VARIATIONAL INTRINSIC CONTROL REVISITED

Anonymous authors
Paper under double-blind review

ABSTRACT

In this paper, we revisit variational intrinsic control (VIC), an unsupervised reinforcement learning method for finding the largest set of intrinsic options available to an agent. In the original work by Gregor et al. (2016), two VIC algorithms were proposed: one that represents the options explicitly, and the other that does it implicitly. We show that the intrinsic reward used in the latter is subject to bias in stochastic environments, causing convergence to suboptimal solutions. To correct this behavior, we propose two methods respectively based on the transitional probability model and Gaussian Mixture Model. We substantiate our claims through rigorous mathematical derivations and experimental analyses.

1 INTRODUCTION

Variational intrinsic control (VIC) proposed by Gregor et al. (2016) is an unsupervised reinforcement learning algorithm that aims to discover as many *intrinsic* options as possible, i.e., the policies with a termination condition that meaningfully affect the world. The main idea of VIC is to maximize the mutual information between the set of options and final states, called *empowerment*. The maximum empowerment is desirable because it maximizes the information about the final states the agent can achieve with the available options. These options are independent of the extrinsic reward of the environment, so they can be considered as the agent's universal knowledge about the environment.

The concept of empowerment has been introduced in (Klyubin et al., 2005; Salge et al., 2014) along with the methods for measuring it (Arimoto, 1972; Blahut, 1972). They defined the option as a sequence of a fixed number of actions. Yeung (2008) proposed to maximize the empowerment using the Blahut & Arimoto (BA) algorithm, but its complexity increase exponentially with the length of sequence, rendering it impractical for high dimensional and long-horizon options. Mohamed & Rezende (2015) adopted techniques from deep learning and variational inference (Barber & Agakov, 2003) and successfully applied empowerment maximization for high dimensional and long-horizon control. However, this method maximizes the empowerment over *open-loop* options, meaning that the sequence of action is chosen in advance and conducted regardless of the (potentially stochastic) environment dynamics. This can severely degrade the empowerment (Gregor et al., 2016).

To overcome this limitation, Gregor et al. (2016) proposed to use the *closed-loop* options where the agent considers transited states while it is generating a sequence of actions. This type of option differs from those in Klyubin et al. (2005), Salge et al. (2014) and Mohamed & Rezende (2015) in that they have a termination condition, instead of a fixed number of actions. They presented two algorithms: VIC with explicit and implicit options (we will call them *explicit* and *implicit VIC* from here on). The explicit VIC defines a fixed number of options. An option is sampled at the beginning of the trajectory, conditioning the policies of an agent until the termination. In other words, both the state and the sampled option are the input to the policy function of the agent. One clear limitation of the explicit VIC is that it requires the preset number of options. This does not only apply to the explicit VIC, but also to some recent unsupervised learning algorithms that adopt a discrete option or skill with a predefined set (Machado et al., 2017; Eysenbach et al., 2018) which hinders the maximal level of learning. Choosing a proper number of options is not straightforward, since the maximum of the objective for a given number of options depends on several unknown environmental factors such as the cardinality of the state space and the transitional model. To overcome this issue, Gregor et al. (2016) proposed the implicit VIC which defines the option as the trajectory until the termination (Gregor et al., 2016). There exist multiple trajectories that lead to the same final state and the implicit VIC learns to maximize the information of the final states by controlling these

options (trajectories). As a result, the number of options is no longer limited by the preset number, and it becomes possible to learn the maximum number of options for the given environment.

In this work, we analyze and overcome the weakness of the implicit VIC under stochastic environment dynamics. Our contributions are summarized as follows:

1. We show that the intrinsic reward in the implicit VIC suffers from the variational bias in stochastic environments, causing convergence to suboptimal solutions (Section 2).
2. To compensate this bias and achieve the maximal empowerment, we suggest two modifications of the implicit VIC: the environment dynamics modeling incorporating the transitional probability (Section 3) and Gaussian Mixture Model (Section 4).

2 VARIATIONAL BIAS OF IMPLICIT VIC IN STOCHASTIC ENVIRONMENTS

In this section, we derive the variational bias of intrinsic reward under stochastic environment dynamics. First, we adopt the definition of termination action and final state from Gregor et al. (2016) for a clear explanation. The termination action terminates the option and yields the final state $s_f = s_t$ independently of the environmental action space. The objective term in VIC, i.e., the mutual information between option Ω and final state s_f, can be written as follows:

$$I(\Omega, s_f|s_0) = -\sum_{\Omega} p(\Omega|s_0) \log p(\Omega|s_0) + \sum_{\Omega, s_f} p(s_f|s_0, \Omega) p(\Omega|s_0) \log p(\Omega|s_0, s_f). \tag{1}$$

Since $p(\Omega|s_0, s_f)$ is intractable, VIC Gregor et al. (2016) derives the variational bound $I^{VB} \leq I$ and optimizes it instead:

$$I^{VB}(\Omega, s_f|s_0) = -\sum_{\Omega} p(\Omega|s_0) \log p(\Omega|s_0) + \sum_{\Omega, s_f} p(s_f|s_0, \Omega) p(\Omega|s_0) \log q(\Omega|s_0, s_f), \tag{2}$$

where $q(\Omega|s_0, s_f)$ is the inference to be trained. When I^{VB} is maximized, we have $p(\Omega|s_0, s_f) = q(\Omega|s_0, s_f)$ and achieve the maximum I.

The implicit VIC defines the option Ω as the trajectory of an agent, i.e., the sequence of states and actions: $\Omega = (s_0, a_0, s_1, a_1, ..., s_{T-1}, a_{T-1} = a_f, s_T = s_f = s_{T-1})$. Hence, using Bayes rule, the probability of an option can be decomposed as

$$p(\Omega|s_0) = \prod_{(\tau_t, a_t, s_{t+1}) \in \Omega} p(a_t|\tau_t) p(s_{t+1}|\tau_t, a_t) \quad \text{with} \quad \tau_t = (s_0, a_0, ..., s_t). \tag{3}$$

Similarly, $p(\Omega|s_0, s_f)$ can be expressed as

$$p(\Omega|s_0, s_{f|\Omega}) = \prod_{(\tau_t, a_t, s_{t+1}) \in \Omega} p(a_t|\tau_t, s_{f|\Omega}) p(s_{t+1}|\tau_t, a_t, s_{f|\Omega}). \tag{4}$$

Note that s_f is replaced by $s_{f|\Omega}$ since it is determined by the given Ω in the following way:

$$p(s_f|\Omega, s_0) = \begin{cases} 1, & \text{if } s_f = s_{f|\Omega} \\ 0, & \text{otherwise} \end{cases}. \tag{5}$$

Using (3), (4) and (5), we can rewrite the mutual information (1) as

$$I(\Omega, s_f|s_0) = \sum_{\Omega} p(\Omega|s_0) \sum_{(\tau_t, a_t, s_{t+1}) \in \Omega} \Big[\log \frac{p(a_t|\tau_t, s_{f|\Omega})}{p(a_t|\tau_t)} + \log \frac{p(s_{t+1}|\tau_t, a_t, s_{f|\Omega})}{p(s_{t+1}|\tau_t, a_t)} \Big]. \tag{6}$$

The intrinsic reward of the implicit VIC (Gregor et al., 2016) is given by

$$r_{\Omega}^{I^{VB}} = \sum_{(\tau_t, a_t, s_{t+1}) \in \Omega} \log \frac{q(a_t|\tau_t, s_{f|\Omega})}{p(a_t|\tau_t)}, \tag{7}$$

where $q(a_t|\tau_t, s_f)$ is inference and $p(a_t|\tau_t)$ is policy of an agent. It can be shown that r_Ω^{IVB} comes from the first part of (6) (see Appendix A for details). Under deterministic environment dynamics, $\log p(s_{t+1}|\tau_t, a_t, s_{f|\Omega})/p(s_{t+1}|\tau_t, a_t)$ is canceled out since both nominator and denominator are always 1, turning (6) into the expectation of (7) with variational inference. However, this is not possible under the stochastic environment dynamics and it yields the variational bias b_Ω^{VIC} in the intrinsic reward:

$$b_\Omega^{VIC} = \sum_{(\tau_t, a_t, s_{t+1})\in\Omega} \log \frac{p(s_{t+1}|\tau_t, a_t, s_{f|\Omega})}{p(s_{t+1}|\tau_t, a_t)}. \tag{8}$$

In Section 5, we provide the experimental evidence that this variational bias leads to a suboptimal training. Even though the original VIC (Gregor et al., 2016) subtracts $b(s_0)$ from r_Ω^{IVB} to reduce to reduce the variance of learning, it cannot compensate this bias since it also depends on Ω. In the next section, we analyze the mutual information (1) in more detail under stochastic environment dynamics and define the variational estimate of (1), I^{VE}, for training.

3 Implicit VIC with transitional probability model

In this section, we analyze $I(\Omega, s_f|s_0)$ under stochastic environment dynamics and propose to explicitly model transitional probabilities. First, for a given option and final state, define $p_\pi(\Omega|s_f, s_0)$, $p_\rho(\Omega|s_f, s_0)$, $p_\pi(\Omega|s_0)$ and $p_\pi(\Omega|s_0)$ as follows:

$$\begin{aligned} p_\pi(\Omega|s_f, s_0) &= \prod_{(\tau_t, a_t, s_{t+1})\in\Omega} p_\pi(a_t|\tau_t, s_f), \quad & p_\rho(\Omega|s_f, s_0) &= \prod_{(\tau_t, a_t, s_{t+1})\in\Omega} p_\rho(s_{t+1}|\tau_t, a_t, s_f), \\ p_\pi(\Omega|s_0) &= \prod_{(\tau_t, a_t, s_{t+1})\in\Omega} p_\pi(a_t|\tau_t), & p_\rho(\Omega|s_0) &= \prod_{(\tau_t, a_t, s_{t+1})\in\Omega} p_\rho(s_{t+1}|\tau_t, a_t). \end{aligned} \tag{9}$$

Notice that $p(\Omega|s_f, s_0) = p_\pi(\Omega|s_f, s_0)p_\rho(\Omega|s_f, s_0)$ where $p_\pi(\Omega|s_f, s_0)$ is policy related part and $p_\rho(\Omega|s_f, s_0)$ is transitional part of $p(\Omega|s_f, s_0)$ and so do $p_\pi(\Omega|s_0)$ and $p_\rho(\Omega|s_0)$. Next, we define $p_\pi^q(\Omega|s_f, s_0)$, $p_\rho^q(\Omega|s_f, s_0)$, $p_\pi^p(\Omega|s_0)$ and $p_\pi^p(\Omega|s_0)$ as follows:

$$\begin{aligned} p_\pi^q(\Omega|s_f, s_0) &= \prod_{(\tau_t, a_t, s_{t+1})\in\Omega} \pi_{\theta_\pi^q}^q(a_t|\tau_t, s_f), \quad & p_\rho^q(\Omega|s_f, s_0) &= \prod_{(\tau_t, a_t, s_{t+1})\in\Omega} \rho_{\theta_\rho^q}^q(s_{t+1}|\tau_t, a_t, s_f), \\ p_\pi^p(\Omega|s_0) &= \prod_{(\tau_t, a_t, s_{t+1})\in\Omega} \pi_{\theta_\pi^p}^p(a_t|\tau_t), & p_\rho^p(\Omega|s_0) &= \prod_{(\tau_t, a_t, s_{t+1})\in\Omega} \rho_{\theta_\rho^p}^p(s_{t+1}|\tau_t, a_t), \end{aligned} \tag{10}$$

where $\pi_{\theta_\pi^q}^q(a_t|\tau_t, s_f)$, $\rho_{\theta_\rho^q}^q(s_{t+1}|\tau_t, a_t, s_f)$, $\pi_{\theta_\pi^p}^p(a_t|\tau_t)$ and $\rho_{\theta_\rho^p}^p(s_{t+1}|\tau_t, a_t)$ are our estimates of $p_\pi(a_t|\tau_t, s_f)$, $p_\rho(s_{t+1}|\tau_t, a_t, s_f)$, $p_\pi(a_t|\tau_t)$ and $p_\rho(s_{t+1}|\tau_t, a_t)$. Since we know the policy of an agent, we have $p_\pi(a_t|\tau_t) = \pi_{\theta_\pi^p}^p(a_t|\tau_t)$. For the other probabilities, they are trained to fit true distributions. Using (9), we can rewrite $I(\Omega, s_f|s_0)$ as:

$$I(\Omega, s_f|s_0) = \sum_{\Omega, s_f} p(\Omega, s_f|s_0)\Big[\log p_\rho(\Omega|s_f, s_0)p_\pi(\Omega|s_f, s_0) - \log p_\rho(\Omega|s_0)p_\pi(\Omega|s_0)\Big]. \tag{11}$$

Now, using (10), define I^{VE} as follows:

$$I^{VE}(\Omega, s_f|s_0) = \sum_{\Omega, s_f} p(\Omega, s_f|s_0)\Big[\log p_\rho^q(\Omega|s_f, s_0)p_\pi^q(\Omega|s_f, s_0) - \log p_\rho^p(\Omega|s_0)p_\pi^p(\Omega|s_0)\Big]. \tag{12}$$

This is our estimate of the mutual information between Ω and s_f. The absolute difference between (11) and (12) is given by

$$\Big|I - I^{VE}\Big| = \left|\sum_{\Omega, s_f} p(\Omega, s_f|s_0)\Big[\log \frac{p_\pi(\Omega|s_f, s_0)p_\rho(\Omega|s_f, s_0)}{p_\pi^q(\Omega|s_f, s_0)p_\rho^q(\Omega|s_f, s_0)} - \log \frac{p_\pi(\Omega|s_0)p_\rho(\Omega|s_0)}{p_\pi^p(\Omega|s_0)p_\rho^p(\Omega|s_0)}\Big]\right|, \tag{13}$$

where I denotes $I(\Omega, s_f|s_0)$ and I^{VE} denotes $I^{VE}(\Omega, s_f|s_0)$. The upper bound on (13), U^{VE} is

$$\begin{aligned} \Big|I - I^{VE}\Big| \le U^{VE} = &\sum_{s_f} p(s_f|s_0) D_{\mathrm{KL}}\big[p_\pi(\cdot|s_f, s_0)p_\rho(\cdot|s_f, s_0)\|p_\pi^q(\cdot|s_f, s_0)p_\rho^q(\cdot|s_f, s_0)\big] \\ &+ D_{\mathrm{KL}}\big[p_\pi^p(\cdot|s_0)p_\rho(\cdot|s_0)\|p_\pi^p(\cdot|s_0)p_\rho^p(\cdot|s_0)\big]. \end{aligned} \tag{14}$$

Algorithm 1 Implicit VIC with transitional probability model

Initialize s_0, η, T_{train}, θ^q_π, θ^q_ρ, θ^p_π and θ^p_ρ.
for i_{train} : 1 to T_{train} **do**
  Follow $\pi^p_{\theta^p_\pi}(a_t|\tau_t)$ result in $\Omega = (s_0, a_0, ..., s_f)$.
  $r^{I^{VE}}_\Omega \leftarrow \sum_t [\log \pi^q_{\theta^q_\pi}(a_t|\tau_t, s_f) - \log \pi^p_{\theta^p_\pi}(a_t|\tau_t)]$ ▷ from (16)
  $r^{I^{VE}}_\Omega \leftarrow r^{I^{VE}}_\Omega + \sum_t [\log \rho^q_{\theta^q_\rho}(s_{t+1}|\tau_t, a_t, s_f) - \log \rho^p_{\theta^p_\rho}(s_{t+1}|\tau_t, a_t)]$ ▷ from (16), (*)
  Update each parameter:
    $\theta^p_\pi \leftarrow \theta^p_\pi + \eta r^{I^{VE}}_\Omega \nabla_{\theta^p_\pi} \sum_t \log \pi^p_{\theta^p_\pi}(a_t|\tau_t)$ ▷ from (16)
    $\theta^q_\pi \leftarrow \theta^q_\pi + \eta \nabla_{\theta^q_\pi} \sum_t \log \pi^q_{\theta^q_\pi}(a_t|\tau_t, s_f)$ ▷ from (15)
    $\theta^p_\rho \leftarrow \theta^p_\rho + \eta \nabla_{\theta^p_\rho} \sum_t \log \rho^p_{\theta^p_\rho}(s_{t+1}|\tau_t, a_t)$ ▷ from (15), (*)
    $\theta^q_\rho \leftarrow \theta^q_\rho + \eta \nabla_{\theta^q_\rho} \sum_t \log \rho^q_{\theta^q_\rho}(s_{t+1}|\tau_t, a_t, s_f)$ ▷ from (15), (*)
  end
end for

See Appendix B for the derivation. This upper bound implies that $I^{VE} \rightarrow I$ as ($p^q_\pi(\Omega|s_f, s_0)$, $p^q_\rho(\Omega|s_f, s_0)$, $p^p_\rho(\Omega|s_0)$) $\rightarrow$ ($p_\pi(\Omega|s_f, s_0)$, $p_\rho(\Omega|s_f, s_0)$, $p_\rho(\Omega|s_0)$) for all Ω. (Note that $p_\pi(\Omega|s_0)$ is omitted since we know the true value of it.) In other words, our estimate of the mutual information converges to the true value as our estimates (10) converge to the true distribution (9). It makes sense that we can estimate the true value of the mutual information if we know the true distribution. Furthermore, $U^{VE} = 0$ if and only if $(p^q_\pi(\Omega|s_f, s_0), p^q_\rho(\Omega|s_f, s_0), p^p_\rho(\Omega|s_0)) = p_\pi(\Omega|s_f, s_0), p_\rho(\Omega|s_f, s_0), p_\rho(\Omega|s_0))$ for all Ω since U^{VE} is sum of positively weighted KL divergences. This means that minimizing U^{VE} will make (10) converge to (9). Now, we can obtain $\nabla_{\theta^q_\pi} U^{VE}$, $\nabla_{\theta^q_\rho} U^{VE}$ and $\nabla_{\theta^p_\rho} U^{VE}$ using (5):

$$\begin{aligned}
\nabla_{\theta^q_\pi} U^{VE} &= -\sum_\Omega p(\Omega|s_0) \nabla_{\theta^q_\pi} \log p^q_\pi(\Omega|s_{f|\Omega}, s_0), \\
\nabla_{\theta^q_\rho} U^{VE} &= -\sum_\Omega p(\Omega|s_0) \nabla_{\theta^q_\rho} \log p^q_\rho(\Omega|s_{f|\Omega}, s_0), \\
\nabla_{\theta^p_\rho} U^{VE} &= -\sum_\Omega p(\Omega|s_0) \nabla_{\theta^p_\rho} \log p^p_\rho(\Omega|s_0),
\end{aligned} \tag{15}$$

which can be estimated from sample mean. Once we have ($p^q_\pi(\Omega|s_f, s_0)$, $p^q_\rho(\Omega|s_f, s_0)$, $p^p_\rho(\Omega|s_0)$) $\approx$ ($p_\pi(\Omega|s_f, s_0)$, $p_\rho(\Omega|s_f, s_0)$, $p_\rho(\Omega|s_0)$) for all Ω, we can update the policy to maximize I. The gradients, $\nabla_{\theta^p_\pi} I$ and $\nabla_{\theta^p_\pi} I^{VE}$ can be obtained in the following form using (5):

$$\begin{aligned}
\nabla_{\theta^p_\pi} I &= \sum_\Omega p(\Omega|s_0) \underbrace{\left[\log p_\pi(\Omega|s_{f|\Omega}, s_0) p_\rho(\Omega|s_{f|\Omega}, s_0) - \log p^p_\pi(\Omega|s_0) p_\rho(\Omega|s_0)\right]}_{r^I_\Omega} \nabla_{\theta^p_\pi} \log p^p_\pi(\Omega|s_0), \\
\nabla_{\theta^p_\pi} I^{VE} &= \sum_\Omega p(\Omega|s_0) \underbrace{\left[\log p^q_\pi(\Omega|s_{f|\Omega}, s_0) p^q_\rho(\Omega|s_{f|\Omega}, s_0) - \log p^p_\pi(\Omega|s_0) p^p_\rho(\Omega|s_0)\right]}_{r^{I^{VE}}_\Omega} \nabla_{\theta^p_\pi} \log p^p_\pi(\Omega|s_0),
\end{aligned} \tag{16}$$

where θ^p_π is parameter of $p^p_\pi(\Omega|s_0)$ and $p_\pi(\Omega|s_0)$ is replaced by $p^p_\pi(\Omega|s_0)$ since we know the true value of policy (see Appendix A for details). From (16), we can see that $\nabla_{\theta^p_\pi} I^{VE} \rightarrow \nabla_{\theta^p_\pi} I$ as ($p^q_\pi(\Omega|s_f, s_0)$, $p^q_\rho(\Omega|s_f, s_0)$, $p^p_\rho(\Omega|s_0)$) $\rightarrow$ ($p_\pi(\Omega|s_f, s_0)$, $p_\rho(\Omega|s_f, s_0)$, $p_\rho(\Omega|s_0)$) for all Ω. Note that for the deterministic environment, we can omit $p^q_\rho(\Omega|s_f, s_0)$ and $p^p_\rho(\Omega|s_0)$ since they are always 1 and it satisfies $I^{VE} = I^{VB}$ of (2) and $\nabla_{\theta^q_\pi} U^{VE} = -\nabla_{\theta^q_\pi} I^{VB}$, which means that maximizing I^{VB} is equivalent to minimizing U^{VE} for θ^q_π (i.e., it is equivalent to the original implicit VIC). Algorithm 1 summarizes the modified implicit VIC with the transitional probability model. The additional steps added to the original implicit VIC are marked with $(*)$ in Algorithm 1. Note that Algorithm 1 is not always practically applicable since it is hard to model $p(s_{t+1}|\tau_t, a_t)$ and $p(s_{t+1}|\tau_t, a_t, s_f)$ when the cardinality of the state space is unknown. In our experiment, we will assume that we know the

cardinality of the state space. This allows us to model $p_\rho(s_{t+1}|\tau_t, a_t)$ and $p_\rho(s_{t+1}|\tau_t, a_t, s_f)$ using $softmax$ and show the convergence of Algorithm 1. In the next section, we propose a practically applicable method that avoids this intractability of the cardinality.

4 Implicit VIC with Gaussian Mixture Model

The alternative method we propose to correct the bias of the implicit VIC is to model the smoothed transitional distributions for estimating the mutual information. First, we smooth $p(s_{t+1}|\tau_t, a_t, s_f)$ and $p(s_{t+1}|\tau_t, a_t)$ into $f_\sigma(x_{t+1}|\tau_t, a_t, s_f)$ and $f_\sigma(x_{t+1}|\tau_t, a_t)$ where $x_{t+1} = s_{t+1} + z_{t+1}$ and $z_{t+1} \sim \mathcal{N}(\mathbf{0}, \sigma^2 \mathrm{I}_n)$:

$$\begin{aligned} f_\sigma(x_{t+1}|\tau_t, a_t, s_f) &= \sum_{s' \in S(\tau_t, a_t, s_f)} p(s'|\tau_t, a_t, s_f) f_\sigma(x_{t+1} - s'; \mathbf{0}, \sigma^2 \mathrm{I}_n), \\ f_\sigma(x_{t+1}|\tau_t, a_t) &= \sum_{s' \in S(\tau_t, a_t)} p(s'|\tau_t, a_t) f_\sigma(x_{t+1} - s'; \mathbf{0}, \sigma^2 \mathrm{I}_n), \end{aligned} \tag{17}$$

where $S(\tau_t, a_t, s_f) = \{s' | p(s'|\tau_t, a_t, s_f) > 0\}$, $S(\tau_t, a_t) = \{s' | p(s'|\tau_t, a_t) > 0\}$ and n is the dimension of the state. Then, using Gaussian Mixture Model (GMM) (Reynolds & Rose, 1995), we model them as $f^q_\sigma(x_{t+1}|\tau_t, a_t, s_f)$ and $f^p_\sigma(x_{t+1}|\tau_t, a_t)$:

$$\begin{aligned} f^q_\sigma(x_{t+1}|\tau_t, a_t, s_f) &= \sum_{i=1}^{n_{gmm}} w_i(\tau_t, a_t, s_f) f_\sigma(x_{t+1}; \boldsymbol{\mu}_i(\tau_t, a_t, s_f), \sigma^2 \mathrm{I}_n), \\ f^p_\sigma(x_{t+1}|\tau_t, a_t) &= \sum_{i=1}^{n_{gmm}} w_i(\tau_t, a_t) f_\sigma(x_{t+1}; \boldsymbol{\mu}_i(\tau_t, a_t), \sigma^2 \mathrm{I}_n). \end{aligned} \tag{18}$$

Note that if we set $n_{gmm} > \max\left(\max_{\tau_t, a_t, s_f} |S(\tau_t, a_t, s_f)|, \max_{\tau_t, a_t} |S(\tau_t, a_t)|\right)|$, (18) can perfectly fit (17). Under this smoothing, we have

$$L_{\sigma|\Omega} \leq \log \frac{p_\rho(\Omega|s_f, s_0)}{p_\rho(\Omega|s_0)} - \log \frac{f_\sigma(\Omega|s_f, s_0)}{f_\sigma(\Omega|s_0)} \leq U_{\sigma|\Omega}, \tag{19}$$

where

$$\begin{aligned} f_\sigma(\Omega|s_f, s_0) &= \prod_{(\tau_t, a_t, s_{t+1}) \in \Omega} f_\sigma(s_{t+1}|\tau_t, a_t, s_{f|\Omega}), \quad f_\sigma(\Omega|s_0) = \prod_{(\tau_t, a_t, s_{t+1}) \in \Omega} f_\sigma(s_{t+1}|\tau_t, a_t), \\ U_{\sigma|\Omega} &= \frac{T_\Omega}{p_{min}} \exp\big(-\frac{d^2_{min}}{2\sigma^{2n}}\big), \quad L_{\sigma|\Omega} = -\frac{T_\Omega}{p_{min,f}} \exp\big(-\frac{d^2_{min}}{2\sigma^{2n}}\big), \end{aligned} \tag{20}$$

with $d_{min} = \min_{s_i \neq s_j} \|s_i - s_j\|_2$, $p_{min} = \min_{s', \tau, a} p(s'|\tau, a) > 0$, $p_{min,f} = \min_{s', \tau, a, s_f} p(s'|\tau, a, s_f) > 0$ and T_Ω is number of transitions in Ω (see Appendix C for the derivation). From (20), we notice that $L_{\sigma|\Omega} \to 0$ and $U_{\sigma|\Omega} \to 0$ as $\sigma \to 0$ for finite T_Ω. By taking the expectation of each side of (19) with respect to $p(\Omega, s_f|s_0)$, we obtain

$$\check{L}_\sigma \leq \overline{L}_\sigma \leq D_\sigma \leq \overline{U}_\sigma \leq \hat{U}_\sigma, \tag{21}$$

where

$$\begin{aligned} D_\sigma &= \sum_{\Omega, s_f} p(\Omega, s_f|s_0) \log\Big[\frac{p_\rho(\Omega|s_f, s_0)}{p_\rho(\Omega|s_0)} \frac{f_\sigma(\Omega|s_0)}{f_\sigma(\Omega|s_f, s_0)}\Big], \\ \check{L}_\sigma &= -\frac{T_{max}}{p_{min,f}} \exp\big(-\frac{d^2_{min}}{2\sigma^2}\big), \quad \overline{L}_\sigma = -\frac{\overline{T}}{p_{min,f}} \exp\big(-\frac{d^2_{min}}{2\sigma^2}\big), \\ \hat{U}_\sigma &= \frac{T_{max}}{p_{min}} \exp\big(-\frac{d^2_{min}}{2\sigma^2}\big), \quad \overline{U}_\sigma = \frac{\overline{T}}{p_{min}} \exp\big(-\frac{d^2_{min}}{2\sigma^2}\big), \end{aligned} \tag{22}$$

with $T_{max} = \max_\Omega T_\Omega$ and $\overline{T} = \sum_{\Omega, s_f} p(\Omega, s_f|s_0) T_\Omega$. Note that $\hat{U}_\sigma$ and $\check{L}_\sigma$ are the worst-case bounds and that $\hat{U}_\sigma \to 0$, $\overline{U}_\sigma \to 0$, $\check{L}_\sigma \to 0$, $\overline{L}_\sigma \to 0$ as $\sigma \to 0$ for finite T_{max} and $\overline{T}$. Next, we define I^{VE}_σ, the variational estimate of the mutual information with smoothing:

$$I^{VE}_\sigma = \sum_{\Omega, s_f} p(\Omega, s_f|s_0) \Big[\log p_\pi(\Omega|s_f, s_0) f^q_\sigma(\Omega|s_f, s_0) - \log p_\pi(\Omega|s_0) f^p_\sigma(\Omega|s_0)\Big]. \tag{23}$$

It can be seen that this is equivalent to (12) if we substitute $p^q_\rho(\Omega|s_f, s_0)$ and $p^p_\rho(\Omega|s_0)$ with $f^q_\sigma(\Omega|s_f, s_0)$ and $f^p_\sigma(\Omega|s_0)$. The upper bound on $|I - I^{VE}_\sigma|$ can be obtained as follows:

$$\begin{aligned}
&\left|I - I^{VE}_\sigma\right| \leq U^{VE}_{\sigma,1} + U^{VE}_{\sigma,2} \quad \text{with} \\
&U^{VE}_{\sigma,1} = \left|\sum_{\Omega, s_f} p(\Omega, s_f|s_0) \log\left[\frac{p_\rho(\Omega|s_f, s_0)}{p_\rho(\Omega|s_0)} \frac{f^p_\sigma(\Omega|s_0)}{f^q_\sigma(\Omega|s_f, s_0)}\right]\right|, \\
&U^{VE}_{\sigma,2} = \sum_{s_f} p(s_f|s_0) KL_\Omega\big[p_\pi(\cdot|s_f, s_0)p_\rho(\cdot|s_f, s_0)||p^q_\pi(\cdot|s_f, s_0)p_\rho(\cdot|s_f, s_0)\big].
\end{aligned} \tag{24}$$

See Appendix D for the derivation. This upper bound implies that $U^{VE}_{\sigma,1} \to 0$ as $f^q_\sigma(\Omega|s_f, s_0)/f^p_\sigma(\Omega|s_0) \to p_\rho(\Omega|s_f, s_0)/p_\rho(\Omega|s_0)$ and $U^{VE}_{\sigma,2} \to 0$ as $p^q_\pi(\Omega|s_f, s_0) \to p_\pi(\Omega|s_f, s_0)$ for all Ω. We can directly minimize $U^{VE}_{\sigma,2}$ using gradient descent as in (15) using (5):

$$\nabla_{\theta^q_\pi} U^{VE}_{\sigma,2} = -\sum_\Omega p(\Omega|s_0) \nabla_{\theta^q_\pi} \log p^q_\pi(\Omega|s_{f|\Omega}, s_0). \tag{25}$$

This can be estimated from the sample mean. Since $U^{VE}_{\sigma,2} = 0$ if and only if $p^q_\pi(\Omega|s_f, s_0) = p_\pi(\Omega|s_f, s_0)$ for all Ω, this update will make $p^q_\pi(\Omega|s_f, s_0)$ converge to $p_\pi(\Omega|s_f, s_0)$.

Unlike $U^{VE}_{\sigma,2}$, it is difficult to directly minimize $U^{VE}_{\sigma,1}$ due to the absolute value function. Therefore, we instead minimize $D_{\mathrm{KL}}[f_\sigma(\cdot|\tau_t, a_t, s_f)\|f^q_\sigma(\cdot|\tau_t, a_t, s_f)]$ and $D_{\mathrm{KL}}[f_\sigma(\cdot|\tau_t, a_t)\|f^p_\sigma(\cdot|\tau_t, a_t)]$ to fit $f^q_\sigma(x_{t+1}|\tau_t, a_t, s_f)$ and $f^p_\sigma(x_{t+1}|\tau_t, a_t)$ to $f_\sigma(x_{t+1}|\tau_t, a_t, s_f)$ and $f_\sigma(x_{t+1}|\tau_t, a_t)$. If $f^q_\sigma(x_{t+1}|\tau_t, a_t, s_f)$ and $f^p_\sigma(x_{t+1}|\tau_t, a_t)$ are parameterized by θ^q_ρ and θ^p_ρ, the gradients of KL divergences can be obtained as

$$\begin{aligned}
\nabla_{\theta^q_\rho} D_{\mathrm{KL}}[f_\sigma(\cdot|\tau_t, a_t, s_f)\|f^q_\sigma(\cdot|\tau_t, a_t, s_f)] &= -\int_{x_{t+1}} f_\sigma(x_{t+1}|\tau_t, a_t, s_f) \nabla_{\theta^q_\rho} \log f^q_\sigma(x_{t+1}|\tau_t, a_t, s_f), \\
\nabla_{\theta^p_\rho} D_{\mathrm{KL}}[f_\sigma(\cdot|\tau_t, a_t)\|f^q_\sigma(\cdot|\tau_t, a_t)] &= -\int_{x_{t+1}} f_\sigma(x_{t+1}|\tau_t, a_t) \nabla_{\theta^p_\rho} \log f^p_\sigma(x_{t+1}|\tau_t, a_t),
\end{aligned} \tag{26}$$

which can be estimated from the sample mean. These updates will make $(f^q_\sigma(\Omega|s_f, s_0), f^p_\sigma(\Omega|s_0))$ converge to $(f_\sigma(\Omega|s_f, s_0), f_\sigma(\Omega|s_0))$. It can be seen from (22) that $U^{VE}_{\sigma,1} \to |D_\sigma|$ as $(f^q_\sigma(\Omega|s_f, s_0), f^p_\sigma(\Omega|s_0)) \to (f_\sigma(\Omega|s_f, s_0), f_\sigma(\Omega|s_0))$. Since $|D_\sigma| \approx 0$ for finite $\overline{T}$ and $\sigma \ll d_{min}$ from (21), we can minimize $U^{VE}_{\sigma,1}$ to nearly zero in this case. Once we have $(p^q_\pi(\Omega|s_f, s_0), f^q_\sigma(\Omega|s_f, s_0), f^p_\sigma(\Omega|s_0)) \approx (p_\pi(\Omega|s_f, s_0), f_\sigma(\Omega|s_f, s_0), f_\sigma(\Omega|s_0))$ after the update, we get $I^{VE}_\sigma \approx I$. Now, we rewrite $\nabla_{\theta^p_\pi} I$ and obtain $\nabla_{\theta^p_\pi} I^{VE}_\sigma$ using (5):

$$\begin{aligned}
\nabla_{\theta^p_\pi} I &= \sum_\Omega p(\Omega|s_0) \underbrace{\left[\log \frac{p_\pi(\Omega|s_f, s_0)}{p^p_\pi(\Omega|s_0)} + \log \frac{p_\rho(\Omega|s_f, s_0)}{p_\rho(\Omega|s_0)}\right]}_{r^I_\Omega} \nabla_{\theta^p_\pi} \log p^p_\pi(\Omega|s_0), \\
\nabla_{\theta^p_\pi} I^{VE}_\sigma &= \sum_\Omega p(\Omega|s_0) \underbrace{\left[\log \frac{p^q_\pi(\Omega|s_f, s_0)}{p^p_\pi(\Omega|s_0)} + \log \frac{f^q_\sigma(\Omega|s_f, s_0)}{f^p_\sigma(\Omega|s_0)}\right]}_{r^{I^{VE}_\sigma}_\Omega} \nabla_{\theta^p_\pi} \log p^p_\pi(\Omega|s_0).
\end{aligned} \tag{27}$$

Note that $\nabla_{\theta^p_\pi} I^{VE}_\sigma \to \nabla_{\theta^p_\pi} I$ as $(p^q_\pi(\Omega|s_f, s_0),\ f^q_\sigma(\Omega|s_f, s_0)/f^p_\sigma(\Omega|s_0)) \to (p_\pi(\Omega|s_{f|\Omega}, s_0),\ p_\rho(\Omega|s_{f|\Omega}, s_0)/p_\rho(\Omega|s_0))$. From (21), we get $\nabla_{\theta^p_\pi} I^{VE}_\sigma \approx \nabla_{\theta^p_\pi} I$ for finite $\overline{T}$ and $\sigma \ll d_{min}$. However, choosing a small enough σ is not straightforward since d_{min}, T_{max} and $\overline{T}$ depend on the environment. Too small σ makes the training of $f^q_\sigma(s_{t+1}|\tau_t, a_t, s_f)$ and $f^p_\sigma(s_{t+1}|\tau_t, a_t)$ unstable due to its extreme gradient near the center of Gaussian distribution. Another issue of GMM is the choice of a proper n_{gmm} of (18). We may choose a very large n_{gmm} for the perfect fit of (18) to (17) but it makes training hard for its complexity. We leave the proper choice of σ and n_{gmm} as future work and use empirically chosen values ($\sigma = 0.25$ and $n_{gmm} = 10$) for the experiments in this paper. Using (25), (26) and (27), we summarize our method in Algorithm 2. Additional steps added to the original implicit VIC are marked with $(*)$.

Algorithm 2 Implicit VIC with Gaussian mixture model

Initialize s_0, η, T_{train}, T_{smooth}, θ^q_π, θ^q_ρ, θ^p_π and θ^p_ρ.
for i_{train} : 1 to T **do**
  Follow $\pi^p_{\theta^p_\pi}(a_t|\tau_t)$ result in $\Omega = (s_0, a_0, ..., s_f)$.
  $r^{I^{VE}_\sigma}_\Omega \leftarrow \sum_t [\log \pi^q_{\theta^q_\pi}(a_t|\tau_t, s_f) - \log \pi^p_{\theta^p_\pi}(a_t|\tau_t)]$ ▷ from (27)
  $r^{I^{VE}_\sigma}_\Omega \leftarrow r^{I^{VE}_\sigma}_\Omega + \sum_t [\log f^q_\sigma(s_{t+1}|\tau_t, a_t, s_f) - \log f^p_\sigma(s_{t+1}|\tau_t, a_t)]$ ▷ from (27), (*)
  Update each parameter:
    $\theta^p_\pi \leftarrow \theta^p_\pi + \eta r^{I^{VE}_\sigma}_\Omega \nabla_{\theta^p_\pi} \sum_t \log \pi^p_{\theta^p_\pi}(a_t|\tau_t)$ ▷ from (27)
    $\theta^q_\pi \leftarrow \theta^q_\pi + \eta \nabla_{\theta^q_\pi} \sum_t \log \pi^q_{\theta^q_\pi}(a_t|\tau_t, s_f)$ ▷ from (25)
    $\Delta\theta^p_\rho \leftarrow 0$ ▷ (*)
    $\Delta\theta^q_\rho \leftarrow 0$ ▷ (*)
    for i_{smooth} : 1 to T_{smooth} **do** ▷ (*)
      Sample $(z_1, z_2, ..., z_f), z_i \sim \mathcal{N}(\mathbf{0}, \sigma^2 I_n)$
      $\Delta\theta^p_\rho \leftarrow \Delta\theta^p_\rho + \eta \nabla_{\theta^p_\rho} \sum_t \log f^p_\sigma(s_{t+1} + z_{t+1}|\tau_t, a_t)$ ▷ from (26)
      $\Delta\theta^q_\rho \leftarrow \Delta\theta^q_\rho + \eta \nabla_{\theta^q_\rho} \sum_t \log f^q_\sigma(s_{t+1} + z_{t+1}|\tau_t, a_t, s_f)$ ▷ from (26)
    end for
    $\theta^p_\rho \leftarrow \theta^p_\rho + \Delta\theta^p_\rho / T_{smooth}$ ▷ (*)
    $\theta^q_\rho \leftarrow \theta^q_\rho + \Delta\theta^q_\rho / T_{smooth}$ ▷ (*)
  end
end for

5 EXPERIMENTS

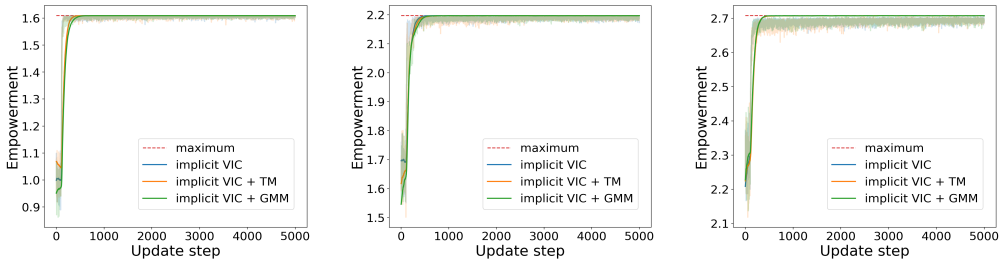
In this section, we evaluate the original implicit VIC (Gregor et al., 2016), the implicit VIC with transitional probability model (Algorithm 1) and with GMM (Algorithm 2). We use LSTM (Hochreiter & Schmidhuber, 1997) to encode $\tau_t = (s_0, a_0, ..., s_t)$ into a vector. We conduct experiments on both deterministic and stochastic environments. We evaluate each experiment by measuring the mutual information I from the samples. To measure I, we rewrite (1) using (5):

$$
\begin{aligned}
I(\Omega, s_f|s_0) &= -\sum_{\Omega, s_f} p(s_f|s_0) \log p(s_f|s_0) + \sum_{\Omega, s_f} p(\Omega|s_0) p(s_f|\Omega, s_0) \log p(s_f|\Omega, s_0) \\
&= -\sum_{s_f} p(s_f|s_0) \log p(s_f|s_0),
\end{aligned}
$$

which is maximized when s_f is distributed uniformly. For each experiment, we plot $\hat{I}$ estimated from the distribution of s_f from the samples, i.e., $\hat{p}(s_f|s_0)$. To estimate $\hat{p}(s_f|s_0)$, each experiment is repeated 5 times and the average of their result is smoothed with the exponential moving average with a smoothing factor of 0.99. We manually set T_{max} for each experiment such that the termination action is the only available action at T_{max}th action. For the training, we have a warm-up phase (100 updates in this paper) which trains the base function $b(s_0)$ in Gregor et al. (2016) and the transitional models. After the warm-up phase, we update the base function, policy and transitional models simultaneously. States are encoded as one-hot vectors for all environments. Note that d_{min} in (22) is $\sqrt{2}$ in this case and $\sigma = 0.25$ results in $\exp(-\frac{d^2_{min}}{2\sigma^2}) = \exp(-16) \approx 1.125e-7$ which is expected to be negligible. Please see Appendix E.1 for details on the hyper-parameter settings.

5.1 DETERMINISTIC ENVIRONMENTS

We compare the algorithms in a deterministic 1D, 2D and tree environment. Please see Appendix E.2 for details on the set-up. Note that although (8) is clearly zero for deterministic environments, we still train the transitional probability model of Algorithm 1 and GMM of Algorithm 2 to show the convergence to the optimum of Algorithm 1 and Algorithm 2 under deterministic environments. We set $T_{max} = 5$ for 1D and 2D environment and $T_{max} = 4$ for the tree environment where it is the maximum level of the tree. Note that having $p_{min} = p_{min,f} = 1$ results in $|D_\sigma| \leq 5.625e-7$ for $T_{max} = 5$ and $|D_\sigma| \leq 5.625e-7$ for $T_{max} = 4$, which is negligible compared to the scale of empowerment. Fig. 1 shows that all three algorithms rapidly achieve the maximal empowerment.



(a) Deterministic 1D environment with $T_{max} = 5$. Theoretical maximum is $\ln 5 \approx 1.609$.

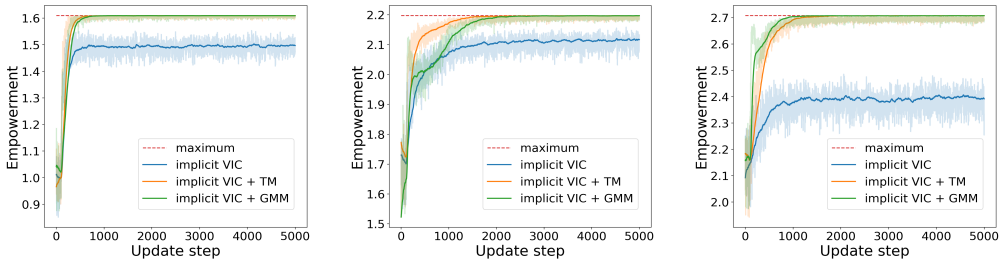
(b) Deterministic 2D environment with $T_{max} = 5$. Theoretical maximum is $\ln 9 \approx 2.197$.

(c) Deterministic tree environment with $T_{max} = 4$. Theoretical maximum is $\ln 16 \approx 2.773$.

Figure 1: Estimated empowerment during the training in deterministic environments.

5.2 Stochastic environments

We compare the algorithms in a stochastic 1D, 2D and tree environment. Please see Appendix E.3 for details on the set-up. As in the previous experiment, we set $T_{max} = 5$ for 1D and 2D environment and $T_{max} = 4$ for the tree environment. Note that $T_{max} = 5$ and $p_{min} = 0.15$ result in $\hat{U}_\sigma = 3.75e-6$ for the 1D environment, $T_{max} = 5$ and $p_{min} = 0.075$ result in $\hat{U}_\sigma = 7.5e-6$ for the 2D environment and $T_{max} = 4$ and $p_{min} = 0.15$ result in $\hat{U}_\sigma = 3.00e-6$ for the tree environment, which are all negligible to the scale of empowerment (it is hard to estimate $\check{L}_\sigma$ since $p_{min,f}$ is intractable). Fig. 2 shows that while the original implicit VIC converges to sub-optimum, our two algorithms achieve the maximal empowerment.



(a) Stochastic 1D environment with $T_{max} = 5$. Theoretical maximum is $\ln 5 \approx 1.609$.

(b) Stochastic 2D environment with $T_{max} = 5$. Theoretical maximum is $\ln 9 \approx 2.197$.

(c) Stochastic tree environment with $T_{max} = 4$. Theoretical maximum is $\ln 16 \approx 2.773$.

Figure 2: Estimated empowerment during the training in stochastic environments.

6 Conclusion

In this work, we revisited the variational intrinsic control. We showed that for the VIC with implicit options, the environment stochasticity induces a variational bias in the intrinsic reward, leading to convergence to sub-optimum. To reduce this bias and achieve maximal empowerment, we proposed to model the environment dynamics using either the transitional probability model or the Gaussian mixture model. Evaluations on stochastic environments demonstrated the superiority of our methods over the original VIC algorithm with implicit options (Gregor et al., 2016).

REFERENCES

Suguru Arimoto. An algorithm for computing the capacity of arbitrary discrete memoryless channels. *IEEE Transactions on Information Theory*, 18(1):14–20, 1972.

David Barber and Felix V Agakov. The im algorithm: a variational approach to information maximization. In *Advances in neural information processing systems*, pp. None, 2003.

Richard Blahut. Computation of channel capacity and rate-distortion functions. *IEEE transactions on Information Theory*, 18(4):460–473, 1972.

Benjamin Eysenbach, Abhishek Gupta, Julian Ibarz, and Sergey Levine. Diversity is all you need: Learning skills without a reward function. *arXiv preprint arXiv:1802.06070*, 2018.

Karol Gregor, Danilo Jimenez Rezende, and Daan Wierstra. Variational intrinsic control. *arXiv preprint arXiv:1611.07507*, 2016.

Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8): 1735–1780, 1997.

Alexander S Klyubin, Daniel Polani, and Chrystopher L Nehaniv. Empowerment: A universal agent-centric measure of control. In *2005 IEEE Congress on Evolutionary Computation*, volume 1, pp. 128–135. IEEE, 2005.

Marlos C Machado, Marc G Bellemare, and Michael Bowling. A laplacian framework for option discovery in reinforcement learning. *arXiv preprint arXiv:1703.00956*, 2017.

Shakir Mohamed and Danilo Jimenez Rezende. Variational information maximisation for intrinsically motivated reinforcement learning. In *Advances in neural information processing systems*, pp. 2125–2133, 2015.

Douglas A Reynolds and Richard C Rose. Robust text-independent speaker identification using gaussian mixture speaker models. *IEEE transactions on speech and audio processing*, 3(1):72–83, 1995.

Christoph Salge, Cornelius Glackin, and Daniel Polani. Empowerment–an introduction. In *Guided Self-Organization: Inception*, pp. 67–114. Springer, 2014.

Raymond W Yeung. The blahut–arimoto algorithms. In *Information Theory and Network Coding*, pp. 211–228. Springer, 2008.

A DERIVATION OF INTRINSIC REWARD FROM MUTUAL INFORMATION

Here we derive the intrinsic reward of VIC by taking the gradient of (1) with respect to the parameter of policy θ. We omit s_0 for simplicity. Note that $p(\Omega), p(s_f)$ and $p(\Omega, s_f)$ can be parameterized by θ since they are all determined by policy. We start by rewriting (1):

$$I(\Omega, s_f) = \sum_{\Omega, s_f} p_\theta(\Omega, s_f) \Big[\log p_\theta(\Omega, s_f) - \log p_\theta(\Omega) p_\theta(s_f) \Big].$$

Then by taking the gradient with respect to θ, we obtain

$$\begin{aligned}
\nabla_\theta I(\Omega, s_f) &= \sum_{\Omega, s_f} p_\theta(\Omega, s_f) \Big[\log p_\theta(\Omega, s_f) - \log p_\theta(\Omega) p_\theta(s_f) \Big] \nabla_\theta \log p_\theta(\Omega, s_f) \\
&\quad + \sum_{\Omega, s_f} p_\theta(\Omega, s_f) \Big[\frac{\nabla_\theta p_\theta(\Omega, s_f)}{p_\theta(\Omega, s_f)} - \frac{\nabla_\theta p_\theta(\Omega)}{p_\theta(\Omega)} - \frac{\nabla_\theta p_\theta(s_f)}{p_\theta(s_f)} \Big] \\
&= \sum_{\Omega, s_f} p_\theta(\Omega, s_f) \Big[\log p_\theta(\Omega, s_f) - \log p_\theta(\Omega) p_\theta(s_f) \Big] \nabla_\theta \log p_\theta(\Omega, s_f) \\
&\quad + \sum_{\Omega, s_f} \Big[\nabla_\theta p_\theta(\Omega, s_f) - p_\theta(s_f|\Omega) \nabla_\theta p_\theta(\Omega) - p_\theta(\Omega|s_f) \nabla_\theta p_\theta(s_f) \Big].
\end{aligned}$$

Using

$$\sum_{\Omega,s_f} \nabla_\theta p_\theta(\Omega, s_f) = 0, \sum_{\Omega,s_f} p_\theta(s_f|\Omega)\nabla_\theta p_\theta(\Omega) = 0, \sum_{\Omega,s_f} p_\theta(\Omega|s_f)\nabla_\theta p_\theta(s_f) = 0,$$

and (5) we have

$$\begin{aligned}\nabla_\theta I(\Omega, s_f) &= \sum_{\Omega,s_f} p_\theta(\Omega, s_f)\Big[\log p_\theta(\Omega, s_f) - \log p_\theta(\Omega)p_\theta(s_f)\Big]\nabla_\theta \log p_\theta(\Omega, s_f) \\ &= \sum_{\Omega} p_\theta(\Omega) r^I_\Omega \nabla_\theta p_\theta(\Omega)\end{aligned} \tag{28}$$

where

$$r^I_\Omega = \log p_\theta(\Omega|s_{f|\Omega}) - \log p_\theta(\Omega).$$

Similarly, we can obtain

$$\begin{aligned}\nabla_\theta I^{VE}(\Omega, s_f) &= \sum_{\Omega,s_f} p_\theta(\Omega, s_f)\Big[\log q(\Omega|s_f) - \log p_\theta(\Omega)\Big]\nabla_\theta \log p_\theta(\Omega, s_f) \\ &= \sum_{\Omega} p_\theta(\Omega) r^{I^{VE}}_\Omega \nabla_\theta p_\theta(\Omega)\end{aligned}$$

where

$$r^{I^{VE}}_\Omega = \log q(\Omega|s_{f|\Omega}) - \log p_\theta(\Omega).$$

Using (3) and (4), we can rewrite r^I_Ω as

$$r^I_\Omega = \sum_{\tau_t,a_t,s_{t+1}\in\Omega} \log \frac{p_\theta(a_t|\tau_t, s_{f|\Omega})p_\theta(s_{t+1}|\tau_t, a_t, s_{f|\Omega})}{p_\theta(a_t|\tau_t)p(s_{t+1}|\tau_t, a_t)}.$$

Since $p_\theta(a_t|\tau_t, s_{f|\Omega})$ is intractable, we may replace it with variational inference $q_\phi(a_t|\tau_t, s_{f|\Omega})$ which result in

$$r^{I^{VE}}_\Omega = \sum_{\tau_t,a_t,s_{t+1}\in\Omega} \log \frac{q_\phi(a_t|\tau_t, s_{f|\Omega})p_\theta(s_{t+1}|\tau_t, a_t, s_{f|\Omega})}{p_\theta(a_t|\tau_t)p(s_{t+1}|\tau_t, a_t)}.$$

For deterministic environment, we have $p_\theta(s_{t+1}|\tau_t, a_t, s_{f|\Omega}) = p(s_{t+1}|\tau_t, a_t) = 1$ and both r^I_Ω and $r^{I^{VE}}_\Omega$ can be reduced into

$$\begin{aligned}r^I_\Omega &= \sum_{\tau_t,a_t,s_{t+1}\in\Omega} \log \frac{p_\theta(a_t|\tau_t, s_{f|\Omega})}{p_\theta(a_t|\tau_t)} \\ r^{I^{VE}}_\Omega &= \sum_{\tau_t,a_t,s_{t+1}\in\Omega} \log \frac{q_\phi(a_t|\tau_t, s_{f|\Omega})}{p_\theta(a_t|\tau_t)} = r^{I^{VB}}_\Omega.\end{aligned}$$

B Derivation of U^{VE}

Here we derive U^{VE} from (13) with omitted s_0 for simplicity:

$$\begin{aligned}\left|I - I^{VE}\right| &= \left|\sum_{\Omega,s_f} p(\Omega, s_f)\Big[\log \frac{p_\pi(\Omega|s_f)p_\rho(\Omega|s_f)}{p^q_\pi(\Omega|s_f)p^q_\rho(\Omega|s_f)} - \log \frac{p_\pi(\Omega)p_\rho(\Omega)}{p^p_\pi(\Omega)p^p_\rho(\Omega)}\Big]\right| \\ &\leq \left|\sum_{\Omega,s_f} p(\Omega, s_f) \log \frac{p_\pi(\Omega|s_f)p_\rho(\Omega|s_f)}{p^q_\pi(\Omega|s_f)p^q_\rho(\Omega|s_f)}\right| + \left|\sum_{\Omega,s_f} p(\Omega, s_f) \log \frac{p_\pi(\Omega)p_\rho(\Omega)}{p^p_\pi(\Omega)p^p_\rho(\Omega)}\right| \\ &= \left|\sum_{\Omega,s_f} p(s_f)p(\Omega|s_f) \log \frac{p_\pi(\Omega|s_f)p_\rho(\Omega|s_f)}{p^q_\pi(\Omega|s_f)p^q_\rho(\Omega|s_f)}\right| + \left|\sum_{\Omega,s_f} p(\Omega)p(s_f|\Omega) \log \frac{p_\pi(\Omega)p_\rho(\Omega)}{p^p_\pi(\Omega)p^p_\rho(\Omega)}\right|.\end{aligned}$$

Using (5), (9) and (10), we obtain

$$\begin{aligned}\left|I - I^{VE}\right| \leq & \left|\sum_{\Omega, s_f} p(s_f)p(\Omega|s_f)\log\frac{p_\pi(\Omega|s_f)p_\rho(\Omega|s_f)}{p^q_\pi(\Omega|s_f)p^q_\rho(\Omega|s_f)}\right| + \left|\sum_{\Omega, s_f} p(\Omega)p(s_f|\Omega)\log\frac{p_\pi(\Omega)p_\rho(\Omega)}{p^p_\pi(\Omega)p^p_\rho(\Omega)}\right| \\ = & \left|\sum_{\Omega, s_f} p(s_f)p(\Omega|s_f)\log\frac{p_\pi(\Omega|s_f)p_\rho(\Omega|s_f)}{p^q_\pi(\Omega|s_f)p^q_\rho(\Omega|s_f)}\right| + \left|\sum_{\Omega} p(\Omega)\log\frac{p_\pi(\Omega)p_\rho(\Omega)}{p^p_\pi(\Omega)p^p_\rho(\Omega)}\right| \quad (\because (5)) \\ = & \sum_{s_f} p(s_f) D_{\mathrm{KL}}\left[p_\pi(\cdot|s_f)p_\rho(\cdot|s_f)||p^q_\pi(\cdot|s_f)p^q_\rho(\cdot|s_f)\right] + D_{\mathrm{KL}}\left[p_\pi(\cdot)p_\rho(\cdot)||p^p_\pi(\cdot)p^p_\rho(\cdot)\right] \\ = & U^{VE}.\end{aligned}$$

C Derivation of $L_{\sigma|\Omega}$ and $U_{\sigma|\Omega}$

Here we derive (19) from (17). First, we derive the upper bound on $f_\sigma(s_{t+1}|\tau_t, a_t)$:

$$\begin{aligned}f_\sigma(s_{t+1}|\tau_t, a_t) = & \sum_{s' \in S(\tau_t, a_t)} p(s'|\tau_t, a_t) f_\sigma(s_{t+1} - s'; \mathbf{0}, \sigma^2 \mathbf{I}_n) \\ = & \frac{1}{\sqrt{(2\pi\sigma^2)^n}}\Big(p(s_{t+1}|\tau_t, a_t) + \sum_{s' \neq s_{t+1} \in S(\tau_t, a_t)} p(s'|\tau_t, a_t)\exp\left(-\frac{\left\|s_{t+1} - s'\right\|_2^2}{2\sigma^2}\right)\Big) \\ \leq & \frac{1}{\sqrt{(2\pi\sigma^2)^n}}\Big(p(s_{t+1}|\tau_t, a_t) + \sum_{s' \neq s_{t+1} \in S(\tau_t, a_t)} p(s'|\tau_t, a_t)\exp\left(-\frac{d^2_{min}}{2\sigma^2}\right)\Big) \\ = & \frac{1}{\sqrt{(2\pi\sigma^2)^n}}\Big(p(s_{t+1}|\tau_t, a_t) + (1 - p(s_{t+1}|\tau_t, a_t))\exp\left(-\frac{d^2_{min}}{2\sigma^2}\right)\Big) \\ \leq & \frac{1}{\sqrt{(2\pi\sigma^2)^n}}\Big(p(s_{t+1}|\tau_t, a_t) + \exp\left(-\frac{d^2_{min}}{2\sigma^2}\right)\Big).\end{aligned}$$

Obviously, we have $\frac{1}{\sqrt{(2\pi\sigma^2)^n}} p(s_{t+1}|\tau_t, a_t) \leq f_\sigma(s_{t+1}|\tau_t, a_t)$ which results in:

$$p(s_{t+1}|\tau_t, a_t) \leq \sqrt{(2\pi\sigma^2)^n} f_\sigma(s_{t+1}|\tau_t, a_t) \leq p(s_{t+1}|\tau_t, a_t) + \exp\left(-\frac{d^2_{min}}{2\sigma^2}\right). \tag{29}$$

Similarly, we can obtain the bounds of $f_\sigma(s_{t+1}|\tau_t, a_t, s_f)$ as follows:

$$p(s_{t+1}|\tau_t, a_t, s_f) \leq \sqrt{(2\pi\sigma^2)^n} f_\sigma(s_{t+1}|\tau_t, a_t, s_f) \leq p(s_{t+1}|\tau_t, a_t, s_f) + \exp\left(-\frac{d^2_{min}}{2\sigma^2}\right). \tag{30}$$

Combining (29) and (30), we can obtain

$$\frac{p(s_{t+1}|\tau_t, a_t, s_f)}{p(s_{t+1}|\tau_t, a_t) + \exp\left(-\frac{d^2_{min}}{2\sigma^2}\right)} \leq \frac{f_\sigma(s_{t+1}|\tau_t, a_t, s_f)}{f_\sigma(s_{t+1}|\tau_t, a_t)} \leq \frac{p(s_{t+1}|\tau_t, a_t, s_f) + \exp\left(-\frac{d^2_{min}}{2\sigma^2}\right)}{p(s_{t+1}|\tau_t, a_t, s_f)}. \tag{31}$$

Taking log and using $\log(a + b) \leq \log a + \frac{b}{a}$ for $a, b > 0$ to (31), we have

$$\begin{aligned}\log\frac{p(s_{t+1}|\tau_t, a_t, s_f)}{p(s_{t+1}|\tau_t, a_t)} - \frac{1}{p_{min}}\exp\left(-\frac{d^2_{min}}{2\sigma^2}\right) \leq & \\ \log\frac{f_\sigma(s_{t+1}|\tau_t, a_t, s_f)}{f_\sigma(s_{t+1}|\tau_t, a_t)} & \\ \leq \log\frac{p(s_{t+1}|\tau_t, a_t, s_f)}{p(s_{t+1}|\tau_t, a_t)} + \frac{1}{p_{min,f}}\exp\left(-\frac{d^2_{min}}{2\sigma^2}\right) & \end{aligned}$$

which results in

$$-\frac{1}{p_{min,f}}\exp\left(-\frac{d^2_{min}}{2\sigma^2}\right) \leq \log\frac{p(s_{t+1}|\tau_t, a_t, s_f)}{p(s_{t+1}|\tau_t, a_t)} - \log\frac{f_\sigma(s_{t+1}|\tau_t, a_t, s_f)}{f_\sigma(s_{t+1}|\tau_t, a_t)} \leq \frac{1}{p_{min}}\exp\left(-\frac{d^2_{min}}{2\sigma^2}\right). \tag{32}$$

Finally, Using (9) and (32) we obtain (19):

$$-\frac{T_\Omega}{p_{min,f}}\exp(-\frac{d_{min}^2}{2\sigma^2}) \leq \log\frac{p_\rho(\Omega|s_f,s_0)}{p_\rho(\Omega|s_0)} - \log\frac{f_\sigma(\Omega|s_f,s_0)}{f_\sigma(\Omega|s_0)} \leq \frac{T_\Omega}{p_{min}}\exp(-\frac{d_{min}^2}{2\sigma^2}).$$

D DERIVATION OF $U_{\sigma,1}^{VE}$ AND $U_{\sigma,2}^{VE}$

Here we derive (24). We also omit s_0 for simplicity here. We start from $\left|I - I_\sigma^{VE}\right|$:

$$\begin{aligned}
\left|I - I_\sigma^{VE}\right| &= \left|\sum_{\Omega,s_f} p(\Omega,s_f)\Big[\log\frac{p_\pi(\Omega|s_f)p_\rho(\Omega|s_f)}{p_\pi^q(\Omega|s_f)f_\sigma^q(\Omega|s_f)} - \log\frac{p_\pi(\Omega)p_\rho(\Omega)}{p_\pi^p(\Omega)f_\sigma^p(\Omega)}\Big]\right| \\
&= \left|\sum_{\Omega,s_f} p(\Omega,s_f)\Big[\log\frac{p_\pi(\Omega|s_f)p_\rho(\Omega|s_f)}{p_\pi^q(\Omega|s_f)f_\sigma^q(\Omega|s_f)} - \log\frac{p_\rho(\Omega)}{f_\sigma^p(\Omega)}\Big]\right| \quad (\because p_\pi^p(\Omega) = p_\pi(\Omega)) \\
&= \left|\sum_{\Omega,s_f} p(\Omega,s_f)\Big[\log\frac{p_\rho(\Omega|s_f)f_\sigma^p(\Omega)}{p_\rho(\Omega)f_\sigma^q(\Omega|s_f)} + \log\frac{p_\pi(\Omega|s_f)p_\rho(\Omega|s_f)}{p_\pi^q(\Omega|s_f)p_\rho(\Omega|s_f)}\Big]\right| \\
&\leq \left|\sum_{\Omega,s_f} p(\Omega,s_f)\log\frac{p_\rho(\Omega|s_f)f_\sigma^p(\Omega)}{p_\rho(\Omega)f_\sigma^q(\Omega|s_f)}\right| + \left|\sum_{\Omega,s_f} p(\Omega,s_f)\log\frac{p_\pi(\Omega|s_f)p_\rho(\Omega|s_f)}{p_\pi^q(\Omega|s_f)p_\rho(\Omega|s_f)}\right| \\
&= \left|\sum_{\Omega,s_f} p(\Omega,s_f)\log\frac{p_\rho(\Omega|s_f)f_\sigma^p(\Omega)}{p_\rho(\Omega)f_\sigma^q(\Omega|s_f)}\right| + \left|\sum_{\Omega,s_f} p(s_f)p(\Omega|s_f)\log\frac{p_\pi(\Omega|s_f)p_\rho(\Omega|s_f)}{p_\pi^q(\Omega|s_f)p_\rho(\Omega|s_f)}\right| \\
&= U_{\sigma,1}^{VE} + U_{\sigma,2}^{VE}.
\end{aligned}$$

E EXPERIMENTAL DETAILS

Here we specify experimental details and environment details.

E.1 HYPER-PARAMETERS

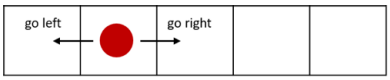
Table 1: Hyper-parameters used for experiments

Hyper-parameter	**Value**
Optimizer	Adam
Learning rate	1e-3
Betas	(0.9, 0.999)
Weight initialization	Gaussian with std. 0.1 and mean 0
Batch size	128
T_{smooth}	128
σ (GMM)	0.25
n_{gmm} (GMM)	10

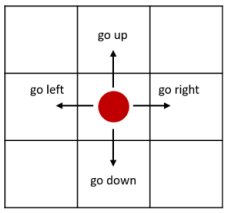
E.2 DETERMINISTIC ENVIRONMENTS DETAIL

Here we specify deterministic environments used for experiments in 5.1. Fig. 3a shows a deterministic 1D environment. Its cardinality is 5 and states are encoded as one-hot vectors. The agent can choose one of 2 actions (go left and go right) and it stays there if the chosen action makes the agent break the bound. Fig. 3b shows deterministic 2D environment. Its cardinality is 9 and states are encoded with one-hot vectors. The agent can choose one of 4 actions (go left, go up, go right and

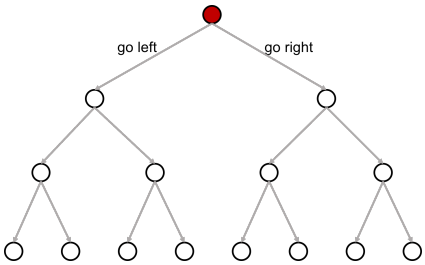
go down) and it stays there if the chosen action makes the agent break the bound. Fig. 3c shows deterministic tree environment. Its cardinality is 16 and states are encoded as one-hot vectors. The agent can choose one of 4 actions (go left, go up, go right and go down).



(a) **Deterministic 1D environment.** The agent starts from the very left and it navigates through the environment in given T_{max} transitions.



(b) **Deterministic 2D environment.** The agent starts from the very left bottom and it navigates through the environment in given T_{max} transitions.



(c) **Deterministic tree world.** The agent starts from the root and it navigates through the environment in given T_{max} transitions. Note that the T_{max} in this environment is set to the maximal level of the tree.

Figure 3: Deterministic environments.

E.3 Stochastic environments detail

Here we specify the stochastic environments used for experiments in 5.2. The environments are identical to deterministic environments except the transition is stochastic. We simply show transition tables of stochastic 1D, 2D and tree environment in table 2, 3, and 4. Note that the agent stays there if the transition makes the agent break the bound.

Table 2: Transition table of stochastic 1D environment

Action \ Transition	go left	go right
go left	0.85	0.15
go right	0.15	0.85

Table 3: Transition table of stochastic 2D environment

Action \ Transition	go left	go up	go right	go down
go left	0.775	0.075	0.075	0.075
go up	0.075	0.775	0.075	0.075
go right	0.075	0.075	0.775	0.075
go down	0.075	0.075	0.075	0.775

Table 4: Transition table of stochastic tree environment

Action \ Transition	go left	go right
go left	0.85	0.15
go right	0.15	0.85