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

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 spare 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, $r(s,a)$, 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:

$$J_\pi^{sp}=d^\intercal G_\pi^{-1}r_\pi^{sp}, \qquad V_\pi^{sp}=G_\pi^{-1}r_\pi^{sp}$$
$$J_\pi^{soft}=d^\intercal G_\pi^{-1}r_\pi^{soft}, \qquad V_\pi^{soft}=G_\pi^{-1}r_\pi^{soft}, \quad \rho_\pi=d^\intercal G_\pi^{-1}$$

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

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

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

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

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, sungjoon.choi, songhwai.oh}@cpslab.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:

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

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:

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

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

be obtained by using the chain rule.

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

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

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

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:

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

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.

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

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 optimality equation of V_π^{sp} is induced.

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

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

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

□

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:

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

□

D. Upper and Lower Bounds for Sparsemax Operation

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

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

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,

$$\max(z) \le \alpha\text{spmax}(\frac{z}{\alpha}) \le \max(z)+\alpha\frac{d-1}{2d}. \tag{9}$$

Note that the proof of lower bound of (9) is provided in [3]. 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 $\max(z) \le \text{spmax}(z)$ and next prove $\text{spmax}(z) \le \max(z)+\frac{d-1}{2d}$. 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 $z \in \mathbb{R}^d$, $\max(z) \le \text{spmax}(z)$ holds.

Proof. We prove that, for all z, $\text{spmax}(z)-z_{(1)} \ge 0$ where $z_{(1)} = \max(z)$ by definition. The proof is done by simply rearranging the terms in (8),

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

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:

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

Then, $K\sum_{i=2}^{K}z_{(i)}^2+2\sum_{i=2}^{K}z_{(i)}+K$ can be replaced with $K\sum_{i=2}^{K}\left(z_{(i)}+1\right)^2-2(K-1)\sum_{i=2}^{K}z_{(i)}$ 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,

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

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,

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

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

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

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

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

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 $(\mathbf{p}^\intercal\mathbf{q})^2 \leq ||\mathbf{p}||^2||\mathbf{q}||^2$. Let $z_{2:K} = [z_{(2)}, \cdots, z_{(K)}]^\intercal$, then, by setting $\mathbf{p} = z_{2:K} + \mathbb{1}$ and $\mathbf{q} = \frac{1}{K-1}\mathbb{1}$ where $\mathbb{1}$ is a $K-1$ dimensional vector of ones, it can be shown that the second term is nonnegative. Therefore, $\text{spmax}(z) - z_{(1)}$ is always nonnegative for all z since three remaining terms are always nonnegative, completing the proof. □

Now, we prove the upper bound of sparsemax operation.

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

Proof. First, we decompose the summation of (8) into two terms as follows:

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

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, $\sum_i^d p_i z_i \leq \max(z)$ 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. □

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 [3], 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,

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

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

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

Finally, we can obtain

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

□

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

Proof. In [3], it is shown that for $c \in \mathbb{R}$ and $x \in \mathbb{R}^{|\mathcal{S}|}$, $\text{spmax}(x+c\mathbb{1}) = \text{spmax}(x) + c\mathbb{1}$. Using this property,

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

□

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 $\phi : \mathbb{R}^{|\mathcal{S}|} \to \mathbb{R}^{|\mathcal{S}|}$ 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 $\mathbb{R}^{|\mathcal{S}|}$ 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. □

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. □

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 vlaues 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 $x \in \mathbb{R}^{|\mathcal{S}|}$,

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

Then following inequalities hold for every positive integer n:

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

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:

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

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

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

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

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

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 $x \in \mathbb{R}^{|\mathcal{S}|}$. Then, when $n = k+1$,

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

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 $n \to \infty$,

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

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. □

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:

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

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

Proof. For $W(\pi)$,

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

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 $H(\pi)$,

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

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. □

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 $\max_\pi \mathbb{E}_\pi(\mathbf{r}(s,a))$.

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

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

$$V_* \leq V_*^{sp}, \tag{10}$$

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:

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

□

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:

$$V_* \leq V_*^{soft} \tag{11}$$

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

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

□

II. Full Experimental Results

In this section, we present the full experimental results of reinforcement learning with a continuous action space. We performe experiments on *Inverted Pendulum* and *Reacher* and 28 algorithms are tested including our sparse exploration method and sparse Bellman update rule.

References

[1] J. Ye, "Constraint qualifications and necessary optimality conditions for optimization problems with variational inequality constraints," *SIAM Journal on Optimization*, vol. 10, no. 4, pp. 943–962, 2000.

[2] W. Wang and M. A. Carreira-Perpinán, "Projection onto the probability simplex: An efficient algorithm with a simple proof, and an application," *arXiv preprint arXiv:1309.1541*, 2013.

[3] A. Martins and R. Astudillo, "From softmax to sparsemax: A sparse model of attention and multi-label classification," in *International Conference on Machine Learning*, June 2016, pp. 1614–1623.

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

TABLE I: Notations and Properties

The Number of Action	3	101	1001	2001	Average
Sparse+SparseBellman-1	1000.0	996.8	1000.0	1000.0	**999.2**
Sparse+SparseBellman-0.1	1000.0	933.1	668.2	1000.0	900.3
Sparse+SparseBellman-0.01	1000.0	992.1	1000.0	1000.0	**998.0**
Sparse+SoftBellman-1	1000.0	1000.0	925.2	1000.0	981.3
Sparse+SoftBellman-0.1	1000.0	1000.0	1000.0	1000.0	**1000.0**
Sparse+SoftBellman-0.01	782.7	988.6	775.8	1000.0	886.8
Sparse+Bellman-1	1000.0	1000.0	919.7	715.3	908.7
Sparse+Bellman-0.1	980.2	745.5	1000.0	1000.0	931.4
Sparse+Bellman-0.01	1000.0	1000.0	1000.0	1000.0	**1000.0**
Soft+SparseBellman-1	673.9	835.5	53.3	1000.0	640.7
Soft+SparseBellman-0.1	688.0	1000.0	938.0	904.2	882.6
Soft+SparseBellman-0.01	993.0	1000.0	736.4	1000.0	932.3
Soft+SoftBellman-1	939.6	738.3	506.3	943.2	781.9
Soft+SoftBellman-0.1	1000.0	1000.0	1000.0	681.6	920.4
Soft+SoftBellman-0.01	1000.0	974.8	1000.0	1000.0	993.7
Soft+Bellman-1	668.9	621.5	668.7	643.2	650.6
Soft+Bellman-0.1	1000.0	1000.0	1000.0	1000.0	**1000.0**
Soft+Bellman-0.01	977.6	1000.0	1000.0	1000.0	994.4
EpsGrdy+SparseBellman-1	479.8	669.0	344.5	678.1	542.9
EpsGrdy+SparseBellman-0.1	668.1	1000.0	351.1	666.6	671.4
EpsGrdy+SparseBellman-0.01	1000.0	124.6	477.5	667.8	567.5
EpsGrdy+SoftBellman-1	940.3	684.9	658.3	505.6	697.3
EpsGrdy+SoftBellman-0.1	338.5	376.8	1000.0	1000.0	678.8
EpsGrdy+SoftBellman-0.01	551.5	652.8	735.2	677.9	654.3
EpsGrdy+Bellman-1	332.7	1000.0	1000.0	369.8	675.6
EpsGrdy+Bellman-0.1	1000.0	618.7	1000.0	771.7	847.6
EpsGrdy+Bellman-0.01	462.6	676.5	698.0	48.1	471.3
DDPG	253.1				253.1

TABLE II: Expected return of *Inverted Pendulum*. Top five performances are marked in bold.

The Number of Action	3	101	1001	2001
Sparse+SparseBellman-1	1164	692	742	864
Sparse+SparseBellman-0.1	1060	2923	3998	599
Sparse+SparseBellman-0.01	685	1431	1010	811
Sparse+SoftBellman-1	863	1316	1698	657
Sparse+SoftBellman-0.1	914	901	857	802
Sparse+SoftBellman-0.01	3342	907	3930	522
Sparse+Bellman-1	879	668	2337	3137
Sparse+Bellman-0.1	937	3925	773	1030
Sparse+Bellman-0.01	999	329	962	962
Soft+SparseBellman-1	3789	3416	3996	2684
Soft+SparseBellman-0.1	3844	2835	1494	2771
Soft+SparseBellman-0.01	854	545	3814	999
Soft+SoftBellman-1	1885	3666	3994	3912
Soft+SoftBellman-0.1	869	780	787	3871
Soft+SoftBellman-0.01	533	1241	2565	3020
Soft+Bellman-1	3898	3947	3978	3758
Soft+Bellman-0.1	876	1923	954	807
Soft+Bellman-0.01	1419	689	755	1265
EpsGrdy+SparseBellman-1	3978	3993	4000	3863
EpsGrdy+SparseBellman-0.1	3895	2449	4000	3910
EpsGrdy+SparseBellman-0.01	3437	4000	3962	3777
EpsGrdy+SoftBellman-1	2959	3919	3715	4000
EpsGrdy+SoftBellman-0.1	3997	3969	3037	2509
EpsGrdy+SoftBellman-0.01	3976	3936	3785	3784
EpsGrdy+Bellman-1	4000	2603	1093	3969
EpsGrdy+Bellman-0.1	2584	3897	3160	3846
EpsGrdy+Bellman-0.01	3891	3699	3905	3993

TABLE III: The number of episodes required to reach the threshold return, 980.

The Number of Action	9	121	961	2601	Average
Sparse+SparseBellman-1	-7.7	-7.8	-10.1	-11.5	-9.3
Sparse+SparseBellman-0.1	-11.3	-5.7	-5.4	-5.5	**-7.0**
Sparse+SparseBellman-0.01	-11.3	-8.7	-8.6	-6.3	-8.7
Sparse+SoftBellman-1	-7.6	-10.5	-11.5	-10.0	-9.9
Sparse+SoftBellman-0.1	-10.4	-5.8	-5.5	-9.3	-7.8
Sparse+SoftBellman-0.01	-11.2	-6.4	-8.9	-6.4	-8.2
Sparse+Bellman-1	-7.6	-7.7	-5.7	-10.2	-7.8
Sparse+Bellman-0.1	-10.8	-5.5	-5.4	-5.8	**-6.9**
Sparse+Bellman-0.01	-11.6	-5.9	-5.9	-9.4	-8.2
Soft+SparseBellman-1	-52.0	-48.0	-29.6	-39.3	-42.2
Soft+SparseBellman-0.1	-7.4	-22.4	-20.8	-25.5	-19.0
Soft+SparseBellman-0.01	-11.1	-5.5	-5.5	-9.2	-7.8
Soft+SoftBellman-1	-52.2	-43.1	-46.8	-44.1	-46.5
Soft+SoftBellman-0.1	-7.5	-22.5	-23.9	-32.9	-21.7
Soft+SoftBellman-0.01	-11.6	-5.7	-5.5	-7.6	-7.6
Soft+Bellman-1	-51.4	-51.7	-44.2	-41.2	-47.1
Soft+Bellman-0.1	-7.1	-10.0	-26.7	-27.5	-17.8
Soft+Bellman-0.01	-11.3	-5.3	-5.3	-10.2	-8.0
EpsGrdy+SparseBellman-1	-11.2	-7.6	-5.6	-6.2	-7.6
EpsGrdy+SparseBellman-0.1	-11.2	-5.9	-5.8	-6.1	-7.2
EpsGrdy+SparseBellman-0.01	-10.9	-5.9	-5.5	-6.0	**-7.1**
EpsGrdy+SoftBellman-1	-10.5	-5.9	-5.7	-6.1	**-7.0**
EpsGrdy+SoftBellman-0.1	-11.3	-5.7	-5.6	-6.2	**-7.2**
EpsGrdy+SoftBellman-0.01	-10.8	-6.2	-12.1	-9.5	-9.6
EpsGrdy+Bellman	-10.8	-6.5	-5.7	-6.5	-7.4
EpsGrdy+Bellman	-11.1	-6.2	-5.9	-5.9	-7.3
EpsGrdy+Bellman	-10.6	-9.4	-8.4	-6.5	-8.7
DDPG	-10.1				-10.1

TABLE IV: Expected return of *Reacher*. Top five performances are marked in bold.

TABLE V: Reacher Required Episodes

The Number of Action	9	121	961	2601
Sparse+SparseBellman-1	9193	9648	7275	9065
Sparse+SparseBellman-0.1	9791	5837	5779	6851
Sparse+SparseBellman-0.01	9783	6456	6631	7941
Sparse+SoftBellman-1	9126	9834	7603	8503
Sparse+SoftBellman-0.1	9779	5449	5642	7509
Sparse+SoftBellman-0.01	9795	5011	7260	7768
Sparse+Bellman-1	9073	9619	5646	8371
Sparse+Bellman-0.1	9756	5366	5338	6936
Sparse+Bellman-0.01	9797	5204	6525	7965
Soft+SparseBellman-1	10000	10000	10000	10000
Soft+SparseBellman-0.1	8801	9998	10000	10000
Soft+SparseBellman-0.01	9783	4988	5934	8774
Soft+SoftBellman-1	10000	10000	10000	10000
Soft+SoftBellman-0.1	8810	9999	10000	10000
Soft+SoftBellman-0.01	9794	4597	5927	7915
Soft+Bellman-1	10000	10000	10000	10000
Soft+Bellman-0.1	8700	9999	10000	10000
Soft+Bellman-0.01	9790	4810	6004	8737
EpsGrdy+SparseBellman-1	9861	6909	6994	7977
EpsGrdy+SparseBellman-0.1	9850	6808	6775	7873
EpsGrdy+SparseBellman-0.01	9847	7079	6850	7923
EpsGrdy+SoftBellman-1	9850	6839	6858	8026
EpsGrdy+SoftBellman-0.1	9844	6918	6752	7849
EpsGrdy+SoftBellman-0.01	9841	7176	9803	8114
EpsGrdy+Bellman	9842	6797	6933	8001
EpsGrdy+Bellman	9846	6680	7051	7845
EpsGrdy+Bellman	9864	7192	6928	7925

TABLE VI: The number of episodes required to reach the threshold return, -6.