TibGM: A Transferable and Information-Based Graphical Model Approach for Reinforcement Learning

Tameem Adel[1] **Adrian Weller**[1 2]

Abstract

One of the challenges to reinforcement learning (RL) is scalable transferability among complex tasks. Incorporating a graphical model (GM), along with the rich family of related methods, as a basis for RL frameworks provides potential to address issues such as transferability, generalisation and exploration. Here we propose a flexible GM-based RL framework which leverages efficient inference procedures to enhance generalisation and transfer power. In our proposed transferable and information-based graphical model framework 'TibGM', we show the equivalence between our mutual information-based objective in the GM, and an RL consolidated objective consisting of a standard reward maximisation target and a generalisation/transfer objective. In settings where there is a sparse or deceptive reward signal, our TibGM framework is flexible enough to incorporate exploration bonuses depicting intrinsic rewards. We empirically verify improved performance and exploration power.

1. Introduction

Reinforcement learning (RL) is a powerful approach yet standard RL algorithms do not scale well for complex or large tasks (Bakker & Schmidhuber, 2004b), and often suffer when facing problems with sparse rewards. Exceptions include a few hierarchical reinforcement learning (HRL) frameworks (Watkins, 1989; Kaelbling et al., 1996; Parr & Russell, 1998; Sutton et al., 1999; Dietterich, 2000; Bakker & Schmidhuber, 2004b), which have the potential to improve over standard RL in terms of scalability, and to handle non-Markovian environments. Other advantages of HRL include the potential to solve subtasks independently, the reuse of learnt representations among subtasks, improved efficiency due to the task decoupling and the reduced search space (Bakker & Schmidhuber, 2004a; Cao & Ray, 2012; Schmidhuber, 2015; Florensa et al., 2017), and efficient exploration at higher levels of the hierarchy (Levy et al., 2017).

Problems related to the notion of a hierarchy, such as how to learn or implement efficient inference queries, pave the way for paradigms like probabilistic graphical models (PGMs, Jordan, 1998; Jordan et al., 1999; Wainwright & Jordan, 2008; Koller & Friedman, 2009). The ability to cast a problem as a PGM facilitates a path between objectives written specifically to match our targets, and automated learning and inference techniques enabling efficient solutions to achieve the objectives.

Graphical models grant the possibility of beginning from precise objectives, expressing them in the interpretable form of a graph, and achieving the objectives via efficient inference. Compared to previous RL frameworks that have cast the problem in a PGM, we aim at taking full advantage of this cycle here. We propose a novel information theoretic objective that aims at both maximising 'local' reward, and facilitating transfer learning (transferability) and exploration, ultimately leading to improved 'global' reward maximisation. We take advantage of a latent space disentangled into components fit for our purpose, and develop a recognition model-based variational inference procedure to achieve our objectives.

Graphical models enable earlier work in approximate inference to be adopted. A seminal example is variational autoencoders (VAEs, Kingma & Welling, 2014; Kingma et al., 2014), which effectively merge two types of models, graphical models and deep learning, into a single comprehensive framework. In VAEs, a generative and a recognition model are integrated for the sake of developing a powerful probabilistic framework utilising the recent advances in both deep generative models and scalable variational inference (Kingma & Welling, 2014; Rezende et al., 2014). Here we use similar techniques and propose an information theoretic objective which expresses our joint goals of local reward maximisation, and high transferability and generalisation power. We construct a graphical model (GM) that

[1] Department of Engineering, University of Cambridge, UK [2] The Alan Turing Institute, UK. Correspondence to: Tameem Adel <tah47@cam.ac.uk>.

Proceedings of the 36^{th} International Conference on Machine Learning, Long Beach, California, PMLR 97, 2019. Copyright 2019 by the author(s).

captures our modeling assumptions then show that inference on the GM to achieve our information theoretic objective corresponds to the quest for a two-fold RL objective targeting both reward maximisation and the enhancement of the generalisation capabilities of the GM-based framework. We show that this framework can also be used in environments with sparse or deceptive extrinsic reward signals by introducing a corresponding exploration strategy.

In our approach, the latent space representing the environment is disentangled into components which (i) aim at maximising the reward at each time step, and (ii) 'global' components corresponding to more generic, time-independent information about the environment. This disentanglement is analogous to functional theories of hierarchy in psychology (Parsons, 1940; Boker, 2002; Moody & White, 2003). To perform inference, we develop an efficient variational inference procedure, involving both a generative and a recognition model. We derive a novel analogy between our proposed information theoretic objective and an RL objective taking into account both reward maximisation and the optimisation for generalisation and transferability. We name our approach based on the Transferable Information-Based Graphical Model `TibGM`. Our proposed two-fold objective contains a term aiming at the maximisation of the extrinsic reward, together with another term which encourages the 'global' subset of the latent space not to depend heavily on the current reward (to aid generalisation).

If the reward signal is sparse, deceptive (Colas et al., 2018; Hong et al., 2018; Khadka & Tumer, 2018) or delayed (van der Pol & Oliehoek, 2016; Andreas et al., 2017), extrinsic rewards are hardly sufficient to convey the modeling objective, stimulating the need for intrinsic rewards and motivations such as exploration bonuses (Bellemare et al., 2016; Fu et al., 2017; Tang et al., 2017; Ostrovski et al., 2018; Burda et al., 2018; 2019). We show that it is possible to augment our `TibGM` framework with an unsupervised pretraining objective that does not count on extrinsic rewards.

We highlight the following contributions: 1) We propose a graphical model based on which an introduced information theoretic objective leads to the solution of RL problems (Section 3); 2) We derive a correspondence between the proposed mutual information-based objective and a two-fold RL objective, where both reward maximisation on the one hand, and generalisation and transferability on the other hand, are optimised (Section 3); 3) The introduced latent space is disentangled into 'local' components focused on maximising the reward at each time step, and 'global' latent components; 4) In cases with sparse, deceptive or very delayed extrinsic rewards, we propose an information theoretic unsupervised pretraining procedure that can further focus on exploration while still exploiting the learning and inference efficiency advantages of the graphical model (Section 4); 5) We verify our approach empirically on 16 benchmark tasks, outperforming recent state-of-the-art methods (Section 5).

2. Related work

Some works have explored links between RL and PGMs, e.g. (Dayan & Hinton, 1997; Kappen, 2005; Manfredi & Mahadevan, 2005; Neumann, 2011; Kappen et al., 2012; Levine, 2014; Abdolmaleki et al., 2018; Haarnoja et al., 2018a;b; Xu et al., 2018; Fellows et al., 2019). Out of these works, the most similar frameworks to ours are SAC (Haarnoja et al., 2018b) and LSP (Haarnoja et al., 2018a). The LSP framework (Haarnoja et al., 2018a) provides an ELBO that corresponds to reward maximisation and a maximum entropy objective. They then develop a hierarchical stochastic policy where learning done at any level can be undone at the higher level, in case the latter deems it more beneficial to do so. In addition to the exploration bonuses and the modeling and inference flexibility provided by `TibGM`, other differences to most of these models, and in particular to (Haarnoja et al., 2018a), include the following: The driving objective that leads to the equivalence with a two-fold RL objective in `TibGM`, i.e. our starting point, is the one based on mutual information, not the ELBO. In addition, the framework in (Haarnoja et al., 2018a) does not contain a recognition (inference) model whose amortisation leads to improvements in efficiency and further modeling flexibility. Another key difference is that, unlike several related frameworks, e.g. (Ziebart et al., 2008; Nachum et al., 2017; Schulman et al., 2017a; Haarnoja et al., 2017; 2018a;b; Levine, 2018; Grau-Moya et al., 2019), the objective proposed by `TibGM` does not depend on maximum entropy. The framework in (Haarnoja et al., 2018a) is focussed on the fact that higher hierarchy levels can undo lower level transformations. This is not what we pursue, as we believe that providing efficient inference in the first place would be better in terms of the computational runtime. Other RL algorithms that have links with probabilistic inference include (Todorov, 2007; Toussaint, 2009; Peters et al., 2010; Rawlik et al.; Hausman et al., 2018; Tang & Agrawal, 2018) .

Regarding pretraining strategies, Florensa et al. (2017) established an information theoretic exploration procedure that maximises the mutual information between states and actions at the top level of their hierarchy. The work in (Goyal et al., 2019) also contains an information theoretic regularisation term. In addition to the architectural ones, differences to the latter include the fact that there is no disentanglement in the learnt latent space to keep a part of it focussed on maximising the reward at each time step; there is single focus on exploration. Also, our derived analogy provides further modeling flexibility, e.g. the unsupervised pretraining procedure.

In the experiments, we compare to several other state-of-the-art algorithms, like the sample-efficient deterministic off-policy algorithm, DDPG (Lillicrap et al., 2015). We also compare to proximal policy optimization (PPO, Schulman et al., 2017b) which mostly converges to nearly deterministic policies as well. Other works involved in the comparisons include algorithms with evolutionary nature like ERL (Khadka & Tumer, 2018) and GEP-PG (Colas et al., 2018), and ProMP (Rothfuss et al., 2019), whose target is to adapt well to new similar environments.

3. Methodology

Our TibGM framework is based on designing a latent space that aims both at modeling the environment with its actions and states with high fidelity, and also at capturing the global, persistent aspects of such environment that can be transferable. The proposed latent space must be capable of accomplishing two goals: i) achieving high rewards on the tasks the RL learner is facing; and ii) constructing a representation with a high generalisation and transfer potential.

Notation and Model

Consider a Markov decision process (MDP) defined as $(\mathbf{S}, \mathbf{A}, \mathbf{P}, \mathbf{r})$ with state space $\mathbf{S}$, action space $\mathbf{A}$, transition dynamics $\mathbf{P}$, and transition reward $\mathbf{r}$. Refer to the current state, next state and current action as $\mathbf{s_t}$, $\mathbf{s_{t+1}}$ and $\mathbf{a_t}$, respectively. The transition dynamics $\mathbf{P}(\mathbf{s_{t+1}}|\mathbf{s_t}, \mathbf{a_t})$ refers to the probability of the transition from state $\mathbf{s_t}$ to state $\mathbf{s_{t+1}}$ by taking action $\mathbf{a_t}$. Without loss of generality, assume the reward $\mathbf{r}(\mathbf{s_t}, \mathbf{a_t})$ is non-positive. Denote by $\mathbf{s}_0$ the initial state, by $\tau = \{\mathbf{s}_0, \mathbf{a}_0, \mathbf{s}_1, \mathbf{a}_1, \ldots, \mathbf{s_T}, \mathbf{a_T}\}$ a trajectory, and refer to the distribution of the trajectory under policy $\pi(\mathbf{a}|\mathbf{s})$ as $\rho_\pi(\tau)$.

We propose an objective motivated by information theory, then show how it corresponds to an RL objective which aims at both (i) maximising the reward of the resulting trajectory, and (ii) constructing a transferable representation with good generalisation power.

The proposed probabilistic model is displayed in Figure 1. Assume there is a binary variable $\mathbf{b}$ (for 'best') denoting whether it is ideal (when $\mathbf{b_t} = 1$) at time step $\mathbf{t}$ to choose action $\mathbf{a_t}$. The probability that $\mathbf{b_t} = 1$ is proportionate to the reward value, as we will see later. We show that the quest for the aforementioned goals in our graphical model via the introduced information theoretic objective is equivalent to a solution to the learning problem that maximises the reward along with the optimisation for an intrinsic goal of generalisation. The latent variables $\mathbf{h_{t=1...T}}$ and $\mathbf{z}$ are responsible for the generation of the sequence of actions $\mathbf{a_t}$. Note that there is one 'local' $\mathbf{h_t}$ per time step $\mathbf{t}$, unlike $\mathbf{z}$ which is 'global'.

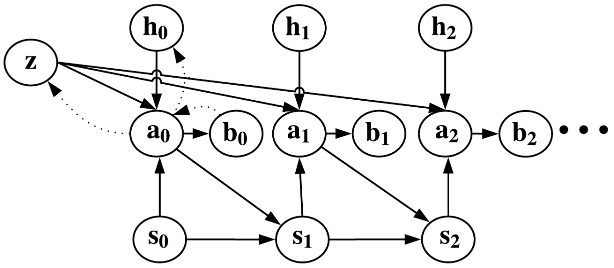


Figure 1. The graphical model depicting the proposed TibGM model. Actions at each time step depend on the states $\mathbf{s_t}$, as well as the disentangled latent space consisting of time-variant 'local' components $\mathbf{h_t}$ and 'global' components $\mathbf{z}$, encouraging generalisation. The variable $\mathbf{b_t}$ conveys information about the reward, see the text. Dotted lines (shown only at the first time step for improved readability) indicate the dependencies of the recognition model, when different from their generative counterparts. Achieving our information theoretic objective by performing variational inference on this graphical model is equivalent to optimising for a two-fold RL objective aiming at maximising the reward as well as achieving an additional generalisation/transfer goal.

In the following, we portray our mutual information-based objective. Afterwards, we derive a correspondence between our objective and an RL objective which considers both a typical reward maximisation objective and an introduced goal of estimating a latent space with high transferability and generalisation power. We then give details about the inference procedure. We aim at achieving a maximum reward, which we initially convey via maximising the mutual information between the actions $\mathbf{a_t}$ and the optimality variable $\mathbf{b_t}$, $\mathbf{I}(\mathbf{a_t}, \mathbf{b_t})$. The time-dependent latent variables $\mathbf{h_t}$ are responsible for maximising this mutual information term, which denotes the aim to choose optimal actions at each time step. Both $\mathbf{h_t}$ and the global latent components, referred to as $\mathbf{z}$, affect the actions $\mathbf{a}$ at each time step, but $\mathbf{z}$ is constrained via the other modeling objective (generalisation) expressed by the second term. This second term, $\mathbf{I}(\mathbf{z}, \mathbf{b_t}|\mathbf{s_t})$, is minimized to encourage a latent global representation with high generalisation and transfer power. The intuition here is that by minimising the mutual information between $\mathbf{z}$ and the reward directed variable $\mathbf{b_t}$, the former is induced to be free to model the global aspects that can potentially lead to a highly generalisable representation. Our overall objective can be described as follows:

$$\max_{\boldsymbol{\theta},\boldsymbol{\phi}} \ \mathbf{I}(\mathbf{a_t}, \mathbf{b_t}) - \beta\, \mathbf{I}(\mathbf{z}, \mathbf{b_t}|\mathbf{s_t}), \tag{1}$$

Where $\boldsymbol{\theta}, \boldsymbol{\phi}$ are the parameters of the generative and recognition model, respectively. The parameter β controls the degree to which the second part of the objective is imposed, i.e. how much we want to enforce our latent space to model general concepts about the environment that are not necessarily related to the extrinsic reward of the current task. A high positive value of β places great emphasis on transferability and generalisation, whereas $\beta = 0$ leads to the standard reward-maximising RL objective. For our experiments, we determine β by cross-validation. The first term in (1) is equivalent to the following:

$$\begin{aligned}
&\mathbf{I}(\mathbf{a_t}, \mathbf{b_t}) = \int_{\mathbf{a_t},\mathbf{b_t}} \mathbf{p}_{\boldsymbol\theta}(\mathbf{a_t}, \mathbf{b_t}) \log \frac{\mathbf{p}_{\boldsymbol\theta}(\mathbf{a_t}, \mathbf{b_t})}{\mathbf{p}_{\boldsymbol\theta}(\mathbf{a_t})\mathbf{p}(\mathbf{b_t})}\, d\mathbf{a_t}\, d\mathbf{b_t} \\
&= \int_{\mathbf{a_t},\mathbf{b_t}} \mathbf{p}_{\boldsymbol\theta}(\mathbf{a_t}, \mathbf{b_t}) \log \frac{\mathbf{p}_{\boldsymbol\theta}(\mathbf{a_t})\mathbf{p}_{\boldsymbol\theta}(\mathbf{b_t}|\mathbf{a_t})}{\mathbf{p}_{\boldsymbol\theta}(\mathbf{a_t})\mathbf{p}(\mathbf{b_t})}\, d\mathbf{a_t}\, d\mathbf{b_t} \\
&= \mathbb{E}_{\mathbf{p}_{\boldsymbol\theta}(\mathbf{a_t},\mathbf{b_t})}[\log \mathbf{p}_{\boldsymbol\theta}(\mathbf{b_t}|\mathbf{a_t}) - \log \mathbf{p}(\mathbf{b_t})] \\
&= \mathbb{E}_{\mathbf{p}_{\boldsymbol\theta}(\mathbf{a_t},\mathbf{b_t})}[\log \mathbf{p}_{\boldsymbol\theta}(\mathbf{b_t}|\mathbf{a_t})] + \mathbf{H}(\mathbf{b_t}) \qquad (2)
\end{aligned}$$

Since the entropy term $\mathbf{H}(\mathbf{b_t})$ in (2) does not have an impact on our optimization objective, we will not take it into account in the remaining part.

Meanwhile, regarding the second term in (1):

$$\begin{aligned}
&\mathbf{I}(\mathbf{z}, \mathbf{b_t}|\mathbf{s_t}) = \\
&\int_{\mathbf{s_t}} \mathbf{p}(\mathbf{s_t}) \int_{\mathbf{z},\mathbf{b_t}} \mathbf{p}_{\boldsymbol\theta}(\mathbf{z}, \mathbf{b_t}|\mathbf{s_t}) \log \frac{\mathbf{p}_{\boldsymbol\theta}(\mathbf{z}, \mathbf{b_t}|\mathbf{s_t})}{\mathbf{p}_{\boldsymbol\theta}(\mathbf{z}|\mathbf{s_t})\mathbf{p}_{\boldsymbol\theta}(\mathbf{b_t}|\mathbf{s_t})}\, d\mathbf{s_t} d\mathbf{z} d\mathbf{b_t} \\
&= \int_{\mathbf{s_t},\mathbf{z},\mathbf{b_t}} \mathbf{p}_{\boldsymbol\theta}(\mathbf{z}, \mathbf{b_t}, \mathbf{s_t}) \log \frac{\mathbf{p}_{\boldsymbol\theta}(\mathbf{z}|\mathbf{b_t}, \mathbf{s_t})\mathbf{p}_{\boldsymbol\theta}(\mathbf{b_t}|\mathbf{s_t})}{\mathbf{p}_{\boldsymbol\theta}(\mathbf{z}|\mathbf{s_t})\mathbf{p}_{\boldsymbol\theta}(\mathbf{b_t}|\mathbf{s_t})}\, d\mathbf{s_t} d\mathbf{z} d\mathbf{b_t} \\
&= \int_{\mathbf{s_t},\mathbf{z},\mathbf{b_t}} \mathbf{p}_{\boldsymbol\theta}(\mathbf{z}, \mathbf{b_t}, \mathbf{s_t})[\log \mathbf{p}_{\boldsymbol\theta}(\mathbf{z}|\mathbf{b_t}, \mathbf{s_t}) - \log \mathbf{p}_{\boldsymbol\theta}(\mathbf{z}|\mathbf{s_t})]\, d\mathbf{s_t} d\mathbf{z} d\mathbf{b_t} \\
&\qquad (3)
\end{aligned}$$

Let $\mathbf{q}_{\boldsymbol\phi}(\mathbf{z}|\mathbf{s_t})$ be the estimated approximation to the ground truth $\mathbf{p}(\mathbf{z}|\mathbf{s_t})$. The KL-divergence between both is:

$$\mathbb{KL}[\mathbf{p}_{\boldsymbol\theta}(\mathbf{z}|\mathbf{s_t}) \| \mathbf{q}_{\boldsymbol\phi}(\mathbf{z}|\mathbf{s_t})] \geq 0 \qquad (4)$$

$$\int \mathbf{p}_{\boldsymbol\theta}(\mathbf{z}|\mathbf{s_t}) \log \mathbf{p}_{\boldsymbol\theta}(\mathbf{z}|\mathbf{s_t}) d\mathbf{z} \geq \int \mathbf{p}_{\boldsymbol\theta}(\mathbf{z}|\mathbf{s_t}) \log \mathbf{q}_{\boldsymbol\phi}(\mathbf{z}|\mathbf{s_t}) d\mathbf{z}$$

Using (4) back in (3), we obtain:

$$\begin{aligned}
&\mathbf{I}(\mathbf{z}, \mathbf{b_t}|\mathbf{s_t}) \leq \\
&\int_{\mathbf{s_t},\mathbf{z},\mathbf{b_t}} \mathbf{p}_{\boldsymbol\theta}(\mathbf{z}, \mathbf{b_t}, \mathbf{s_t})[\log \mathbf{p}_{\boldsymbol\theta}(\mathbf{z}|\mathbf{b_t}, \mathbf{s_t}) - \log \mathbf{q}_{\boldsymbol\phi}(\mathbf{z}|\mathbf{s_t})] d\mathbf{s_t} d\mathbf{z} d\mathbf{b_t} \\
&= \int_{\mathbf{s_t},\mathbf{z},\mathbf{b_t}} \mathbf{p}_{\boldsymbol\theta}(\mathbf{b_t}, \mathbf{s_t}) \mathbf{p}_{\boldsymbol\theta}(\mathbf{z}|\mathbf{b_t}, \mathbf{s_t}) \log \frac{\mathbf{p}_{\boldsymbol\theta}(\mathbf{z}|\mathbf{b_t}, \mathbf{s_t})}{\mathbf{q}_{\boldsymbol\phi}(\mathbf{z}|\mathbf{s_t})}\, d\mathbf{s_t} d\mathbf{z} d\mathbf{b_t} \\
&\qquad (5) \\
&= \mathbb{E}_{\mathbf{p}_{\boldsymbol\theta}(\mathbf{b_t},\mathbf{s_t})} \mathbb{KL}[\mathbf{p}_{\boldsymbol\theta}(\mathbf{z}|\mathbf{b_t}, \mathbf{s_t}) \| \mathbf{q}_{\boldsymbol\phi}(\mathbf{z}|\mathbf{s_t})] \qquad (6)
\end{aligned}$$

By combining the bounds in (2) and (6), our information theoretic objective is equivalent to:

$$\begin{aligned}
&\max \quad \mathbf{I}(\mathbf{a_t}, \mathbf{b_t}) - \beta\, \mathbf{I}(\mathbf{z}, \mathbf{b_t}|\mathbf{s_t}) \\
&\equiv \max_{\boldsymbol\theta,\boldsymbol\phi} \\
&\mathbb{E}_{\mathbf{p}_{\boldsymbol\theta}(\mathbf{a_t},\mathbf{b_t})} \log \mathbf{p}_{\boldsymbol\theta}(\mathbf{b_t}|\mathbf{a_t}) - \beta \mathbb{E}_{\mathbf{p}_{\boldsymbol\theta}(\mathbf{b_t},\mathbf{s_t})} \mathbb{KL}[\mathbf{p}_{\boldsymbol\theta}(\mathbf{z}|\mathbf{b_t}, \mathbf{s_t}) \| \mathbf{q}_{\boldsymbol\phi}(\mathbf{z}|\mathbf{s_t})] \\
&\qquad (7)
\end{aligned}$$

For specific values of $\mathbf{s_t}$ and $\mathbf{a_t}$, the probability that $\mathbf{b_t} = 1$ is equal to $\mathbf{p}(\mathbf{b_t}|\mathbf{a_t}, \mathbf{s_t}) = \mathrm{e}^{\mathbf{r}(\mathbf{a_t},\mathbf{s_t})}$, in order to keep it as a probability value between 0 and 1, similar to (Haarnoja et al., 2018a). The rewards $\mathbf{r}$ are assumed to have non-positive values as in (Haarnoja et al., 2018a). The first term in (7) is equivalent to the following:

$$\begin{aligned}
&\mathbb{E}_{\mathbf{p}_{\boldsymbol\theta}(\mathbf{a_t},\mathbf{b_t})} \log \mathbf{p}_{\boldsymbol\theta}(\mathbf{b_t}|\mathbf{a_t}) = \\
&\int_{\mathbf{a_t},\mathbf{b_t},\mathbf{s_t}} \mathbf{p}_{\boldsymbol\theta}(\mathbf{a_t}, \mathbf{b_t}, \mathbf{s_t}) \log \mathbf{p}_{\boldsymbol\theta}(\mathbf{b_t}|\mathbf{a_t}, \mathbf{s_t}) d\mathbf{a_t} d\mathbf{b_t} d\mathbf{s_t}, \qquad (8)
\end{aligned}$$

where $\mathbf{p}_{\boldsymbol\theta}(\mathbf{b_t}|\mathbf{a_t}, \mathbf{s_t})$ can be used since $\mathbf{b_t}$ and $\mathbf{s_t}$ are independent given $\mathbf{a_t}$. From (7) and (8), we can then see the equivalence between the objective introduced here by `TibGM` and an RL objective that typically maximises the reward -first term in (7)- along with an additional goal. The second term in (7) aims at guaranteeing that some components $\mathbf{z}$ of the latent space aim at capturing generalised concepts of the environment, i.e. concepts and aspects not particularly related to the specific optimality conditions and targets of the current task.

Inference

Recall that the observed variables at each time step are the current state $\mathbf{s_t}$ and the optimality variable $\mathbf{b_t}$. To perform inference on the model proposed and described above, we construct a recognition model (Stuhlmuller et al., 2013; Kingma & Welling, 2014; Rezende et al., 2014) whose parameters $\boldsymbol\phi$ approximate the true posterior, along with the generative model with parameters $\boldsymbol\theta$. Our inference procedure aims at approximately inferring an optimal policy with high fidelity. As such, similar to other works portraying the reinforcement learning problem in a graphical model setting (Strehl et al., 2006; 2009; Fu et al., 2018; Haarnoja et al., 2018a), we constrain the dynamics, that will be estimated by sampling, to be equal to the true dynamics. The generative and recognition models developed here are the components of the parameterised distribution estimation of the policy.

In the recognition model, the dependency between the potentially generic (and rather constant throughout time steps) $\mathbf{z}$ and $\mathbf{a_t}$, $\mathbf{q}_{\boldsymbol\phi}(\mathbf{z}|\mathbf{a_t})$, is modeled via orthogonal Sylvester normalizing flows (van den Berg et al., 2018). On the other hand, the dependency between each $\mathbf{h_t}$ and $\mathbf{a_t}$, $\mathbf{q}_{\boldsymbol\phi}(\mathbf{h_t}|\mathbf{a_t})$ is modeled via a Gaussian transformation whose parameters are computed via a neural network. Sylvester normalizing flows (SNFs, van den Berg et al., 2018) are a state-of-the-art generalisation of planar flows and they lead to much more flexible density transformations in addition to improved overall performance. The intuition (which is empirically supported) behind using a Gaussian, whose parameters are estimated via a neural network, to estimate $\mathbf{q}_{\boldsymbol\phi}(\mathbf{h_t}|\mathbf{a_t})$ is to try and reserve the modeling power of $\mathbf{h_t}$ to estimate solely time-dependent aspects, while keeping the rest of the features into the more powerful, SNF based $\mathbf{q_\phi(z|a_t)}$. A rather similar trick has been successfully imposed in (Li & Mandt, 2018), where they assigned a low-dimensional representation to model the time-dependent latent component.

SNFs (van den Berg et al., 2018), like other normalizing flows (NFs) (Rezende & Mohamed, 2015; Kingma et al., 2016) apply a chain of invertible parameterized transformations, $\mathbf{f_t}$, $\mathbf{t} = \mathbf{1}, \dots, \mathbf{T}$, to its input $\mathbf{a}$ such that the outcome of the last iteration, $\mathbf{z} = \mathbf{z_T}$ has a more flexible distribution that in our algorithm is optimized for high fidelity in representing the environment aspects preserved through time. Each transformation step is indexed by $\mathbf{t}$, where $\mathbf{z_0}$ is an initial random variable with (Gaussian) density $\mathbf{q_0(z_0)}$ that is successively transformed through a chain of transformations $\mathbf{f_1}, \dots, \mathbf{f_T}$ (Rezende & Mohamed, 2015):

$$\mathbf{z_t} = \mathbf{f_t}(\mathbf{z_{t-1}}, \mathbf{a}) \quad \forall \mathbf{t} = \mathbf{1} \dots \mathbf{T} \tag{9}$$

The probability density function of the ultimate latent representation, $\mathbf{z} = \mathbf{z_T}$, can be computed provided that the determinant of the Jacobian of each of the transformations, $\mathbf{det}(\mathbf{f_t})$, can be computed. The probability density $\mathbf{q_\phi(z|a)} = \mathbf{q_T(z_T|a)}$ can be expressed as follows:

$$\log \mathbf{q_T}(\mathbf{z_T}|\mathbf{a}) = \log \mathbf{q_0}(\mathbf{z_0}|\mathbf{a}) - \sum_{\mathbf{t}=\mathbf{1}}^{\mathbf{T}} \log \left| \det \tfrac{d\mathbf{z_t}}{d\mathbf{z_{t-1}}} \right|, \\ \mathbf{z} = \mathbf{z_T}. \tag{10}$$

SNFs have the following form:

$$\mathbf{f_t}(\mathbf{z_{t-1}}) = \mathbf{z_{t-1}} + \mathbf{A}\mathbf{u}(\mathbf{B}\mathbf{z_{t-1}} + \mathbf{c}), \tag{11}$$

where $\mathbf{A} \in \mathbb{R}^{D \times M}, \mathbf{B} \in \mathbb{R}^{M \times D}$, $\mathbf{c} \in \mathbb{R}^{M}$, and $M < D$, and $\mathbf{u}$ is a nonlinearity, for a single layer MLP with M hidden units (van den Berg et al., 2018). In addition to their aforementioned advantages, SNFs provide an efficient solution to the single-unit bottleneck problem of planar flows, which as argued by Kingma et al. (2016), is due to the fact that (without an SNF) the impact of the second term in (11) would saturate to be similar to that of a single-neuron MLP. SNFs lead to a flexible bottleneck since the Jacobian determinant of the transformation can now be obtained using $M < D$ dimensions[1]. Each map from $\mathbf{a}$ up to $\mathbf{z}$ has the form given in (11). Thus:

$$\mathbf{z} = \mathbf{f_T} \circ \mathbf{f_{T-1}} \circ \dots \circ \mathbf{f_1}(\mathbf{a}). \tag{12}$$

Other versions of normalizing flows have been used before in similar contexts, such as (Haarnoja et al., 2018a). However, unlike there, which was fully based on a generative model with no recognition model, we use a different type of normalizing flows (SNFs), and we use them in our recognition model.

We now describe how to implement the objective. At time step $\mathbf{t}$, the first term in (7) is:

$$\mathbb{E}_{\mathbf{p_\theta(a_t,b_t)}} \log \mathbf{p_\theta(b_t|a_t)} = \int_{\mathbf{a_t,b_t}} \mathbf{p_\theta(a_t, b_t)} \log \mathbf{p_\theta(b_t|a_t)} d\mathbf{a_t} d\mathbf{b_t} \\ = \int_{\mathbf{z,h_t,s_t,a_t,b_t}} \mathbf{p_\theta(z, h_t, s_t, a_t, b_t)} \log \mathbf{p_\theta(b_t|a_t)} d\mathbf{z} d\mathbf{h_t} d\mathbf{s_t} d\mathbf{a_t} d\mathbf{b_t} \tag{13}$$

The joint probability in (13) can be obtained through samples from the joint distribution of the generative model at the current time step. Regarding $\mathbf{p_\theta(b_t|a_t)}$, it is expressed using the sampled $\mathbf{s_t}$ and the reward, $\mathrm{e}^{\mathbf{r(a_t,s_t)}}$.

Now consider the second term in (7). From (5) it is equivalent to:

$$\mathbf{I(z, b_t|s_t)} = \int_{\mathbf{z,s_t,b_t}} \mathbf{p_\theta(z, s_t, b_t)} \log \frac{\mathbf{p_\theta(z|b_t, s_t)}}{\mathbf{q_\phi(z|s_t)}} d\mathbf{z} d\mathbf{s_t} d\mathbf{b_t} \\ = \int_{\mathbf{z,s_t,b_t}} \mathbf{p_\theta(z, s_t, b_t)} [\log \mathbf{p_\theta(z|b_t, s_t)} - \mathbf{q_\phi(z|s_t)}] d\mathbf{z} d\mathbf{s_t} d\mathbf{b_t} \\ = \int_{\mathbf{z,h_t,s_t,a_t,b_t}} \mathbf{p_\theta(z, h_t, s_t, a_t, b_t)} [\log \mathbf{p_\theta(z, b_t, s_t)} - \mathbf{p_\theta(b_t, s_t)} \\ - \mathbf{q_\phi(z|s_t)}] d\mathbf{z} d\mathbf{h_t} d\mathbf{s_t} d\mathbf{a_t} d\mathbf{b_t} \tag{14}$$

The generative terms -all except $\mathbf{q_\phi(z|s_t)}$- are obtained by sampling from the joint distribution of the generative model. Regarding the recognition model, the reparameterisation trick (Kingma & Welling, 2014) is used in estimating $\mathbf{a_t}$. The action variable $\mathbf{a_t}$ is expressed as: $\mathbf{a_t} = \mathbf{f_\phi(s_t, b_t, \epsilon)}$. The function $\mathbf{f_\phi}$ is a deterministic function of $\mathbf{s_t}$, $\mathbf{b_t}$ and the Gaussian random variable ϵ. One of the advantages of the reparameterisation trick is that the noise term ϵ becomes independent of the model parameters, which facilitates taking gradients (Kingma & Welling, 2014; Alemi et al., 2017). Afterwards, we can proceed with computing $\mathbf{q_\phi(z|s_t)}$, where, as mentioned above, $\mathbf{z|a_t}$ is modeled via an SNF.

$$\mathbf{q_\phi(z|s_t)} \propto \\ \mathbf{q_\phi}\Big(\mathbf{z}|\mu_\phi(\mathbf{f_\phi(s_t, b_t, \epsilon)}), \sigma_\phi(\mathbf{f_\phi(s_t, b_t, \epsilon)})\Big) \mathbf{q_\phi(a_t|s_t, b_t)} = \\ \mathbf{q_\phi}\Big(\mathbf{z}|\mu_\phi(\mathbf{f_\phi(s_t, b_t, \epsilon)}), \sigma_\phi(\mathbf{f_\phi(s_t, b_t, \epsilon)})\Big) \mathcal{N}(\epsilon|\mathbf{0}, \mathbf{I}) \tag{15}$$

4. Sparse or Deceptive Rewards

There are cases where there is hardly any reward signal, and therefore the variable $\mathbf{b}$ is not prevalent anymore, e.g. in environments with rare or sparse rewards. In addition, there are also environments with deceptive reward signals.

[1] For the proof and full details, see van den Berg et al. (2018).

In such cases, the dependence on $\mathbf{b}$ is either not feasible, misleading or both. The flexibility provided by the graphical modeling nature of `TibGM` enables us to model these cases via the following diversification procedure. This procedure aims to enhance exploration power, encouraging the learner to visit new states, which becomes more important when it is not possible to count on the reward signal, and also still to provide a level of generalisation as a foundation of `TibGM`. As we will show in the experiments, this procedure can be used for unsupervised pretraining prior to the supervised stage involving rewards, in cases when such rewards are available.

Standing along the mutual information based objectives, we propose to handle the sparse-reward case with an information bottleneck based objective which encourages exploration when there is no $\mathbf{b}$ (hence the name `ExTibGM`). The intuition is that, with no extrinsic reward available, the latent $\mathbf{z}$ shall now be further focused on exploration, and it should hence be made maximally informative about the states $\mathbf{s}$ at the expense of being compressive about $\mathbf{a}$. Being maximally informative about the states $\mathbf{s}$ is key to the emphasis on exploring new states. On the other hand, being compressive about the actions $\mathbf{a}$ provides a level of generalisation as promised by the latent space $\mathbf{z}$. This is similar, but not identical, to the DIAYN algorithm in (Eysenbach et al., 2019), where they also base their reasoning on an information theoretic objective. However, in addition to other differences in the objectives, they base their reasoning in DIAYN on diversifying policies via maximum entropy, while we base ours on generalisation via a different objective.

With no reward available for our approach, the latent space, now consisting only of $\mathbf{z}$, aims at enhancing the exploration power by maximising the opportunity of visiting new states.[2] Other information bottleneck objectives have been introduced before within VAE-based frameworks, e.g. (Alemi et al., 2017; Adel et al., 2018; Alemi et al., 2018). Our objective at time step $\mathbf{t}$ is thus defined as:

$$\mathbf{IB}(\mathbf{z}, \mathbf{s_t}, \mathbf{a_t}) = \mathbf{I}(\mathbf{z}, \mathbf{s_t}) - \boldsymbol{\alpha}\mathbf{I}(\mathbf{z}, \mathbf{a_t}), \tag{16}$$

where $\boldsymbol{\alpha}$ is a parameter that controls the level of generalisation provided by the objective. Similar to β, we set $\boldsymbol{\alpha}$ by cross-validation in our experiments.

We begin by analyzing the first term in (16), $\mathbf{I}(\mathbf{z}, \mathbf{s_t})$:

$$\begin{aligned}\mathbf{I}(\mathbf{z}, \mathbf{s_t}) &= \int_{\mathbf{z},\mathbf{s_t}} \mathbf{p}_{\boldsymbol{\theta}}(\mathbf{z}, \mathbf{s_t}) \log \frac{\mathbf{p}_{\boldsymbol{\theta}}(\mathbf{s_t})\mathbf{p}_{\boldsymbol{\theta}}(\mathbf{z}|\mathbf{s_t})}{\mathbf{p}(\mathbf{s_t})\mathbf{p}_{\boldsymbol{\theta}}(\mathbf{z})}\, d\mathbf{z}\, d\mathbf{s_t} \\ &\geq \int_{\mathbf{z},\mathbf{s_t}} \mathbf{p}_{\boldsymbol{\theta}}(\mathbf{z}, \mathbf{s_t}) \log \frac{\mathbf{q}_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{s_t})}{\mathbf{p}_{\boldsymbol{\theta}}(\mathbf{z})}\, d\mathbf{z}\, d\mathbf{s_t} \end{aligned} \tag{17}$$

due to the non-negativity of the KL-divergence between $\mathbf{p}_{\boldsymbol{\theta}}(\mathbf{z}|\mathbf{s_t})$ and $\mathbf{q}_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{s_t})$, similar to step (4). Therefore, $\mathbf{I}(\mathbf{z}, \mathbf{s_t})$ is equivalent to:

$$\mathbf{I}(\mathbf{z}, \mathbf{s_t}) = \mathbb{E}_{\mathbf{p}_{\boldsymbol{\theta}}(\mathbf{z},\mathbf{s_t})}[\log \mathbf{q}_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{s_t}) - \log \mathbf{p}_{\boldsymbol{\theta}}(\mathbf{z})] \tag{18}$$

Regarding the second term $\mathbf{I}(\mathbf{z}, \mathbf{a_t})$, and using the non-negativity of $\mathbb{KL}[\mathbf{p}_{\boldsymbol{\theta}}(\mathbf{z})\|\mathbf{q}_{\boldsymbol{\phi}}(\mathbf{z})]$:

$$\begin{aligned}\mathbf{I}(\mathbf{z}, \mathbf{a_t}) &\leq \int_{\mathbf{z},\mathbf{a_t}} \mathbf{p}(\mathbf{a_t})\mathbf{p}_{\boldsymbol{\theta}}(\mathbf{z}|\mathbf{a_t}) \log \frac{\mathbf{p}_{\boldsymbol{\theta}}(\mathbf{z}|\mathbf{a_t})}{\mathbf{q}_{\boldsymbol{\phi}}(\mathbf{z})}\, d\mathbf{z}\, d\mathbf{a_t} \\ &= \int_{\mathbf{a_t}} \mathbf{p}(\mathbf{a_t})\mathbb{KL}[\mathbf{p}_{\boldsymbol{\theta}}(\mathbf{z}|\mathbf{a_t})\|\mathbf{q}_{\boldsymbol{\phi}}(\mathbf{z})]\, d\mathbf{a_t} \end{aligned} \tag{19}$$

Using the bounds in (18) and (19) in the objective in (16):

$$\begin{aligned}\mathbf{IB}(\mathbf{z}, \mathbf{s_t}, \mathbf{a_t}) \geq& \mathbb{E}_{\mathbf{p}_{\boldsymbol{\theta}}(\mathbf{z},\mathbf{s_t})}[\log \mathbf{q}_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{s_t}) - \log \mathbf{p}_{\boldsymbol{\theta}}(\mathbf{z})] \\ &- \boldsymbol{\alpha}\mathbb{E}_{\mathbf{p}(\mathbf{a_t})}\mathbb{KL}[\mathbf{p}_{\boldsymbol{\theta}}(\mathbf{z}|\mathbf{a_t})\|\mathbf{q}_{\boldsymbol{\phi}}(\mathbf{z})] \end{aligned} \tag{20}$$

The first term in (18) -which can also be found in DIAYN (Eysenbach et al., 2019), though with a different implementation using no normalizing flows- is enforcing exploration since maximising this term leads to gaining further reward from visiting states that are easy to discriminate. A sample is taken from the Gaussian $\mathbf{p}_{\boldsymbol{\theta}}(\mathbf{z})$, whereas the corresponding computation of $\mathbf{q}_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{s_t})$ proceeds using SNFs in the recognition model as described in Section 3, but with no $\mathbf{b}$ in the model anymore. The second term, which is a novel way of enforcing generalisation in environments with sparse rewards to the best of our knowledge, induces $\mathbf{z}$ to be general by being less dependent on $\mathbf{a}$. Here we use another SNF in the generative model to allow for an analytical computation of the KL-divergence.

5. Experiments

We evaluate empirically the effectiveness of our `TibGM` and `ExTibGM` frameworks, comparing to several state-of-the-art algorithms: DDPG (Lillicrap et al., 2015), LSP (Haarnoja et al., 2018a), SAC[3] (Haarnoja et al., 2018b), PPO (Schulman et al., 2017b), ERL (Khadka & Tumer, 2018), DIAYN (Eysenbach et al., 2019), VIREL (Fellows et al., 2019), GEP-PG (Colas et al., 2018) and ProMP (Rothfuss et al., 2019).

We conduct our experiments on six standard benchmark tasks from across OpenAI Gym (Brockman et al., 2016) and rllab (Duan et al., 2016): Swimmer (rllab), Hopper (v1), Walker2d (v1), HalfCheetah (v1), Ant (rllab) and Humanoid (rllab). We also consider the Continuous Mountain Car (CMC) environment, a control benchmark with deceptive reward properties (Lehman et al., 2017; Colas et al., 2018).

[2] We suggest that there is not enough supervision in such case to disentangle our latent space into time-dependent and time-invariant subsets.

[3] We build our code on the top of SAC, https://github.com/haarnoja/sac.

We evaluate the following aspects: 1) Performance resulting from our disentangled latent space-based approach TibGM and its inference model; 2) Impact of the exploration-focused version of our framework ExTibGM on a deceptive-reward problem; 3) Impact of using the pretraining of ExTibGM and the resulting diversity on computational run-time; and 4) How TibGM fares in a policy transfer-learning setup.

Confidence intervals are shown in all the plots. Unless noted otherwise, each experiment was repeated 50 times and significance has been tested via a paired t-test with significance level at 5%. Cross-validation was used to estimate optimum Values of $\beta = 0.3$ and $\boldsymbol{\alpha} = 0.2$. In all plots, we run a sufficient number of time steps until we observe that all prior methods appear to reach saturation (Haarnoja et al. (2018a) used a similar number of steps). Other experimental details are given in the Appendix.

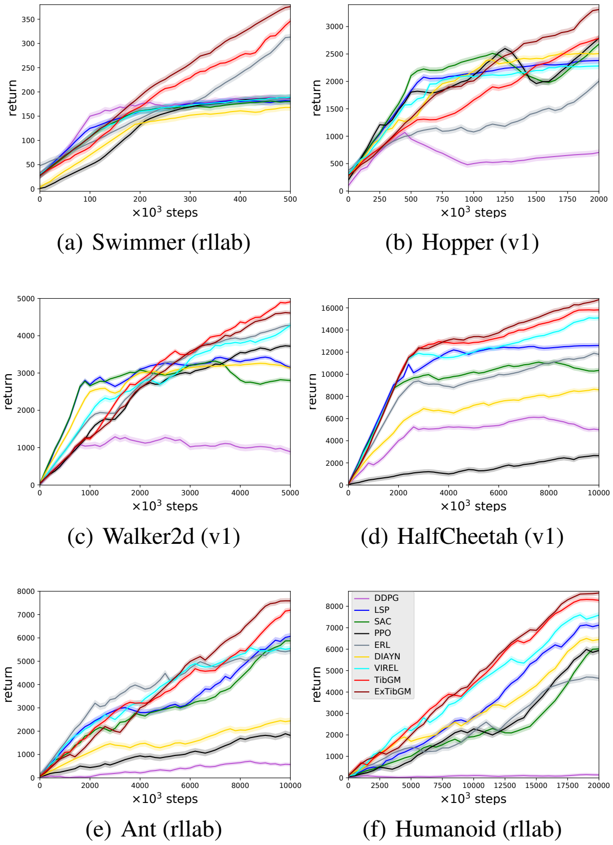
5.1. Average Reward on Benchmark Tasks

We compare our two algorithms TibGM and ExTibGM, where the latter refers to adopting the introduced exploration strategy as an unsupervised pretraining procedure. The results of the total expected return are displayed in Figure 2. Most of the algorithms we compare to here, e.g. DDPG and PPO are single-minded on the expected return, so this comparison is favourable to them against algorithms like TibGM, ExTibGM and maximum entropy based algorithms that inherently optimise for a two-fold objective.

As can be seen in Figure 2, the proposed ExTibGM, which corresponds to TibGM preceded with the introduced pretraining procedure, achieves significantly better results on all the 6 tasks. On the other hand, TibGM, i.e. with no pretraining, achieves better results than the rest of the competitors in 5 out of the 6 tasks. The latter achieves a significantly higher return than ExTibGM on the Walker2d task. In such case, it can be that too much exploration has confused the algorithm, and the vanilla TibGM is more suitable to the task. As such, the proposed ExTibGM and TibGM, especially the former, achieve state-of-the-art performance on the 6 tasks.

5.2. Exploration Efficiency on CMC

We evaluate the exploration potential of ExTibGM on a task with deceptive rewards. The standard CMC benchmark, which has been used in (Lehman et al., 2017; Colas et al., 2018), attains special deceptive reward characteristics. CMC is a control benchmark where an object (vehicle) whose the set of actions are $\{-1, 0, 1\}$ must reach its goal at the top of a hill by gaining momentum and accelerating from another hill. One of the interesting exploration issues raised by CMC is that large accelerations are necessary to reach the goal, but larger accelerations cause larger



(a) Swimmer (rllab)

(b) Hopper (v1)

(c) Walker2d (v1)

(d) HalfCheetah (v1)

(e) Ant (rllab)

(f) Humanoid (rllab)

Figure 2. Total expected return on 6 benchmark tasks. Thick lines in the middle of each curve indicate the average performance, while the standard deviations over 50 random seeds are shown by the shaded regions. The proposed algorithm ExTibGM, which corresponds to TibGM preceded with the introduced pretraining procedure, achieves significantly better results on all the 6 tasks. On the other hand, TibGM, i.e. with no pretraining, achieves better results than the rest of the competitors in 5 out of the 6 tasks. The latter achieves a significantly higher return than ExTibGM on the Walker2d task. The proposed framework therefore achieves state-of-the-art performance on the 6 tasks.

penalties too (Colas et al., 2018). As such, the agent should figure out how to perform the smallest sufficient accelerations (Lehman et al., 2017). Also, the CMC environment may mislead the agent since it provides a deceptive reward signal (Colas et al., 2018). Thus, the ability to wisely explore the environment in CMC is of paramount importance.

As described in (Colas et al., 2018), one of the fundamental exploration issues raised by CMC is the time at which the goal is first reached, which is very critical. In Figure 3, we display the histograms resulting from performing 1000 trials and showing the number of steps required before the goal is reached for the first time. We compare to two variants of GEP (Colas et al., 2018). The proposed ExTibGM reaches the goal much earlier. In 632 out of the 1000 trials, the goal was reached in the first 5,000 steps by ExTibGM,

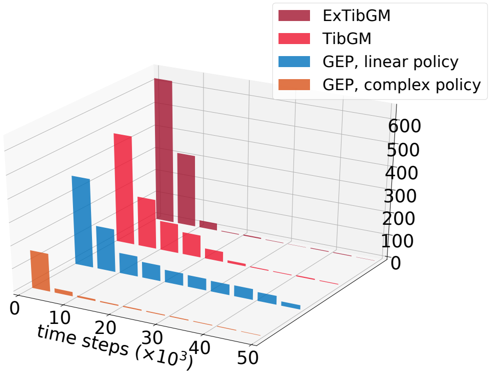
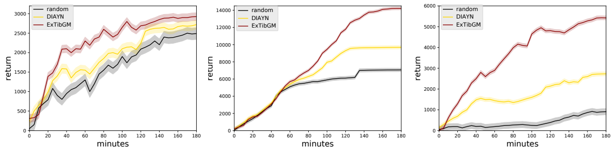


Figure 3. Histograms showing the number of steps required to reach the goal for the first time in CMC. On average, the proposed `ExTibGM` reaches the goal much earlier. In 632 out of the 1000 trials, the goal was reached in the first 5,000 steps by `ExTibGM`, compared to 394 by the best GEP version.



(a) Swimmer (rllab) (b) Hopper (v1) (c) Ant (rllab)

Figure 4. Impact of the pretraining procedure of `ExTibGM` and DIAYN on the run-time of 3 benchmarks. `ExTibGM` significantly accelerates the learning procedure in 2 of the 3.

compared to 394 by the best GEP version.

5.3. Accelerating Learning with Pretraining

We evaluate the impact of the pretraining performed in `ExTibGM` and its exploration on the computational run-time. Figure 4 displays the results showing the impact of pretraining performed via DIAYN (Eysenbach et al., 2019) and by `ExTibGM`, along with a random initialisation baseline, on the run-time of the Hopper, HalfCheetah and Ant benchmarks. Similar to the setting of an identical experiment in (Eysenbach et al., 2019), the times spent during our amortised unsupervised pretraining are omitted from the plot, since the assumption is that the bottleneck is in the supervised training. We are doing the same for both DIAYN and `ExTibGM` anyway, so this does not favour the proposed framework. As can be seen in Figure 4, the pretraining in `ExTibGM` (significantly) accelerates the learning procedure in (two of) the three benchmarks.

5.4. Transfer Learning

We evaluate `TibGM` on 6 transfer learning tasks that require adaptation (Rothfuss et al., 2019). In two of the tasks, the HalfCheetah and the Walker agents need to keep switching between walking forward and backward. The Ant and Humanoid should adapt to run in various directions. In the final two tasks, the Hopper and Walker have to adapt to different dynamic configurations (Rothfuss et al., 2019). In Table 1, we compare to two state-of-the-art transfer learning algorithms, ProMP (Rothfuss et al., 2019) and InfoBot (Goyal et al., 2019), in terms of the total expected return after 2×10^7 steps. The proposed `TibGM` achieves significant improvements and clearly leads to state-of-the-art results on the 6 adaptation tasks. This demonstrates that the the latent components $\mathbf{z}$ seem to have managed to capture the common aspects, which stand across the changes in the domain. Moreover the variance resulting from `TibGM` is lower than the competitors in 5 out of the 6 tasks.

Table 1. Total expected return on 6 adaptation tasks. `TibGM` achieves significant improvements and clearly leads to state-of-the-art results on the 6 adaptation tasks. Moreover, the variance resulting from `TibGM` is the lowest among competitors in 5 out of the 6 tasks. Bold refers to an expected return value that is significantly better than its competitors. Significance is tested using the same paired t-test described above.

	`TibGM`	ProMP	InfoBot
HopperRandParams	**912** ± **36**	438 ± 60	679 ± 54
WalkerRandParams	**848** ± **49**	525 ± 72	458 ± 44
HalfCheetahFwdBack	**1187** ± **77**	735 ± 93	714 ± 84
WalkerFwdBack	**953** ± **26**	542 ± 51	410 ± 58
AntRandDir	**684** ± **44**	218 ± 48	311 ± 89
HumanoidRandDir	**1186** ± **82**	527 ± 113	334 ± 89

6. Conclusion

We introduced an RL framework which leverages the expressiveness and power of graphical models. We defined a novel information theoretic objective, and showed its correspondence to an RL objective aiming at both maximising the reward, and facilitating transfer learning and exploration. We developed an inference procedure based on state-of-the-art advances in variational inference. The latent space representing the policy is disentangled into local components focused on reward maximisation, and global components capturing information which is useful across different environment settings. We also introduced an unsupervised information theoretic pretraining strategy, demonstrating the flexibility of our framework, which further focuses on exploration, and performs well in environments with sparse or deceptive reward signals.

Acknowledgements

Both authors thank the anonymous reviewers for helpful comments, and acknowledge support from the Leverhulme Trust via the CFI. AW acknowledges support from the

David MacKay Newton research fellowship at Darwin College and The Alan Turing Institute under EPSRC grant EP/N510129/1 & TU/B/000074.

References

Abdolmaleki, A., Springenberg, J., Tassa, Y., Munos, R., Heess, N., and Riedmiller, M. Maximum a posteriori policy optimisation. *International Conference on Learning Representations (ICLR)*, 2018.

Adel, T., Ghahramani, Z., and Weller, A. Discovering interpretable representations for both deep generative and discriminative models. *International Conference on Machine Learning (ICML)*, 2018.

Alemi, A., Fischer, I., Dillon, J., and Murphy, K. Deep variational information bottleneck. *International Conference on Learning Representations (ICLR)*, 2017.

Alemi, A., Poole, B., , Fischer, I., Dillon, J., Saurous, R., and Murphy, K. Fixing a broken ELBO. *International Conference on Machine Learning (ICML)*, 2018.

Andreas, J., Klein, D., and Levine, S. Modular multitask reinforcement learning with policy sketches. *International Conference on Machine Learning (ICML)*, 2017.

Bakker, B. and Schmidhuber, J. Hierarchical reinforcement learning with subpolicies specializing for learned subgoals. *Neural Networks and Computational Intelligence*, 2004a.

Bakker, B. and Schmidhuber, J. Hierarchical reinforcement learning based on subgoal discovery and subpolicy specialization. *Conf. on Intelligent Autonomous Systems*, 8: 438–445, 2004b.

Bellemare, M., Srinivasan, S., Ostrovski, G., Schaul, T., Saxton, D., and Munos, R. Unifying count-based exploration and intrinsic motivation. *Advances in neural information processing systems (NIPS)*, 2016.

Boker, S. Consequences of continuity: The hunt for intrinsic properties within parameters of dynamics in psychological processes. *Multivariate Behavioral Research*, 37: 405–422, 2002.

Brockman, G., Cheung, V., Pettersson, L., Schneider, J., Schulman, J., Tang, J., and Zaremba, W. OpenAI Gym. *arXiv preprint arXiv:1606.01540*, 2016.

Burda, Y., Edwards, H., Pathak, D., Storkey, A., Darrell, T., and Efros, A. Large-scale study of curiosity-driven learning. *arXiv preprint arXiv:1808.04355*, 2018.

Burda, Y., Edwards, H., Storkey, A., and Klimov, O. Exploration by random network distillation. *International Conference on Learning Representations (ICLR)*, 2019.

Cao, F. and Ray, S. Bayesian hierarchical reinforcement learning. *Advances in neural information processing systems (NIPS)*, 2012.

Colas, C., Sigaud, O., and Oudeyer, P. GEP-PG: Decoupling exploration and exploitation in deep reinforcement learning algorithms. *International Conference on Machine Learning (ICML)*, 2018.

Dayan, P. and Hinton, G. Using expectation-maximization for reinforcement learning. *Neural Computation*, 9:271–278, 1997.

Dietterich, T. Hierarchical reinforcement learning with the MAXQ value function decomposition. *Journal of Artificial Intelligence Research*, 13:227–303, 2000.

Duan, Y., Chen, X., Houthooft, R., Schulman, J., and Abbeel, P. Benchmarking deep reinforcement learning for continuous control. *International Conference on Machine Learning (ICML)*, pp. 1329–1338, 2016.

Eysenbach, B., Gupta, A., Ibarz, J., and Levine, S. Diversity is all you need: Learning skills without a reward function. *International Conference on Learning Representations (ICLR)*, 2019.

Fellows, M., Mahajan, A., Rudner, T., and Whiteson, S. VIREL: A variational inference framework for reinforcement learning. *Artificial Intelligence and Statistics (AISTATS)*, 2019.

Florensa, C., Duan, Y., and Abbeel, P. Stochastic neural networks for hierarchical reinforcement learning. *International Conference on Learning Representations (ICLR)*, 2017.

Fu, J., Co-Reyes, J., and Levine, S. Exploration with exemplar models for deep reinforcement learning. *Advances in neural information processing systems (NIPS)*, 2017.

Fu, J., Singh, A., Ghosh, D., Yang, L., and Levine, S. Variational inverse control with events: A general framework for data-driven reward definition. *Advances in neural information processing systems (NIPS)*, 2018.

Goyal, A., Islam, R., Strouse, D., Ahmed, Z., Larochelle, H., Botvinick, M., Levine, S., and Bengio, Y. Transfer and exploration via the information bottleneck. *International Conference on Learning Representations (ICLR)*, 2019.

Grau-Moya, J., Leibfried, F., and Vrancx, P. Soft Q-learning with mutual-information regularization. *International Conference on Learning Representations (ICLR)*, 2019.

Haarnoja, T., Tang, H., Abbeel, P., and Levine, S. Reinforcement learning with deep energy-based policies. *International Conference on Machine Learning (ICML)*, 2017.

Haarnoja, T., Hartikainen, K., Abbeel, P., and Levine, S. Latent space policies for hierarchical reinforcement learning. *International Conference on Machine Learning (ICML)*, 2018a.

Haarnoja, T., Zhou, A., Abbeel, P., and Levine, S. Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. *International Conference on Machine Learning (ICML)*, 2018b.

Hausman, K., Springenberg, J., Wang, Z., Heess, N., and Riedmiller, M. Learning an embedding space for transferable robot skills. *International Conference on Learning Representations (ICLR)*, 2018.

Hong, Z., Shann, T., Su, S., Chang, Y., Fu, T., and Lee, C. Diversity-driven exploration strategy for deep reinforcement learning. *Advances in neural information processing systems (NIPS)*, 2018.

Jordan, M. *Learning in graphical models*. Springer Science and Business Media, 1998.

Jordan, M., Ghahramani, Z., Jaakkola, T., and Saul, L. An introduction to variational methods for graphical models. *Machine Learning*, pp. 183–233, 1999.

Kaelbling, L., Littman, M., and Moore, A. Reinforcement learning: A survey. *Journal of Artificial Intelligence Research*, 4:237–285, 1996.

Kappen, H. Path integrals and symmetry breaking for optimal control theory. *Journal of Statistical Mechanics: Theory And Experiment*, 11, 2005.

Kappen, H., Gomez, V., and Opper, M. Optimal control as a graphical model inference problem. *Machine Learning*, 87:159–182, 2012.

Khadka, S. and Tumer, K. Evolution-guided policy gradient in reinforcement learning. *Advances in neural information processing systems (NIPS)*, 2018.

Kingma, D. and Welling, M. Auto-encoding variational Bayes. *International Conference on Learning Representations (ICLR)*, 2014.

Kingma, D., Rezende, D., Mohamed, S., and Welling, M. Semi-supervised learning with deep generative models. *Advances in neural information processing systems (NIPS)*, 28:3581–3589, 2014.

Kingma, D., Salimans, T., Jozefowicz, R., Chen, X., Sutskever, I., and Welling, M. Improved variational inference with inverse autoregressive flow. *Advances in neural information processing systems (NIPS)*, 30, 2016.

Koller, D. and Friedman, N. *Probabilistic graphical models: Principles and techniques*. MIT Press, 2009.

Lehman, J., Chen, J., Clune, J., and Stanley, K. ES is more than just a traditional finite-difference approximator. *arXiv preprint arXiv:1712.06568*, 2017.

Levine, S. Motor skill learning with local trajectory methods. *PhD thesis, Stanford University*, 2014.

Levine, S. Reinforcement learning and control as probabilistic inference: Tutorial and review. *arXiv preprint arXiv:1805.00909*, 2018.

Levy, A., Platt, R., and Saenko, K. Hierarchical actor-critic. *arXiv preprint arXiv:1712.00948*, 2017.

Li, Y. and Mandt, S. Disentangled sequential autoencoder. *International Conference on Machine Learning (ICML)*, 35, 2018.

Lillicrap, T., Hunt, J., Pritzel, A., Heess, N., Erez, T., Tassa, Y., Silver, D., and Wierstra, D. Continuous control with deep reinforcement learning. *arXiv preprint arXiv:1509.02971*, 2015.

Manfredi, V. and Mahadevan, S. Hierarchical reinforcement learning using graphical models. *ICML workshop on Rich Representations for Reinforcement Learning*, 2005.

Moody, J. and White, D. Structural cohesion and embeddedness: A hierarchical concept of social groups. *American Sociological Review*, pp. 103–127, 2003.

Nachum, O., Norouzi, M., Xu, K., and Schuurmans, D. Bridging the gap between value and policy based reinforcement learning. *Advances in neural information processing systems (NIPS)*, 2017.

Neumann, G. Variational inference for policy search in changing situations. *International Conference on Machine Learning (ICML)*, 2011.

Ostrovski, G., Bellemare, M., van den Oord, A., and Munos, A. Count-based exploration with neural density models. *International Conference on Machine Learning (ICML)*, 2018.

Parr, R. and Russell, S. Reinforcement learning with hierarchies of machines. *Advances in neural information processing systems (NIPS)*, pp. 1043–1049, 1998.

Parsons, T. An analytical approach to the theory of social stratification. *The American Journal of Sociology*, 45: 841–862, 1940.

Peters, J., Mulling, K., and Altun, Y. Relative entropy policy search. *AAAI Conference on Artificial Intelligence*, pp. 1607–1612, 2010.

Rawlik, K., Toussaint, M., and S. Vijayakumar, title = On stochastic optimal control and reinforcement learning by approximate inference, j. . R. v. . . p. . . y. . .

Rezende, D. and Mohamed, S. Variational inference with normalizing flows. *International Conference on Machine Learning (ICML)*, 32:1530–1538, 2015.

Rezende, D., Mohamed, S., and Wierstra, D. Stochastic backpropagation and approximate inference in deep generative models. *International Conference on Machine Learning (ICML)*, 31, 2014.

Rothfuss, J., Lee, D., Clavera, I., Asfour, T., and Abbeel, P. ProMP: Proximal meta-policy search. *International Conference on Learning Representations (ICLR)*, 2019.

Schmidhuber, J. On learning to think: Algorithmic information theory for novel combinations of reinforcement learning controllers and recurrent neural world models. *arXiv preprint arXiv:1511.09249*, 2015.

Schulman, J., Abbeel, P., and Chen, X. Equivalence between policy gradients and soft Q-learning. *arXiv preprint arXiv:1704.06440*, 2017a.

Schulman, J., Wolski, F., Dhariwal, P., Radford, A., and Klimov, O. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017b.

Strehl, A., Li, L., and Littman, M. Incremental model-based learners with formal learning-time guarantees. *Uncertainty in Aritifical Intelligence (UAI)*, 2006.

Strehl, A., Li, L., and Littman, M. Reinforcement learning in finite MDPs: PAC analysis. *Journal of Machine Learning Research (JMLR)*, pp. 2413–2444, 2009.

Stuhlmuller, A., Taylor, J., and Goodman, N. Learning stochastic inverses. *Advances in neural information processing systems (NIPS)*, 27:3048–3056, 2013.

Sutton, R., Precup, D., and Singh, S. Between MDPs and semi-MDPs: A framework for temporal abstraction in reinforcement learning. *Artificial intelligence*, 112:181–211, 1999.

Tang, H., Houthooft, R., Foote, D., Stooke, A., Chen, X., Duan, Y., Schulman, J., Turck, F. D., and Abbeel, P. #Exploration: A study of count-based exploration for deep reinforcement learning. *Advances in neural information processing systems (NIPS)*, 2017.

Tang, Y. and Agrawal, S. Implicit Policy for Reinforcement Learning. *arXiv preprint arXiv:1806.06798*, 2018.

Todorov, E. Linearly-solvable Markov decision problems. *Advances in neural information processing systems (NIPS)*, pp. 1369–1376, 2007.

Toussaint, M. Robot trajectory optimization using approximate inference. *International Conference on Machine Learning (ICML)*, 2009.

van den Berg, R., Hasenclever, L., Tomczak, J., and Welling, M. Sylvester normalizing flows for variational inference. *Uncertainty in Aritifical Intelligence (UAI)*, 2018.

van der Pol, E. and Oliehoek, F. Coordinated Deep Reinforcement Learners for Traffic Light Control. *NIPS 2016 workshop on Learning, Inference and Control of Multi-agent Systems*, 2016.

Wainwright, M. and Jordan, M. Graphical models, exponential families, and variational inference. *Foundations and trends in Machine Learning*, 1:1–305, 2008.

Watkins, C. Learning from delayed rewards. *PhD thesis, Cambridge University*, 1989.

Xu, K., Ratner, E., Dragan, A., Levine, S., and Finn, C. Learning a prior over intent via meta-inverse reinforcement learning. *arXiv preprint arXiv:1805.12573*, 2018.

Ziebart, B., Maas, A., Bagnell, J., and Dey, A. Maximum entropy inverse reinforcement learning. *AAAI Conference on Artificial Intelligence*, 2008.