State Abstraction as Compression in Apprenticeship Learning Supplementary Material

David Abel[1], **Dilip Arumugam**[2], **Kavosh Asadi**[1]
Yuu Jinnai[1], **Michael L. Littman**[1], **Lawson L.S. Wong**[3]

1: Department of Computer Science, Brown University.
2: Department of Computer Science, Stanford University.
3: College of Computer and Information Science, Northeastern University.

We here include proofs of the introduced theorems and lemmas and provide further discussion about extensions to more general settings.

Proofs

We begin with the two central Lemmas, which serve as the main deductive steps for the proof of Theorem 2.

.....

Lemma 1. *Consider a discrete random variable X, with alphabet $\mathcal{X}$ and some pmf $p(x)$. For a given threshold $\delta_{min} \in (0, 1)$, the pmf-used alphabet size of the alphabet is bounded:*

$$|\mathcal{X}|^{\delta_{min}}_{p(x)} \leq \frac{H(X)}{\delta_{min} \log\left(\frac{1}{\delta_{min}}\right)}. \tag{18}$$

Proof. Recall the entropy of a random variable $H(X)$:

$$H(X) := -\sum_{x \in \mathcal{X}} p(x) \log_2 p(x). \tag{19}$$

Then, for a given maximum entropy $H(X) \leq N$, we seek to understand the largest possible $|\mathcal{X}|^{\delta_{min}}_{p(x)}$. That is:

$$\max_{p(x): H(X) \leq N} |\mathcal{X}|^{\delta_{min}}_{p(x)}. \tag{20}$$

Note that this is maximized at the uniform distribution, where each element has δ_{min} probability, across the largest alphabet such that $H(X) = N$:

$$\begin{aligned} H(X) &= -\sum_{x \in \mathcal{X}} p(x) \log_2 p(x) && (21)\\ &= -\sum_{x \in \mathcal{X}} \delta_{min} \log_2 \delta_{min} && (22)\\ &= -|\mathcal{X}| \delta_{min} \log_2 \delta_{min} && (23)\\ &= |\mathcal{X}| \delta_{min} \log_2 \frac{1}{\delta_{min}}. && (24) \end{aligned}$$

Therefore, for a given pmf $p(x)$ with entropy $H(X)$, and a minimum threshold of probability, the minimum size of the alphabet $\mathcal{X}$ is upper bounded:

$$|\mathcal{X}| \leq \frac{H(X)}{\delta_{min} \log_2 \frac{1}{\delta_{min}}}. \qquad \square$$

.....

Lemma 2. *Consider two stochastic policies, π_1 and π_2 on state space $\mathcal{S}$, and a fixed probability distribution over $\mathcal{S}$, $p(s)$. If, for some $k \in \mathbb{R}_{\geq 0}$:*

$$\mathop{\mathbb{E}}_{p(s)} \left[D_{KL}(\pi_1(a \mid s) \mid\mid \pi_2(a \mid s)) \right] \leq k, \tag{25}$$

then:

$$\mathop{\mathbb{E}}_{p(s)} \left[V^{\pi_1}(s) - V^{\pi_2}(s) \right] \leq \sqrt{2k} \textsc{VMax}. \tag{26}$$

Proof. Recall the total variation distance (TVD) between our two policies for a given state s is defined as:

$$TV(\pi_E, \pi_\phi) := \sup_{a \in \mathcal{A}} |\pi_E(a \mid s) - \pi_\phi(a \mid s)|. \tag{27}$$

Furthermore, recall that TVD relates to the L_1 norm and the KL divergence:

$$TV(\pi_E, \pi_\phi) = \frac{1}{2} \sum_{a \in \mathcal{A}} |\pi_E(a) - \pi_\phi(a)| \leq \sqrt{\frac{1}{2} D_{KL}(\pi_E \mid\mid \pi_\phi)}, \tag{28}$$

where the inequality in Equation 28 is formally known as Pinsker's inequality. With this inequality in place, we expand the expectation in the value bound:

$$\begin{aligned} &\mathbb{E}_{p(s)}[V^{\pi_E}(s) - V^{\pi_\phi}(s)] \qquad (29)\\ &\leq \sum_s p(s) \left(\sum_a |\pi_E(a \mid s) - \pi_\phi(a \mid s)| (\mathcal{R}(s,a) + \gamma \sum_{s'} \mathcal{T}(s,a,s') |V^{\pi_E}(s') - V^{\pi_\phi}(s')| \right) \\ &= \sum_s \sum_a p(s) |\pi_E(a \mid s) - \pi_\phi(a \mid s)| \left(\mathcal{R}(s,a) + \gamma \sum_{s'} \mathcal{T}(s,a,s') |V^{\pi_E}(s') - V^{\pi_\phi}(s')| \right) \end{aligned}$$

Then, applying the upper bound on the possible value $\textsc{VMax} = \textsc{RMax}/(1-\gamma)$ to Equation 29:

$$\mathbb{E}_{p(s)}[V^{\pi_E}(s) - V^{\pi_\phi}(s)] \leq \textsc{VMax}\, \mathbb{E}_{p(s)} \left[|\pi_E(a \mid s) - \pi_\phi(a \mid s)| \right]. \tag{30}$$

Then, by Pinsker's inequality, we conclude:

$$\begin{aligned} \mathbb{E}_{p(s)}[V^{\pi_E}(s) - V^{\pi_\phi}(s)] &\leq 2\textsc{VMax}\, \mathbb{E}_{p(s)} \left[\sqrt{\frac{1}{2} D_{KL}(\pi_E(a \mid s) \mid\mid \pi_\phi(a \mid s))} \right] \qquad (31)\\ &\leq \sqrt{2k} \textsc{VMax} \qquad \square \end{aligned}$$

.....

With these two Lemmas in place, we now prove our main Theorem.

Theorem 2. *A function f of the DIB objective $\hat{\mathcal{J}}_{DIB}$ is an upper bound for the CVA Objective, $\mathcal{J}$, where state space size is treated as $|\mathcal{S}_\phi|^{\delta_{min}}_{\rho_\phi(s)}$:*

$$\forall_\phi : \mathcal{J}[\phi] \leq f\left(\hat{\mathcal{J}}_{DIB}[\phi] \right). \tag{32}$$

Proof. With Lemma 1 and Lemma 2, the proof is straightforward. Consider the ϕ that minimizes $\hat{\mathcal{J}}_{\text{DIB}}$, yielding the value of at most $N + \beta k$, where:

$$N := H(S_\phi) \tag{33}$$

$$k := \mathop{\mathbb{E}}_{s\sim\rho_E}\left[D_{KL}(\pi_E(a \mid s) \mid\mid \pi_\phi(a \mid \phi(s))\right]. \tag{34}$$

Then, by Lemma 1, we know:

$$|\mathcal{S}_\phi|^{\delta_{min}}_{\rho_\phi(s)} \leq \frac{N}{\delta_{min}\log_2 \frac{1}{\delta_{min}}}. \tag{35}$$

By Lemma 2, we know:

$$\mathbb{E}_{\rho(s)}\left[V^{\pi_E}(s) - V^{\pi_\phi}(s)\right] \leq \sqrt{2k}\text{VMAX}. \tag{36}$$

Therefore, since both quantities are non-negative, we conclude:

$$\mathcal{J}[\phi] = |\mathcal{S}_\phi|^{\delta_{min}}_{\rho_\phi(s)} + \beta\mathbb{E}_{\rho(s)}\left[V^{\pi_E}(s) - V^{\pi_\phi}(s)\right] \tag{37}$$

$$\leq \frac{N}{\delta_{min}\log_2 \frac{1}{\delta_{min}}} + \beta\sqrt{2k}\text{VMAX}. \tag{38}$$

Thus, we can upper bound the quantities in $\mathcal{J}$ as a function of the quantities in $\hat{\mathcal{J}}_{\text{DIB}}$. □

.....

Extensions

The results we present in this work offer a first step toward understanding state abstraction as compression in sequential decision making. Looking forward, we hope to extend this work into a general abstraction-as-compression learning paradigm for RL. To this end, we introduce two concrete avenues for extending the work to more general settings.

Agent Has Control

Relaxing the assumption that π_E controls the MDP is essential for extending this work to traditional RL. We here propose a path toward removing this restriction by focusing on an intermediate goal: define an algorithm with the same properties as DIBS, but with the learning agent's *non-stationary* policy controlling the underlying MDP instead of π_E. Ultimately, we seek an algorithm that, after $T < \infty$ iterations, can produce an abstraction–policy pair such that, for some state distribution $p(s)$:

$$\mathbb{E}_{p(s)}\left[V^d(s) - V^{\pi^T_\phi}(s)\right] \leq f(\beta, T). \tag{39}$$

The most challenging aspect of this setup is that the source distribution is no longer fixed, since the agent's policy will change over time as the agent learns and updates both ϕ and π_ϕ. To this end, we prove the following result that suggests a route to defining a convergent algorithm for the case where the agent controls the MDP.

Lemma 3. *Given two policies π_1 and π_2, if:*

$$\sup_s \sum_a |\pi_1(a \mid s) - \pi_2(a \mid s)| \leq \Delta, \tag{40}$$

then:

$$\sum_s |\rho_{\pi_1,s_0}(s) - \rho_{\pi_2,s_0}(s)| \leq \frac{\Delta\gamma}{1-\gamma}, \tag{41}$$

where ρ_{π,s_0} denotes the stationary distribution over states under π, starting in state s_0.

Proof. We first bound the difference between the two state distributions after t steps:

$$\begin{aligned}
& \sum_{s'} |\rho^t_{\pi_1,s_0}(s') - \rho^t_{\pi_2,s_0}(s')| \\
= & \sum_{s'} |p(s_t = s'|s_0,\pi_1) - p(s_t = s'|s_0,\pi_2)| \\
= & \sum_{s'} |\sum_s p(s_{t-1} = s \mid s_0,\pi_1) \sum_a \pi_1(a \mid s) p(s' \mid s,a) \\
- & \sum_s p(s_{t-1} = s \mid s_0,\pi_2) \sum_a \pi_2(a \mid s) p(s' \mid s,a)| \\
\leq & \sum_{s'} |\sum_s p(s_{t-1} = s \mid s_0,\pi_1) \sum_a \Big(\pi_1(a \mid s) - \pi_2(a \mid s)\Big) p(s' \mid s,a)| \\
+ & \sum_{s'} |\sum_s \Big(p(s_{t-1} = s \mid s_0,\pi_1) - p(s_{t-1} = s \mid s_0,\pi_2)\Big) \sum_a \pi_2(a \mid s) p(s' \mid s,a)| \\
\leq & \sum_s p(s_{t-1} = s \mid s_0,\pi_1) \sum_a \Big|\pi_1(a \mid s) - \pi_2(a \mid s)\Big| \sum_{s'} p(s' \mid s,a) \\
+ & \sum_s \Big|p(s_{t-1} = s \mid s_0,\pi_1) - p(s_{t-1} = s \mid s_0,\pi_2)\Big| \sum_a \pi_2(a \mid s) \sum_{s'} p(s' \mid s,a) \\
\leq & \Delta + \sum_s \Big|p(s_{t-1} = s \mid s_0,\pi_1) - p(s_{t-1} = s \mid s_0,\pi_2)\Big| \\
= & \Delta + \sum_{s'} |\rho^{t-1}_{\pi_1,s_0}(s') - \rho^{t-1}_{\pi_2,s_0}(s')|
\end{aligned}$$

From the above bound, and using induction, we have:

$$\sum_{s'} |\rho^t_{\pi_1,s_0}(s') - \rho^t_{\pi_2,s_0}(s')| \leq t\Delta$$

$$\begin{aligned}
& \sum_s |\rho_{\pi_1,s_0}(s) - \rho_{\pi_2,s_0}(s)| \\
= & \sum_s |(1-\gamma) \sum_t \gamma^t \rho^t_{\pi_1,s_0}(s) - (1-\gamma) \sum_t \gamma^t \rho^t_{\pi_2,s_0}(s)| \\
\leq & (1-\gamma) \sum_t \gamma^t \sum_s |\rho^t_{\pi_1,s_0}(s) - \rho^t_{\pi_2,s_0}(s)| \\
\leq & (1-\gamma) \sum_t \gamma^t t\Delta = (1-\gamma) \frac{\gamma\Delta}{(1-\gamma)^2} = \frac{\gamma\Delta}{1-\gamma} \quad \square
\end{aligned}$$

.....

Corollary 1. *As a simple corollary of Proposition 1 we get:*

$$\begin{aligned}
|V^{\pi_1}(s) - V^{\pi_2}(s)| &= |\sum_t \sum_s \rho^t_{\pi_1,s_0}(s) R(s) - \sum_t \sum_s \rho^t_{\pi_2,s_0}(s) R(s)| \\
&\leq \textsc{RMax} \sum_t (1-\gamma) \sum_s |\rho^t_{\pi_1,s_0}(s) - \rho^t_{\pi_2,s_0}(s)| \\
&\leq \textsc{RMax}(1-\gamma) \sum_t \Delta\gamma^t t \\
&= \textsc{RMax} \frac{\gamma\Delta}{1-\gamma} = \Delta\gamma \textsc{VMax}
\end{aligned}$$

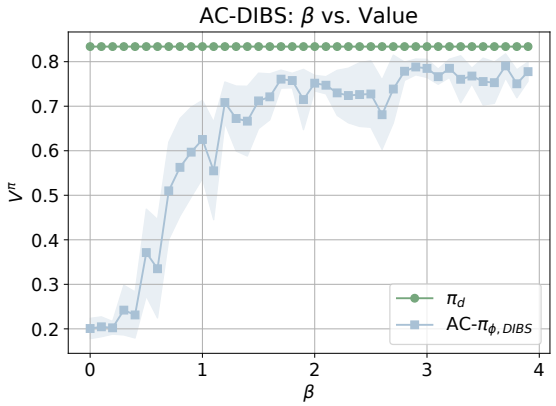


Figure 1: The value of the abstract policy found by Agent Controlled DIBSfor values of β between 0 and 4. The line indicates the average over 100 trials, with 95% confidence intervals.

The above Lemma suggests that two policies that deviate by a bounded amount are guaranteed to share similar stationary distributions. So, with the right intermediate updates that encourage the abstract policy to mimic π_E more faithfully, we can construct a convergent algorithm for the agent-in-control setting.

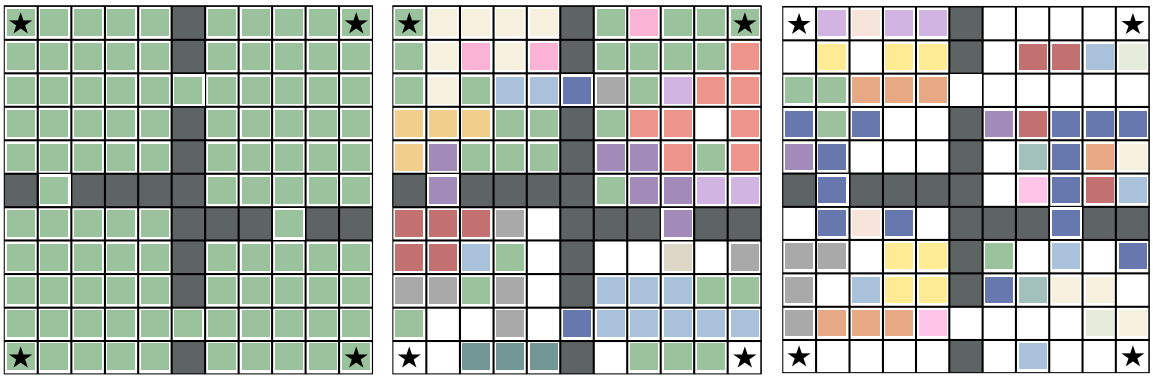
We offer an initial variant of this algorithm, which we called agent-controlled DIBS (AC-DIBS). We conduct a Four Rooms experiment similar to that of the previous section. Here, we now run the entire process of DIBS for T rounds, each time to convergence, but letting the agent's *initial* policy for that round define the stationary state distribution. Results are presented in Figure 1. Surprisingly, we find AC-DIBS *always* converges quickly to a point where ρ_ϕ^T is sufficiently close to ρ_E, when $\beta > 1$. This finding supports our supposition, along with Lemma 3, there is a route to defining a convergent form of DIBSwhen the agent's policy controls the underlying MDP.

Multiple MDPs

Second, we use our framework to compute a single abstraction sufficient for representing the demonstrator policy across different, but potentially related tasks. We suppose we are given a set of MDPs $\mathscr{M}$, each sharing a state and action space, but are allowed to vary in $\mathcal{T}$, $\mathcal{R}$, and γ. We conduct an experiment in which $|\mathscr{M}| = 4$, each with a goal in one of the four corners of the world. We run DIBS for each MDP in $\mathscr{M}$, for a fixed β, and form a master abstraction $\phi_{\mathscr{M}}$ by taking the intersection across each computed state abstraction. That is, for any state pair (s_1, s_2), for ϕ_i computed by DIBS on each MDP, we define $\phi_{\mathscr{M}}$:

$$\phi_{\mathscr{M}}(s_1) = \phi_{\mathscr{M}}(s_2) \equiv \bigwedge_{i=1}^{|\mathscr{M}|} \{\phi_i(s_1) = \phi_i(s_2)\}. \tag{42}$$

Figure 2 shows $\phi_{\mathscr{M}}$ for different values of β. All cells with the same color are grouped into the same state, except for white: all white states are each treated as their true ground state. Note that the abstraction becomes far more detailed as β increases: when β is close to 0, the algorithm prioritizes *compression*, as is reflected by Figure 2a, which only has a single state. Conversely, as β increases, we find the algorithm adds more distinctions between states, only grouping those that are close to one another or the near the same wall. When $\beta = 1$, we find that the abstraction groups the top and bottom hallways together, and the left and right hallways together. Similarly, it groups large regions of contiguous states together, such as the central group of states in the bottom right room, and the right most states in the top right room. When $\beta = 10$, we find less compression, but still find contiguous regions that have some structural similarities. For instance, the dark blue regions extending above and below the left and right doorways are grouped together, as across all four goals it is important to move up and down between the rooms. The orange groups in the top left and bottom right rooms suggest a similar structure for moving left and right. Critically, none of the abstractions are perfect: we do not know what constitutes the optimal abstraction in the case, nor how well our proposed algorithm approximates this optimal. In future work we hope to bring clarifying answers to these questions.



(a) $\beta = 0$ (b) $\beta = 1$ (c) $\beta = 10$

Figure 2: State abstractions computed by DIBS for a collection of MDPs for different values of β.

.....

Lastly, we briefly discuss the variational upper bound introduced as part of our VAE extension. In Equation 14, we reference a known result relating the KL-divergence to mutual information in the context of variational autoencoders (VAEs) Kingma and Welling (2013). Recall that a VAE is concerned with learning a probabilistic encoder-decoder pair, $q_\psi(z|x)$ and $p_\theta(x \mid z)$ respectively, for compressing raw data into a latent representation and then reconstructing the original input. During the course of training a VAE via the evidence lower bound objective (ELBO), the KL-divergence between $q_\psi(z|x)$ and some prior distribution over latent codes, $p(z)$, is minimized in expectation over the data distribution, $p(x)$; that is, $\mathbb{E}_{p(x)}[D_{KL}(q_\psi(z \mid x) \mid\mid p(z))]$ is minimized.

To define our variational upper bound to $\hat{J}$ (the stochastic IB objective for state abstraction), we leverage a known result Makhzani and Frey (2017); Kim and Mnih (2018); Dupont (2018) whose proof we replicate here for completeness with $q(z, x) = p(x)q_\psi(z|x)$ and $q(z) = \mathbb{E}_{p(x)}[q_\psi(z|x)]$:

$$
\begin{aligned}
\mathbb{E}_{p(x)}[D_{KL}(q_\psi(z \mid x) \mid\mid p(z))] &= \mathbb{E}_{p(x)}\left[\mathbb{E}_{q_\psi(z|x)}[\log \frac{q_\psi(z|x)}{p(z)}]\right] \\
&= \mathbb{E}_{q(z,x)}\left[\log\left(\frac{q_\psi(z \mid x)}{p(z)}\frac{q(z)}{q(z)}\right)\right] \\
&= \mathbb{E}_{q(z,x)}\left[\log \frac{q_\psi(z \mid x)}{q(z)}\right] + \mathbb{E}_{q(z,x)}\left[\log \frac{q(z)}{p(z)}\right] \\
&= \mathbb{E}_{q(z,x)}\left[\log \frac{q(z,x)}{q(z)p(x)}\right] + \mathbb{E}_{q(z)}\left[\log \frac{q(z)}{p(z)}\right] \\
&= I(X;Z) + D_{KL}(q(z) \mid\mid p(z)) \\
&\geq I(X;Z)
\end{aligned}
$$

.....

References

Dupont, E. 2018. Joint-vae: Learning disentangled joint continuous and discrete representations. *arXiv preprint arXiv:1804.00104*.

Kim, H., and Mnih, A. 2018. Disentangling by Factorising. *arXiv preprint arXiv:1802.05983*.

Kingma, D. P., and Welling, M. 2013. Auto-Encoding Variational Bayes. *Proceedings of the International Conference on Learning Representations*.

Makhzani, A., and Frey, B. J. 2017. Pixelgan autoencoders. In *Advances in Neural Information Processing Systems*, 1975–1985.