} \sum_{i=1}^K p_i^*(z) + \frac{1}{2} \\
&= \frac{1}{2} \sum_{i=1}^K p_i^*(z) z_{(i)} + \frac{\tau(z)}{2} + \frac{1}{2} \\
&= \frac{1}{2} \sum_{i=1}^K p_i^*(z) z_{(i)} + \frac{1}{2} \sum_{i=1}^K \frac{z_{(i)}}{K} - \frac{1}{2K} + \frac{1}{2}
\end{aligned}$$

where $p_i^* = \max(z_{(i)} - \tau(z), 0)$ which is the optimal solution of the simplex projection problem and $\sum_{i=1}^K p_i^*(z) = 1$ by definition. Now, we use the fact that, for every p on $d-1$ dimensional simplex, $\sum_{i=1}^d p_i z_i \leq \max(z)$ for all $z \in \mathbb{R}^d$. By using this property, as $p^*(z)$ and $\frac{1}{K}\mathbb{1}$ are on the probability simplex, following inequality is induced,

$$\begin{aligned}
\text{spmax}(z) &= \frac{1}{2} \sum_{i=1}^K p_i^*(z) z_{(i)} + \frac{1}{2} \sum_{i=1}^K \frac{z_{(i)}}{K} - \frac{1}{2K} + \frac{1}{2} \\
&\leq \frac{1}{2} \max(z) + \frac{1}{2} \max(z) - \frac{1}{2K} + \frac{1}{2} \leq \max(z) - \frac{1}{2K} + \frac{1}{2} \\
&\leq \max(z) - \frac{1}{2d} + \frac{1}{2}
\end{aligned}$$

where $d \geq K$ by definition of K . Therefore, $\text{spmax}(z) \leq \max(z) + \frac{d-1}{2d}$ holds. $\square$

E. Comparison to Log-Sum-Exp

We explain the error bounds for the *log-sum-exp* operation and compare it to the bounds of the *sparsemax* operation. The *log-sum-exp* operation has widely known bounds,

$$\max(z) \leq \text{logsumexp}(z) \leq \max(z) + \log(d).$$

We would like to note that *sparsemax* has tighter bounds than *log-sum-exp* as it is always satisfied that, for all $d > 1$, $\frac{d-1}{2d} \leq \log(d)$. Intuitively, the approximation error of *log-sum-exp* increases as the dimension of input space increases. However, the approximation error of *sparsemax* approaches to $\frac{1}{2}$ as the dimension of input space goes infinity. This fact plays a crucial role in comparing performance error bounds of the sparse MDP and soft MDP.

F. Convergence and Optimality of Sparse Value Iteration

In this section, the monotonicity, discounting property, contraction of sparse Bellman operation U^{sp} are proved.

Lemma 1. U^{sp} is monotone: if $x \leq y$, $U^{sp}(x) \leq U^{sp}(y)$, where $\leq$ indicates the element-wise inequality.

Proof. In [1], the monotonicity of (8) is proved. Then, the monotonicity of U^{sp} can be proved using (8). Let x and y are given such that $x \leq y$. Then,

$$\frac{\mathbf{r}(s, a) + \gamma \sum_{s'} x(s') T(s'|s, a)}{\alpha} \leq \frac{\mathbf{r}(s, a) + \gamma \sum_{s'} y(s') T(s'|s, a)}{\alpha}$$

where $T(s'|s, a)$ is a transition probability which is always nonnegative. Since the *sparsemax* operation is monotone, the following inequality is induced

$$\begin{aligned}
&\alpha \text{spmax} \left(\frac{\mathbf{r}(s, a) + \gamma \sum_{s'} x(s') T(s'|s, a)}{\alpha} \right) \\
&\leq \alpha \text{spmax} \left(\frac{\mathbf{r}(s, a) + \gamma \sum_{s'} y(s') T(s'|s, a)}{\alpha} \right).
\end{aligned}$$

Finally, we can obtain

$$\therefore U^{sp}(x) \leq U^{sp}(y).$$

$\square$

Lemma 2. For any constant $c \in \mathbb{R}$, $U^{sp}(x + c\mathbb{1}) = U^{sp}(x) + \gamma c\mathbb{1}$ where $\mathbb{1} \in \mathbb{R}^{|S|}$.

Proof. In [1], it is shown that for $c \in \mathbb{R}$ and $x \in \mathbb{R}^{|S|}$, $\text{spmax}(x + c\mathbb{1}) = \text{spmax}(x) + c\mathbb{1}$. Using this property,

$$\begin{aligned}
U^{sp}(x + c\mathbb{1})(s) &= \alpha \text{spmax} \left(\frac{\mathbf{r}(s, a) + \gamma \sum_{s'} (x(s') + c) T(s'|s, a)}{\alpha} \right) \\
&= \alpha \text{spmax} \left(\frac{\mathbf{r}(s, a) + \gamma \sum_{s'} x(s') T(s'|s, a) + \gamma c \sum_{s'} T(s'|s, a)}{\alpha} \right) \\
&= \alpha \text{spmax} \left(\frac{\mathbf{r}(s, a) + \gamma \sum_{s'} x(s') T(s'|s, a)}{\alpha} + \frac{\gamma c}{\alpha} \right) \\
&= \alpha \text{spmax} \left(\frac{\mathbf{r}(s, a) + \gamma \sum_{s'} x(s') T(s'|s, a)}{\alpha} \right) + \gamma c \\
&\therefore U^{sp}(x + c\mathbb{1}) = U^{sp}(x) + \gamma c\mathbb{1}.
\end{aligned}$$

$\square$

Lemma 3. U^{sp} is a γ -contraction mapping with respect to the infinite norm d_{max} and has a unique fixed point.

Proof. First, we prove that U^{sp} is a γ -contraction mapping with respect to d_{max} . Without loss of generality, the proof is discussed for a general function $\phi : \mathbb{R}^{|S|} \rightarrow \mathbb{R}^{|S|}$ with discounting and monotone properties.

Let $d_{max}(x, y) = M$. Then, $y - M\mathbb{1} \leq x \leq y + M\mathbb{1}$ is satisfied. By monotone and discounting properties, the following inequality between mappings $\phi(x)$ and $\phi(y)$ is established.

$$\phi(y) - \gamma M\mathbb{1} \leq \phi(x) \leq \phi(y) + \gamma M\mathbb{1},$$

where γ is a discounting factor of ϕ . From this inequality, $d_{max}(\phi(x), \phi(y)) \leq \gamma M = \gamma d_{max}(x, y)$ and $\gamma \in (0, 1)$. Therefore, ϕ is a γ -contraction mapping. In our case, U^{sp} is a γ -contraction mapping.

As $\mathbb{R}^{|S|}$ and $d_{max}(x, y)$ are a non-empty complete metric space, by Banach fixed-point theorem, a γ -contraction mapping U^{sp} has a unique fixed point. $\square$

Using Lemma 1, Lemma 2, and Lemma 3, we can prove the convergence and optimality of sparse value iteration.

Theorem 3. Sparse value iteration converges into the optimal value of (1).

Proof. Sparse value iteration converges into a fixed point of U^{sp} by the contraction property. Let x_* be a fixed point of U^{sp} and, by definition of U^{sp} , x_* is the point that satisfies the sparse Bellman equation, i.e. $x_* = U^{sp}(x_*)$. Hence, by Theorem 1, x_* satisfies necessity conditions of the optimal solution. By the Banach fixed point theorem, x_* is a unique point which satisfies necessity conditions of optimal solution. In particular, $x_* = U^{sp}(x_*)$ is precisely equivalent to the sparse Bellman equation. In other words, there is no other point that satisfies the sparse Bellman equation. Therefore, x_* is the optimal value of a sparse MDP. $\square$

G. Performance Error Bounds for Sparse Value Iteration

In this section, we prove the performance error bounds for sparse value iteration and soft value iteration. We first show that the optimal values of a sparse MDP and a soft MDP are greater than that of the original MDP.

Lemma 4. Let U and U^{soft} be the Bellman operations of the original MDP and a soft MDP, respectively, such that, for state s and $x \in \mathbb{R}^{|S|}$,

$$U(x)(s) = \max_{a'} \left(\mathbf{r}(s, a') + \gamma \sum_{s'} x(s') T(s'|s, a') \right)$$

$$U^{soft}(x)(s) = \alpha \log \sum_{a'} \exp \left(\frac{\mathbf{r}(s, a') + \gamma \sum_{s'} x(s') T(s'|s, a')}{\alpha} \right).$$

Then following inequalities hold for every positive integer n :

$$U^n(x) \leq (U^{sp})^n(x), \quad U^n(x) \leq (U^{soft})^n(x),$$

where U^n (resp., $(U^{sp})^n$) is the result after applying U (resp., U^{sp}) n times. In addition, let x_* , x_*^{sp} and x_*^{soft} be the fixed point of U , U^{sp} and U^{soft} , respectively. Then, following inequalities also hold:

$$x_* \leq x_*^{sp}, \quad x_* \leq x_*^{soft}.$$

Proof. We first prove the inequality of the sparse Bellman operation

$$U^n(x) \leq (U^{sp})^n(x), \quad x_* \leq x_*^{sp}.$$

This inequality can be proven by the mathematical induction. When $n = 1$, the inequality is proven as follows:

$$\begin{aligned} & \max_{a'} (\mathbf{r}(s, a') + \gamma \sum_{s'} x(s') T(s'|s, a')) \\ & \leq \text{spmax}(\mathbf{r}(s, \cdot) + \gamma \sum_{s'} x(s') T(s'|s, \cdot)) \\ & (\because \max(z) \leq \text{spmax}(z)). \end{aligned}$$

Therefore,

$$U(x) \leq U^{sp}(x).$$

For some positive integer k , let us assume that $U^k(x) \leq (U^{sp})^k(x)$ holds for every $x \in \mathbb{R}^{|S|}$. Then, when $n = k + 1$,

$$\begin{aligned} U^{k+1}(x) &= U^k(U(x)) \\ &\leq (U^{sp})^k(U(x)) \quad (\because U^k(x) \leq (U^{sp})^k(x)) \\ &\leq (U^{sp})^k(U^{sp}(x)) \quad (\because U(x) \leq U^{sp}(x)) \\ &= (U^{sp})^{k+1}(x). \end{aligned}$$

Therefore, by mathematical induction, it is satisfied $U^n(x) \leq (U^{sp})^n(x)$ for every positive integer n . Then, the inequality of the fixed points of U and U^{sp} can be obtained by $n \rightarrow \infty$,

$$x_* \leq x_*^{sp}$$

where $*$ indicates the fixed point. The above arguments also hold when U^{sp} and *sparsemax* are replaced with U^{soft} and *log-sum-exp* operation, respectively. $\square$

Before showing the performance error bounds, the upper bounds of $W(\pi)$ and $H(\pi)$ are proved first.

Lemma 5. $W(\pi)$ and $H(\pi)$ have following upper bounds:

$$W(\pi) \leq \frac{1}{1-\gamma} \frac{|\mathcal{A}| - 1}{2|\mathcal{A}|}, \quad H(\pi) \leq \frac{\log(|\mathcal{A}|)}{1-\gamma},$$

where $|\mathcal{A}|$ is the cardinality of the action space $\mathcal{A}$.

Proof. For $W(\pi)$,

$$\begin{aligned} W(\pi) &= \sum_s \rho_\pi(s) \sum_a \frac{1}{2} (1 - \pi(a|s)) \pi(a|s) \\ &\leq \sum_s \rho_\pi(s) \frac{|\mathcal{A}| - 1}{2|\mathcal{A}|} \quad (\because \sum_a \frac{1}{2} (1 - \pi(a|s)) \pi(a|s) \leq \frac{|\mathcal{A}| - 1}{2|\mathcal{A}|}) \\ &= \frac{1}{1-\gamma} \frac{|\mathcal{A}| - 1}{2|\mathcal{A}|} \quad (\because \sum_s \rho_\pi(s) = \frac{1}{1-\gamma}). \end{aligned}$$

The inequality that $\sum_a \frac{1}{2} (1 - \pi(a|s)) \pi(a|s) \leq \frac{|\mathcal{A}| - 1}{2|\mathcal{A}|}$ can be obtained by finding the point where the derivative of $\frac{1}{2} (1 - x)x$ is zero. Similarly, for $H(\pi)$,

$$\begin{aligned} H(\pi) &= \mathbb{E} \left[\sum_{t=0}^{\infty} \gamma^t - \log(\pi(a_t|s_t)) \middle| \pi, d, T \right] \\ &= \sum_{s,a} -\log(\pi(a|s)) \mathbb{E} \left[\sum_{t=0}^{\infty} \gamma^t \mathbb{1}_{\{s_t=s, a_t=a\}} \middle| \pi, d, T \right] \\ &= \sum_{s,a} -\log(\pi(a|s)) \rho_\pi(s, a) \\ &= \sum_s \rho_\pi(s) \sum_a -\log(\pi(a|s)) \pi(a|s) \\ &\leq \sum_s \rho_\pi(s) \log(|\mathcal{A}|) \quad (\because \sum_a -\log(\pi(a|s)) \pi(a|s) \leq \log(|\mathcal{A}|)) \\ &= \frac{1}{1-\gamma} \log(|\mathcal{A}|) \quad (\because \sum_s \rho_\pi(s) = \frac{1}{1-\gamma}). \end{aligned}$$

The inequality that $\sum_a -\log(\pi(a|s)) \pi(a|s) \leq \log(|\mathcal{A}|)$ also can be obtained by finding the point where the derivative of $-x \log(x)$ is zero. $\square$

Using Lemma 4 and Lemma 5, the error bounds of sparse and soft value iterations can be proved.

Theorem 4. Following inequalities hold:

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

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

Proof. Let π_* be the optimal policy of the original MDP, where the problem is defined as $\max_{\pi} \mathbb{E}_{\pi}(\mathbf{r}(s, a))$.

$$\mathbb{E}_{\pi_*^{sp}}(\mathbf{r}(s, a)) \leq \max_{\pi} \mathbb{E}_{\pi}(\mathbf{r}(s, a)) = \mathbb{E}_{\pi_*}(\mathbf{r}(s, a)).$$

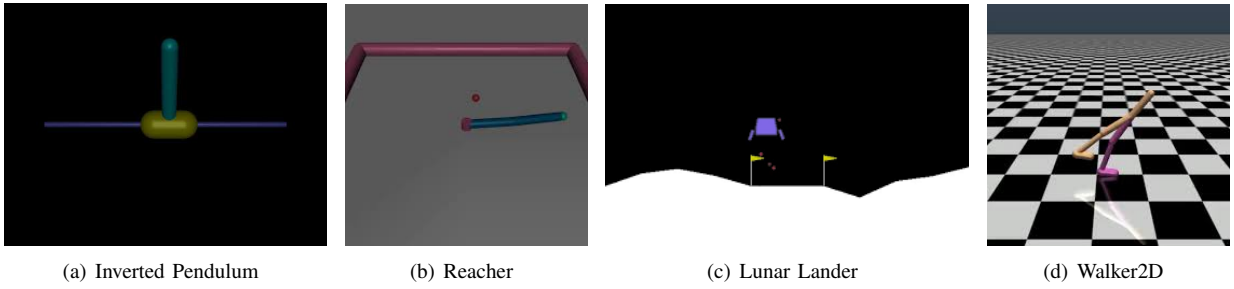


Fig. 1: (a) An inverted pendulum is mounted on a cart. The cart can only move horizontally. The goal of this task is to control a cart to keep the pendulum upright. (b) Two joint arm is fixed at the center point. The goal of this task is to control the end effector of arm to reach the goal point. (c) The point mass is dropped by gravity. The goal of this task is to land safely inside the landing pad (between two flags) and use less fuel. (d) 7 joint bipedal robot should move forward as quickly as possible.

The rightside inequality is by the definition of optimality. Before proving the leftside inequality, we first derive the following inequality from Lemma 4:

$$V_* \leq V_*^{sp}, \quad (10)$$

where $*$ indicates an optimal value. Since the fixed points of U and U^{sp} are the optimal solutions of the original MDP and sparse MDP, respectively, (10) can be derived from Lemma 4. The leftside inequality is proved using (10) as follows:

$$\begin{aligned} \mathbb{E}_{\pi_*}(\mathbf{r}(s, a)) &= d^T V_* \\ &\leq d^T V_*^{sp} = J_*^{sp} = \mathbb{E}_{\pi_*^{sp}}(\mathbf{r}(s, a)) + \alpha W(\pi_*^{sp}) \\ &\leq \mathbb{E}_{\pi_*^{sp}}(\mathbf{r}(s, a)) + \frac{\alpha}{1-\gamma} \frac{|\mathcal{A}| - 1}{2|\mathcal{A}|} \quad (\because \text{Lemma 5}). \end{aligned}$$

□

Theorem 5. Following inequalities hold:

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

where π_* and π_*^{soft} are the optimal policies obtained by the original MDP and a soft MDP, respectively, and $|\mathcal{A}|$ is the cardinality of the action space.

Proof. Let π_* be the optimal policy of the original MDP which is defined as $\max_{\pi} \mathbb{E}_{\pi}(\mathbf{r}(s, a))$. The rightside inequality is by the definition of optimality.

$$\mathbb{E}_{\pi_*^{soft}}(\mathbf{r}(s, a)) \leq \max_{\pi} \mathbb{E}_{\pi}(\mathbf{r}(s, a)) = \mathbb{E}_{\pi_*}(\mathbf{r}(s, a)).$$

Before proving the leftside inequality, we first derive following inequality from Lemma 4:

$$V_* \leq V_*^{soft} \quad (11)$$

where $*$ indicates an optimal solution. Then, the proof of the leftside inequality is done by using (11) as follows:

$$\begin{aligned} \mathbb{E}_{\pi_*}(\mathbf{r}(s, a)) &= d^T V_* \\ &\leq d^T V_*^{soft} = J_*^{soft} = \mathbb{E}_{\pi_*^{soft}}(\mathbf{r}(s, a)) + \alpha H(\pi_*^{soft}) \\ &\leq \mathbb{E}_{\pi_*^{soft}}(\mathbf{r}(s, a)) + \frac{\alpha}{1-\gamma} \log(|\mathcal{A}|) \quad (\because \text{Lemma 5}). \end{aligned}$$

□

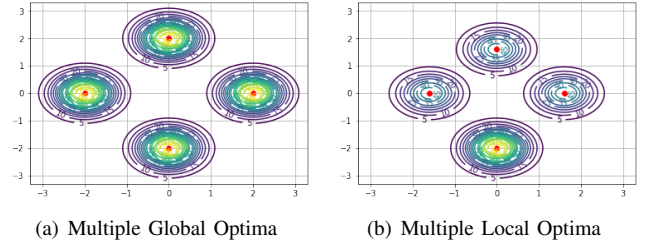


Fig. 2: Reward maps with multiple global optima (a) and multiple local optima (b). Red dot indicates goal points. The agent always starts at the center.

II. FULL EXPERIMENTAL RESULTS

A. Reinforcement Learning with Continuous Action Space

In this section, we explain the full experimental settings for reinforcement learning with a continuous action space. All problems we have tested are shown in Figure 1. All parameters are found by trial and error. For α , 0.01, 0.5, 0.1, and 1.0 are tested and the best performed value is selected. For the decaying rate for ϵ , 0.99, 0.995, 0.999, and 0.9995 are tested and the best performed value is selected and the minimum ϵ is set to 0.001. All selected parameters are shown in Table II. The network structures used for experiments are also reported in Table III.

B. Multi-Objective Exploration

In order to verify that sparsemax exploration can successfully learn multi-modal optimal actions, we designed a simple multi-objective environment where an agent follows point mass dynamics and tries to reach one of equally distributed multiple modes. The reward function is defined as a mixture of squared exponential functions whose centers are placed at the goal positions (see Figure 2(a)). If the exploration method successfully explores the environment, then the resulting policy distribution will equally reach the multiple modes. We compare the sparsemax, softmax, and ϵ -greedy methods and measure the average return and ratio of reached modes to given modes while changing the number of global optima with five different random seeds. The values of α and ϵ are found by a brute force search. α is set to 3 and 5 for softmax and sparsemax policy, respectively, and the decay

Task	Sps+SpsB	Sps+SftB	Sps+B	Stf+SpsB	Stf+StfB	Stf+B	Eps+SpsB	Eps+StfB	Eps+B
Inv. Pend.	$\alpha : 0.1$ $\epsilon : 0.995$	$\alpha : 0.01$ $\epsilon : 0.995$	$\alpha : 0.01$ $\epsilon : 0.995$	$\alpha : 0.01$ -	$\alpha : 0.1$ -	$\alpha : 0.1$ -	$\alpha : 0.1$ $\epsilon : 0.995$	$\alpha : 1$ $\epsilon : 0.995$	- $\epsilon : 0.995$
Reacher	$\alpha : 0.1$ $\epsilon : 0.99$	$\alpha : 0.1$ $\epsilon : 0.99$	$\alpha : 0.1$ $\epsilon : 0.99$	$\alpha : 0.01$ -	$\alpha : 0.01$ -	$\alpha : 0.01$ -	$\alpha : 0.01$ $\epsilon : 0.995$	$\alpha : 1$ $\epsilon : 0.995$	- $\epsilon : 0.995$
Lun. Lan.	$\alpha : 0.1$ $\epsilon : 0.999$	$\alpha : 1$ $\epsilon : 0.999$	$\alpha : 1$ $\epsilon : 0.999$	$\alpha : 0.1$ -	$\alpha : 0.1$ -	$\alpha : 0.1$ -	$\alpha : 1$ $\epsilon : 0.99$	$\alpha : 0.1$ $\epsilon : 0.99$	- $\epsilon : 0.99$
Walker2D	$\alpha : 0.01$ $\epsilon : 0.9995$	$\alpha : 0.1$ $\epsilon : 0.9995$	$\alpha : 0.5$ $\epsilon : 0.9995$	$\alpha : 0.1$ -	$\alpha : 0.1$ -	$\alpha : 0.1$ -	$\alpha : 1.0$ $\epsilon : 0.9995$	$\alpha : 1.0$ $\epsilon : 0.9995$	- $\epsilon : 0.9995$

TABLE II: The best performed parameters for each algorithm in each problem.

Task	Deep Q learning	DDPG
Inv. Pend. Walker2D	Two 512 ReLU layers	Two 128 ReLU layers for both actor and critic networks
Reacher Lun. Lan. Cont.	Four 256 ReLU layers	Two 64 ReLU layers for both actor and critic networks

TABLE III: Network Structures. ReLU indicates a rectified linear unit.

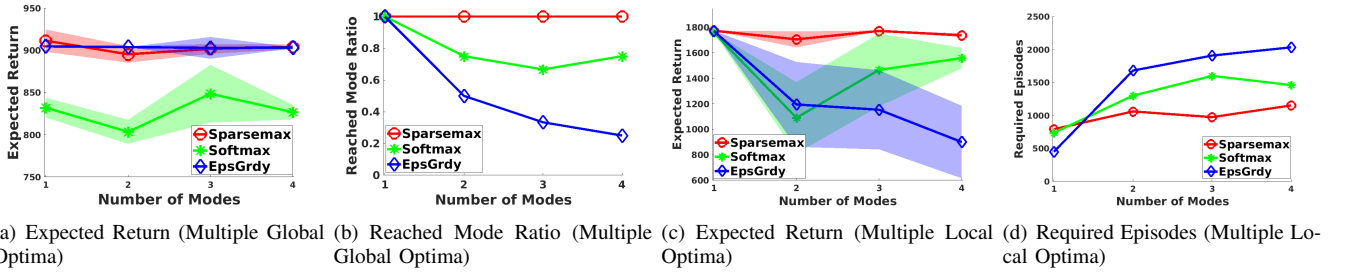


Fig. 3: (a) The average performance of each algorithm averaging over 500 test episodes after training with 3000 episodes. (b) The ratio of the number of given goals to the number of the goals reached by a trained policy. (c) The average performance of each algorithm averaging over 500 test episodes sampled from a greedy policy after training with 3000 episodes. (d) The number of episodes to reach the threshold value (800) during the exploration phase.

rate of ϵ is set to 0.9995. The results are shown in Figure 3. The average performance and the reached mode ratio are shown in Figure 3(a). The resulting sparsemax policy can reach every modes while maintaining its performance when the number of global optima increases. However, the softmax policy shows a performance drop since it assigns nonzero probability to non-optimal actions to explore the every optimal points during the exploration phase and this effect hampers the convergence of Q network to the multiple modes. The example of sampled trajectories are shown in Figure 4.

We test our algorithm on a more difficult problem to verify that our method can find the global optimum when multiple local optima exist (see Figure 2(b)). We designed a reward function with single global optimum and multiple local optima. The local optima are located near the initial state and the global optimum is farther from the initial state than the local optima are. Hence, in order to reach the global optimum, an agent should keep exploring with wide directions. We train a Q network with sparsemax, softmax, and ϵ -greedy explorations and evaluate the trained Q network with the greedy policy, i.e., $\arg \max Q(s, a)$. If exploration method can search the global optimum within limited episodes and the Q network converges into the global optimum, then the greedy policy will reach the global optimal point. For each algorithm, α is selected from the best value among

$[0.1, 1, 5, 10, 100]$ and ϵ decaying rate is also selected from the best value among $[0.99, 0.999, 0.9995, 0.9999]$ with the minimum ϵ at 0.001. The experiments are repeated with five different random seeds and the test average return and required episodes to reach the threshold average return are shown in Figure 3. To compute the number of episodes to reach the threshold average return, we measure the average return over the consecutive 100 episodes during the exploration phase and find the first point to cross the specific threshold average return which is set to 800. In the given problem, the expected value of local optima and global optimum are 600 and 1800, respectively. Therefore, the threshold average return, 800, indicates that some of 100 episodes reach the global optimum. In terms of the number of episodes, sparsemax exploration shows the fastest convergence to the global optima than the other methods as shown in Figure 3(d). As a result, it can be shown that sparsemax exploration escapes the local optima faster than the other explorations. When it comes to performance evaluation, the Q network trained by sparsemax exploration outperforms softmax and ϵ -greedy exploration as shown in Figure 3(c), since sparsemax exploration reaches the global optima faster than the others.

REFERENCES

- [1] A. Martins and R. Astudillo, "From softmax to sparsemax: A sparse model of attention and multi-label classification," in *Proc. of the*

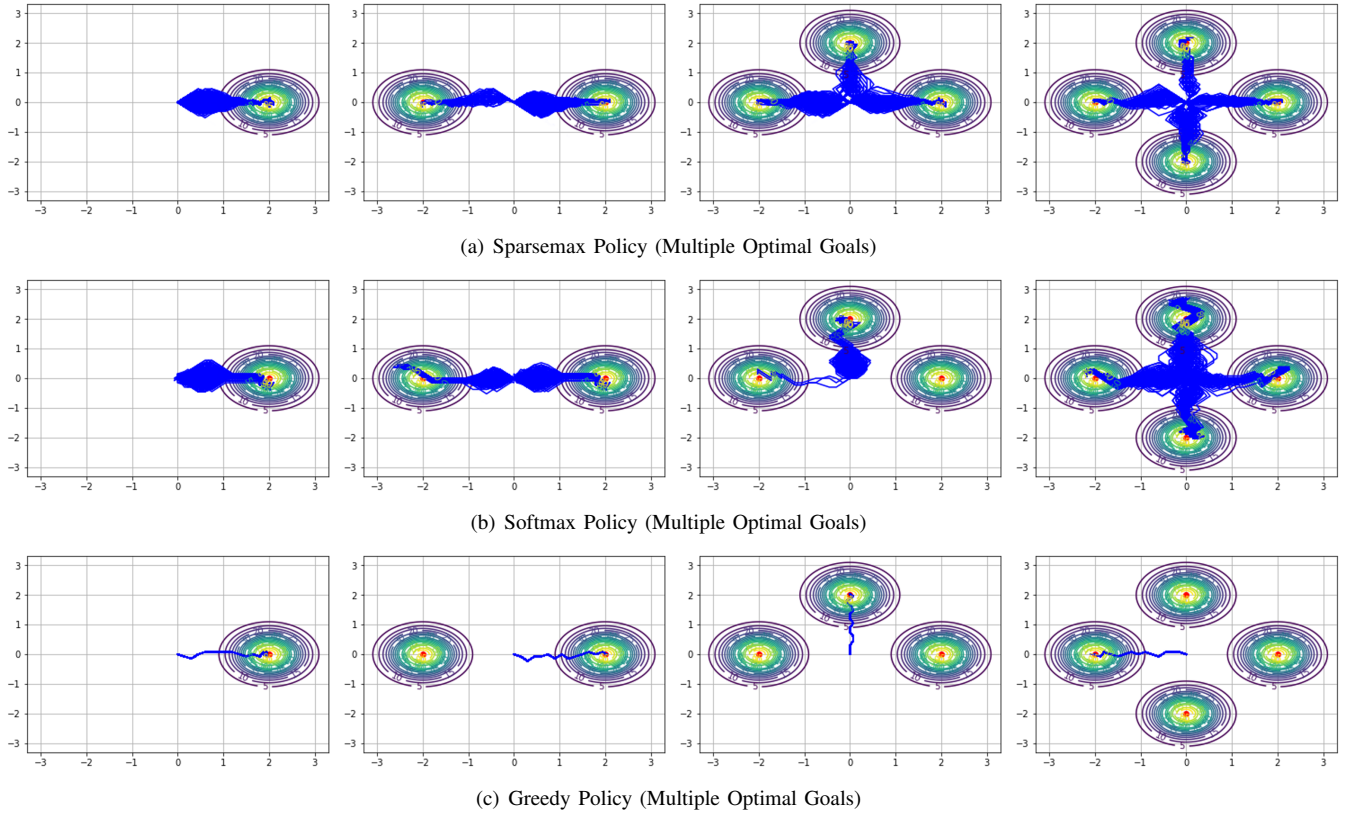


Fig. 4: (a) Example trajectories sampled from the sparsemax policy distribution trained by sparsemax exploration. (b) Example trajectories sampled from the softmax policy distribution trained by softmax exploration. (c) Example trajectories sampled from the greedy policy distribution trained by ϵ -greedy exploration (when we sample the trajectory, ϵ is set to zero).

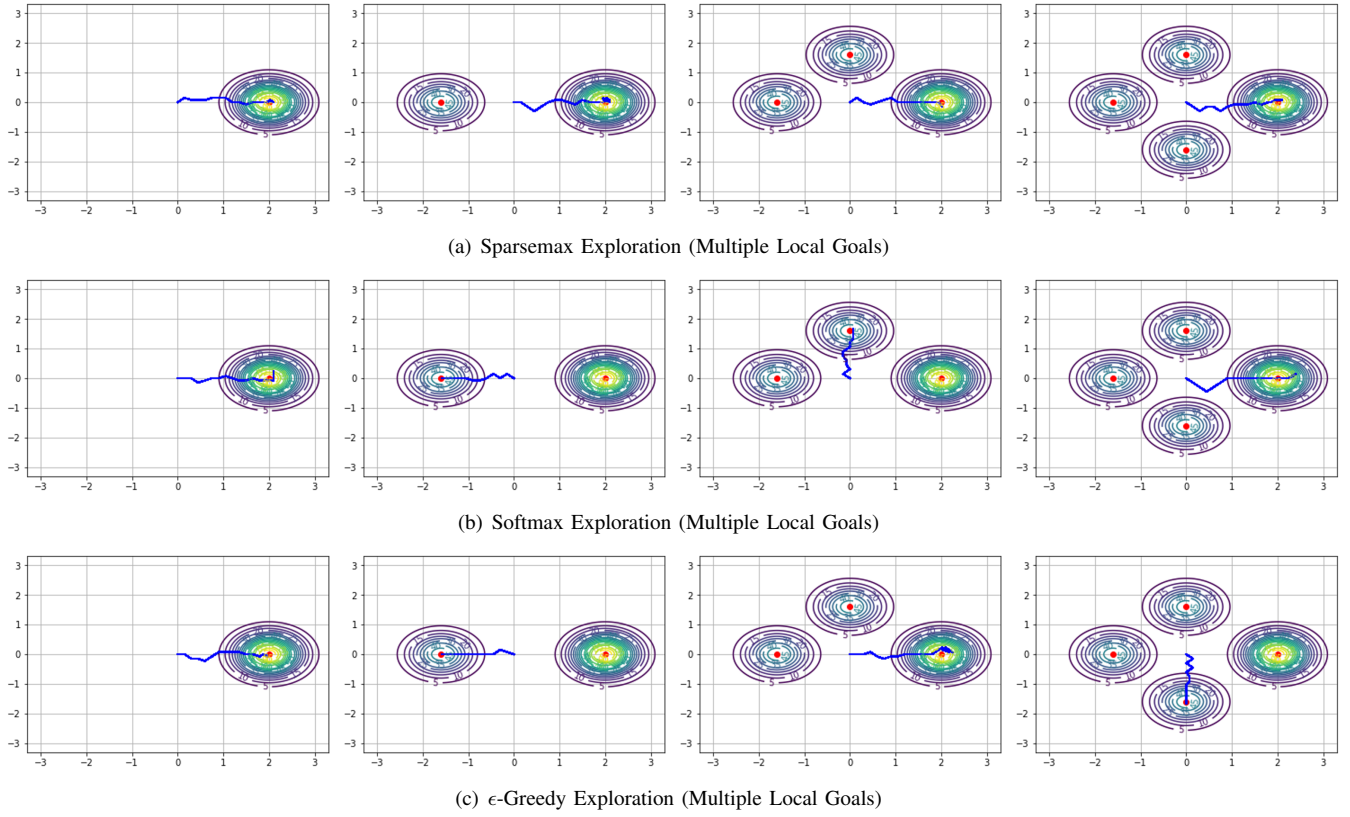


Fig. 5: (a) Example trajectories sampled by greedy policy trained by sparsemax exploration. (b) Example trajectories sampled by greedy policy trained by softmax exploration. (c) Example trajectories sampled by greedy policy trained by ϵ -greedy exploration.