Generative Intrinsic Optimization: Intrinsic Control with Model Learning

Jianfei Ma
School of Mathematics and Statistics
Northwestern Polytechnical University
`matrixfeeney@gmail.com`

Abstract

Future sequence represents the outcome after executing the action into the environment (i.e. the trajectory onwards). When driven by the information-theoretic concept of mutual information, it seeks maximally informative consequences. Explicit outcomes may vary across state, return, or trajectory serving different purposes such as credit assignment or imitation learning. However, the inherent nature of incorporating intrinsic motivation with reward maximization is often neglected. In this work, we propose a policy iteration scheme that seamlessly incorporates the mutual information, ensuring convergence to the optimal policy. Concurrently, a variational approach is introduced, which jointly learns the necessary quantity for estimating the mutual information and the dynamics model, providing a general framework for incorporating different forms of outcomes of interest. While we mainly focus on theoretical analysis, our approach opens the possibilities of leveraging intrinsic control with model learning to enhance sample efficiency and incorporate uncertainty of the environment into decision-making.

1 Introduction

Deep reinforcement learning (RL) aims to improve an agent's policy with a task-specific reward, showing promise in solving complex tasks such as video games [24] and robot locomotion [14]. However, in many cases, obtaining a task-specific reward can be challenging, hindering the learning process. Intrinsic motivation, on the other hand, offers an alternative approach where the agent is driven by internal rewards to achieve goals or complete tasks. Its effectiveness has been shown in RL, including skill discovery [11], curiosity-driven exploration [16], and representation learning [3]. However, existing methods often treat intrinsic reward as an additional component to the task-specific reward, optimizing them using standard RL algorithms, without fully considering its unique nature in the agent's decision-making process. Furthermore, these methods often rely on specific variational approaches tailored to particular applications, lacking a unified perspective. In this work, we propose a novel approach that transforms the standard RL objective into a mutual information maximization framework, which employs a variational approach, enabling simultaneous approximation of the posterior and the transition model. This unified approach facilitates efficient intrinsic control combined with model learning.

In this paper, we present a comprehensive intrinsic control framework called Generative Intrinsic Optimization (GIO) that integrates a policy iteration scheme and a variational approach, enabling effective policy optimization by incorporating intrinsic motivation as a fundamental component of the agent's decision-making process. Our method is applicable to various future sequence forms, from one-step future sequences $\mathcal{F} = (s', r)$ to multi-step transitions, as well as compressed future sequences, offering potential synergies with existing approaches for further improvement. We provide
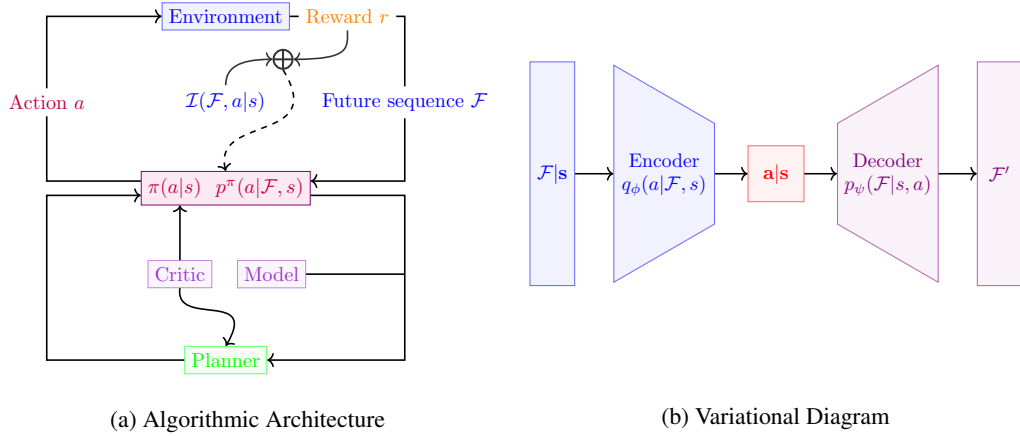
(a) Algorithmic Architecture

(b) Variational Diagram

Figure 1: **Left**: GIO combines both intrinsic and extrinsic rewards for policy learning and utilizes the learned model for planning purposes; **Right**: The variational model comprises an inference model and a generative model for posterior approximation and model learning.

a theoretical analysis of the convergence of our proposed scheme, ensuring monotonicity, and derive variational lower bounds for both one-step and multi-step scenarios.

2 Preliminaries

2.1 Notation

Consider a regularized infinite-horizon discounted MDP, defined by a tuple $(\mathcal{S}, \mathcal{A}, P, r, \rho_0, \gamma, \Delta)$, where $\mathcal{S}$ is the state space, $\mathcal{A}$ is the action space, $P : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to \mathbb{R}$ is the transition probability distribution, $r : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ is the reward function, $\rho_0 : \mathcal{S} \to \mathbb{R}$ is the distribution of the initial state s_0, $\gamma \in [0, 1)$ is the discount factor, and the additional term Δ represents other rewards such as intrinsic reward. We aim to maximize the objective function $\mathbb{E}_\tau \left[\sum_{t=0}^{\infty} \gamma^t (r_t + \eta \Delta_t) \right]$ with a temperature parameter η, where τ represents the trajectory generated by a stochastic policy $\pi : \mathcal{S} \times \mathcal{A} \to [0, 1]$. We denote the entropy of a distribution as $H(\cdot)$.

Information Seeking RL In the context of information-seeking RL, we introduce $\Delta = \mathcal{I}^\pi(\mathcal{F}, a|s)$ as the state-conditional mutual information between the current action and a future sequence $\mathcal{F}$ beyond the action execution. Our goal is to maximize the expected augmented reward by incorporating this mutual information term.

$$\eta(\pi) = \mathbb{E}_\tau \left[\sum_{t=0}^{\infty} \gamma^t (r_t + \eta \mathcal{I}^\pi(\mathcal{F}_t, a_t|s_t)) \right] \tag{1}$$

This formulation captures the uncertainty reduction between the current policy and the posterior, providing a flexible framework for various RL formulations. For instance, adopting an optimistic perspective, where the future sequence fully explains the executed action, the mutual information reduces to the entropy $\Delta = H(\pi)$, encouraging pure exploration [37]. In contrast, standard RL [33] takes a pessimistic stance, assuming the future sequence reveals no information about the executed action, that is, $\Delta = 0$.

In this paper, our focus will primarily be on the more general form of mutual information, allowing the incorporation of different choices of future sequences within a unified framework.

3   Mutual Information

Mutual information depicts mutual dependence between two random variables. Being an information-theoretic measure, it can be used to quantify the amount of information contained in the future $\mathcal{F}$ that explains the action a given the current state s

$$\mathcal{I}^{\pi}(\mathcal{F}, a|s) = \mathbb{E}_{\pi(a|s)p(\mathcal{F}|s,a)} \left[\log \frac{p(\mathcal{F}, a|s)}{p(\mathcal{F}|s)\pi(a|s)} \right] \tag{2}$$

where $\mathcal{F}$ can be any successor outcomes ahead of (s, a), for instance one-step transition $\mathcal{F} = (s', r)$, or multi-steps subsequence $\mathcal{F} \subseteq (s'_>, r_>)$. This quantity is compelling as it encourages the agent to seek maximally informative future outcomes and thereby reduce the uncertainty of the decisions.

$$\mathcal{I}^{\pi}(\mathcal{F}, a|s) = H(\pi(a|s)) - \mathbb{E}_{p^{\pi}(\mathcal{F}|s)} [H(p^{\pi}(a|\mathcal{F}, s))] \tag{3}$$

where $p^{\pi}(a|\mathcal{F}, s)$ is the posterior distribution corresponds to the prior π after observing new outcomes. The mutual information quantifies the reduction in uncertainty between the prior and posterior. In what will follow, we present a policy iteration scheme that helps the agent pursue a policy that seeks maximum information about the future.

4   Intrinsic Policy Iteration

We start by deriving an intrinsic Bellman operator and proposing a policy iteration scheme. We then present a general convergence result for all valid future sequences $\mathcal{F}$.

It is useful to define the following operator

$$\mathcal{T}^{\pi}Q(s_t, a_t) = r(s_t, a_t) + \gamma \mathbb{E}_{s_{t+1}}[V(s_{t+1})], \tag{4}$$

where

$$V(s_t) = \mathbb{E}_{a_t \sim \pi, \mathcal{F}_t}[Q(s_t, a_t) + \eta(\log p^{\pi}(a_t|s_t, \mathcal{F}_t) - \log \pi(a_t|s_t))] \tag{5}$$

where η is a hyperparameter that controls the relative strength of the augmentation against the reward.

It is not difficult to see that $\mathcal{T}^{\pi}$ is a contraction by modifying the reward as $r(s, a) + \gamma \mathbb{E}_{s'}[\mathcal{I}(a', \mathcal{F}'|s')]$. It indicates that if we repeatedly apply the intrinsic Bellman operator, we will get the intrinsic action-value function Q^{π}.

Proposition 4.1. *If $\mathcal{I}(\mathcal{F}, a|s)$ is bounded for any $s \in \mathcal{S}$, then $\lim_{k\to\infty}(\mathcal{T}^{\pi})^k Q = Q^{\pi}$ for any initial function Q, and specifically Q^{π} is the unique solution of* (4).

Although the intricate relationship between the posterior p^{π} and the policy π makes a direct improvement over Q^{π} infeasible, it is possible to follow an alternating optimization procedure when both Q^{π} and p^{π} are fixed. In such cases, under certain conditions, this approach can still ensure optimality.

We can solve for the one-step optimal policy when the intrinsic action-value function is attained as follows

$$\mathcal{G}(Q^{\pi}, p^{\pi}) = \frac{\exp \frac{1}{\eta}\left(Q^{\pi} + \eta \mathbb{E}_{\mathcal{F}}[\log p^{\pi}(a|s, \mathcal{F})]\right)}{Z^{\pi}(s)} \tag{6}$$

where $Z^{\pi}(s)$ is a partition function dependent only on state s.

Repeated application of the intrinsic Bellman operator and the softmax operator, we can produce a sequence of $Q^{\pi_k}, k = 0, 1, \cdots$ by starting from arbitrary policy π_0. Unsurprisingly, under some mild condition, for any future sequence $\mathcal{F}$ of interest, it is guaranteed to converge to the optimal policy $\pi^{\star} \triangleq \arg\max_{\pi} V^{\pi}$ (where V^{π} can be obtained by inserting Q^{π} into Equation (5)) and the optimal action-value function $Q^{\pi^{\star}}$.

Assumption 4.2. The entropy $H(\pi^{\star})$ is bounded.

Assumption 4.3. The initial policy π_0 is non-zero everywhere.

Assumption 4.4. The limit of

$$\sum_{k=0}^{n} \mathbb{E}_{s', (a', \mathcal{F}') \sim (p^{\pi^{\star}}(\mathcal{F}', a'|s') - p^{\pi_{k+1}}(\mathcal{F}', a'|s'))} \left[\eta(\log p^{\pi^{\star}}(a'|\mathcal{F}', s') - \log \pi^{\star}(a'|s')) + Q^{\pi_k} \right] \tag{7}$$

exists[1] for any $(s, a) \in \mathcal{S} \times \mathcal{A}$.

---

[1] We denote $\mathbb{E}_{p-q} = \mathbb{E}_p - \mathbb{E}_q$ for less verbatim repetition.

Theorem 4.5. *Under assumptions 4.2–4.4, for any future sequence $\mathcal{F}$, it holds that*

$$\lim_{k \to \infty} Q^{\pi_k} = Q^{\pi^\star} \tag{8}$$

However, without knowing the transition model and the posterior, it may be difficult to utilize this general convergence result. In the next section, we unify the model learning and posterior approximation into a single model, considering one-step transitions.

5 Inference with Model Learning

5.1 Variational Inference

Due to the intractability of the marginal distribution, obtaining the posterior can be challenging. Therefore, we employ variational inference [20] using an inference model $q_\phi(a|s, s', r)$ to approximate the true posterior. The dynamic model is parameterized as $p_\psi(s', r|s, a)$. For a given policy π, we can derive a variational lower bound on the conditional marginal distribution of $\mathcal{F} = (s', r)$

$$
\begin{aligned}
\log p^\pi(s', r|s) &\geq \mathcal{L}(\phi, \psi; s, s', r) \\
&= -D_{\text{KL}}(q_\phi(a|s, s', r)\|\pi(a|s)) + \mathbb{E}_{q_\phi(a|s,s',r)}[\log p_\psi(s', r|s, a)]
\end{aligned} \tag{9}
$$

where the action space is naturally treated as a latent inference target, for which the policy contains the necessary prior knowledge. The recognition model encodes the sequential experiences to infer the true posterior and the generative model constructs environment dynamics. This allows efficient posterior approximation for any future sequence $\mathcal{F}$, capturing complex dynamics.

In practice, we make a common assumption of factorization for the transition model p_ψ

$$p_\psi(s', r|s, a) = p_\psi(s'|s, a)p_\psi(r|s, a) \tag{10}$$

5.2 Policy Improvement

After observing new outcomes emitted from the environment, the agent will update its belief over the current policy based on both extrinsic and intrinsic rewards. We project the policy onto the one-step optimal policy $\mathcal{G}(Q^\pi, p^\pi)$ as shown in Equation (6) for each state $s \in \mathcal{S}$

$$\arg\min_{\pi' \in \Pi} D_{\text{KL}}\left(\pi'(\cdot|s) \middle\| \mathcal{G}(Q^\pi, p^\pi)\right) \tag{11}$$

Once we arrive at our new policy, by reevaluating the corresponding posterior, we can guarantee a monotonic improvement.

Theorem 5.1. *If $\tilde{\pi}$ minimizes the projection loss against π for any $s \in \mathcal{S}$ and the corresponding posterior is attained. Then $Q^{\tilde{\pi}}(s_t, a_t) \geq Q^\pi(s_t, a_t)$ for all $(s_t, a_t) \in \mathcal{S} \times \mathcal{A}$.*

5.3 Trajectory-Wise Lower Bound

Alternatively, we can consider a lower bound upon a trajectory τ with a finite horizon T, enabling the capture of longer horizon information with a more accurate prediction of the transition and wider coverage of the trajectory surprise. Denote $\mathbf{x}_{1:T} = (s_{t+1}, r_t|s_t)_{t=1}^T$ and $\mathbf{z}_{1:T} = (a_t|s_t)_{t=1}^T$, then we have

$$\log p^\pi(\mathbf{x}_{1:T}) \geq \sum_{\tau=1}^T \mathbb{E}_{\mathbf{z}_{1:\tau}} \left[\log p_\psi(s'_\tau, r_\tau|s_\tau, a_\tau) - D_{KL}(q_\phi(a_\tau|s_{1:\tau+1}, r_{1:\tau}, a_{1:\tau-1}) \| \pi(a_\tau|s_\tau))\right] \tag{12}$$

where we assume factorization of the recognition and generative model and also use the Markov property of the policy distribution.

If we further assume conditional independence of the past information i.e. $t < \tau$ for q_ϕ, then it reduces to a compact formulation being a summation of a series of one-step lower bounds (Equation (9)). It can be useful for inferring multi-step posteriors simultaneously and fitting a transition model with a longer temporal dependence. Although it is promising to combine with techniques such as RNN [30] or transformer [35], it is beyond our scope and can be a further enhancement of our method.

6 Related Work

Intrinsic Motivation Intrinsic motivation is the drive to seek out and engage in activities that promote learning, exploration, and curiosity-driven behavior [27]. Mutual information has proven effective in diverse domains, including curiosity-driven exploration [16, 19], options discovery [11, 7], and empowerment maximization [25, 17]. Our method differs from empowerment maximization as we do not learn an open-loop distribution to maximize channel capability [21]. Additionally, there exist other techniques for motivating agents from different perspectives, such as model uncertainty [28, 31, 3, 9], count-based exploration [26, 4, 34], and surprise or novelty [2, 32, 5]. While entropy is commonly used in model-free algorithms, we treat it as an intrinsic motivation solely encouraging exploration, in contrast to our nature of uncertainty reduction from the environment. [22] unifies reward and empowerment maximization, but requires extensive application of the Blahut-Arimoto algorithm, whose complexity has posed a challenge to scale to the continuous domain [25]. Another essential difference is that our method focuses on an efficient policy iteration approach analogous to [13], whereas [22] manipulates the optimality operator similar to [12], but with far less flexibility on $\mathcal{F}$.

Incorporating the Future [18] incorporates information from future observations and actions using a bidirectional recurrent network in an autoregressive manner. To address credit assignment, [15] introduces the importance ratio between a state-conditional posterior and the policy, measuring the relevance of past decisions to the trajectory return or future state. Compressing the sequence into a compact representation reduces the challenges of long sequence modeling [36] [18]. The same ideas can also be drawn from, particularly when $\mathcal{F}$ is too long to capture useful information, we can instead employ an additional variational model to construct a compact representation $\mathcal{F}_z$, which is then used for posterior approximation and model learning. RL Upside Down [29] predicts actions using reward signals and states, resembling our posterior formulation. However, it diverges by redefining the policy, while our method solely informs it. Our approach provides a unified perspective, accommodating different forms of $\mathcal{F}$ within a policy iteration scheme, ensuring convergence.

7 Conclusion

In this paper, we introduce a novel learning framework that integrates intrinsic control with model learning. Our algorithm adapts to different types of future sequences, focusing on maximizing the informativeness of future outcomes given executed actions. It guarantees convergence and monotonicity. Our approach opens up possibilities for various algorithmic formulations, including trajectory-wise methods, imitation learning, and direct probabilistic control, with the full utilization of function approximations as future work.

References

[1] Abbas Abdolmaleki, Jost Tobias Springenberg, Yuval Tassa, Rémi Munos, Nicolas Heess, and Martin A. Riedmiller. Maximum a posteriori policy optimisation. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net, 2018.

[2] Joshua Achiam and Shankar Sastry. Surprise-based intrinsic motivation for deep reinforcement learning. *CoRR*, abs/1703.01732, 2017.

[3] Mohammad Gheshlaghi Azar, Bilal Piot, Bernardo A. Pires, Jean-Bastien Grill, Florent Altché, and Rémi Munos. World discovery models. *CoRR*, abs/1902.07685, 2019.

[4] Marc G. Bellemare, Sriram Srinivasan, Georg Ostrovski, Tom Schaul, David Saxton, and Rémi Munos. Unifying count-based exploration and intrinsic motivation. In Daniel D. Lee, Masashi Sugiyama, Ulrike von Luxburg, Isabelle Guyon, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016, December 5-10, 2016, Barcelona, Spain*, pages 1471–1479, 2016.

[5] Yuri Burda, Harrison Edwards, Deepak Pathak, Amos J. Storkey, Trevor Darrell, and Alexei A. Efros. Large-scale study of curiosity-driven learning. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net, 2019.

[6] Damien Ernst, Pierre Geurts, and Louis Wehenkel. Tree-based batch mode reinforcement learning. *Journal of Machine Learning Research*, 6, 2005.

[7] Benjamin Eysenbach, Abhishek Gupta, Julian Ibarz, and Sergey Levine. Diversity is all you need: Learning skills without a reward function. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net, 2019.

[8] Jianqing Fan, Zhaoran Wang, Yuchen Xie, and Zhuoran Yang. A theoretical analysis of deep q-learning. In Alexandre M. Bayen, Ali Jadbabaie, George J. Pappas, Pablo A. Parrilo, Benjamin Recht, Claire J. Tomlin, and Melanie N. Zeilinger, editors, *Proceedings of the 2nd Annual Conference on Learning for Dynamics and Control, L4DC 2020, Online Event, Berkeley, CA, USA, 11-12 June 2020*, volume 120 of *Proceedings of Machine Learning Research*, pages 486–489. PMLR, 2020.

[9] Justin Fu, John D. Co-Reyes, and Sergey Levine. EX2: exploration with exemplar models for deep reinforcement learning. In Isabelle Guyon, Ulrike von Luxburg, Samy Bengio, Hanna M. Wallach, Rob Fergus, S. V. N. Vishwanathan, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 2577–2587, 2017.

[10] Scott Fujimoto, Herke van Hoof, and David Meger. Addressing function approximation error in actor-critic methods. In Jennifer G. Dy and Andreas Krause, editors, *Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018*, volume 80 of *Proceedings of Machine Learning Research*, pages 1582–1591. PMLR, 2018.

[11] Karol Gregor, Danilo Jimenez Rezende, and Daan Wierstra. Variational intrinsic control. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Workshop Track Proceedings*. OpenReview.net, 2017.

[12] Tuomas Haarnoja, Haoran Tang, Pieter Abbeel, and Sergey Levine. Reinforcement learning with deep energy-based policies. In Doina Precup and Yee Whye Teh, editors, *Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017*, volume 70 of *Proceedings of Machine Learning Research*, pages 1352–1361. PMLR, 2017.

[13] Tuomas Haarnoja, Aurick Zhou, Pieter Abbeel, and Sergey Levine. Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. In Jennifer G. Dy and Andreas Krause, editors, *Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018*, volume 80 of *Proceedings of Machine Learning Research*, pages 1856–1865. PMLR, 2018.

[14] Tuomas Haarnoja, Aurick Zhou, Kristian Hartikainen, George Tucker, Sehoon Ha, Jie Tan, Vikash Kumar, Henry Zhu, Abhishek Gupta, Pieter Abbeel, and Sergey Levine. Soft actor-critic algorithms and applications. *CoRR*, abs/1812.05905, 2018.

[15] Anna Harutyunyan, Will Dabney, Thomas Mesnard, Mohammad Gheshlaghi Azar, Bilal Piot, Nicolas Heess, Hado van Hasselt, Gregory Wayne, Satinder Singh, Doina Precup, and Rémi Munos. Hindsight credit assignment. In Hanna M. Wallach, Hugo Larochelle, Alina Beygelzimer, Florence d'Alché-Buc, Emily B. Fox, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages 12467–12476, 2019.

[16] Rein Houthooft, Xi Chen, Yan Duan, John Schulman, Filip De Turck, and Pieter Abbeel. VIME: variational information maximizing exploration. In Daniel D. Lee, Masashi Sugiyama, Ulrike von Luxburg, Isabelle Guyon, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016, December 5-10, 2016, Barcelona, Spain*, pages 1109–1117, 2016.

[17] Maximilian Karl, Philip Becker-Ehmck, Maximilian Soelch, Djalel Benbouzid, Patrick van der Smagt, and Justin Bayer. Unsupervised real-time control through variational empowerment. In Tamim Asfour, Eiichi Yoshida, Jaeheung Park, Henrik Christensen, and Oussama Khatib, editors, *Robotics Research - The 19th International Symposium ISRR 2019, Hanoi, Vietnam, October 6-10, 2019*, volume 20 of *Springer Proceedings in Advanced Robotics*, pages 158–173. Springer, 2019.

[18] Nan Rosemary Ke, Amanpreet Singh, Ahmed Touati, Anirudh Goyal, Yoshua Bengio, Devi Parikh, and Dhruv Batra. Modeling the long term future in model-based reinforcement learning. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net, 2019.

[19] Hyoungseok Kim, Jaekyeom Kim, Yeonwoo Jeong, Sergey Levine, and Hyun Oh Song. EMI: exploration with mutual information. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA*, volume 97 of *Proceedings of Machine Learning Research*, pages 3360–3369. PMLR, 2019.

[20] Diederik P. Kingma and Max Welling. Auto-encoding variational bayes. In Yoshua Bengio and Yann LeCun, editors, *2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Conference Track Proceedings*, 2014.

[21] Alexander S. Klyubin, Daniel Polani, and Chrystopher L. Nehaniv. Empowerment: a universal agent-centric measure of control. In *Proceedings of the IEEE Congress on Evolutionary Computation, CEC 2005, 2-4 September 2005, Edinburgh, UK*, pages 128–135. IEEE, 2005.

[22] Felix Leibfried, Sergio Pascual-Diaz, and Jordi Grau-Moya. A unified bellman optimality principle combining reward maximization and empowerment. In Hanna M. Wallach, Hugo Larochelle, Alina Beygelzimer, Florence d'Alché-Buc, Emily B. Fox, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages 7867–7878, 2019.

[23] Timothy P. Lillicrap, Jonathan J. Hunt, Alexander Pritzel, Nicolas Heess, Tom Erez, Yuval Tassa, David Silver, and Daan Wierstra. Continuous control with deep reinforcement learning. In Yoshua Bengio and Yann LeCun, editors, *4th International Conference on Learning Representations, ICLR 2016, San Juan, Puerto Rico, May 2-4, 2016, Conference Track Proceedings*, 2016.

[24] Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Alex Graves, Ioannis Antonoglou, Daan Wierstra, and Martin A. Riedmiller. Playing atari with deep reinforcement learning. *CoRR*, abs/1312.5602, 2013.

[25] Shakir Mohamed and Danilo Jimenez Rezende. Variational information maximisation for intrinsically motivated reinforcement learning. In Corinna Cortes, Neil D. Lawrence, Daniel D. Lee, Masashi Sugiyama, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 28: Annual Conference on Neural Information Processing Systems 2015, December 7-12, 2015, Montreal, Quebec, Canada*, pages 2125–2133, 2015.

[26] Georg Ostrovski, Marc G. Bellemare, Aäron van den Oord, and Rémi Munos. Count-based exploration with neural density models. In Doina Precup and Yee Whye Teh, editors, *Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017*, volume 70 of *Proceedings of Machine Learning Research*, pages 2721–2730. PMLR, 2017.

[27] Pierre-Yves Oudeyer and Frédéric Kaplan. What is intrinsic motivation? A typology of computational approaches. *Frontiers Neurorobotics*, 1:6, 2007.

[28] Deepak Pathak, Dhiraj Gandhi, and Abhinav Gupta. Self-supervised exploration via disagreement. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA*, volume 97 of *Proceedings of Machine Learning Research*, pages 5062–5071. PMLR, 2019.

[29] Jürgen Schmidhuber. Reinforcement learning upside down: Don't predict rewards - just map them to actions. *CoRR*, abs/1912.02875, 2019.

[30] Alex Sherstinsky. Fundamentals of recurrent neural network (RNN) and long short-term memory (LSTM) network. *CoRR*, abs/1808.03314, 2018.

[31] Pranav Shyam, Wojciech Jaskowski, and Faustino Gomez. Model-based active exploration. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA*, volume 97 of *Proceedings of Machine Learning Research*, pages 5779–5788. PMLR, 2019.

[32] Yi Sun, Faustino J. Gomez, and Jürgen Schmidhuber. Planning to be surprised: Optimal bayesian exploration in dynamic environments. In Jürgen Schmidhuber, Kristinn R. Thórisson, and Moshe Looks, editors, *Artificial General Intelligence - 4th International Conference, AGI 2011, Mountain View, CA, USA, August 3-6, 2011. Proceedings*, volume 6830 of *Lecture Notes in Computer Science*, pages 41–51. Springer, 2011.

[33] Richard S. Sutton and Andrew G. Barto. *Reinforcement Learning: An Introduction*. The MIT Press, second edition, 2018.

[34] Haoran Tang, Rein Houthooft, Davis Foote, Adam Stooke, Xi Chen, Yan Duan, John Schulman, Filip De Turck, and Pieter Abbeel. #exploration: A study of count-based exploration for deep reinforcement learning. In Isabelle Guyon, Ulrike von Luxburg, Samy Bengio, Hanna M. Wallach, Rob Fergus, S. V. N. Vishwanathan, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 2753–2762, 2017.

[35] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. *CoRR*, abs/1706.03762, 2017.

[36] David Venuto, Elaine Lau, Doina Precup, and Ofir Nachum. Policy gradients incorporating the future. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net, 2022.

[37] Brian D. Ziebart. *Modeling Purposeful Adaptive Behavior with the Principle of Maximum Causal Entropy*. PhD thesis, Carnegie Mellon University, USA, 2010.

A   Pseudocode of GIO

Algorithm 1 Generative Intrinsic Optimization

Input: η, τ
Initial Parameter: $\{q_\phi, p_\psi\}, Q_{\bar{w}_i}, Q_{w_i}, \pi_\theta$
   for step $t \leftarrow 0, 1, \ldots, M-1$ **do**
      Execute policy π_θ in the environment
      Store transition (s_t, a_t, r_t, s_{t+1}) to the replay buffer $\mathcal{D}$
      Sample mini-batch of n transitions (s, a, r, s') from $\mathcal{D}$
      Train VAE based on the variational lower bound upon (p_ϕ, p_ψ)
      Evaluate $\Delta = (\log q_\phi(a'|s', s'', r') - \log \pi_\theta(a'|s'))$ with new sampled action $a' \sim \pi_\theta$ and experience $(s'', r') \sim p_\psi$
      Compute value target $y = r(s, a) + \gamma(\min_i Q_{\bar{w}_i}(s', a') + \eta\Delta)$
      Update critic with $\nabla_w \mathcal{J}(w)$ (Equation (13))
      Update actor with $\nabla_\theta \mathcal{J}(\theta)$ (Equation (14))
      if it is time to plan **then**
         Update actor with simulated policy gradient
      end if
      Update target net $\bar{w}_i \leftarrow (1-\tau)\bar{w}_i + \tau w_i, i = 1, 2$
   end for

We provide a potential learning procedure that utilizes the clipped double-Q technique [10] and the reparameterized policy gradient. The action-value function and the policy are parameterized as Q_w and π_θ respectively.

Denote $\Delta = (\log q_\phi(a'|s', s'', r') - \log \pi_\theta(a'|s'))$, the critic is updated by following fitted Q-iteration [6] [8]

$$\mathcal{J}(w) = \mathbb{E}_{(s,a,s',r)\sim\mathcal{D}}\left[(Q_{w_i}(s, a) - (r + \gamma(\min_i Q_{\bar{w}_i}(s', a') + \eta\Delta)))^2\right], i = 1, 2 \tag{13}$$

where a new action $a' \sim \pi_\theta(\cdot|s')$ and experience $(s'', r') \sim p_\psi$ are sampled for evaluating the log ratio. The target value network $Q_{\bar{w}_i}$ is utilized to stablize the behavior of the neural networks, which is commonly used in off-policy algorithms [23] [13] [10].

And the parameterized policy orients itself to the softmax policy w.r.t. the approximate action-value function and posterior.

$$\mathcal{J}(\theta) = \mathbb{E}_{s\sim\mathcal{D}}\left[D_{\text{KL}}\left(\pi_\theta(\cdot|s)\middle\|\mathcal{G}(Q_w, q_\phi)\right)\right] \tag{14}$$

which can also utilize the reparametrization trick [20], resulting in a potential lower variance gradient estimator.

B   Proof of Proposition 4.1

Proof. By plugging Equation (5) into (4), we have

$$\mathcal{T}^\pi Q(s_t, a_t) = r(s_t, a_t) + \gamma\mathbb{E}_{s_{t+1}, a_{t+1}, \mathcal{F}_{t+1}}[Q(s_{t+1}, a_{t+1}) + \eta(\log p^\pi(a_{t+1}|s_{t+1}, \mathcal{F}_{t+1}) - \log \pi(a_{t+1}|s_{t+1}))] \tag{15}$$

If we merge the log ratio into reward such that $\tilde{r}_t = r_t + \gamma\mathbb{E}_{s_{t+1}, a_{t+1}, \mathcal{F}_{t+1}}[\eta(\log p^\pi(a_{t+1}|s_{t+1}, \mathcal{F}_{t+1}) - \log \pi(a_{t+1}|s_{t+1}))]$, we alternatively have

$$\mathcal{T}^\pi Q(s_t, a_t) = \tilde{r}(s_t, a_t) + \gamma\mathbb{E}_{s_{t+1}, a_{t+1}}[Q(s_{t+1}, a_{t+1})] \tag{16}$$

For any Q_1, Q_2 in the action-value space $\mathcal{Q}$

$$\begin{aligned}
\|\mathcal{T}^\pi Q_1 - \mathcal{T}^\pi Q_2\|_\infty &= \sup_{s,a}\left|\tilde{r}(s, a) + \gamma\mathbb{E}_{s',a'}[Q_1(s', a')] - \tilde{r}(s, a) + \gamma\mathbb{E}_{s',a'}[Q_2(s', a')]\right| \\
&= \gamma\sup_{s,a}\left|\mathbb{E}_{s',a'}[Q_1(s', a') - Q_2(s', a')]\right| \\
&\leq \gamma\sup_{s,a}\left|\mathbb{E}_{s',a'}[\sup_{s',a'}|Q_1(s', a') - Q_2(s', a')|]\right| \\
&= \gamma\sup_{s,a}\sup_{s',a'}|Q_1(s', a') - Q_2(s', a')| \\
&= \gamma\sup_{s',a'}|Q_1(s', a') - Q_2(s', a')| \\
&= \gamma\|Q_1 - Q_2\|_\infty
\end{aligned} \tag{17}$$

This implies $\mathcal{T}^\pi$ is a contraction mapping in the metric space $\mathcal{Q}$. From the Banach fixed-point theorem, we know that starting from any initial point Q, the sequence $Q_{k+1} = \mathcal{T}^\pi Q_k$ converges to a unique fixed point $Q^\star$. Since Q^π solves for $\mathcal{T}^\pi$ by definition, it implies $Q^\pi = Q^\star$. $\qquad\square$

C Proof of Equation (6)

Proof. Considering the one-step optimization problem[2] with the posterior p^π and the value function V^π being fixed

$$V^{\tilde{\pi}, p_\pi}(s) \triangleq \sup_\pi \mathbb{E}_{a \sim \pi, \mathcal{F}} \left[r(s, a) + \eta(\log p^\pi(a|s, \mathcal{F}) - \log \pi(a|s)) + \gamma \mathbb{E}_{s'}[V^\pi(s')] \right]$$

$$= \sup_\pi \mathbb{E}_{a \sim \pi, \mathcal{F}} \left[\eta(\log p^\pi(a|s, \mathcal{F}) - \log \pi(a|s)) + Q^\pi(s, a) \right] \tag{18}$$

Define the Lagrangian function $\mathcal{L}(s; \lambda) : \mathcal{S} \to \mathbb{R}$

$$\mathcal{L}(s; \lambda) = \mathbb{E}_{a \sim \pi, \mathcal{F}} \left[\eta(\log p^\pi(a|s, \mathcal{F}) - \log \pi(a|s)) + Q^\pi(s, a) \right] - \lambda \left(\sum_{a \in \mathcal{A}} \pi(a|s) - 1 \right) \tag{19}$$

Solving for the first-order equation

$$0 = \frac{\partial \mathcal{L}(s; \lambda)}{\partial \pi(a|s)} = Q^\pi(s, a) + \eta \mathbb{E}_{\mathcal{F}} \left[\log p^\pi(a|s, \mathcal{F}) \right] - \eta \log \pi(a|s) - \eta - \lambda \tag{20}$$

Rearranging

$$\tilde{\pi} = \exp\left(-\frac{\lambda}{\eta} - 1\right) \exp \frac{1}{\eta} \left(Q^\pi + \eta \mathbb{E}_{\mathcal{F}} \left[\log p^\pi(a|s, \mathcal{F}) \right] \right) \tag{21}$$

With the equality constraint

$$\sum_{a \in \mathcal{A}} \tilde{\pi}(a|s) = 1 \tag{22}$$

by applying log transformation on both sides, we can solve for the multiplier as

$$\tilde{\lambda} = \eta \log \sum_{a \in \mathcal{A}} \exp \frac{1}{\eta} \left(Q^\pi + \eta \mathbb{E}_{\mathcal{F}} \left[\log p^\pi(a|s, \mathcal{F}) \right] \right) - \eta \tag{23}$$

inserting which into Equation (21), we get

$$\tilde{\pi}(a|s) = \frac{\exp \frac{1}{\eta} \left(Q^\pi + \eta \mathbb{E}_{\mathcal{F}} \left[\log p^\pi(a|s, \mathcal{F}) \right] \right)}{\sum_{a \in \mathcal{A}} \exp \frac{1}{\eta} \left(Q^\pi + \eta \mathbb{E}_{\mathcal{F}} \left[\log p^\pi(a|s, \mathcal{F}) \right] \right)} \tag{24}$$

For the optimal policy, there must exist a multiplier that jointly satisfy KKT condition. Since $(\tilde{\pi}, \tilde{\lambda})$ uniquely satisfies the KKT condition as above, it implies $\tilde{\pi}$ is the optimal policy. Denote the denominator as $Z^\pi(s)$, it completes the proof. $\qquad \square$

D Proof of Theorem 4.5

Lemma D.1. *Let $p(x, y)$ be the joint distribution, and $p(x)$ and $p(y)$ be the marginal distribution correspondingly, then for any distribution $q(y)$, it holds that*

$$D_{KL}(p(x, y) \| p(y)p(x)) \leq D_{KL}(p(x, y) \| q(y)p(x)) \tag{25}$$

Proof. Denote $\Gamma_{x,y}$ as either $\int_{x,y}$ for continuous case or $\sum_{x,y}$ for discrete case, by non-negativity of KL divergence, it follows that

$$D_{\text{KL}}(p(x, y) \| q(y)p(x)) - D_{\text{KL}}(p(x, y) \| p(y)p(x))$$

$$= \Gamma_{x,y} p(x, y) \log \frac{p(x, y)}{p(x)p(y)} - \Gamma_{x,y} p(x, y) \log \frac{p(x, y)}{p(x)q(y)}$$

$$= \Gamma_{x,y} p(x, y) \log \frac{p(y)}{q(y)}$$

$$= \Gamma_y p(y) \log \frac{p(y)}{q(y)} \tag{26}$$

$$= D_{\text{KL}}(p(y) \| q(y))$$

$$\geq 0$$

which completes the proof. $\qquad \square$

---

[2]For simplicity, our derivation is based on the discrete case, however, the same procedure also applies for the continuous case likewise [1]. Thus, the claims will not degenerate.

Corollary D.2. *For any distribution $q(x|y)$, it holds that*

$$\mathbb{E}_{p(x,y)}\left[\frac{\log p(x|y)}{p(x)}\right] \geq \mathbb{E}_{p(x,y)}\left[\frac{\log q(x|y)}{p(x)}\right] \tag{27}$$

Proof. The proof is similar to that of the previous lemma, by non-negativity of KL divergence, it follows that

$$
\begin{aligned}
&\mathbb{E}_{p(x,y)}\left[\frac{\log p(x|y)}{p(x)}\right] - \mathbb{E}_{p(x,y)}\left[\frac{\log q(x|y)}{p(x)}\right] \\
&= \Gamma_{x,y} p(x,y) \log \frac{p(x|y)}{q(x|y)} \\
&= \Gamma_{x,y} p(y)p(x|y) \log \frac{p(x|y)}{q(x|y)} \\
&= \mathbb{E}_{p(y)}\left[D_{\mathrm{KL}}(p(x|y)\|q(x|y))\right] \\
&\geq 0
\end{aligned}
\tag{28}
$$

$\square$

We will formally give a proof of the theorem.

Proof. We first investigate the optimal intrinsic Bellman operator $\mathcal{T}^\star$, and then relate it with any intermediate operator $\mathcal{T}^{\pi_k}$.

As defined previously, the optimal policy is $\pi^\star = \arg\max_\pi V^\pi$, whose corresponding optimal value function thereby is $V^{\pi^\star}$. It should satisfy the intrinsic Bellman equation, therefore $Q^{\pi^\star}$ is defined as follows

$$Q^{\pi^\star}(s,a) = r(s,a) + \gamma \mathbb{E}_{s',a'\sim\pi^\star,\mathcal{F}'}\left[Q^{\pi^\star}(s',a') + \eta(\log p^{\pi^\star}(a'|\mathcal{F}',s') - \log\pi(a'|s'))\right] \tag{29}$$

By Proposition 4.1, it turns out that $\mathcal{T}^\star Q^{\pi^\star} = Q^{\pi^\star}$.

Now we relate it to $\mathcal{T}^{\pi_k}$, considering $\mathcal{T}^\star Q^{\pi_k}, \forall k \geq 0$, which can be bounded as

$$
\begin{aligned}
\mathcal{T}^\star Q^{\pi_k} &= r(s,a) + \gamma \mathbb{E}_{s',a'\sim\pi^\star,\mathcal{F}'}\left[Q^{\pi_k}(s',a') + \eta(\log p^{\pi^\star}(a'|\mathcal{F}',s') - \log\pi^\star(a'|s'))\right] \\
&= r(s,a) + \gamma \mathbb{E}_{s',a'\sim\pi^\star,\mathcal{F}'}\left[Q^{\pi_k}(s',a') + \eta(\log\frac{p(\mathcal{F}'|s',a')\pi^\star(a'|s')}{\sum_{a'}p(\mathcal{F}'|s',a')\pi^\star(a'|s')} - \log\pi^\star(a'|s'))\right] \\
&= r(s,a) + \gamma \mathbb{E}_{s',a'\sim\pi^\star,\mathcal{F}'}\left[Q^{\pi_k}(s',a') + \eta(\log\frac{p(\mathcal{F}'|s',a')}{\sum_{a'}p(\mathcal{F}'|s',a')\pi^\star(a'|s')})\right] \\
&= r(s,a) + \gamma \mathbb{E}_{s',a'\sim\pi^\star,\mathcal{F}'}\left[Q^{\pi_k}(s',a') + \eta(\log\frac{p(\mathcal{F}'|s',a')}{p^{\pi^\star}(\mathcal{F}'|s')})\right] \\
&= r(s,a) + \gamma \mathbb{E}_{s',a'\sim\pi^\star,\mathcal{F}'}\left[Q^{\pi_k}(s',a') + \eta(\log\frac{p(\mathcal{F}'|s',a')\pi^\star(a'|s')}{p^{\pi^\star}(\mathcal{F}'|s')\pi^\star(a'|s')})\right] \\
&\leq r(s,a) + \gamma \mathbb{E}_{s',a'\sim\pi^\star,\mathcal{F}'}\left[Q^{\pi_k}(s',a') + \eta(\log\frac{p(\mathcal{F}'|s',a')\pi^\star(a'|s')}{\underbrace{\sum_{a'}p(\mathcal{F}'|s',a')\pi_k(a'|s')}_{q(\mathcal{F}'|s')}\pi^\star(a'|s')})\right] \quad \triangleright\text{ by Lemma D.1} \\
&= r(s,a) + \gamma \mathbb{E}_{s',a'\sim\pi^\star,\mathcal{F}'}\left[Q^{\pi_k}(s',a') + \eta(\log\frac{p(\mathcal{F}'|s',a')}{\sum_{a'}p(\mathcal{F}'|s',a')\pi_k(a'|s')})\right] \\
&= r(s,a) + \gamma \mathbb{E}_{s',a'\sim\pi^\star,\mathcal{F}'}\left[Q^{\pi_k}(s',a') + \eta(\log\frac{p(\mathcal{F}'|s',a')\pi_k(a'|s')}{\sum_{a'}p(\mathcal{F}'|s',a')\pi_k(a'|s')} - \log\pi_k(a'|s'))\right] \\
&= r(s,a) + \gamma \mathbb{E}_{s',a'\sim\pi^\star,\mathcal{F}'}\left[Q^{\pi_k}(s',a') + \eta(\log p^{\pi_k}(a'|\mathcal{F}',s') - \log\pi_k(a'|s'))\right]
\end{aligned}
\tag{30}
$$

11

where $q(\mathcal{F}'|s')$ is a well-defined probability since

$$\begin{aligned}
\sum_{\mathcal{F}'} q(\mathcal{F}'|s') &= \sum_{\mathcal{F}'}\sum_{a'} p(\mathcal{F}'|s',a')\pi_k(a'|s') \\
&= \sum_{a'}\sum_{\mathcal{F}'} p(\mathcal{F}'|s',a')\pi_k(a'|s') \\
&= \sum_{a'} 1\cdot\pi_k(a'|s') \\
&= 1
\end{aligned} \tag{31}$$

By plugging Equation (24) into Equation (18) in proof C, we can get

$$\begin{aligned}
\mathcal{T}^{\pi_k}Q^{\pi_k}(s,a) &= r(s,a)+\gamma\mathbb{E}_{s'}[V^{\pi_k}(s')] \\
&\leq r(s,a)+\gamma\mathbb{E}_{s'}[V^{\pi_{k+1},p^{\pi_k}}(s')] \\
&= r(s,a)+\gamma\mathbb{E}_{s'}\left[\eta\log\sum_{a'}\exp\frac{1}{\eta}\left(Q^{\pi_k}(s',a')+\eta\mathbb{E}_{\mathcal{F}'}\left[\log p^{\pi_k}(a'|s',\mathcal{F}')\right]\right)\right] \\
&= r(s,a)+\gamma\mathbb{E}_{s'}\left[\eta\log Z^{\pi_k}(s')\right] \\
&\triangleq \mathcal{T}^{\pi_{k+1},p^{\pi_k}}Q^{\pi_k}(s,a)
\end{aligned} \tag{32}$$

With a useful identity from taking logarithm of both sides of $\pi_{k+1}(a'|s')$ (Equation (24))

$$Q^{\pi_k}(s',a')+\eta\log p^{\pi_k}(a'|\mathcal{F}',s') = \eta(\log\pi_{k+1}(a'|s')+\log Z^{\pi_k}(s')) \tag{33}$$

we have an upper bound between $\mathcal{T}^{\star}Q^{\pi_k}$ and $\mathcal{T}^{\pi_{k+1},p^{\pi_k}}Q^{\pi_k}$ for $\forall k$

$$\begin{aligned}
&\mathcal{T}^{\star}Q^{\pi_k}(s,a)-\mathcal{T}^{\pi_{k+1},p^{\pi_k}}Q^{\pi_k}(s,a) \\
&\leq r(s,a)+\gamma\mathbb{E}_{s',a'\sim\pi^{\star},\mathcal{F}'}\left[Q^{\pi_k}(s',a')+\eta(\log p^{\pi_k}(a'|\mathcal{F}',s')-\log\pi_k(a'|s'))\right]-\left(r(s,a)+\gamma\mathbb{E}_{s'}\left[\eta\log Z^{\pi_k}(s')\right]\right) \\
&= \gamma\mathbb{E}_{s',a'\sim\pi^{\star},\mathcal{F}'}\left[\eta(\log\pi_{k+1}(a'|s')+\log Z^{\pi_k}(s'))-\eta\log\pi_k(a'|s')\right]-\gamma\mathbb{E}_{s'}\left[\eta\log Z^{\pi_k}(s')\right] \\
&= \gamma\mathbb{E}_{s',a'\sim\pi^{\star}}\left[\eta(\log\pi_{k+1}(a'|s')-\log\pi_k(a'|s'))\right]
\end{aligned} \tag{34}$$

Therefore, for an integer $n\geq 1$

$$\begin{aligned}
\frac{1}{n}\sum_{k=0}^{n-1}\mathcal{T}^{\star}Q^{\pi_k}(s,a)-\mathcal{T}^{\pi_{k+1},p^{\pi_k}}Q^{\pi_k}(s,a) &\leq \frac{\eta\gamma}{n}\sum_{k=0}^{n-1}\mathbb{E}_{s',a'\sim\pi^{\star}}\left[\log\pi_{k+1}(a'|s')-\log\pi_k(a'|s')\right] \\
&= \frac{\eta\gamma}{n}\mathbb{E}_{s',a'\sim\pi^{\star}}\left[\log\frac{\pi_n(a'|s')}{\pi_0(a'|s')}\right] \\
&\leq \frac{\eta\gamma}{n}\mathbb{E}_{s',a'\sim\pi^{\star}}\left[\log\frac{\pi^{\star}(a'|s')}{\pi_0(a'|s')}\right]
\end{aligned} \tag{35}$$

where the last inequality is from that cross entropy is always greater than the entropy i.e. $H(\pi^{\star},\pi_n)\geq H(\pi^{\star}),\forall n$, due to non-negativity of KL divergence, and reverse the sign, it follows. By assumption of $H(\pi^{\star})$ being bounded and π_0 non-zero everywhere, the upper bound approaches to zero as $n\to\infty$.

In the next step, we will find a lower bound on the Equation (35). By Corollary D.2, we have

$$\begin{aligned}
\mathcal{T}^{\pi_{k+1},p^{\pi_k}}Q^{\pi_k}(s,a) &= r(s,a)+\gamma\mathbb{E}_{s',a'\sim\pi_{k+1},\mathcal{F}'}\left[Q^{\pi_k}(s',a')+\eta(\log p^{\pi_k}(a'|\mathcal{F}',s')-\log\pi_{k+1}(a'|s'))\right] \\
&\leq r(s,a)+\gamma\mathbb{E}_{s',a'\sim\pi_{k+1},\mathcal{F}'}\left[Q^{\pi_k}(s',a')+\eta(\log p^{\pi_{k+1}}(a'|\mathcal{F}',s')-\log\pi_{k+1}(a'|s'))\right] \\
&= \mathcal{T}^{\pi_{k+1}}Q^{\pi_k}(s,a)
\end{aligned} \tag{36}$$

Re-implementing the same justifications of Equation (30), we further have

$$\begin{aligned}
\mathcal{T}^{\pi_{k+1},p^{\pi_k}}Q^{\pi_k}(s,a) &\leq \mathcal{T}^{\pi_{k+1}}Q^{\pi_k}(s,a) \\
&\leq r(s,a)+\gamma\mathbb{E}_{s',a'\sim\pi_{k+1},\mathcal{F}'}\left[Q^{\pi_k}(s',a')+\eta(\log p^{\pi^{\star}}(a'|\mathcal{F}',s')-\log\pi^{\star}(a'|s'))\right]
\end{aligned} \tag{37}$$

Therefore

$$\begin{aligned}
&\mathcal{T}^{\star}Q^{\pi_k}(s,a)-\mathcal{T}^{\pi_{k+1},p^{\pi_k}}Q^{\pi_k}(s,a) \\
&\geq r(s,a)+\gamma\mathbb{E}_{s',a'\sim\pi^{\star},\mathcal{F}'}\left[Q^{\pi_k}(s',a')+\eta(\log p^{\pi^{\star}}(a'|\mathcal{F}',s')\log\pi^{\star}(a'|s'))\right]- \\
&\quad (r(s,a)+\gamma\mathbb{E}_{s',a'\sim\pi_{k+1},\mathcal{F}'}\left[Q^{\pi_k}(s',a')+\eta(\log p^{\pi^{\star}}(a'|\mathcal{F}',s')-\log\pi^{\star}(a'|s'))\right]) \\
&= \gamma\mathbb{E}_{s',(a',\mathcal{F}')\sim((p^{\pi^{\star}}(\mathcal{F}',a'|s')-p^{\pi_{k+1}}(\mathcal{F}',a'|s')))}\left[\eta(\log p^{\pi^{\star}}(a'|\mathcal{F}',s')-\log\pi^{\star}(a'|s'))+Q^{\pi_k}\right]^3
\end{aligned} \tag{38}$$

12

Summing together, we have

$$\frac{1}{n}\sum_{k=0}^{n-1}\mathcal{T}^{\star}Q^{\pi_k}(s,a)-\mathcal{T}^{\pi_{k+1},p^{\pi_k}}Q^{\pi_k}(s,a) \tag{39}$$

$$\geq \gamma\frac{1}{n}\sum_{k=0}^{n-1}\mathbb{E}_{s',(a',\mathcal{F}')\sim(p^{\pi^\star}(\mathcal{F}',a'|s')-p^{\pi_{k+1}}(\mathcal{F}',a'|s'))}\left[\eta(\log p^{\pi^\star}(a'|\mathcal{F}',s')-\log\pi^\star(a'|s'))+Q^{\pi_k}\right]$$

Since $\lim_{n\to\infty}\sum_{k=0}^{n}\mathbb{E}_{s',(a',\mathcal{F}')\sim((p^{\pi^\star}(\mathcal{F}',a'|s')-p^{\pi_{k+1}}(\mathcal{F}',a'|s')))}\left[\eta(\log p^{\pi^\star}(a'|\mathcal{F}',s')-\log\pi^\star(a'|s'))+Q^{\pi_k}\right]$
exists, the lower bound approaches to zero as $n\to\infty$.

Combining those two ends, we conclude that $\frac{1}{n}\sum_{k=0}^{n-1}\mathcal{T}^{\star}Q^{\pi_k}(s,a)-\mathcal{T}^{\pi_{k+1},p^{\pi_k}}Q^{\pi_k}(s,a)$ approaches

zero as $n\to\infty$, which implies $\lim_{n\to\infty}\sum_{k=0}^{n-1}\mathcal{T}^{\star}Q^{\pi_k}(s,a)-\mathcal{T}^{\pi_{k+1},p^{\pi_k}}Q^{\pi_k}(s,a)$ exists. It im-

mediately follows that $\lim_{n\to\infty}\left(\mathcal{T}^{\star}Q^{\pi_k}(s,a)-\mathcal{T}^{\pi_{k+1},p^{\pi_k}}Q^{\pi_k}(s,a)\right)=0$. It is also held for $\lim_{n\to\infty}\left(\mathcal{T}^{\star}Q^{\pi_k}(s,a)-\mathcal{T}^{\pi_{k+1}}Q^{\pi_k}(s,a)\right)=0$, since $\mathcal{T}^{\pi_{k+1}}$ is bounded below by $\mathcal{T}^{\pi_{k+1},p^{\pi_k}}$. We also note

$$\lim_{n\to\infty}\mathcal{T}^{\pi_{k+1}}Q^{\pi_k}(s,a)=\mathcal{T}^{\pi_\infty}Q^{\pi_\infty}(s,a)=Q^{\pi_\infty}(s,a) \tag{40}$$

And it follows that $\|\mathcal{T}^{\star}(Q^{\pi_k}-Q^{\pi_\infty})\|\leq\|\mathcal{T}^{\star}\|\|Q^{\pi_k}-Q^{\pi_\infty}\|$. Since $\mathcal{T}^{\pi^\star}$ is a bounded linear operator, and $Q^{\pi_k}\to Q^{\pi_\infty}$, it implies that $\lim_{n\to\infty}\mathcal{T}^{\star}Q^{\pi_k}(s,a)=\mathcal{T}^{\star}Q^{\pi_\infty}(s,a)$. Comparing those terms, we have $\mathcal{T}^{\star}Q^{\pi_\infty}(s,a)=Q^{\pi_\infty}(s,a)$. However, since $\mathcal{T}^{\pi^\star}$ has a unique fixed point, it implies that $Q^{\pi_\infty}(s,a)=Q^{\pi^\star}(s,a)$. $\qquad\square$

E Proof of Theorem 5.1

Proof. Since $\tilde{\pi}$ minimizes the projection loss, then it follows that

$$\mathbb{E}_{a_t\sim\tilde{\pi},\mathcal{F}_t}\left[\eta(\log\tilde{\pi}(a_t|s_t)-\log p^{\pi}(a_t|\mathcal{F}_t,s_t))-Q^{\pi}(s_t,a_t)+\eta\log Z^{\pi}(s_t)\right]$$
$$\leq\mathbb{E}_{a_t\sim\pi,\mathcal{F}_t}\left[\eta(\log\pi(a_t|s_t)-\log p^{\pi}(a_t|\mathcal{F}_t,s_t))-Q^{\pi}(s_t,a_t)+\eta\log Z^{\pi}(s_t)\right] \tag{41}$$

Since the partition function is dependent only on state and not relies on $\tilde{\pi}$, thus it can be canceled out from both sides. Rearranging, we have

$$V^{\pi}(s_t)\leq\mathbb{E}_{a_t\sim\tilde{\pi},\mathcal{F}_t}\left[Q^{\pi}(s_t,a_t)+\eta(\log p^{\pi}(a_t|\mathcal{F}_t,s_t)-\log\tilde{\pi}(a_t|s_t))\right] \tag{42}$$

Define $\mathcal{I}^{\tilde{\pi},p^{\pi}}(a,\mathcal{F}|s)$ as follows

$$\mathcal{I}^{\tilde{\pi},p^{\pi}}(a,\mathcal{F}|s)=\mathbb{E}_{\tilde{\pi}(a|s)p(\mathcal{F}|s,a)}\left[\log\frac{p^{\pi}(a|\mathcal{F},s)}{\tilde{\pi}(a|s)}\right] \tag{43}$$

By repeatedly applying above inequality, we have

$$Q^{\pi}(s_t,a_t)=r(s_t,a_t)+\gamma\mathbb{E}_{s_{t+1}}[V^{\pi}(s_{t+1})]$$
$$\leq r(s_t,a_t)+\gamma\mathbb{E}_{s_{t+1}}[\mathbb{E}_{a_{t+1}\sim\tilde{\pi},\mathcal{F}_{t+1}}[Q^{\pi}(s_{t+1},a_{t+1})+\eta(\log p^{\pi}(a_{t+1}|\mathcal{F}_{t+1},s_{t+1})-\log\tilde{\pi}(a_{t+1}|s_{t+1}))]$$
$$=r(s_t,a_t)+\gamma\eta\mathbb{E}_{s_{t+1}}\left[\mathcal{I}^{\tilde{\pi},p^{\pi}}(a_{t+1},\mathcal{F}_{t+1}|s_{t+1})\right]+\gamma\mathbb{E}_{s_{t+1}}\left[\mathbb{E}_{a_{t+1}\sim\tilde{\pi},,\mathcal{F}_{t+1}}\left[r(s_{t+1},a_{t+1})+\gamma\mathbb{E}_{s_{t+2}}[V^{\pi}(s_{t+2})]\right]\right]$$
$$\vdots$$
$$\leq\mathbb{E}_{s_{t+1},a_{t+1},\dots|\tilde{\pi}}\left[\sum_{l=0}^{\infty}\gamma^l(r(s_{t+l},a_{t+l})+\gamma\alpha\eta\mathbb{E}_{s_{t+l+1}}\left[\mathcal{I}^{\tilde{\pi},p^{\pi}}(a_{t+l+1},\mathcal{F}_{t+l+1}|s_{t+l+1})\right])\right]$$
$$\triangleq Q^{\tilde{\pi},p^{\pi}}(s_t,a_t) \tag{44}$$

By Corollary D.2, we have

$$\mathcal{I}^{\tilde{\pi},p^{\pi}}(a,\mathcal{F}|s)\leq\mathcal{I}^{\tilde{\pi}}(a,\mathcal{F}|s) \tag{45}$$

Therefore

$$Q^{\pi}(s_t,a_t)\leq Q^{\tilde{\pi},p^{\pi}}(s_t,a_t)\leq Q^{\tilde{\pi}}(s_t,a_t) \tag{46}$$
$$\square$$

---

[3]We merge $(a',\mathcal{F}')$ together for the reason that $p(\mathcal{F}'|s',a')$ may have a complex dependency on π as $\mathcal{F}$ becomes longer.

F Derivation of Lower Bounds

F.1 One-step Lower Bound

We will present a more general lower bound considering future sequence $\mathcal{F}$ by using importance sampling and Jensen's inequality

$$
\begin{aligned}
\log p^\pi(\mathcal{F}|s) &= \log \int_a p^\pi(\mathcal{F}, a|s)\,da \\
&= \log \mathbb{E}_{a\sim q_\phi(a|\mathcal{F},s)}\left[\frac{p^\pi(\mathcal{F},a|s)}{q_\phi(a|\mathcal{F},s)}\right] \\
&\geq \mathbb{E}_{a\sim q_\phi(a|\mathcal{F},s)}\left[\log \frac{p^\pi(\mathcal{F},a|s)}{q_\phi(a|\mathcal{F},s)}\right] \\
&= \mathbb{E}_{q_\phi(a|\mathcal{F},s)}\left[\log p_\psi(\mathcal{F}|s,a)\right] - D_{\mathrm{KL}}(q_\phi(a|\mathcal{F},s)||\pi(a|s))
\end{aligned}
\tag{47}
$$

When $\mathcal{F} = (s', r)$, we can get the one-step variational lower bound

$$
\begin{aligned}
\log p^\pi(s',r|s) &\geq \mathcal{L}(\phi,\psi; s, s', r) \\
&= -D_{\mathrm{KL}}(q_\phi(a|s,s',r)||\pi(a|s)) + \mathbb{E}_{q_\phi(a|s,s',r)}[\log p_\psi(s',r|s,a)]
\end{aligned}
\tag{48}
$$

F.2 Trajectory-Wise Lower Bound

Denote $\mathbf{x}_{1:T} = (s_{t+1}, r_t|s_t)_{t=1}^T$ and $\mathbf{z}_{1:T} = (a_t|s_t)_{t=1}^T$, we assume the joint distribution $p^\pi(\mathbf{x}_{1:T}, \mathbf{z}_{1:T})$ and $q_\phi(\mathbf{z}_{1:T}|\mathbf{x}_{1:T})$ can be factorized as follows

$$
\begin{aligned}
p^\pi(\mathbf{x}_{1:T}, \mathbf{z}_{1:T}) &= \prod_{\tau=1}^T p_\psi(s'_\tau, r_\tau|s_\tau, a_\tau) p^\pi(a_\tau|s_{1:\tau}, a_{1:\tau-1}) \\
&= \prod_{\tau=1}^T p_\psi(s'_\tau, r_\tau|s_\tau, a_\tau)\pi(a_\tau|s_\tau) \qquad \triangleright \text{ by Markov property}
\end{aligned}
\tag{49}
$$

$$
q_\phi(\mathbf{z}_{1:T}|\mathbf{x}_{1:T}) = \prod_{\tau=1}^T q_\phi(a_\tau|s_{1:\tau+1}, r_{1:\tau}, a_{1:\tau-1})
\tag{50}
$$

In a similar fashion

$$
\begin{aligned}
\log p^\pi(\mathbf{x}_{1:T}) &= \int_{\mathbf{z}_{1:T}} p(\mathbf{x}_{1:T}, \mathbf{z}_{1:T})\,d\mathbf{z}_{1:T} \\
&= \log \mathbb{E}_{\mathbf{z}_{1:T}\sim q_\phi(\mathbf{z}_{1:T}|\mathbf{x}_{1:T})}\left[\frac{p(\mathbf{x}_{1:T}, \mathbf{z}_{1:T})}{q_\phi(\mathbf{z}_{1:T}|\mathbf{x}_{1:T})}\right] \\
&\geq \mathbb{E}_{\mathbf{z}_{1:T}}\left[\log \frac{p(\mathbf{x}_{1:T}, \mathbf{z}_{1:T})}{q_\phi(\mathbf{z}_{1:T}|\mathbf{x}_{1:T})}\right] \\
&= \mathbb{E}_{\mathbf{z}_{1:T}}\left[\log \frac{\prod_{\tau=1}^T p_\psi(s'_\tau, r_\tau|s_\tau, a_\tau)\pi(a_\tau|s_\tau)}{\prod_{\tau=1}^T q_\phi(a_\tau|s_{1:\tau+1}, r_{1:\tau}, a_{1:\tau-1})}\right] \\
&= \mathbb{E}_{\mathbf{z}_{1:T}}\left[\sum_{\tau=1}^T \log p_\psi(s'_\tau, r_\tau|s_\tau, a_\tau) + \log \pi(a_\tau|s_\tau) - \log q_\phi(a_\tau|s_{1:\tau+1}, r_{1:\tau}, a_{1:\tau-1})\right] \\
&= \sum_{\tau=1}^T \mathbb{E}_{\mathbf{z}_{1:\tau}}\left[\log p_\psi(s'_\tau, r_\tau|s_\tau, a_\tau) - D_{KL}(q_\phi(a_\tau|s_{1:\tau+1}, r_{1:\tau}, a_{1:\tau-1})\|\pi(a_\tau|s_\tau))\right]
\end{aligned}
\tag{51}
$$

If we further assume conditional independence of the past information i.e. $t < \tau$ for q_ϕ, then we have $q_\phi(a_\tau|s_{1:\tau+1}, r_{1:\tau}, a_{1:\tau-1}) = q_\phi(a_\tau|s_{\tau+1}, r_\tau, s_\tau)$. The above formulation then deduces to

$$
\sum_{\tau=1}^T \mathbb{E}_{\mathbf{z}_{1:\tau}}\left[\log p_\psi(s'_\tau, r_\tau|s_\tau, a_\tau) - D_{KL}(q_\phi(a_\tau|s_{\tau+1}, r_\tau, s_\tau)\|\pi(a_\tau|s_\tau))\right]
\tag{52}
$$

which is simply a summation of a series of one-step lower bounds as derived earlier. This is helpful since we can employ the same model architecture while explore different training procedures, such as being more on-policy to capture trajectory's information.