A Variational Formulation of Reinforcement Learning in Infinite-Horizon Markov Decision Processes

Anonymous Author(s)
Affiliation
Address
email

Abstract

Reinforcement learning in infinite-horizon Markov decision processes (MDPs) is typically framed as expected discounted return maximization. In this paper, we formulate an alternative principle for optimal sequential decision-making in infinite-horizon MDPs: variational Bayesian inference in transdimensional probabilistic models. In particular, we specify a probabilistic model of random size and consider the variational problem of finding an approximation to the posterior distribution over state–action trajectories conditioned on state–action trajectories that reflect a desired behavior. We derive a tractable variational objective for infinite-horizon settings, prove a variational dynamic-discount policy iteration theorem, show that fixed discount factor KL-regularized reinforcement learning objectives are special cases of dynamic-discount variational objectives, and prove that learning dynamic discount factors is optimal.

1 Introduction

We provide a Bayesian framework for deriving behavior-driven optimal decision rules for sequential decision problems. In particular, we provide a mathematical justification for learned, dynamic discount factors in KL-regularized reinforcement learning, which have been proposed as an empirically useful tool in recently developed reinforcement learning algorithms [6, 8, 4], and establish a rigorous foundation for framing modern reinforcement learning methods as probabilistic inference. Although control as inference has gained in popularity, the treatment of infinite-horizon settings in previous works is ad-hoc and not probabilistically well-motivated. With this work, we hope to address this shortcoming and provide a clear formulation of control as inference that carefully disambiguates modeling and inference assumptions.

Levine [3] and Haarnoja et al. [1] presented a framework for framing maximum-entropy reinforcement learning as Bayesian inference in probabilistic models over finite-horizon state–action trajectories. However, most modern reinforcement learning problems are not formulated as finite but as *infinite-horizon* problems [5, 9]. To apply their probabilistic probabilistic formulation of reinforcement learning to infinite-horizon problems, Levine [3] and Haarnoja et al. [1] introduce a fixed discount-factor into their formulation post-hoc and without providing a probabilistic justification for doing so. In this paper, we show that including a (fixed) discount factor as proposed by Levine [3] and Haarnoja et al. [1] is a special case of a more general probabilistic framing of the problem, leads to a variational formulation with a loose evidence lower bound, and can provably be improved upon by framing Bayesian variational inference in infinite-horizon MDPs as variational inference in a *transdimensional* probabilistic model.

To derive a learning algorithm that allows us to infer a policy that reflects the behavior encoded in desired state trajectories, we frame the problem of finding an optimal policy as computing an approximation to the conditional distribution over state–action trajectories given state–action

Submitted to the ICML 2024 Workshop on Foundations of Reinforcement Learning and Control. Do not distribute.

trajectories that reflect a desired behavior. We formulate a corresponding probabilistic model and derive tractable variational objectives for finite- and infinite-horizon settings. Based on these results, we define a novel Bellman backup operator and show that for tabular settings, the repeated application of the operator converges to an optimal policy and an optimal dynamic discount factor. Building on this result, we show that fixed discount factor KL-regularized reinforcement learning objectives are special cases of the dynamic-discount objectives derived here and demonstrate that variationally learned, dynamic discount factors are optimal in KL-regularized RL.

2 Preliminaries

Standard reinforcement learning (RL) addresses reward maximization in a Markov decision process (MDP) defined by the tuple $(\mathcal{S}, \mathcal{A}, p_{\mathbf{S}_0}, p_d, r, \gamma)$ [10, 11], where $\mathcal{S}$ and $\mathcal{A}$ denote the state and action space, respectively, p_0 denotes the initial state distribution, p_d is a state transition distribution, r is an immediate reward function, and γ is a discount factor. To sample trajectories, an initial state is sampled according to $p_{\mathbf{S}_0}$, and successive states are sampled from the state transition distribution $\mathbf{S}_{t+1} \sim p_d(\cdot \,|\, \mathbf{s}_t, \mathbf{a}_t)$ and actions from a policy $\mathbf{A}_t \sim \pi(\cdot \,|\, \mathbf{s}_t)$. We will write $\boldsymbol{\mathcal{T}}_{0:t} = \{\mathbf{S}_0, \mathbf{A}_0, \mathbf{S}_1, \ldots, \mathbf{S}_t, \mathbf{A}_t\}$ to represent a finite-horizon and $\boldsymbol{\mathcal{T}}_0 \doteq \{\mathbf{S}_t, \mathbf{A}_t\}_{t=0}^{\infty}$ to represent an infinite-horizon stochastic state–action trajectory, and write $\boldsymbol{\tau}_{0:t} = \{\mathbf{s}_0, \mathbf{a}_0, \mathbf{s}_1, \ldots, \mathbf{s}_t, \mathbf{a}_t\}$ and $\boldsymbol{\tau}_0 \doteq \{\mathbf{s}_t, \mathbf{a}_t\}_{t=0}^{\infty}$ for the respective trajectory realizations. Given a reward function $r : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ and discount factor $\gamma \in (0, 1)$, the objective in reinforcement learning is to find a policy π that maximizes the returns, defined as $\mathbb{E}_{p_\pi}\left[\sum_{t=0}^{\infty} \gamma^t r(\mathbf{s}_t, \mathbf{a}_t)\right]$, where p_π denotes the distribution of states induced by a policy π.

3 A Variational Formulation of RL in Infinite-Horizon MDPs

Desired behaviors for artificial agents are often abstract and hard to encode into reward functions. However, in practice, it is often easy to represent desired behaviors via demonstrations. Such demonstrations can be thought of as sample state–action trajectories from a distribution over optimal state–action trajectories. In the remainder of this paper, we will demonstrate how to use variational Bayesian inference to infer an optimal policy from a set of optimal state–action demonstrations.

To start the exposition, we note that for every state in the environment, there exists a desired, or optimal, behavior that an agent *could* take. We denote this optimal behavior for any given state as the set of state–action trajectories by $\boldsymbol{\tau}^{\Omega}$. Throughout, we will use the index Ω to denote optimality. Hence, for any state $\mathbf{s} \in \mathcal{S}$, assuming the MDP is ergodic and transition dynamics are deterministic, there exists a set of actions that will set an agent on an optimal state–action trajectory, that is, $\mathcal{A}^{\Omega} \doteq \{\mathbf{a} \in \mathcal{A} \,|\, \mathbf{s}' \sim p_d(\mathbf{s}' \,|\, \mathbf{s}, \mathbf{a}) : \mathbf{s}' \in \boldsymbol{\tau}^{\Omega}\}$, meaning there exists a set of actions that will set an agent on the optimal state–action trajectory with probability one.

If the transition dynamics are stochastic, each state–action pair will have some probability less than one of transitioning the agent onto an optimal state–action trajectory. Denoting the event of a state being in the optimal state–action trajectory by $\mathbf{s} \in \mathcal{S}^{\Omega}$, where $\mathcal{S}^{\Omega} \doteq \{\mathbf{s} \in \boldsymbol{\tau}^{\Omega}\}$, we can define a random variable $\xi(\mathcal{S}^{\Omega}) \doteq \mathbb{I}\{\mathbf{s}' \in \mathcal{S}^{\Omega}\}$. We then have that $\xi = 1$ if the state $\mathbf{s}'$ into which an agent transitioned after taking action $\mathbf{a}$ in state $\mathbf{s}$ is in the optimal trajectory and $\xi(\mathcal{S}^{\Omega}) = 0$ otherwise. The probability of transitioning into a state on the optimal state–action trajectory at time step $t+1$ is then given by

$$\mathbb{P}(\xi_{t+1}(\mathcal{S}^{\Omega}) = 1 \,|\, \mathbf{s}_t, \mathbf{a}_t) = \int_{\mathcal{S}^{\Omega}} p_d(\mathbf{s}_{t+1} \,|\, \mathbf{s}_t, \mathbf{a}_t) d\mathbf{s}_{t+1} = \int_{\mathcal{S}} p_d(\mathbf{s}_{t+1} \,|\, \mathbf{s}_t, \mathbf{a}_t)\, \mathbb{I}\{\mathbf{s}_{t+1} \in \mathcal{S}^{\Omega}\} d\mathbf{s}_{t+1}. \quad (1)$$

In other words, the probability of transitioning into a state on the optimal state–action trajectory corresponds to marginalization over the set of optimal states $\mathcal{S}^{\Omega}$. Equation (1) is a likelihood function.

Similarly, by the Markov property, the joint probability of transitioning into a state on the optimal state–action trajectory and staying on it from time step 1 to time step $t^\star \doteq t+1$, given a state–action trajectory, factorizes and is given by

$$\begin{aligned} &\mathbb{P}(\xi_{1:t+1}(\mathcal{S}^{\Omega}) = 1 \,|\, \mathbf{s}_0, \mathbf{a}_0, ..., \mathbf{s}_t, \mathbf{a}_t) \\ &= \prod_{t'=0}^{t} \int_{\mathcal{S}^{\Omega}} p_d(\mathbf{s}_{t'+1} \,|\, \mathbf{s}_{t'}, \mathbf{a}_{t'})\, d\mathbf{s}_{t'+1} = \prod_{t'=0}^{t} \int_{\mathcal{S}} p_d(\mathbf{s}_{t'+1} \,|\, \mathbf{s}_{t'}, \mathbf{a}_{t'})\, \mathbb{I}\{\mathbf{s}_{t'+1} \in \mathcal{S}^{\Omega}\}\, d\mathbf{s}_{t'+1}, \end{aligned} \quad (2)$$

where we start from $t' = 0$ without loss of generality.

3.1 Warm-Up: Finite-Horizon Reinforcement Learning as Variational Inference

First, we consider the finite-horizon setting. This formulation only diverges slightly from prior work but will help us transition to the transdimensional model formulation for the infinite-horizon setting.

With the notion of trajectory-dependent optimality described in the previous section, we can now specify a model over finite-horizon state–action trajectories and $\xi_{1:t+1}(\mathcal{S}^\Omega)$,

$$p(\boldsymbol{\tau}_{0:t}, \xi^\star_{1:t+1}(\mathcal{S}^\Omega)) \doteq p_{\mathbf{S}_0}(\mathbf{s}_0) \prod_{t'=0}^{t} \mathbb{P}(\xi_{t'+1}(\mathcal{S}^\Omega) = 1 \,|\, \mathbf{s}_{t'}, \mathbf{a}_{t'}) p(\mathbf{a}_t \,|\, \mathbf{s}_t) \prod_{t'=0}^{t-1} p_d(\mathbf{s}_{t'+1} \,|\, \mathbf{s}_{t'}, \mathbf{a}_{t'}) p(\mathbf{a}_{t'} \,|\, \mathbf{s}_{t'}),$$

where $\tilde{\boldsymbol{\mathcal{T}}}_{0:t}$ is a state–action trajectory starting at state $\mathbf{S}_0$ and ending at state $\mathbf{S}_t$, $p(\mathbf{a}_t \,|\, \mathbf{s}_t)$ is a conditional action prior, $p_d(\mathbf{s}_{t+1} \,|\, \mathbf{s}_t, \mathbf{a}_t)$ is the environment's state transition distribution, and $\xi^\star_{1:t+1}(\mathcal{S}^\Omega) \doteq \{\xi_{t'}(\mathcal{S}^\Omega) = 1\}_{t'=1}^{t+1}$ is the set of events corresponding to transitioning onto an optimal trajectory. By extension, the probability of transitioning onto an optimal state–action trajectory and remaining on it for $t^\star$ time steps given a state and a prior policy is given by the marginal likelihood

$$\mathbb{P}(\xi_{1:t+1}(\mathcal{S}^\Omega) = 1 \,|\, \mathbf{s}_0) = \iint_{\mathcal{A}^{t+1}\mathcal{S}^t} p_{\tilde{\boldsymbol{\mathcal{T}}}_{0:t}|\mathbf{S}_0}(\tilde{\boldsymbol{\tau}}_{0:t} \,|\, \mathbf{s}_0) \left(\prod_{t'=0}^{t} \int_{\mathcal{S}^\Omega} p_d(\mathbf{s}_{t'+1} \,|\, \mathbf{s}_{t'}, \mathbf{a}_{t'}) \, d\mathbf{s}_{t'+1} \right) d\mathbf{s}_{1:t} \, d\mathbf{a}_{0:t} \tag{3}$$

$$\text{where} \qquad p_{\tilde{\boldsymbol{\mathcal{T}}}_{0:t}|\mathbf{S}_0}(\tilde{\boldsymbol{\tau}}_{0:t} \,|\, \mathbf{s}_0) \doteq p(\mathbf{a}_t \,|\, \mathbf{s}_t) \prod_{t'=0}^{t-1} p_d(\mathbf{s}_{t'+1} \,|\, \mathbf{s}_{t'}, \mathbf{a}_{t'}) \, p(\mathbf{a}_{t'} \,|\, \mathbf{s}_{t'}) \tag{4}$$

is a prior distribution over state–action trajectories. Using an indicator function $\mathbb{I}\{\mathbf{s}_{t+1} \in \mathcal{S}^\Omega\}$ denoting whether the next state is on the desired state–action trajectory, the marginal likelihood in Equation (3) can equivalently be expressed as

$$\begin{aligned} &\mathbb{P}(\xi_{1:t+1}(\mathcal{S}^\Omega) = 1 \,|\, \mathbf{s}_0) \\ &= \iint_{\mathcal{A}^{t+1}\mathcal{S}^t} p_{\tilde{\boldsymbol{\mathcal{T}}}_{0:t}|\mathbf{S}_0}(\tilde{\boldsymbol{\tau}}_{0:t} \,|\, \mathbf{s}_0) \left(\prod_{t'=0}^{t} \int_{\mathcal{S}} p_d(\mathbf{s}_{t'+1} \,|\, \mathbf{s}_{t'}, \mathbf{a}_{t'}) \, \mathbb{I}\{\mathbf{s}_{t'+1} \in \mathcal{S}^\Omega\} \, d\mathbf{s}_{t'+1} \right) d\mathbf{s}_{1:t} \, d\mathbf{a}_{0:t}. \end{aligned} \tag{5}$$

This marginalization establishes the connection between the full joint distribution in Equation (3) and the likelihood of remaining on an optimal state–action trajectory under a state–action trajectory prior and the likelihood function defined in Equation (1).

$\mathbb{P}(\xi_{1:t+1}(\mathcal{S}^\Omega) = 1 \,|\, \mathbf{s}_0)$ is the marginal likelihood of remaining on the optimal state trajectory from time step 1 to time step $t+1$ under the prior policy $p(\mathbf{a}_t \,|\, \mathbf{s}_t)$ and the dynamics model $p_d(\mathbf{s}_{t+1} \,|\, \mathbf{s}_t, \mathbf{a}_t)$. Using Bayes' Theorem, we could use the marginal likelihood to compute the posterior distribution over state–action trajectories, $p_{\tilde{\boldsymbol{\mathcal{T}}}_{0:t}|\xi_{1:t+1}}(\cdot \,|\, \xi^\star_{1:t+1}(\mathcal{S}^\Omega))$. Unfortunately, the marginal likelihood in Equation (5) is intractable for all but the simplest probabilistic models.

To infer an approximate posterior distribution over state–action trajectories instead, we express posterior inference as the variational minimization problem

$$\min_{q_{\tilde{\boldsymbol{\mathcal{T}}}_{0:t}} \in \hat{\mathcal{Q}}} \mathbb{D}_{\mathrm{KL}}(q_{\tilde{\boldsymbol{\mathcal{T}}}_{0:t}}(\cdot) \parallel p_{\tilde{\boldsymbol{\mathcal{T}}}_{0:t}|\xi_{1:t+1}}(\cdot \,|\, \xi^\star_{1:t+1}(\mathcal{S}^\Omega))), \tag{6}$$

where $\mathbb{D}_{\mathrm{KL}}(\cdot \parallel \cdot)$ is the KL divergence, and $\hat{\mathcal{Q}}$ denotes the variational family over which to optimize. We consider a family of distributions parameterized by a policy π and defined by

$$q_{\tilde{\boldsymbol{\mathcal{T}}}_{0:t}}(\tilde{\boldsymbol{\tau}}_{0:t}) \doteq p_{\mathbf{S}_0}(\mathbf{s}_0) \pi(\mathbf{a}_t \,|\, \mathbf{s}_t) \prod\nolimits_{t'=0}^{t-1} p_d(\mathbf{s}_{t'+1} \,|\, \mathbf{s}_{t'}, \mathbf{a}_{t'}) \pi(\mathbf{a}_{t'} \,|\, \mathbf{s}_{t'}), \tag{7}$$

where $\pi \in \Pi$, a family of policy distributions, and where $p_{\mathbf{S}_0}(\mathbf{s}_0) \prod_{t'=0}^{t-1} p_d(\mathbf{s}_{t'+1} \,|\, \mathbf{s}_{t'}, \mathbf{a}_{t'})$ is the true state transition distribution up to and including the state transition at t. In Proposition 1 (Fixed-Time Variational Objective), we show that under this variational family, the inference problem in Equation (6) can be equivalently stated as the problem of maximizing an entropy-regularized expected reward function at every time step, where the reward function is given by the log-likelihood of transitioning onto an optimal state–action trajectory given a state–action pair. This is effectively the result obtained by Ziebart et al. [13], Levine [3], and Haarnoja et al. [1].

3.2 Infinite-Horizon Reinforcement Learning as Variational Bayesian Inference

To derive an infinite-horizon objective, we modify the probabilistic model used above. To represent the possibility that an agent may stay on the optimal state trajectory for *any* number of time steps, that is, for state–action trajectories of varying lengths, we treat the length of the trajectory itself as a random variable, T, and define the model

$$p(\tilde{\boldsymbol{\tau}}_{0:t}, \xi^\star_{1:t+1}(\mathcal{S}^\Omega), t) \doteq p_T(t) p_{\mathbf{S}_0}(\mathbf{s}_0) \prod_{t'=0}^{t} \mathbb{P}(\xi_{t'+1}(\mathcal{S}^\Omega) = 1 \mid \mathbf{s}_{t'}, \mathbf{a}_{t'}) p_d(\mathbf{s}_{t'+1} \mid \mathbf{s}_{t'}, \mathbf{a}_{t'}) p(\mathbf{a}_{t'} \mid \mathbf{s}_{t'}),$$

where $p_T(t)$ is the probability of remaining on the optimal state trajectory for $t+1$ time steps. Since the trajectory length is itself a random variable, the joint distribution is a *transdimensional* distribution defined on $\biguplus_{t=0}^{\infty} \{t\} \times \mathcal{S}^t \times \mathcal{A}^t$ [2].

Unlike in the fixed-horizon setting, the variational Bayesian inference problem in the infinite-horizon setting corresponds to finding the posterior distribution over both state–action trajectories *and* the length of the optimal state trajectory T conditioned on the desired behavior $\xi^\star_{1:t+1}(\mathcal{S}^\Omega)$. Analogously to the steps above, we can express this inference problem variationally as

$$\min_{q_{\tilde{\boldsymbol{\tau}}_{0:T},T} \in \mathcal{Q}} \mathbb{D}_{\mathrm{KL}}(q_{\tilde{\boldsymbol{\tau}}_{0:T},T}(\cdot) \parallel p_{\tilde{\boldsymbol{\tau}}_{0:T},T \mid \xi_{1:T+1}}(\cdot \mid \xi^\star_{1:t+1}(\mathcal{S}^\Omega))), \tag{8}$$

where t denotes the time step immediately *before* the outcome is achieved, $\mathcal{Q}$ denotes the variational family. Under this variational distribution, we can obtain an unfactorized variational objective that does in general not lend itself to stochastic gradient-based optimization (and off-policy reinforcement learning). The variational objective is given in Proposition 3, but we omit it here for brevity.

To obtain a variational objective amenable to stochastic variational inference and off-policy reinforcement learning, we define the variational family as follows: $q_{\tilde{\boldsymbol{\tau}}_{0:T},T}(\tilde{\boldsymbol{\tau}}_{0:t}, t) = q_{\tilde{\boldsymbol{\tau}}_{0:T}|T}(\tilde{\boldsymbol{\tau}}_{0:t} \mid t) q_T(t)$, where q_T is a distribution over T in some variational family $\mathcal{Q}_T$ parameterized by

$$q_T(t) = q_{\Delta_{t+1}}(\Delta_{t+1} = 1) \prod_{t'=1}^{t} q_{\Delta_{t'}}(\Delta_{t'} = 0), \tag{9}$$

with Bernoulli random variables Δ_t denoting the event of "remaining on the optimal state trajectory from time step 1 to time step $t+1$," we can equivalently express the variational problem in Equation (8) recursively in a way that is tractable and amenable to off-policy optimization:

Theorem 1 (Dynamic-Discount Behavior-Driven RL as Variational Inference)**.** *Let $q_T(t)$ and $q_{\tilde{\boldsymbol{\tau}}_{0:t}|T}(\tilde{\boldsymbol{\tau}}_{0:t} \mid t, \mathbf{s}_0)$ be as defined in Equation (7) and Equation (9), and define a behavior-driven state value function,*

$$V^\pi(\mathbf{s}_t, \mathcal{S}^\Omega; q_T) \doteq \mathbb{E}_{\pi(\mathbf{a}_t \mid \mathbf{s}_t)} \left[Q^\pi(\mathbf{s}_t, \mathbf{a}_t, \mathcal{S}^\Omega; q_T) \right] - \mathbb{D}_{\mathrm{KL}}(\pi(\cdot \mid \mathbf{s}_t) \parallel p(\cdot \mid \mathbf{s}_t)), \tag{10}$$

a behavior-driven state–action value function

$$Q^\pi(\mathbf{s}_t, \mathbf{a}_t, \mathcal{S}^\Omega; q_T) \doteq r(\mathbf{s}_t, \mathbf{a}_t, \mathcal{S}^\Omega; q_\Delta) + q(\Delta_{t+1} = 0) \, \mathbb{E}_{p_d(\mathbf{s}_{t+1} \mid \mathbf{s}_t, \mathbf{a}_t)} \left[V^\pi(\mathbf{s}_{t+1}, \mathcal{S}^\Omega; \pi, q_T) \right], \tag{11}$$

and a behavior-driven reward-like function

$$r(\mathbf{s}_t, \mathbf{a}_t, \mathcal{S}^\Omega; q_\Delta) \doteq \log \mathbb{P}(\xi_{t+1}(\mathcal{S}^\Omega) = 1 \mid \mathbf{s}_t, \mathbf{a}_t) - q_{\Delta_{t+1}}(\Delta_{t+1} = 1) \mathbb{D}_{\mathrm{KL}}(q_{\Delta_{t+1}} \parallel p_{\Delta_{t+1}}). \tag{12}$$

Then given an optimal state trajectory $\mathcal{S}^\Omega$,

$$\mathbb{D}_{\mathrm{KL}}(q_{\tilde{\boldsymbol{\tau}}_{0:T},T}(\cdot) \parallel p_{\tilde{\boldsymbol{\tau}}_{0:T},T \mid \xi_{1:t+1}(\mathcal{S}^\Omega)}(\cdot \mid \xi^\star_{1:t+1}(\mathcal{S}^\Omega))) = -\mathcal{F}(\pi, q_T, \mathcal{S}^\Omega) + C = -V^\pi(\mathbf{s}_0, \mathcal{S}^\Omega; q_T) + C,$$

where $C \doteq \log p(\xi^\star_{1:T+1})$ is independent of π and q_T, and hence maximizing $V^\pi(\mathbf{s}_0, \mathcal{S}^\Omega; \pi, q_T)$ is equivalent to minimizing Equation (8) and hence, the following holds:

$$\operatorname*{arg\,min}_{\pi \in \Pi, q_T \in \mathcal{Q}_T} \mathbb{D}_{\mathrm{KL}}(q_{\tilde{\boldsymbol{\tau}}_{0:T},T}(\cdot) \parallel p_{\tilde{\boldsymbol{\tau}}_{0:T},T \mid \xi_{1:T+1}}(\cdot \mid \xi^\star_{1:T+1})) = \operatorname*{arg\,max}_{\pi \in \Pi, q_T \in \mathcal{Q}_T} V^\pi(\mathbf{s}_0, \mathcal{S}^\Omega; q_T).$$

Theorem 1 tells us that the solution to the variational problem we started out with (Equation (8)), is in fact the solution to an infinite-horizon reinforcement learning problem with a reward function determined by the likelihood of transitioning onto an optimal trajectory, a learned, dynamic discount factor, and KL divergence regularization. In Appendix A, we prove that dynamic-discount factor RL is optimal and preferred over fixed discount factors. For detailed proofs, see the appendix.

4 Conclusion

Using a variational framing of the inference problem, we showed that optimized, dynamic discount factors are optimal in KL-regularized RL and that fixed discount factor methods are a special (less optimal) case of this formulation. We hope that this work contributes to bridging the gap between reinforcement learning and probabilistic inference research and helps establish a mutual reference point from which to derive new insights and methods.

References

[1] Tuomas Haarnoja, Aurick Zhou, Pieter Abbeel, and Sergey Levine. Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. In *International Conference on Machine Learning*, 2018.

[2] Matthew Hoffman, Nando Freitas, Arnaud Doucet, and Jan Peters. An expectation maximization algorithm for continuous markov decision processes with arbitrary reward. In *Artificial intelligence and statistics*, pages 232–239, 2009.

[3] Sergey Levine. Reinforcement learning and control as probabilistic inference: Tutorial and review. 2018.

[4] Cong Lu, Philip J. Ball, Tim G. J. Rudner, Jack Parker-Holder, Michael A. Osborne, and Yee Whye Teh. Challenges and Opportunities in Offline Reinforcement Learning from Visual Observations. 2022.

[5] Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Alex Graves, Ioannis Antonoglou, Daan Wierstra, and Martin Riedmiller. Playing atari with deep reinforcement learning. In *NIPS Workshop on Deep Learning*, pages 1–9, 2013.

[6] Rafael Rafailov, Tianhe Yu, Aravind Rajeswaran, and Chelsea Finn. Offline reinforcement learning from images with latent space models. In Ali Jadbabaie, John Lygeros, George J. Pappas, Pablo A. Parrilo, Benjamin Recht, Claire J. Tomlin, and Melanie N. Zeilinger, editors, *Proceedings of the 3rd Conference on Learning for Dynamics and Control*, volume 144 of *Proceedings of Machine Learning Research*, pages 1154–1168. PMLR, 07 – 08 June 2021.

[7] Tim G. J. Rudner, Cong Lu, Michael A. Osborne, Yarin Gal, and Yee Whye Teh. On pathologies in KL-regularized reinforcement learning from expert demonstrations. In *Advances in Neural Information Processing Systems 34*. 2021.

[8] Tim G. J. Rudner, Vitchyr H. Pong, Rowan McAllister, Yarin Gal, and Sergey Levine. Outcome-driven reinforcement learning via variational inference. In *Advances in Neural Information Processing Systems 34*. 2021.

[9] David Silver, Aja Huang, Chris J. Maddison, Arthur Guez, Laurent Sifre, George van den Driessche, Julian Schrittwieser, Ioannis Antonoglou, Veda Panneershelvam, Marc Lanctot, Sander Dieleman, Dominik Grewe, John Nham, Nal Kalchbrenner, Ilya Sutskever, Timothy Lillicrap, Madeleine Leach, Koray Kavukcuoglu, Thore Graepel, and Demis Hassabis. Mastering the game of go with deep neural networks and tree search. *Nature*, 529(7587):484–489, jan 2016. ISSN 0028-0836.

[10] Richard S Sutton and Andrew G Barto. *Reinforcement Learning: An Introduction*. 1998.

[11] Csaba Szepesvári. *Algorithms for Reinforcement Learning*, volume 4. 2010.

[12] Martha White. Unifying task specification in reinforcement learning. In *International Conference on Machine Learning*, pages 3742–3750. PMLR, 2017.

[13] Brian D Ziebart, Andrew Maas, J Andrew Bagnell, and Anind K Dey. Maximum entropy inverse reinforcement learning. In *AAAI Conference on Artificial Intelligence*, pages 1433–1438, 2008.

Supplementary Material

Table of Contents

Appendix A Dynamic-Discount Behavior-Driven Reinforcement Learning

Building on Theorem 1, we will now define a dynamic-discount behavior-driven Bellman backup operator and use it to derive a policy iteration theorem for variational, dynamic-discount reinforcement learning. In particular, we define:

Definition 1 (Dynamic-Discount Behavior-Driven Bellman Backup Operator)**.** *Given a function $Q : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to \mathbb{R}$, define the operator $\mathcal{T}^\pi$ as*

$$\mathcal{T}^\pi Q(\mathbf{s}_t, \mathbf{a}_t, \mathcal{S}^\Omega; q_T) \doteq r(\mathbf{s}_t, \mathbf{a}_t, \mathcal{S}^\Omega; q_\Delta) + q_{\Delta_{t+1}}(\Delta_{t+1} = 0)\, \mathbb{E}_{p_d(\mathbf{s}_{t+1} \mid \mathbf{s}_t, \mathbf{a}_t)} \left[V(\mathbf{s}_{t+1}, \mathcal{S}^\Omega; q_T) \right], \quad \text{(A.1)}$$

where $r(\mathbf{s}_t, \mathbf{a}_t, \mathcal{S}^\Omega; q_\Delta)$ is from Theorem 1 (Dynamic-Discount Behavior-Driven RL as Variational Inference) and

$$V(\mathbf{s}_t, \mathcal{S}^\Omega; q_T) \doteq \mathbb{E}_{\pi(\mathbf{a}_t \mid \mathbf{s}_t)} \left[Q(\mathbf{s}_t, \mathbf{a}_t, \mathcal{S}^\Omega; q_T) \right] + \mathbb{D}_{\mathrm{KL}}(\pi(\cdot \mid \mathbf{s}_t) \,\|\, p(\cdot \mid \mathbf{s}_t)). \quad \text{(A.2)}$$

This dynamic-discount, behavior-driven Bellman backup operator is identical to the Bellman backup operator for KL-regularized reinforcement learning [7] except for the learned, dynamic discount factor, $q_{\Delta_{t+1}}(\Delta_{t+1} = 0)$.

In tabular settings, repeated application of this Bellman operator will result in an optimal policy and an optimal dynamic discount factor. More specifically, alternating between policy evaluation and optimization of the variational distribution over the state–action trajectory and the trajectory length converges to an optimal policy.

Theorem 2 (Variational Dynamic-Discount Behavior-Driven Policy Iteration)**.** *Assume $|\mathcal{A}| < \infty$ and that the MDP is ergodic.*

1. *Dynamic-Discount Behavior-Driven Policy Evaluation (D2BD-PE): Given policy π and a function $Q^0 : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to \mathbb{R}$, define $Q^{i+1} = \mathcal{T}^\pi Q^i$. Then the sequence Q^i converges to the lower bound in Theorem 1.*

2. *Dynamic-Discount Behavior-Driven Policy Improvement (D2BD-PI): The policy*

$$\pi^+ = \arg\max_{\pi' \in \Pi} \left\{ \mathbb{E}_{\pi'(\mathbf{a}_t \mid \mathbf{s}_t)} \left[Q^\pi(\mathbf{s}_t, \mathbf{a}_t, \mathcal{S}^\Omega; q_T) \right] - \mathbb{D}_{\mathrm{KL}}(\pi'(\cdot \mid \mathbf{s}_t) \,||\, p(\cdot \mid \mathbf{s}_t)) \right\} \quad \text{(A.3)}$$

   and the variational distribution over T recursively defined in terms of

$$\begin{aligned} &q^+(\Delta_{t+1} = 0 \mid \mathbf{s}_0; \pi, Q^\pi) \\ &\quad = \sigma\left(\mathbb{E}_{\pi(\mathbf{a}_{t+1} \mid \mathbf{s}_{t+1}) p_d(\mathbf{s}_{t+1} \mid \mathbf{s}_t, \mathbf{a}_t)} [Q^\pi(\mathbf{s}_{t+1}, \mathbf{a}_{t+1}, \mathcal{S}^\Omega; q_T)] + \sigma^{-1}\left(p_{\Delta_{t+1}}(\Delta_{t+1} = 0) \right) \right) \end{aligned} \quad \text{(A.4)}$$

   improve the variational objective. In other words, $V^{\pi^+}(\mathbf{s}_0, \mathcal{S}^\Omega; q_T) \geq V^\pi(\mathbf{s}_0, \mathcal{S}^\Omega; q_T)$ and $V^\pi(\mathbf{s}_0, \mathcal{S}^\Omega; q_T^+) \geq V^\pi(\mathbf{s}_0, \mathcal{S}^\Omega; q_T)$ for all $\mathbf{s}_0 \in \mathcal{S}$.

3. *Alternating between D2BD-PE and D2BD-PI converges to a policy $\pi^\star$ and a variational distribution over T, $q_T^\star$, such that $Q^{\pi^\star}(\mathbf{s}, \mathbf{a}, \mathcal{S}^\Omega; q_T^\star) \geq Q^\pi(\mathbf{s}, \mathbf{a}, \mathcal{S}^\Omega; q_T)$ for all $(\pi, q_T) \in \Pi \times \mathcal{Q}_T$ and any $(\mathbf{s}, \mathbf{a}) \in \mathcal{S} \times \mathcal{A}$.*

An implication of this result is that an optimal policy found via dynamic-discount behavior-driven policy iteration has at least as high a state value at $\mathbf{S}_0 = \mathbf{s}_0$ as it would under a fixed discount factor. That is, for p_T given by a fixed geometric distribution with parameter γ, the state–action value function simplifies to the standard Bellman backup operator,

$$Q^\pi(\mathbf{s}_t, \mathbf{a}_t, \mathcal{S}^\Omega; p_T) \doteq \log \mathbb{P}(\xi_{t+1}(\mathcal{S}^\Omega) = 1 \mid \mathbf{s}_t, \mathbf{a}_t) + \gamma\, \mathbb{E}_{p_d(\mathbf{s}_{t+1} \mid \mathbf{s}_t, \mathbf{a}_t)} \left[V^\pi(\mathbf{s}_{t+1}, \mathcal{S}^\Omega; \pi, p_T) \right], \quad \text{(A.5)}$$

and

$$Q^\pi(\mathbf{s}_0, \mathbf{a}_0, \mathcal{S}^\Omega; q_T^*) \geq Q^\pi(\mathbf{s}_0, \mathbf{a}_0, \mathcal{S}^\Omega; p_T). \quad \text{(A.6)}$$

In other words, dynamic discount factors are optimal in KL-regularized reinforcement learning and can be justified using the variational Bayesian inference formulation described here.

Appendix B Proofs

B.1 Finite- and Infinite-Horizon Variational Objectives

In this section, we present detailed derivations and proofs for the results in the main text.

Proposition 1 (Fixed-Time Variational Objective)**.** *Let the variational distribution $q_{\tilde{\boldsymbol{\mathcal{T}}}_{0:t}}(\tilde{\boldsymbol{\tau}}_{0:t})$ be as defined in Equation (7). Then, given a horizon length $t^\star$ and optimal state trajectory $\mathcal{S}^\Omega$,*

$$\mathbb{D}_{\mathrm{KL}}(q_{\tilde{\boldsymbol{\mathcal{T}}}_{0:t}}(\cdot)\,\|\,p_{\tilde{\boldsymbol{\mathcal{T}}}_{0:t}|\Delta_{1:t+1}}(\cdot\,|\,\Delta^\star_{1:t+1})) = \log p(\Delta^\star_{1:t+1}) - \bar{\mathcal{F}}(\pi, \mathcal{S}^\Omega), \tag{B.7}$$

where

$$\bar{\mathcal{F}}(\pi, \mathcal{S}^\Omega) \doteq \mathbb{E}_{q_{\tilde{\boldsymbol{\mathcal{T}}}_{0:t}}(\tilde{\boldsymbol{\tau}}_{0:t})}\left[\sum_{t'=0}^{t} \log \mathbb{P}(\xi_{t'+1}(\mathcal{S}^\Omega) = 1\,|\,\mathbf{s}_{t'}, \mathbf{a}_{t'}) - \mathbb{D}_{\mathrm{KL}}(\pi(\cdot\,|\,\mathbf{s}_{t'})\,||\,p(\cdot\,|\,\mathbf{s}_{t'}))\right], \tag{B.8}$$

and since $\log p(\Delta^\star_{1:t+1})$ is constant in π,

$$\arg\min_{\pi\in\Pi} \mathbb{D}_{\mathrm{KL}}(q_{\tilde{\boldsymbol{\mathcal{T}}}_{0:t}}(\cdot)\,\|\,p_{\tilde{\boldsymbol{\mathcal{T}}}_{0:t}|\Delta_{1:t+1}}(\cdot\,|\,\Delta^\star_{1:t+1})) = \arg\max_{\pi\in\Pi} \bar{\mathcal{F}}(\pi, \mathcal{S}^\Omega). \tag{B.9}$$

Proof. To find an approximation to the posterior $p_{\tilde{\boldsymbol{\mathcal{T}}}_{0:t}|\Delta_{1:t+1}}(\cdot\,|\,\Delta^\star_{1:t+1}))$, we can use variational inference. To do so, we consider the trajectory distribution under $p_{\tilde{\boldsymbol{\mathcal{T}}}_{0:t}|\Delta_{1:t+1}}(\cdot\,|\,\Delta^\star_{1:t+1}))$, which by Bayes' Theorem is given by

$$p_{\tilde{\boldsymbol{\mathcal{T}}}_{0:t}|\Delta_{1:t+1}}(\cdot\,|\,\Delta^\star_{1:t+1})) = \frac{p_{\mathbf{S}_0}(\mathbf{s}_0)\prod_{t'=0}^{t}\mathbb{P}(\xi_{t'+1}(\mathcal{S}^\Omega)=1\,|\,\mathbf{s}_{t'},\mathbf{a}_{t'})p(\mathbf{a}_t\,|\,\mathbf{s}_t)\prod_{t'=0}^{t-1}p_d(\mathbf{s}_{t'+1}\,|\,\mathbf{s}_{t'},\mathbf{a}_{t'})p(\mathbf{a}_{t'}\,|\,\mathbf{s}_{t'})}{p(\Delta^\star_{1:t+1})}. \tag{B.10}$$

Inferring an approximation to the posterior distribution $p_{\tilde{\boldsymbol{\mathcal{T}}}_{0:t}|\Delta_{1:t+1}}(\cdot\,|\,\Delta^\star_{1:t+1})$ then becomes equivalent to finding a variational distribution $q_{\tilde{\boldsymbol{\mathcal{T}}}_{0:T}|\mathbf{S}_0}(\cdot\,|\,\mathbf{s}_0)$, which induces a trajectory distribution $q_{\tilde{\boldsymbol{\mathcal{T}}}_{0:t}}(\cdot)$ that minimizes the KL divergence from $q_{\tilde{\boldsymbol{\mathcal{T}}}_{0:t}}(\cdot)$ to $p_{\tilde{\boldsymbol{\mathcal{T}}}_{0:t}|\Delta_{1:t+1}}(\cdot\,|\,\Delta^\star_{1:t+1})$:

$$\min_{q_{\tilde{\boldsymbol{\mathcal{T}}}_{0:t}}\in\hat{\mathcal{Q}}} \mathbb{D}_{\mathrm{KL}}(q_{\tilde{\boldsymbol{\mathcal{T}}}_{0:t}}(\cdot)\,\|\,p_{\tilde{\boldsymbol{\mathcal{T}}}_{0:t}|\Delta_{1:t+1}}(\cdot\,|\,\Delta^\star_{1:t+1})). \tag{B.11}$$

If we find a distribution $q_{\tilde{\boldsymbol{\mathcal{T}}}_{0:t}}(\cdot)$ for which the resulting KL divergence is zero, then $q_{\tilde{\boldsymbol{\mathcal{T}}}_{0:t}}(\cdot)$ is the exact posterior. If the KL divergence is positive, then $q_{\tilde{\boldsymbol{\mathcal{T}}}_{0:t}}(\cdot)$ is an approximate posterior. To solve the variational problem in Equation (B.11), we can define a factorized variational family

$$q_{\tilde{\boldsymbol{\mathcal{T}}}_{0:t}}(\tilde{\boldsymbol{\tau}}_{0:t}) \doteq p_{\mathbf{S}_0}(\mathbf{s}_0)\pi(\mathbf{a}_t\,|\,\mathbf{s}_t)\prod_{t'=0}^{t-1} q_{\mathbf{S}_{t'+1}|\mathbf{S}_{t'},\mathbf{A}_{t'}}(\mathbf{s}_{t'+1}\,|\,\mathbf{s}_{t'},\mathbf{a}_{t'})\pi(\mathbf{a}_{t'}\,|\,\mathbf{s}_{t'}), \tag{B.12}$$

where $\mathbf{A}_{0:t}$ and $\mathbf{S}_{0:t}$ are latent variables over which to infer an approximate posterior distribution. Returning to the variational problem in Equation (B.11), we can now write

$$\begin{aligned}\mathbb{D}_{\mathrm{KL}}(q_{\tilde{\boldsymbol{\mathcal{T}}}_{0:t}}(\cdot)\,\|\,p_{\tilde{\boldsymbol{\mathcal{T}}}_{0:t}|\Delta_{1:t+1}}(\cdot\,|\,\Delta^\star_{1:t+1})) &= \int_{\mathcal{A}^{t+1}}\int_{\mathcal{S}^{t+1}} q_{\tilde{\boldsymbol{\mathcal{T}}}_{0:t}}(\tilde{\boldsymbol{\tau}}_{0:t})\log\frac{q_{\tilde{\boldsymbol{\mathcal{T}}}_{0:t}}(\tilde{\boldsymbol{\tau}}_{0:t})}{p_{\tilde{\boldsymbol{\mathcal{T}}}_{0:t}|\Delta_{1:t+1}}(\tilde{\boldsymbol{\tau}}_{0:t}\,|\,\Delta^\star_{1:t+1})}d\mathbf{s}_{0:t}d\mathbf{a}_{0:t} \\ &= -\bar{\mathcal{F}}(\pi,\mathcal{S}^\Omega) + \log p(\Delta^\star_{1:t+1}),\end{aligned} \tag{B.13}$$

where

$$\begin{aligned}\bar{\mathcal{F}}(\pi,\mathcal{S}^\Omega) \doteq \mathbb{E}_{q_{\tilde{\boldsymbol{\mathcal{T}}}_{0:t}}(\tilde{\boldsymbol{\tau}}_{0:t})}\Bigg[&\sum_{t'=0}^{t}\log\mathbb{P}(\xi_{t'+1}(\mathcal{S}^\Omega)=1\,|\,\mathbf{s}_{t'},\mathbf{a}_{t'}) \\ &+\log p(\mathbf{a}_t\,|\,\mathbf{s}_t) - \log\pi(\mathbf{a}_t\,|\,\mathbf{s}_t) + \sum_{t'=0}^{t-1}\log p(\mathbf{a}_{t'}\,|\,\mathbf{s}_{t'}) - \log\pi(\mathbf{a}_{t'}\,|\,\mathbf{s}_{t'}) \\ &+\sum_{t'=0}^{t-1}\log p_d(\mathbf{s}_{t'+1}\,|\,\mathbf{s}_{t'},\mathbf{a}_{t'}) - \log q_{\mathbf{S}_{t'+1}|\mathbf{S}_{t'},\mathbf{A}_{t'}}(\mathbf{s}_{t'+1}\,|\,\mathbf{s}_{t'},\mathbf{a}_{t'})\Bigg]\end{aligned} \tag{B.14}$$

and

$$\log p(\Delta^\star_{1:t+1}) = \log\int_{\mathcal{A}^{t+1}}\int_{\mathcal{S}^{t+1}}\mathbb{P}(\Delta_{1:t+1}=1\,|\,\mathbf{s}_0,\mathbf{a}_0,...,\mathbf{s}_t,\mathbf{a}_t)p_{\tilde{\boldsymbol{\mathcal{T}}}_{0:t}}(\tilde{\boldsymbol{\tau}}_{0:t})d\mathbf{s}_{0:t}d\mathbf{a}_{0:t} \tag{B.15}$$

$$= \log\int_{\mathcal{A}^{t+1}}\int_{\mathcal{S}^{t+1}}\left(\prod_{t'=0}^{t}\int_{\mathcal{S}} p_d(\mathbf{s}_{t'+1}\,|\,\mathbf{s}_{t'},\mathbf{a}_{t'})\,\mathbb{I}\{\mathbf{s}_{t'+1}\in\mathcal{S}^\Omega\}\,d\mathbf{s}_{t'+1}\right)p_{\tilde{\boldsymbol{\mathcal{T}}}_{0:t}}(\tilde{\boldsymbol{\tau}}_{0:t})d\mathbf{s}_{0:t}d\mathbf{a}_{0:t} \tag{B.16}$$

is a log-marginal likelihood. Following Haarnoja et al. [1], we define the variational distribution over next states as the true transition dynamics, that is,

$$q_{\mathbf{S}_{t+1}|\mathbf{S}_t,\mathbf{A}_t}(\mathbf{s}_{t+1} \mid \mathbf{s}_t, \mathbf{a}_t) = p_d(\mathbf{s}_{t+1} \mid \mathbf{s}_t, \mathbf{a}_t), \tag{B.17}$$

so that

$$q_{\tilde{\boldsymbol{\tau}}_{0:t}}(\tilde{\boldsymbol{\tau}}_{0:t}) \doteq p_{\mathbf{S}_0}(\mathbf{s}_0)\pi(\mathbf{a}_t \mid \mathbf{s}_t) \prod_{t'=0}^{t-1} p_d(\mathbf{s}_{t'+1} \mid \mathbf{s}_{t'}, \mathbf{a}_{t'})\pi(\mathbf{a}_{t'} \mid \mathbf{s}_{t'}). \tag{B.18}$$

We can then simplify $\bar{\mathcal{F}}(\pi, \mathcal{S}^\Omega)$ to

$$\bar{\mathcal{F}}(\pi, \mathcal{S}^\Omega) = \mathbb{E}_{q_{\tilde{\boldsymbol{\tau}}_{0:t}}(\tilde{\boldsymbol{\tau}}_{0:t})}\left[\sum_{t'=0}^{t} \log \mathbb{P}(\xi_{t'+1}(\mathcal{S}^\Omega) = 1 \mid \mathbf{s}_{t'}, \mathbf{a}_{t'}) - \mathbb{D}_{\mathrm{KL}}(\pi(\cdot \mid \mathbf{s}_{t'}) \parallel p(\cdot \mid \mathbf{s}_{t'}))\right]. \tag{B.19}$$

Since $\log p(\Delta^\star_{1:t+1})$ is constant in π, solving the variational optimization problem in Equation (B.11) is equivalent to maximizing the variational objective with respect to $\pi \in \Pi$, where Π is a family of policy distributions. □

Corollary 1 . *The objective in Equation (B.19) corresponds to KL-regularized reinforcement learning with a time-varying reward function given by*

$$r(\mathbf{s}_{t'}, \mathbf{a}_{t'}, \Delta_{t'+1}) \doteq \log \mathbb{P}(\xi_{t'+1}(\mathcal{S}^\Omega) = 1 \mid \mathbf{s}_{t'}, \mathbf{a}_{t'}).$$

Proof. Let

$$r(\mathbf{s}_{t'}, \mathbf{a}_{t'}, \Delta_{t'+1}) \doteq \log \mathbb{P}(\xi_{t'+1}(\mathcal{S}^\Omega) = 1 \mid \mathbf{s}_{t'}, \mathbf{a}_{t'}). \tag{B.20}$$

and note that the objective

$$\bar{\mathcal{F}}(\pi, \mathcal{S}^\Omega) = \mathbb{E}_{q_{\tilde{\boldsymbol{\tau}}_{0:t}}(\tilde{\boldsymbol{\tau}}_{0:t})}\left[\sum_{t'=0}^{t} \log \mathbb{P}(\xi_{t'+1}(\mathcal{S}^\Omega) = 1 \mid \mathbf{s}_{t'}, \mathbf{a}_{t'}) - \mathbb{D}_{\mathrm{KL}}(\pi(\cdot \mid \mathbf{s}_{t'}) \parallel p(\cdot \mid \mathbf{s}_{t'}))\right]. \tag{B.21}$$

can equivalently written as

$$\bar{\mathcal{F}}(\pi, \mathcal{S}^\Omega) = \mathbb{E}_{q_{\tilde{\boldsymbol{\tau}}_{0:t}}(\tilde{\boldsymbol{\tau}}_{0:t})}\left[\sum_{t'=0}^{t} r(\mathbf{s}_{t'}, \mathbf{a}_{t'}, \Delta_{t'+1}) + \mathbb{D}_{\mathrm{KL}}(\pi(\cdot \mid \mathbf{s}_{t'}) \parallel p(\cdot \mid \mathbf{s}_{t'}))\right], \tag{B.22}$$

which, as shown in Haarnoja et al. [1], can be written in the form of Equation (11). □

Proposition 3 (Unfactorized Dynamic-Discount Behavior-Driven RL as Variational Inference)**.** *Let*

$$q_{\tilde{\boldsymbol{\mathcal{T}}}_{0:T},T}(\tilde{\boldsymbol{\tau}}_{0:t}, t) = q_{\tilde{\boldsymbol{\mathcal{T}}}_{0:T}|T}(\tilde{\boldsymbol{\tau}}_{0:t} \mid t) q_T(t), \tag{B.23}$$

let $q_T(t)$ be a variational distribution defined on $t \in \mathbb{N}_0$, and let $q_{\tilde{\boldsymbol{\mathcal{T}}}_{0:T}|T}(\tilde{\boldsymbol{\tau}}_{0:t} \mid t)$ be as defined in Equation (7). Then, given an optimal state trajectory $\mathcal{S}^\Omega$, we have that

$$\mathbb{D}_{\mathrm{KL}}(q_{\tilde{\boldsymbol{\mathcal{T}}}_{0:T},T}(\cdot) \,\|\, p_{\tilde{\boldsymbol{\mathcal{T}}}_{0:T},T|\Delta_{1:T+1}}(\cdot \mid \Delta^\star_{1:T+1})) = \log p(\Delta^\star_{1:T+1}) - \mathcal{F}(\pi, q_T, \mathcal{S}^\Omega), \tag{B.24}$$

where

$$\mathcal{F}(\pi, q_T, \mathcal{S}^\Omega) \doteq \sum_{t=0}^{\infty} q_T(t)\, \mathbb{E}_{q_{\tilde{\boldsymbol{\mathcal{T}}}_{0:T}|T}(\tilde{\boldsymbol{\tau}}_{0:t} \mid t)}\left[\sum_{t'=0}^{t} \log \mathbb{P}(\xi_{t'+1}(\mathcal{S}^\Omega) = 1 \mid \mathbf{s}_{t'}, \mathbf{a}_{t'}) - \mathbb{D}_{\mathrm{KL}}(q_{\tilde{\boldsymbol{\mathcal{T}}}_{0:T},T}(\cdot) \,||\, p_{\tilde{\boldsymbol{\mathcal{T}}}_{0:T},T}(\cdot))\right] \tag{B.25}$$

and $\log p(\Delta^\star_{1:T+1})$ is constant in π and q_T.

Proof. In general, solving the variational problem

$$\min_{q \in \mathcal{Q}} \mathbb{D}_{\mathrm{KL}}(q_{\tilde{\boldsymbol{\mathcal{T}}}_{0:T},T}(\cdot) \parallel p_{\tilde{\boldsymbol{\mathcal{T}}}_{0:T},T|\Delta_{1:T+1}}(\cdot \mid \Delta^\star_{1:T+1})) \tag{B.26}$$

is challenging, but as in the fixed-time setting, we can take advantage of the fact that, by choosing a variational family parameterized by

$$q_{\tilde{\boldsymbol{\mathcal{T}}}_{0:T}|T}(\tilde{\boldsymbol{\tau}}_{0:t} \mid t) \doteq \pi(\mathbf{a}_t \mid \mathbf{s}_t) \prod_{t'=0}^{t-1} p_d(\mathbf{s}_{t'+1} \mid \mathbf{s}_{t'}, \mathbf{a}_{t'})\, \pi(\mathbf{a}_{t'} \mid \mathbf{s}_{t'}), \tag{B.27}$$

with $\pi \in \Pi$, we can follow the same steps as in the proof for Proposition 1 (Fixed-Time Variational Objective) and show that given an an optimal state trajectory $\mathcal{S}^\Omega$,

$$\mathbb{D}_{\mathrm{KL}}(q_{\tilde{\boldsymbol{\mathcal{T}}}_{0:T},T}(\cdot) \parallel p_{\tilde{\boldsymbol{\mathcal{T}}}_{0:T},T|\Delta_{1:T+1}}(\cdot \mid \Delta^\star_{1:T+1}))) = \log p(\Delta^\star_{1:T+1}) - \mathcal{F}(\pi, q_T, \mathcal{S}^\Omega), \tag{B.28}$$

where

$$\mathcal{F}(\pi, q_T, \mathcal{S}^\Omega)$$
$$\doteq \sum_{t=0}^{\infty} q_T(t)\, \mathbb{E}_{q_{\tilde{\boldsymbol{\tau}}_{0:T}|T}(\tilde{\boldsymbol{\tau}}_{0:t}\,|\,t)} \Big[\sum_{t'=0}^{t} \log \mathbb{P}(\xi_{t'+1}(\mathcal{S}^\Omega) = 1 \,|\, \mathbf{s}_{t'}, \mathbf{a}_{t'}) - \mathbb{D}_{\mathrm{KL}}(q_{\tilde{\boldsymbol{\tau}}_{0:T},T|\mathbf{S}_0}(\cdot\,|\,\mathbf{s}_0) \parallel p_{\tilde{\boldsymbol{\tau}}_{0:T},T|\mathbf{S}_0}(\cdot\,|\,\mathbf{s}_0)) \Big], \tag{B.29}$$

where $q_{\tilde{\boldsymbol{\tau}}_{0:T},T}(\tilde{\boldsymbol{\tau}}_{0:t}, t) \doteq q_{\tilde{\boldsymbol{\tau}}_{0:T}|T}(\tilde{\boldsymbol{\tau}}_{0:t}\,|\,t) q_T(t)$, and hence, solving the variational problem in Equation (8) is equivalent to maximizing $\mathcal{F}(\pi, q_T, \mathcal{S}^\Omega)$ with respect to π and q_T. □

B.2 Recursive Variational Objective & Bellman Backup Operator

Proposition 4 (Factorized Dynamic-Discount Behavior-Driven RL as Variational Inference)**.** *Let the variational distribution factorize as*

$$q_{\tilde{\boldsymbol{\tau}}_{0:T},T}(\tilde{\boldsymbol{\tau}}_{0:t}, t) = q_{\tilde{\boldsymbol{\tau}}_{0:T}|T}(\tilde{\boldsymbol{\tau}}_{0:t}\,|\,t) q_T(t), \tag{B.30}$$

let

$$q_T(t) = q_{\Delta_{t+1}}(\Delta_{t+1} = 1) \prod_{t'=1}^{t} q_{\Delta_{t'}}(\Delta_{t'} = 0) \tag{B.31}$$

be a variational distribution defined on $t \in \mathbb{N}_0$, and let $q_{\tilde{\boldsymbol{\tau}}_{0:T}|T}(\tilde{\boldsymbol{\tau}}_{0:t}\,|\,t)$ be as defined in Equation (7). Then, given an optimal state trajectory $\mathcal{S}^\Omega$, Equation (B.25) can be rewritten as

$$\mathcal{F}(\pi, q_T, \mathcal{S}^\Omega) = \mathbb{E}_{q_{\tilde{\boldsymbol{\tau}}_0}(\tilde{\boldsymbol{\tau}}_0)} \left[\sum_{t=0}^{\infty} \left(\prod_{t'=1}^{t} q_{\Delta_{t'}}(\Delta_{t'} = 0) \right) \Big(r(\mathbf{s}_t, \mathbf{a}_t, \mathcal{S}^\Omega; q_\Delta) - \mathbb{D}_{\mathrm{KL}}(\pi(\cdot\,|\,\mathbf{s}_t) \,\|\, p(\cdot\,|\,\mathbf{s}_t)) \Big) \right] \tag{B.32}$$

where

$$r(\mathbf{s}_t, \mathbf{a}_t, \mathcal{S}^\Omega; q_\Delta) \doteq \log \mathbb{P}(\xi_{t+1}(\mathcal{S}^\Omega) = 1 \,|\, \mathbf{s}_t, \mathbf{a}_t) - q_{\Delta_{t+1}}(\Delta_{t+1} = 1) \mathbb{D}_{\mathrm{KL}}(q_{\Delta_{t+1}} \,\|\, p_{\Delta_{t+1}}), \tag{B.33}$$

Proof. Consider the variational objective $\mathcal{F}(\pi, q_T, \mathcal{S}^\Omega)$ in Equation (B.25):

$$\mathcal{F}(\pi, q_T, \mathcal{S}^\Omega)$$
$$= \sum_{t=0}^{\infty} q_T(t)\, \mathbb{E}_{q_{\tilde{\boldsymbol{\tau}}_{0:T}|T}(\tilde{\boldsymbol{\tau}}_{0:t}\,|\,t)} \Big[\sum_{t'=0}^{t} \log \mathbb{P}(\xi_{t'+1}(\mathcal{S}^\Omega) = 1 \,|\, \mathbf{s}_{t'}, \mathbf{a}_{t'}) - \mathbb{D}_{\mathrm{KL}}(q_{\tilde{\boldsymbol{\tau}}_{0:T},T}(\cdot) \parallel p_{\tilde{\boldsymbol{\tau}}_{0:T},T}(\cdot)) \Big] \tag{B.34}$$
$$= \sum_{t=0}^{\infty} q_T(t)\, \mathbb{E}_{q_{\tilde{\boldsymbol{\tau}}_{0:T}|T}(\tilde{\boldsymbol{\tau}}_{0:t}\,|\,t)} \left[\sum_{t'=0}^{t} \log \mathbb{P}(\xi_{t'+1}(\mathcal{S}^\Omega) = 1 \,|\, \mathbf{s}_{t'}, \mathbf{a}_{t'}) - \log \frac{q_{\tilde{\boldsymbol{\tau}}_{0:T}|T}(\tilde{\boldsymbol{\tau}}_{0:t}\,|\,t) q_T(t)}{p_{\tilde{\boldsymbol{\tau}}_{0:T}|T}(\tilde{\boldsymbol{\tau}}_{0:t}\,|\,t) p_T(t)} d\tilde{\boldsymbol{\tau}}_{0:t} \right] \tag{B.35}$$
$$= \sum_{t=0}^{\infty} q_T(t)\, \mathbb{E}_{q_{\tilde{\boldsymbol{\tau}}_{0:T}|T}(\tilde{\boldsymbol{\tau}}_{0:t}\,|\,t)} \left[\sum_{t'=0}^{t} \log \mathbb{P}(\xi_{t'+1}(\mathcal{S}^\Omega) = 1 \,|\, \mathbf{s}_{t'}, \mathbf{a}_{t'}) - \log \frac{q_{\tilde{\boldsymbol{\tau}}_{0:T}|T}(\tilde{\boldsymbol{\tau}}_{0:t}\,|\,t)}{p_{\tilde{\boldsymbol{\tau}}_{0:T}|T}(\tilde{\boldsymbol{\tau}}_{0:t}\,|\,t)} \right] - \sum_{t=0}^{\infty} q_T(t) \log \frac{q_T(t)}{q_T(t)}. \tag{B.36}$$

Noting that $\sum_{t=0}^{\infty} q_T(t) \log \frac{q_T(t)}{q_T(t)} = \mathbb{D}_{\mathrm{KL}}(q_T \parallel p_T)$, we can write

$$\mathcal{F}(\pi, q_T, \mathcal{S}^\Omega)$$
$$= \sum_{t=0}^{\infty} q_T(t)\, \mathbb{E}_{q_{\tilde{\boldsymbol{\tau}}_{0:T}|T}(\tilde{\boldsymbol{\tau}}_{0:t}\,|\,t)} \left[\sum_{t'=0}^{t} \log \mathbb{P}(\xi_{t'+1}(\mathcal{S}^\Omega) = 1 \,|\, \mathbf{s}_{t'}, \mathbf{a}_{t'}) - \log \frac{q_{\tilde{\boldsymbol{\tau}}_{0:T}|T}(\tilde{\boldsymbol{\tau}}_{0:t}\,|\,t)}{p_{\tilde{\boldsymbol{\tau}}_{0:T}|T}(\tilde{\boldsymbol{\tau}}_{0:t}\,|\,t)} \right] - \mathbb{D}_{\mathrm{KL}}(q_T \parallel p_T) \tag{B.37}$$
$$= \sum_{t=0}^{\infty} q_T(t)\, \mathbb{E}_{q_{\tilde{\boldsymbol{\tau}}_{0:T}|T}(\tilde{\boldsymbol{\tau}}_{0:t}\,|\,t)} \Big[\sum_{t'=0}^{t} \log \mathbb{P}(\xi_{t'+1}(\mathcal{S}^\Omega) = 1 \,|\, \mathbf{s}_{t'}, \mathbf{a}_{t'}) \Big] - \sum_{t=0}^{\infty} q_T(t)\, \mathbb{E}_{q_{\tilde{\boldsymbol{\tau}}_{0:T}|T}(\tilde{\boldsymbol{\tau}}_{0:t}\,|\,t)} \left[\log \frac{q_{\tilde{\boldsymbol{\tau}}_{0:T}|T}(\tilde{\boldsymbol{\tau}}_{0:t}\,|\,t)}{p_{\tilde{\boldsymbol{\tau}}_{0:T}|T}(\tilde{\boldsymbol{\tau}}_{0:t}\,|\,t)} \right] - \mathbb{D}_{\mathrm{KL}}(q_T \parallel p_T). \tag{B.38}$$

Further noting that for an infinite-horizon trajectory distribution

$$q_{\tilde{\boldsymbol{\mathcal{T}}}_{t'}|\mathbf{s}_{t'}}(\tilde{\boldsymbol{\tau}}_{t'} \mid \mathbf{s}_{t'}) \doteq \prod_{t=t'}^{\infty} p_d(\mathbf{s}_{t+1} \mid \mathbf{s}_t, \mathbf{a}_t)\pi(\mathbf{a}_t \mid \mathbf{s}_t), \tag{B.39}$$

trajectory realization $\tilde{\boldsymbol{\tau}}_{t+1} \doteq \{\boldsymbol{\tau}_{t'}\}_{t'=t+1}^{\infty}$, and any joint probability density $f(\mathbf{s}_t, \mathbf{a}_t)$,

$$\begin{aligned}
&\sum_{t=0}^{\infty} q_T(t)\, \mathbb{E}_{q_{\tilde{\boldsymbol{\mathcal{T}}}_{0:T}|T}(\tilde{\boldsymbol{\tau}}_{0:t} \mid t)} \Big[f(\mathbf{s}_t, \mathbf{a}_t) \Big] && \text{(B.40)}\\
&= \sum_{t=0}^{\infty} \left(\int q_{\tilde{\boldsymbol{\mathcal{T}}}_{T+1}}(\tilde{\boldsymbol{\tau}}_{t+1}) \left(\int_{\mathcal{S}^t \times \mathcal{A}^t} q_{\tilde{\boldsymbol{\mathcal{T}}}_{0:t}}(\tilde{\boldsymbol{\tau}}_{0:t}) q_T(t) f(\mathbf{s}_t, \mathbf{a}_t) d\tilde{\boldsymbol{\tau}}_{0:t} \right) d\tilde{\boldsymbol{\tau}}_{t+1} \right) \\
&= \sum_{t=0}^{\infty} \left(\mathbb{E}_{q_{\tilde{\boldsymbol{\mathcal{T}}}_{0:T}|T}(\tilde{\boldsymbol{\tau}}_{0:t} \mid t)} \Big[q_T(t) f(\mathbf{s}_t, \mathbf{a}_t) \Big] \cdot \underbrace{\left(\int q_{\tilde{\boldsymbol{\mathcal{T}}}_{T+1}}(\tilde{\boldsymbol{\tau}}_{t+1}) d\tilde{\boldsymbol{\tau}}_{t+1} \right)}_{=1} \right) && \text{(B.41)}\\
&= \sum_{t=0}^{\infty} \left(\left(\int_{\mathcal{S}^t \times \mathcal{A}^t} q(\tilde{\boldsymbol{\tau}}_{0:t}) q_T(t) f(\mathbf{s}_t, \mathbf{a}_t) d\tilde{\boldsymbol{\tau}}_{0:t} \right) \cdot \underbrace{\left(\int q_{\tilde{\boldsymbol{\mathcal{T}}}_{T+1}}(\tilde{\boldsymbol{\tau}}_{t+1}) d\tilde{\boldsymbol{\tau}}_{t+1} \right)}_{=1} \right) && \text{(B.42)}\\
&= \sum_{t=0}^{\infty} \int q_{\tilde{\boldsymbol{\mathcal{T}}}_0}(\tilde{\boldsymbol{\tau}}_0) q_T(t) f(\mathbf{s}_t, \mathbf{a}_t) d\tilde{\boldsymbol{\tau}}_0 && \text{(B.43)}\\
&= \int q_{\tilde{\boldsymbol{\mathcal{T}}}_0}(\tilde{\boldsymbol{\tau}}_0) \sum_{t=0}^{\infty} q_T(t) f(\mathbf{s}_t, \mathbf{a}_t) d\tilde{\boldsymbol{\tau}}_0, && \text{(B.44)}
\end{aligned}$$

we can express Equation (B.38) in terms of the infinite-horizon state–action trajectory

$$q_{\tilde{\boldsymbol{\mathcal{T}}}_0}(\tilde{\boldsymbol{\tau}}_0) \doteq \prod_{t=0}^{\infty} p_d(\mathbf{s}_{t+1} \mid \mathbf{s}_t, \mathbf{a}_t)\pi(\mathbf{a}_t \mid \mathbf{s}_t) \tag{B.45}$$

as

$$\begin{aligned}
\mathcal{F}(\pi, q_T, \mathcal{S}^{\Omega}) &= \int q_{\tilde{\boldsymbol{\mathcal{T}}}_0}(\tilde{\boldsymbol{\tau}}_0) \sum_{t=0}^{\infty} q_T(t) \sum_{t'=0}^{t} \log \mathbb{P}(\xi_{t'+1}(\mathcal{S}^{\Omega}) = 1 \mid \mathbf{s}_{t'}, \mathbf{a}_{t'}) d\tilde{\boldsymbol{\tau}} \\
&\quad - \sum_{t=0}^{\infty} q_T(t) \mathbb{D}_{\mathrm{KL}}(q_{\tilde{\boldsymbol{\mathcal{T}}}_{0:T}|T}(\cdot \mid t) \parallel p_{\tilde{\boldsymbol{\mathcal{T}}}_{0:T}|T}(\cdot \mid t)) - \mathbb{D}_{\mathrm{KL}}(q_T \parallel p_T) && \text{(B.46)}\\
&= \mathbb{E}_{q_{\tilde{\boldsymbol{\mathcal{T}}}_0}(\tilde{\boldsymbol{\tau}}_0)} \Bigg[\sum_{t=0}^{\infty} q_T(t) \Big(\sum_{t'=0}^{t} \log \mathbb{P}(\xi_{t'+1}(\mathcal{S}^{\Omega}) = 1 \mid \mathbf{s}_{t'}, \mathbf{a}_{t'}) \\
&\quad - \mathbb{D}_{\mathrm{KL}}(q_{\tilde{\boldsymbol{\mathcal{T}}}_{0:T}|T}(\cdot \mid t) \parallel p_{\tilde{\boldsymbol{\mathcal{T}}}_{0:T}|T}(\cdot \mid t)) \Big) \Bigg] - \mathbb{D}_{\mathrm{KL}}(q_T \parallel p_T). && \text{(B.47)}
\end{aligned}$$

Using Lemma 5 and the definition of $q_T(t)$ in Equation (9), we can rewrite this objective as

$$\begin{aligned}
&\mathcal{F}(\pi, q_T, \mathcal{S}^{\Omega}) \\
&= \mathbb{E}_{q_{\tilde{\boldsymbol{\mathcal{T}}}_0}(\tilde{\boldsymbol{\tau}}_0)} \Bigg[\sum_{t=0}^{\infty} \Big(\prod_{t'=1}^{t} q_{\Delta_{t'}}(\Delta_{t'} = 0) \Big) q_{\Delta_{t'}}(\Delta_{t'} = 1) \Big(\sum_{t'=0}^{t} \log \mathbb{P}(\xi_{t'+1}(\mathcal{S}^{\Omega}) = 1 \mid \mathbf{s}_{t'}, \mathbf{a}_{t'}) \\
&\quad - \mathbb{D}_{\mathrm{KL}}(q_{\tilde{\boldsymbol{\mathcal{T}}}_{0:T}|T}(\cdot \mid t) \parallel p_{\tilde{\boldsymbol{\mathcal{T}}}_{0:T}|T}(\cdot \mid t)) \Big) \Bigg] - \sum_{t=0}^{\infty} \Big(\prod_{t'=1}^{t} q_{\Delta_{t'}}(\Delta_{t'} = 0) \Big) \mathbb{D}_{\mathrm{KL}}(q_{\Delta_{t+1}} \,||\, p_{\Delta_{t+1}}) && \text{(B.48)}\\
&= \mathbb{E}_{q_{\tilde{\boldsymbol{\mathcal{T}}}_0}(\tilde{\boldsymbol{\tau}}_0)} \Bigg[\sum_{t=0}^{\infty} \left(\prod_{t'=1}^{t} q(\Delta_{t'} = 0) \right) \\
&\quad \cdot \Big(q(\Delta_{t+1} = 1) \big(\sum_{t'=0}^{t} \log \mathbb{P}(\xi_{t'+1}(\mathcal{S}^{\Omega}) = 1 \mid \mathbf{s}_{t'}, \mathbf{a}_{t'}) - \mathbb{D}_{\mathrm{KL}}(q_{\tilde{\boldsymbol{\mathcal{T}}}_{0:T}|T}(\cdot \mid t) \parallel p_{\tilde{\boldsymbol{\mathcal{T}}}_{0:T}|T}(\cdot \mid t)) \big) && \text{(B.49)}\\
&\quad - \mathbb{D}_{\mathrm{KL}}(q_{\Delta_{t+1}} \parallel p_{\Delta_{t+1}}) \Big) \Bigg],
\end{aligned}$$

with

$$\begin{aligned}&\mathbb{D}_{\mathrm{KL}}(q_{\Delta_{t+1}} \parallel p_{\Delta_{t+1}}) \\ &= q_{\Delta_{t+1}}(\Delta_{t+1}=0)\log\frac{q_{\Delta_{t+1}}(\Delta_{t+1}=0)}{p_{\Delta_{t+1}}(\Delta_{t+1}=0)} + (1-q_{\Delta_{t+1}}(\Delta_{t+1}=0))\log\frac{1-q_{\Delta_{t+1}}(\Delta_{t+1}=0)}{1-p_{\Delta_{t+1}}(\Delta_{t+1}=0)}.\end{aligned} \tag{B.50}$$

Next, to re-express $\mathbb{D}_{\mathrm{KL}}(q_{\tilde{\boldsymbol{\mathcal{T}}}_{0:T}|T}(\cdot\,|\,t) \parallel p_{\tilde{\boldsymbol{\mathcal{T}}}_{0:T}|T}(\cdot\,|\,t))$ as a sum over Kullback-Leibler divergences between distributions over single action random variables, we note that

$$\begin{aligned}\mathbb{D}_{\mathrm{KL}}(q_{\tilde{\boldsymbol{\mathcal{T}}}_{0:T}|T}(\cdot\,|\,t) \parallel p_{\tilde{\boldsymbol{\mathcal{T}}}_{0:T}|T}(\cdot\,|\,t)) &= \int_{\mathcal{S}^t\times\mathcal{A}^t} q_{\tilde{\boldsymbol{\mathcal{T}}}_{0:T}|T}(\tilde{\boldsymbol{\tau}}_{0:t}\,|\,t)\log\frac{q_{\tilde{\boldsymbol{\mathcal{T}}}_{0:T}|T}(\tilde{\boldsymbol{\tau}}_{0:t}\,|\,t)}{p_{\tilde{\boldsymbol{\mathcal{T}}}_{0:T}|T}(\tilde{\boldsymbol{\tau}}_{0:t}\,|\,t)}d\tilde{\boldsymbol{\tau}}_{0:t} && \text{(B.51)}\\ &= \int_{\mathcal{S}^t\times\mathcal{A}^t} q_{\tilde{\boldsymbol{\mathcal{T}}}_{0:T}|T}(\tilde{\boldsymbol{\tau}}_{0:t}\,|\,t)\log\frac{\prod_{t'=1}^{t}\pi(\mathbf{a}_{t'}\,|\,\mathbf{s}_{t'})}{\prod_{t'=1}^{t}p(\mathbf{a}_{t'}\,|\,\mathbf{s}_{t'})}d\tilde{\boldsymbol{\tau}}_{0:t} && \text{(B.52)}\\ &= \int_{\mathcal{S}^t\times\mathcal{A}^t} q_{\tilde{\boldsymbol{\mathcal{T}}}_{0:T}|T}(\tilde{\boldsymbol{\tau}}_{0:t}\,|\,t)\sum_{t'=0}^{t}\log\frac{\pi(\mathbf{a}_{t'}\,|\,\mathbf{s}_{t'})}{p(\mathbf{a}_{t'}\,|\,\mathbf{s}_{t'})}d\tilde{\boldsymbol{\tau}}_{0:t} && \text{(B.53)}\\ &= \mathbb{E}_{q_{\tilde{\boldsymbol{\mathcal{T}}}_0}(\tilde{\boldsymbol{\tau}}_0)}\left[\sum_{t'=0}^{t}\int_{\mathcal{A}}\pi(\mathbf{a}_{t'}\,|\,\mathbf{s}_{t'})\log\frac{\pi(\mathbf{a}_{t'}\,|\,\mathbf{s}_{t'})}{p(\mathbf{a}_{t'}\,|\,\mathbf{s}_{t'})}d\mathbf{a}_{t'}\right] && \text{(B.54)}\\ &= \mathbb{E}_{q_{\tilde{\boldsymbol{\mathcal{T}}}_0}(\tilde{\boldsymbol{\tau}}_0)}\left[\sum_{t'=0}^{t}\mathbb{D}_{\mathrm{KL}}(\pi(\cdot\,|\,\mathbf{s}_{t'}) \parallel p(\cdot\,|\,\mathbf{s}_{t'}))\right], && \text{(B.55)}\end{aligned}$$

where we have used the same marginalization trick as above to express the expression in terms of an infinite-horizon trajectory distribution, which allows us to express Equation (B.49) as

$$\begin{aligned}&\mathcal{F}(\pi, q_T, \mathcal{S}^\Omega) \\ &= \mathbb{E}_{q_{\tilde{\boldsymbol{\mathcal{T}}}_0}(\tilde{\boldsymbol{\tau}}_0)}\left[\sum_{t=0}^{\infty}\left(\prod_{t'=1}^{t}q(\Delta_{t'}=0)\right)\cdot\left(q_{\Delta_{t+1}}(\Delta_{t+1}=1)\left(\sum_{t'=0}^{t}\log\mathbb{P}(\xi_{t'+1}(\mathcal{S}^\Omega)=1\,|\,\mathbf{s}_{t'},\mathbf{a}_{t'})\right.\right.\right. \\ &\qquad\left.\left.\left. - \mathbb{E}_{q_{\tilde{\boldsymbol{\mathcal{T}}}_0}(\tilde{\boldsymbol{\tau}}_0)}\left[\sum_{t'=0}^{t}\mathbb{D}_{\mathrm{KL}}(\pi(\cdot\,|\,\mathbf{s}_{t'}) \parallel p(\cdot\,|\,\mathbf{s}_{t'}))\right]\right) - \mathbb{D}_{\mathrm{KL}}(q_{\Delta_{t+1}} \parallel p_{\Delta_{t+1}})\right)\right] \\ &= \mathbb{E}_{q_{\tilde{\boldsymbol{\mathcal{T}}}_0}(\tilde{\boldsymbol{\tau}}_0)}\left[\sum_{t=0}^{\infty}\left(\prod_{t'=1}^{t}q(\Delta_{t'}=0)\right)\cdot\left(q_{\Delta_{t+1}}(\Delta_{t+1}=1)\left(\mathbb{E}_{q_{\tilde{\boldsymbol{\mathcal{T}}}_0}(\tilde{\boldsymbol{\tau}}_0)}\left[\sum_{t'=0}^{t}\log\mathbb{P}(\xi_{t'+1}(\mathcal{S}^\Omega)=1\,|\,\mathbf{s}_{t'},\mathbf{a}_{t'})\right.\right.\right.\right. \\ &\qquad\left.\left.\left.\left. - \mathbb{D}_{\mathrm{KL}}(\pi(\cdot\,|\,\mathbf{s}_{t'}) \parallel p(\cdot\,|\,\mathbf{s}_{t'}))\right]\right) - \mathbb{D}_{\mathrm{KL}}(q_{\Delta_{t+1}} \parallel p_{\Delta_{t+1}})\right)\right].\end{aligned} \tag{B.56}$$

Rearranging and dropping redundant expectation operators, we can now express the objective as

$$\begin{aligned}&\mathcal{F}(\pi, q_T, \mathcal{S}^\Omega) \\ &= \mathbb{E}_{q_{\tilde{\boldsymbol{\mathcal{T}}}_0}(\tilde{\boldsymbol{\tau}}_0)}\left[-\sum_{t=0}^{\infty}\left(\prod_{t'=1}^{t}q_{\Delta_{t'}}(\Delta_{t'}=0)\right)\left(q_{\Delta_{t+1}}(\Delta_{t+1}=1)\mathbb{D}_{\mathrm{KL}}(q_{\Delta_{t+1}} \parallel p_{\Delta_{t+1}})\right)\right] \\ &\qquad + \sum_{t=0}^{\infty}\underbrace{\left(\prod_{t'=1}^{t}q_{\Delta_{t'}}(\Delta_{t'}=0)q_{\Delta_{t+1}}(\Delta_{t+1}=1)\right)}_{=q_T(t)} \\ &\qquad\qquad \cdot\,\mathbb{E}_{q_{\tilde{\boldsymbol{\mathcal{T}}}_0}(\tilde{\boldsymbol{\tau}}_0)}\left[\sum_{t'=0}^{t}\log\mathbb{P}(\xi_{t'+1}(\mathcal{S}^\Omega)=1\,|\,\mathbf{s}_{t'},\mathbf{a}_{t'}) - \mathbb{D}_{\mathrm{KL}}(\pi(\cdot\,|\,\mathbf{s}_{t'}) \parallel p(\cdot\,|\,\mathbf{s}_{t'}))\right],\end{aligned} \tag{B.57}$$

whereupon we note that the last term can be expressed as

$$\sum_{t=0}^{\infty} q_T(t)\, \mathbb{E}_{q_{\tilde{\boldsymbol{\tau}}_0}(\tilde{\boldsymbol{\tau}}_0)} \left[\sum_{t'=0}^{t} \log \mathbb{P}(\xi_{t'+1}(\mathcal{S}^\Omega) = 1 \,|\, \mathbf{s}_{t'}, \mathbf{a}_{t'}) - \mathbb{D}_{\mathrm{KL}}(\pi(\cdot \,|\, \mathbf{s}_{t'}) \,\|\, p(\cdot \,|\, \mathbf{s}_{t'})) \right] \tag{B.58}$$

$$= \mathbb{E}_{q_{\tilde{\boldsymbol{\tau}}_0}(\tilde{\boldsymbol{\tau}}_0)} \left[\sum_{t=0}^{\infty} \sum_{t'=0}^{t} q_T(t) (\log \mathbb{P}(\xi_{t'+1}(\mathcal{S}^\Omega) = 1 \,|\, \mathbf{s}_{t'}, \mathbf{a}_{t'}) - \mathbb{D}_{\mathrm{KL}}(\pi(\cdot \,|\, \mathbf{s}_{t'}) \,\|\, p(\cdot \,|\, \mathbf{s}_{t'}))) \right]$$

$$= \mathbb{E}_{q_{\tilde{\boldsymbol{\tau}}_0}(\tilde{\boldsymbol{\tau}}_0)} \left[\sum_{t=0}^{\infty} q(T \geq t) (\log \mathbb{P}(\xi_{t+1}(\mathcal{S}^\Omega) = 1 \,|\, \mathbf{s}_t, \mathbf{a}_t) - \mathbb{D}_{\mathrm{KL}}(\pi(\cdot \,|\, \mathbf{s}_t) \,\|\, p(\cdot \,|\, \mathbf{s}_t))) \right] \tag{B.59}$$

$$= \mathbb{E}_{q_{\tilde{\boldsymbol{\tau}}_0}(\tilde{\boldsymbol{\tau}}_0)} \left[\sum_{t=0}^{\infty} \underbrace{\left(\prod_{t'=1}^{t} q_{\Delta_{t'}}(\Delta_{t'} = 0) \right)}_{\text{(by Lemma 2)}} (\log \mathbb{P}(\xi_{t+1}(\mathcal{S}^\Omega) = 1 \,|\, \mathbf{s}_t, \mathbf{a}_t) - \mathbb{D}_{\mathrm{KL}}(\pi(\cdot \,|\, \mathbf{s}_t) \,\|\, p(\cdot \,|\, \mathbf{s}_t))) \right], \tag{B.60}$$

where the second line follows from expanding the sums and regrouping terms. By substituting the expression in Equation (B.60) into Equation (B.57), we obtain an objective expressed entirely in terms of distributions over single-index random variables:

$$\mathcal{F}(\pi, q_T, \mathcal{S}^\Omega)$$

$$= \mathbb{E}_{q_{\tilde{\boldsymbol{\tau}}_0}(\tilde{\boldsymbol{\tau}}_0)} \left[\sum_{t=0}^{\infty} \left(\prod_{t'=1}^{t} q_{\Delta_{t'}}(\Delta_{t'} = 0) \right) \right.$$

$$\left. \cdot \left(\log \mathbb{P}(\xi_{t+1}(\mathcal{S}^\Omega) = 1 \,|\, \mathbf{s}_t, \mathbf{a}_t) - q_{\Delta_{t+1}}(\Delta_{t+1} = 1) \mathbb{D}_{\mathrm{KL}}(q_{\Delta_{t+1}} \,\|\, p_{\Delta_{t+1}}) - \mathbb{D}_{\mathrm{KL}}(\pi(\cdot \,|\, \mathbf{s}_t) \,\|\, p(\cdot \,|\, \mathbf{s}_t)) \right) \right] \tag{B.61}$$

$$= \mathbb{E}_{q_{\tilde{\boldsymbol{\tau}}_0}(\tilde{\boldsymbol{\tau}}_0)} \left[\sum_{t=0}^{\infty} \left(\prod_{t'=1}^{t} q_{\Delta_{t'}}(\Delta_{t'} = 0) \right) \left(r(\mathbf{s}_t, \mathbf{a}_t, \mathcal{S}^\Omega; q_\Delta) - \mathbb{D}_{\mathrm{KL}}(\pi(\cdot \,|\, \mathbf{s}_t) \,\|\, p(\cdot \,|\, \mathbf{s}_t)) \right) \right], \tag{B.62}$$

where we defined

$$r(\mathbf{s}_t, \mathbf{a}_t, \mathcal{S}^\Omega; q_\Delta) \doteq \log \mathbb{P}(\xi_{t+1}(\mathcal{S}^\Omega) = 1 \,|\, \mathbf{s}_t, \mathbf{a}_t) - q_{\Delta_{t+1}}(\Delta_{t+1} = 1) \mathbb{D}_{\mathrm{KL}}(q_{\Delta_{t+1}} \,\|\, p_{\Delta_{t+1}}), \tag{B.63}$$

which concludes the proof. □

Theorem 1 (Recursive Dynamic-Discount Behavior-Driven RL as Variational Inference)**.** *Let $q_T(t)$ and $q_{\tilde{\boldsymbol{\mathcal{T}}}_{0:t}|T}(\tilde{\boldsymbol{\tau}}_{0:t} \,|\, t)$ be as defined in Equation (7) and Equation (9), and define*

$$V^\pi(\mathbf{s}_t, \mathcal{S}^\Omega; q_T) \doteq \mathbb{E}_{\pi(\mathbf{a}_t \,|\, \mathbf{s}_t)} \left[Q^\pi(\mathbf{s}_t, \mathbf{a}_t, \mathcal{S}^\Omega; q_T) \right] - \mathbb{D}_{\mathrm{KL}}(\pi(\cdot \,|\, \mathbf{s}_t) \,\|\, p(\cdot \,|\, \mathbf{s}_t)), \tag{B.64}$$

$$Q^\pi(\mathbf{s}_t, \mathbf{a}_t, \mathcal{S}^\Omega; q_T) \doteq r(\mathbf{s}_t, \mathbf{a}_t, \mathcal{S}^\Omega; q_\Delta) + q(\Delta_{t+1} = 0)\, \mathbb{E}_{p_d(\mathbf{s}_{t+1} \,|\, \mathbf{s}_t, \mathbf{a}_t)} \left[V^\pi(\mathbf{s}_{t+1}, \mathcal{S}^\Omega; \pi, q_T) \right], \tag{B.65}$$

$$r(\mathbf{s}_t, \mathbf{a}_t, \mathcal{S}^\Omega; q_\Delta) \doteq \log \mathbb{P}(\xi_{t+1}(\mathcal{S}^\Omega) = 1 \,|\, \mathbf{s}_t, \mathbf{a}_t) - q_{\Delta_{t+1}}(\Delta_{t+1} = 1) \mathbb{D}_{\mathrm{KL}}(q_{\Delta_{t+1}} \,\|\, p_{\Delta_{t+1}}). \tag{B.66}$$

Then given an optimal state trajectory $\mathcal{S}^\Omega$,

$$\mathbb{D}_{\mathrm{KL}}(q_{\tilde{\boldsymbol{\mathcal{T}}}_{0:T}, T}(\cdot) \,\|\, p_{\tilde{\boldsymbol{\mathcal{T}}}_{0:T}, T | \Delta_{1:t+1}(\mathcal{S}^\Omega)}(\cdot \,|\, \Delta^\star_{1:t+1}(\mathcal{S}^\Omega))) = -\mathcal{F}(\pi, q_T, \mathcal{S}^\Omega) + C = -V^\pi(\mathbf{s}_0, \mathcal{S}^\Omega; q_T) + C,$$

where $C \doteq \log p(\Delta^\star_{1:T+1})$ is independent of π and q_T, and hence maximizing $V^\pi(\mathbf{s}_0, \mathcal{S}^\Omega; \pi, q_T)$ is equivalent to minimizing Equation (8). In other words,

$$\begin{aligned} &\arg\min_{\pi \in \Pi, q_T \in \mathcal{Q}_T} \{\mathbb{D}_{\mathrm{KL}}(q_{\tilde{\boldsymbol{\mathcal{T}}}_{0:T}, T}(\cdot) \,\|\, p_{\tilde{\boldsymbol{\mathcal{T}}}_{0:T}, T | \mathbf{s}_0, \mathbf{s}_{T^\star}}(\cdot \,|\, \Delta^\star_{1:T+1}))\} \\ = &\arg\max_{\pi \in \Pi, q_T \in \mathcal{Q}_T} \mathcal{F}(\pi, q_T, \mathcal{S}^\Omega) = \arg\max_{\pi \in \Pi, q_T \in \mathcal{Q}_T} V^\pi(\mathbf{s}_0, \mathcal{S}^\Omega; q_T). \end{aligned}$$

Proof. The proof follows directly from the proof of Theorem 1 in Rudner et al. [8]. □

B.3 Optimal Variational Posterior over T

Proposition 2 (Optimal Variational Distribution over T)**.** *The optimal variational distribution $q^\star_T$ with respect to Equation (10) is defined recursively in terms of $q^\star_{\Delta_{t+1}}(\Delta_{t+1} = 0) \forall t \in \mathbb{N}_0$ by*

$$q^\star_{\Delta_{t+1}}(\Delta_{t+1} = 0; \pi, Q^\pi) = \sigma\left(\Lambda(\mathbf{s}_t, \pi, q_T, Q^\pi) + \sigma^{-1}\left(p_{\Delta_{t+1}}(\Delta_{t+1} = 0)\right)\right), \tag{B.67}$$

where

$$\Lambda(\mathbf{s}_t, \pi, q_T, Q^\pi) \doteq \mathbb{E}_{\pi(\mathbf{a}_{t+1} \mid \mathbf{s}_{t+1}) p_d(\mathbf{s}_{t+1} \mid \mathbf{s}_t, \mathbf{a}_t) \pi(\mathbf{a}_t \mid \mathbf{s}_t)}[Q^\pi(\mathbf{s}_{t+1}, \mathbf{a}_{t+1}, \mathcal{S}^\Omega; q_T) - \log \mathbb{P}(\xi_{t+1}(\mathcal{S}^\Omega) = 1 \mid \mathbf{s}_t, \mathbf{a}_t)]$$

and $\sigma(\cdot)$ is the sigmoid function, that is, $\sigma(x) = \frac{1}{e^{-x}+1}$ and $\sigma^{-1}(x) = \log \frac{x}{1-x}$.

Proof. Consider $\mathcal{F}(\pi, q_T, \mathcal{S}^\Omega)$:

$$\begin{aligned}\mathcal{F}(\pi, q_T, \mathbf{s}_t, \mathcal{S}^\Omega) &= \mathbb{E}_{\pi(\mathbf{a}_t \mid \mathbf{s}_t)}[Q^\pi(\mathbf{s}_t, \mathbf{a}_t, \mathcal{S}^\Omega; q_T)] \\ &= \mathbb{E}_{\pi(\mathbf{a}_t \mid \mathbf{s}_t)}[r(\mathbf{s}_t, \mathbf{a}_t, \mathcal{S}^\Omega; q_\Delta) + q_{\Delta_{t+1}}(\Delta_{t+1} = 0)\, \mathbb{E}\left[V(\mathbf{s}_{t+1}, \mathcal{S}^\Omega; q_T)\right]].\end{aligned} \tag{B.68}$$

Since the variational objective $\mathcal{F}(\pi, q_T, \mathbf{s}_t, \mathcal{S}^\Omega)$ can be expressed recursively as

$$V^\pi(\mathbf{s}_t, \mathcal{S}^\Omega; q_T) \doteq \mathbb{E}_{\pi(\mathbf{a}_t \mid \mathbf{s}_t)}\Big[Q(\mathbf{s}_t, \mathbf{a}_t, \mathcal{S}^\Omega; q_T)\Big] - \mathbb{D}_{\mathrm{KL}}(\pi(\cdot \mid \mathbf{s}_t) \parallel p(\cdot \mid \mathbf{s}_t)),$$

with

$$\begin{aligned}Q^\pi(\mathbf{s}_t, \mathbf{a}_t, \mathcal{S}^\Omega; q_T) &\doteq r(\mathbf{s}_t, \mathbf{a}_t, \mathcal{S}^\Omega; q_\Delta) + q_{\Delta_{t+1}}(\Delta_{t+1} = 0)\, \mathbb{E}_{p_d(\mathbf{s}_{t+1} \mid \mathbf{s}_t, \mathbf{a}_t)}\Big[V^\pi(\mathbf{s}_{t+1}, \mathcal{S}^\Omega; q_T)\Big], \\ r(\mathbf{s}_t, \mathbf{a}_t, \mathcal{S}^\Omega; q_\Delta) &\doteq \log \mathbb{P}(\xi_{t+1}(\mathcal{S}^\Omega) = 1 \mid \mathbf{s}_t, \mathbf{a}_t) - q_{\Delta_{t+1}}(\Delta_{t+1} = 1) \mathbb{D}_{KL}(q_{\Delta_{t+1}} \,\|\, p_{\Delta_{t+1}}),\end{aligned}$$

and since $\mathbb{D}_{\mathrm{KL}}(q_{\Delta_{t+1}} \parallel p_{\Delta_{t+1}})$ is strictly convex in $q_{\Delta_{t+1}}(\Delta_{t+1} = 0)$, we can find the globally optimal Bernoulli distribution parameters $q_{\Delta_{t+1}}(\Delta_{t+1} = 0)$ for all $t \in \mathbb{N}_0$ recursively. That is, it is sufficient to solve the problem

$$q^\star_{\Delta_{t+1}}(\Delta_{t+1} = 0) \doteq \underset{q_{\Delta_{t+1}}(\Delta_{t+1}=0)}{\arg\max} \Big\{\mathcal{F}(\pi, q_T, \mathcal{S}^\Omega)\Big\} = \underset{q_{\Delta_{t+1}}(\Delta_{t+1}=0)}{\arg\max} \Big\{\mathcal{F}(\pi, q_{\Delta_1}, \ldots, q_{\Delta_{t+1}}, \ldots, \mathbf{s}_0, \mathcal{S}^\Omega)\Big\} \tag{B.69}$$

for a fixed $t+1$. To do so, we take the derivative of $\mathcal{F}(\pi, q_{\Delta_1}, \ldots, q_{\Delta_{t+1}}, \ldots, \mathbf{s}_0, \mathcal{S}^\Omega)$, which—defined recursively—is given by

$$\begin{aligned}&\mathbb{E}_{\pi(\mathbf{a}_t \mid \mathbf{s}_t)}\Big[Q(\mathbf{s}_t, \mathbf{a}_t, \mathcal{S}^\Omega; q_T\Big] - \mathbb{D}_{\mathrm{KL}}(\pi(\cdot \mid \mathbf{s}_t) \parallel p(\cdot \mid \mathbf{s}_t)) \\ =&\mathbb{E}_{\pi(\mathbf{a}_t \mid \mathbf{s}_t)}\Big[r(\mathbf{s}_t, \mathbf{a}_t, \mathcal{S}^\Omega; q_\Delta) + q_{\Delta_{t+1}}(\Delta_{t+1} = 0)\, \mathbb{E}_{p_d(\mathbf{s}_{t+1} \mid \mathbf{s}_t, \mathbf{a}_t)}\Big[V^\pi(\mathbf{s}_{t+1}, \mathcal{S}^\Omega; q_T)\Big]\Big] - \mathbb{D}_{\mathrm{KL}}(\pi(\cdot \mid \mathbf{s}_t) \parallel p(\cdot \mid \mathbf{s}_t))\end{aligned} \tag{B.70}$$

$$\begin{aligned}=&\mathbb{E}_{\pi(\mathbf{a}_t \mid \mathbf{s}_t)}\Bigg[\log \mathbb{P}(\xi_{t+1}(\mathcal{S}^\Omega) = 1 \mid \mathbf{s}_t, \mathbf{a}_t) - \mathbb{D}_{\mathrm{KL}}(q_{\Delta_{t+1}} \parallel p_{\Delta_{t+1}}) \\ &+ q_{\Delta_{t+1}}(\Delta_{t+1} = 0)\, \mathbb{E}_{p_d(\mathbf{s}_{t+1} \mid \mathbf{s}_t, \mathbf{a}_t)}\Big[V^\pi(\mathbf{s}_{t+1}, \mathcal{S}^\Omega; q_T)\Big]\Bigg] - \mathbb{D}_{\mathrm{KL}}(\pi(\cdot \mid \mathbf{s}_t) \parallel p(\cdot \mid \mathbf{s}_t))\end{aligned} \tag{B.71}$$

$$\begin{aligned}=&\mathbb{E}_{\pi(\mathbf{a}_t \mid \mathbf{s}_t)}\Bigg[\log \mathbb{P}(\xi_{t+1}(\mathcal{S}^\Omega) = 1 \mid \mathbf{s}_t, \mathbf{a}_t) - \mathbb{D}_{\mathrm{KL}}(q_{\Delta_{t+1}} \parallel p_{\Delta_{t+1}}) \\ &+ q_{\Delta_{t+1}}(\Delta_{t+1} = 0)\, \mathbb{E}_{p_d(\mathbf{s}_{t+1} \mid \mathbf{s}_t, \mathbf{a}_t)}\Big[V^\pi(\mathbf{s}_{t+1}, \mathcal{S}^\Omega; q_T)\Big]\Bigg] - \mathbb{D}_{\mathrm{KL}}(\pi(\cdot \mid \mathbf{s}_t) \parallel p(\cdot \mid \mathbf{s}_t)),\end{aligned} \tag{B.72}$$

with respect to $q_{\Delta_{t+1}}(\Delta_{t+1} = 0)$ and set it to zero, which yields

$$\begin{aligned}0 = &- \mathbb{E}_{\pi(\mathbf{a}_t \mid \mathbf{s}_t)}\Big[\mathbb{E}_{\pi(\mathbf{a}_{t+1} \mid \mathbf{s}_{t+1}) p_d(\mathbf{s}_{t+1} \mid \mathbf{s}_t, \mathbf{a}_t)}[Q^\pi(\mathbf{s}_{t+1}, \mathbf{a}_{t+1}, \mathcal{S}^\Omega; q_T)]\Big] \\ &+ \log \frac{1 - q^\star_{\Delta_{t+1}}(\Delta_{t+1} = 0)}{1 - p_{\Delta_{t+1}}(\Delta_{t+1} = 0)} - \log \frac{q^\star_{\Delta_{t+1}}(\Delta_{t+1} = 0)}{p_{\Delta_{t+1}}(\Delta_{t+1} = 0)}.\end{aligned} \tag{B.73}$$

Rearranging, we get

$$\frac{q^\star_{\Delta_{t+1}}(\Delta_{t+1} = 0)}{1 - q^\star_{\Delta_{t+1}}(\Delta_{t+1} = 0)} = \exp\left(\mathbb{E}[Q^\pi(\mathbf{s}_{t+1}, \mathbf{a}_{t+1}, \mathcal{S}^\Omega; q_T)] + \log \frac{p_{\Delta_{t+1}}(\Delta_{t+1} = 0)}{1 - p_{\Delta_{t+1}}(\Delta_{t+1} = 0)}\right), \tag{B.74}$$

where the expectation is taken with respect to $\pi(\mathbf{a}_{t+1} \mid \mathbf{s}_{t+1}) p_d(\mathbf{s}_{t+1} \mid \mathbf{s}_t, \mathbf{a}_t) \pi(\mathbf{a}_t \mid \mathbf{s}_t)$ and the Q-function depends on $q(\Delta_{t'})$ with $t' > t$, but not on $q^\star_{\Delta_{t+1}}(\Delta_{t+1} = 0)$. Solving for $q^\star_{\Delta_{t+1}}(\Delta_{t+1} = 0)$. Solving for $q^\star_{\Delta_{t+1}}(\Delta_{t+1} = 0)$, we obtain

$$q^\star_{\Delta_{t+1}}(\Delta_{t+1} = 0) = \frac{\exp(\mathbb{E}_{p_{\pi p_d} \pi(\mathbf{a}_t \mid \mathbf{s}_t)}[Q^\pi(\mathbf{s}_{t+1}, \mathbf{a}_{t+1}, \mathcal{S}^\Omega; q_T)] + \log \frac{p_{\Delta_{t+1}}(\Delta_{t+1}=0)}{1-p_{\Delta_{t+1}}(\Delta_{t+1}=0)})}{1 + \exp(\mathbb{E}_{p_{\pi p_d} \pi(\mathbf{a}_t \mid \mathbf{s}_t)}[Q^\pi(\mathbf{s}_{t+1}, \mathbf{a}_{t+1}, \mathcal{S}^\Omega; q_T)] + \log \frac{p_{\Delta_{t+1}}(\Delta_{t+1}=0)}{1-p_{\Delta_{t+1}}(\Delta_{t+1}=0)})} \tag{B.75}$$

$$= \sigma\Big(\mathbb{E}_{p_{\pi p_d}}[Q^\pi(\mathbf{s}_{t+1}, \mathbf{a}_{t+1}, \mathcal{S}^\Omega; q_T)] + \sigma^{-1}\left(p_{\Delta_{t+1}}(\Delta_{t+1} = 0)\right)\Big), \tag{B.76}$$

where $p_{\pi p_d} \doteq \pi(\mathbf{a}_{t+1} \mid \mathbf{s}_{t+1}) p_d(\mathbf{s}_{t+1} \mid \mathbf{s}_t, \mathbf{a}_t)$, $\sigma(\cdot)$ is the sigmoid function with $\sigma(x) = \frac{1}{e^{-x}+1}$ and $\sigma^{-1}(x) = \log \frac{x}{1-x}$. This concludes the proof. □

Remark 1. *As can be seen from Proposition 2 (Optimal Variational Distribution over T), the optimal approximation to the posterior over T trades off short-term rewards via $\mathbb{E}_{\pi(\mathbf{a}_t \mid \mathbf{s}_t)}[r(\mathbf{s}_t, \mathbf{a}_t, \mathcal{S}^\Omega; q_\Delta)]$, long-term rewards via $\mathbb{E}_{\pi(\mathbf{a}_{t+1} \mid \mathbf{s}_{t+1}) p_d(\mathbf{s}_{t+1} \mid \mathbf{s}_t, \mathbf{a}_t)}[Q^\pi(\mathbf{s}_{t+1}, \mathbf{a}_{t+1}, \mathcal{S}^\Omega; q_T)]$, and the prior log-odds of not achieving the outcome at a given point in time conditioned on the outcome not having been achieved yet, $\frac{p_{\Delta_{t+1}}(\Delta_{t+1}=0)}{1-p_{\Delta_{t+1}}(\Delta_{t+1}=0)}$.*

B.4 Dynamic-Discount Behavior-Driven Policy Iteration

Theorem 2 (Variational Dynamic-Discount Behavior-Driven Policy Iteration)**.** *Assume $|\mathcal{A}| < \infty$ and that the MDP is ergodic.*

1. *Dynamic-Discount Behavior-Driven Policy Evaluation (D2BD-PE): Given policy π and a function $Q^0 : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to \mathbb{R}$, define $Q^{i+1} = \mathcal{T}^\pi Q^i$. Then the sequence Q^i converges to the lower bound in Theorem 1 (Dynamic-Discount Behavior-Driven RL as Variational Inference).*

2. *Dynamic-Discount Behavior-Driven Policy Improvement (D2BD-PI): The policy*

$$\pi^+ = \arg\max_{\pi' \in \Pi} \left\{ \mathbb{E}_{\pi'(\mathbf{a}_t \mid \mathbf{s}_t)} \left[Q^\pi(\mathbf{s}_t, \mathbf{a}_t, \mathcal{S}^\Omega; q_T) \right] - \mathbb{D}_{\mathrm{KL}}(\pi'(\cdot \mid \mathbf{s}_t) \,||\, p(\cdot \mid \mathbf{s}_t)) \right\} \tag{B.77}$$

and the variational distribution over T recursively defined in terms of

$$\begin{aligned} &q^+(\Delta_{t+1} = 0; \pi, Q^\pi) \\ &= \sigma\Big(\mathbb{E}_{\pi(\mathbf{a}_{t+1} \mid \mathbf{s}_{t+1}) p_d(\mathbf{s}_{t+1} \mid \mathbf{s}_t, \mathbf{a}_t)}[Q^\pi(\mathbf{s}_{t+1}, \mathbf{a}_{t+1}, \mathcal{S}^\Omega; q_T)] + \sigma^{-1}\left(p_{\Delta_{t+1}}(\Delta_{t+1} = 0)\right) \Big) \end{aligned} \tag{B.78}$$

improve the variational objective. In other words, we have that $V^{\pi^+}(\mathbf{s}_0, \mathcal{S}^\Omega; q_T) \geq V^\pi(\mathbf{s}_0, \mathcal{S}^\Omega; q_T)$ and $V^\pi(\mathbf{s}_0, \mathcal{S}^\Omega; q_T^+) \geq V^\pi(\mathbf{s}_0, \mathcal{S}^\Omega; q_T)$ for all $\mathbf{s}_0 \in \mathcal{S}$.

3. *Alternating between D2BD-PE and D2BD-PI converges to a policy $\pi^\star$ and a variational distribution over T, $q_T^\star$, such that $Q^{\pi^\star}(\mathbf{s}, \mathbf{a}, \mathcal{S}^\Omega; q_T^\star) \geq Q^\pi(\mathbf{s}, \mathbf{a}, \mathcal{S}^\Omega; q_T)$ for all $(\pi, q_T) \in \Pi \times \mathcal{Q}_T$ and any $(\mathbf{s}, \mathbf{a}) \in \mathcal{S} \times \mathcal{A}$.*

Proof. Parts of this proof are adapted from the proof given in Haarnoja et al. [1], modified for the Bellman operator proposed in Definition 1.

1. Dynamic-Discount Behavior-Driven Policy Evaluation (D2BD-PE): Instead of absorbing the entropy term into the Q-function, we can define an entropy-augmented reward as

$$\begin{aligned} r^\pi(\mathbf{s}_t, \mathbf{a}_t, \mathcal{S}^\Omega; q_\Delta) \doteq\ &\log \mathbb{P}(\xi_{t+1}(\mathcal{S}^\Omega) = 1 \,|\, \mathbf{s}_t, \mathbf{a}_t) - \mathbb{D}_{\mathrm{KL}}(q_{\Delta_{t+1}} \parallel p_{\Delta_{t+1}}) \\ &+ q_{\Delta_{t+1}}(\Delta_{t+1} = 0)\, \mathbb{E}_{p_d(\mathbf{s}_{t+1} \mid \mathbf{s}_t, \mathbf{a}_t)}[\mathbb{D}_{\mathrm{KL}}(\pi(\cdot \mid \mathbf{s}_{t+1}) \parallel p(\cdot \mid \mathbf{s}_{t+1}))]. \end{aligned} \tag{B.79}$$

   We can then write an update rule according to Definition 1 as

$$\tilde{Q}(\mathbf{s}_t, \mathbf{a}_t, \mathcal{S}^\Omega; q_T) \leftarrow r^\pi(\mathbf{s}_t, \mathbf{a}_t, \mathcal{S}^\Omega; q_\Delta) + q_{\Delta_{t+1}}(\Delta_{t+1} = 0)\, \mathbb{E}_{\pi(\mathbf{a}_{t+1} \mid \mathbf{s}_{t+1}) p_d(\mathbf{s}_{t+1} \mid \mathbf{s}_t, \mathbf{a}_t)}[\tilde{Q}(\mathbf{s}_{t+1}, \mathbf{a}_{t+1}, \mathcal{S}^\Omega; q_T)], \tag{B.80}$$

   where $q_{\Delta_{t+1}}(\Delta_{t+1} = 0) \leq 1$. This update is similar to a Bellman update [10], but with a discount factor given by $q_{\Delta_{t+1}}(\Delta_{t+1} = 0)$. In general, this discount factor $q_{\Delta_{t+1}}(\Delta_{t+1} = 0)$ can be computed dynamically based on the current state and action, such as in Equation (B.67). As discussed in White [12], this Bellman operator is still a contraction mapping so long as the Markov chain induced by the current policy is ergodic and there exists a state such that $q_{\Delta_{t+1}}(\Delta_{t+1} = 0) < 1$. The first condition is true by assumption. The second condition is true since $q_{\Delta_{t+1}}(\Delta_{t+1} = 0)$ is given by Equation (B.67), which is always strictly between 0 and 1. Therefore, we apply convergence results for policy evaluation with transition-dependent discount factors [12] to this contraction mapping, and the result immediately follows.

2. Dynamic-Discount Behavior-Driven Policy Improvement (D2BD-PI): Let $\pi_{\text{old}} \in \Pi$ and let $Q^{\pi_{\text{old}}}$ and $V^{\pi_{\text{old}}}$ be the behavior-driven state and state-action value functions from Definition 1, let q_T be some variational distribution over T, and let π_{new} be given by

$$\pi_{\text{new}}(\mathbf{a}_t|\mathbf{s}_t) = \arg\max_{\pi' \in \Pi} \left\{ \mathbb{E}_{\pi'(\mathbf{a}_t \mid \mathbf{s}_t)} \left[Q^{\pi_{\text{old}}}(\mathbf{s}_t, \mathbf{a}_t, \mathcal{S}^\Omega; q_T) \right] - \mathbb{D}_{\mathrm{KL}}(\pi'(\cdot \mid \mathbf{s}_t) \parallel p(\cdot \mid \mathbf{s}_t)) \right\} \tag{B.81}$$

$$= \arg\max_{\pi' \in \Pi} \mathcal{J}_{\pi_{\text{old}}}(\pi'(\mathbf{a}_t, \mathbf{s}_t), q_T). \tag{B.82}$$

   Then, it must be true that $\mathcal{J}_{\pi_{\text{old}}}(\pi_{\text{old}}(\mathbf{a}_t|\mathbf{s}_t); q_T) \leq \mathcal{J}_{\pi_{\text{old}}}(\pi_{\text{new}}(\mathbf{a}_t|\mathbf{s}_t); q_T)$, since one could set $\pi_{\text{new}} = \pi_{\text{old}} \in \Pi$. Thus,

$$\begin{aligned} &\mathbb{E}_{\pi_{\text{new}}(\mathbf{a}_t|\mathbf{s}_t)} \left[Q^{\pi_{\text{old}}}(\mathbf{s}_t, \mathbf{a}_t, \mathcal{S}^\Omega; q_T) \right] - \mathbb{D}_{\mathrm{KL}}(\pi_{\text{new}}(\cdot \mid \mathbf{s}_t) \parallel p(\cdot \mid \mathbf{s}_t)) \\ &\quad \geq \mathbb{E}_{\pi_{\text{old}}(\mathbf{a}_t|\mathbf{s}_t)} \left[Q^{\pi_{\text{old}}}(\mathbf{s}_t, \mathbf{a}_t, \mathcal{S}^\Omega; q_T) \right] - \mathbb{D}_{\mathrm{KL}}(\pi_{\text{old}}(\cdot \mid \mathbf{s}_t) \parallel p(\cdot \mid \mathbf{s}_t)), \end{aligned} \tag{B.83}$$

and since

$$V^{\pi_{\text{old}}}(\mathbf{s}_t, \mathcal{S}^{\Omega}; q_T) = \mathbb{E}_{\pi_{\text{old}}(\mathbf{a}_t|\mathbf{s}_t)}\left[Q^{\pi_{\text{old}}}(\mathbf{s}_t, \mathbf{a}_t, \mathcal{S}^{\Omega}; q_T)\right] - \mathbb{D}_{\text{KL}}(\pi_{\text{old}}(\cdot \,|\, \mathbf{s}_t) \parallel p(\cdot \,|\, \mathbf{s}_t)), \tag{B.84}$$

we get

$$\mathbb{E}_{\pi_{\text{new}}(\mathbf{a}_t|\mathbf{s}_t)}\left[Q^{\pi_{\text{old}}}(\mathbf{s}_t, \mathbf{a}_t, \mathcal{S}^{\Omega}; q_T)\right] - \mathbb{D}_{\text{KL}}(\pi_{\text{new}}(\cdot \,|\, \mathbf{s}_t) \parallel p(\cdot \,|\, \mathbf{s}_t)) \geq V^{\pi_{\text{old}}}(\mathbf{s}_t, \mathcal{S}^{\Omega}; q_T). \tag{B.85}$$

We can now write the Bellman equation as

$$\begin{aligned}
&Q^{\pi_{\text{old}}}(\mathbf{s}_t, \mathbf{a}_t, \mathcal{S}^{\Omega}; q_T) \\
&= q_{\Delta_{t+1}}(\Delta_{t+1} = 1) \log p_d(\mathbf{g} \,|\, \mathbf{s}_t, \mathbf{a}_t) + q_{\Delta_{t+1}}(\Delta_{t+1} = 0)\, \mathbb{E}_{p_d(\mathbf{s}_{t+1} \,|\, \mathbf{s}_t, \mathbf{a}_t)}[V^{\pi_{\text{old}}}(\mathbf{s}_{t+1}, \mathcal{S}^{\Omega}; q_T)] && \text{(B.86)} \\
&\leq q_{\Delta_{t+1}}(\Delta_{t+1} = 1) \log p_d(\mathbf{g} \,|\, \mathbf{s}_t, \mathbf{a}_t) + q_{\Delta_{t+1}}(\Delta_{t+1} = 0)\, \mathbb{E}_{p(\mathbf{s}_{t'}|\mathbf{s}_t, \mathbf{a}_t)}[\mathbb{E}_{\pi_{\text{new}}(\mathbf{a}_{t'}|\mathbf{s}_{t'})}\left[Q^{\pi_{\text{old}}}(\mathbf{s}_{t'}, \mathbf{a}_{t'}, \mathcal{S}^{\Omega}; q_T)\right] \\
&\qquad - \mathbb{D}_{\text{KL}}(\pi_{\text{new}}(\cdot \,|\, \mathbf{s}_{t'}) \parallel p(\cdot \,|\, \mathbf{s}_{t'}))], && \text{(B.87)} \\
&\vdots \\
&\leq Q^{\pi_{\text{new}}}(\mathbf{s}_t, \mathbf{a}_t, \mathcal{S}^{\Omega}; q_T) && \text{(B.88)}
\end{aligned}$$

where we defined $t' \doteq t + 1$, repeatedly applied the Bellman backup operator defined in Definition 1 and used the bound in Equation (B.85). Convergence follows from Dynamic-Discount Behavior-Driven Policy Evaluation above.

3. Locally Optimal Variational Dynamic-Discount Behavior-Driven Policy Iteration: Define π^i to be a policy at iteration i. By ODPI for a given q_T, the sequence of state-action value functions $\{Q^{\pi^i}(q_T)\}_{i=1}^{\infty}$ is monotonically increasing in i. Since the reward is finite and the negative KL divergence is upper bounded by zero, $Q^{\pi}(q_T)$ is upper bounded for $\pi \in \Pi$ and the sequence $\{\pi^i\}_{i=1}^{\infty}$ converges to some $\pi^\star$. To see that $\pi^\star$ is an optimal policy, note that it must be the case that $\mathcal{J}_{\pi^\star}(\pi^\star(\mathbf{a}_t|\mathbf{s}_t); q_T) > \mathcal{J}_{\pi^\star}(\pi(\mathbf{a}_t \,|\, \mathbf{s}_t); q_T)$ for any $\pi \in \Pi$ with $\pi \neq \pi^\star$. By the argument used in ODPI above, it must be the case that the behavior-driven state-action value of the converged policy is higher than that of any other non-converged policy in Π, that is, $Q^{\pi^\star}(\mathbf{s}_t, \mathbf{a}_t; q_T) > Q^{\pi}(\mathbf{s}_t, \mathbf{a}_t; q_T)$ for all $\pi \in \Pi$ and any $q_T^i \in \mathcal{Q}_T$ and $(\mathbf{s}, \mathbf{a}) \in \mathcal{S} \times \mathcal{A}$. Therefore, given q_T, $\pi^\star$ must be optimal in Π, which concludes the proof.

4. Globally Optimal Variational Dynamic-Discount Behavior-Driven Policy Iteration: Let π^i be a policy and let q_T^i be variational distributions over T at iteration i. By Locally Optimal Variational Dynamic-Discount Behavior-Driven Policy Iteration, for a *fixed* q_T^i with $q_T^i = q_T^j \forall i, j \in \mathbb{N}_0$, the sequence of $\{(\pi^i, q_T^i)\}_{i=1}^{\infty}$ increases the objective Equation (B.24) at each iteration and converges to a stationary point in π^i, where $Q^{\pi^\star}(\mathbf{s}_t, \mathbf{a}_t; q_T^i) > Q^{\pi}(\mathbf{s}_t, \mathbf{a}_t; q_T^i)$ for all $\pi \in \Pi$ and any $q_T^i \in \mathcal{Q}_T$ and $(\mathbf{s}, \mathbf{a}) \in \mathcal{S} \times \mathcal{A}$. Since the objective in Equation (B.24) is concave in q_T, it must be the case that for, $q_T^{\star^i} \in \mathcal{Q}_T$, the optimal variational distribution over T at iteration i, defined recursively by

$$\begin{aligned}
&q^{\star^i}(\Delta_{t+1} = 0; \pi^i, Q^{\pi^i}) \\
&= \sigma\Big(\mathbb{E}_{\pi(\mathbf{a}_{t+1} \,|\, \mathbf{s}_{t+1}) p_d(\mathbf{s}_{t+1} \,|\, \mathbf{s}_t, \mathbf{a}_t)}[Q^{\pi^i}(\mathbf{s}_{t+1}, \mathbf{a}_{t+1}, \mathcal{S}^{\Omega}; q_T(\pi^i, Q^{\pi^i}))] + \sigma^{-1}(p_{\Delta_{t+1}}(\Delta_{t+1} = 0))\Big),
\end{aligned}$$

for $t \in \mathbb{N}_0$, $Q^{\pi}(\mathbf{s}_t, \mathbf{a}_t; q_T^\star) > Q^{\pi}(\mathbf{s}_t, \mathbf{a}_t; q_T)$ for all $\pi \in \Pi$ and any $(\mathbf{s}, \mathbf{a}) \in \mathcal{S} \times \mathcal{A}$. Note that q_T is defined implicitly in terms of π^i and Q^{π^i}, that is, the optimal variational distribution over T at iteration i is defined as a function of the policy and Q-function at iteration i. Hence, it must then be true that for $Q^{\pi^\star}(\mathbf{s}_t, \mathbf{a}_t; q_T^\star) > Q^{\pi^\star}(\mathbf{s}_t, \mathbf{a}_t; q_T)$ for all $q_T^\star(\pi^\star, Q^{\pi^\star}) \in \mathcal{Q}_T$ and for any $\pi^\star \in \Pi$ and $(\mathbf{s}, \mathbf{a}) \in \mathcal{S} \times \mathcal{A}$. In other words, for an optimal policy and corresponding Q-function, there exists an optimal variational distribution over T that maximizes the Q-function, given the optimal policy. Repeating locally optimal variational behavior-driven policy iteration under the new variational distribution $q_T^\star(\pi^\star, Q^{\pi^\star})$ will yield an optimal policy $\pi^{\star\star}$ and computing the corresponding optimal variational distribution, $q_T^{\star\star}(\pi^{\star\star}, Q^{\pi^{\star\star}})$ will further increase the variational objective such that for $\pi^{\star\star}) \in \Pi$ and $q_T^{\star\star}(\pi^{\star\star}, Q^{\pi^{\star\star}}) \in \mathcal{Q}_T$, we have that

$$Q^{\pi^{\star\star}}(\mathbf{s}_t, \mathbf{a}_t; q_T^{\star\star}) > Q^{\pi^{\star\star}}(\mathbf{s}_t, \mathbf{a}_t; q_T^\star) > Q^{\pi^\star}(\mathbf{s}_t, \mathbf{a}_t; q_T^\star) > Q^{\pi^\star}(\mathbf{s}_t, \mathbf{a}_t; q_T) \tag{B.89}$$

for any $\pi^\star \in \Pi$ and $(\mathbf{s}, \mathbf{a}) \in \mathcal{S} \times \mathcal{A}$. Hence, global optimal variational behavior-driven policy iteration increases the variational objective at every step. Since the objective is upper bounded (by virtue of the rewards being finite and the negative KL divergence being upper bounded by zero) and the sequence of $\{(\pi^i, q_T^i)\}_{i=1}^{\infty}$ increases the objective Equation (B.24) at each iteration, by the monotone convergence theorem, the objective value converges to a supremum and since the objective function is concave the

supremum is unique. Hence, since the supremum is unique and obtained via global optimal variational outcome–driven policy iteration on $(\pi, q_T) \in \Pi \times \mathcal{Q}_T$, the sequence of $\{(\pi^i, q_T^i)\}_{i=1}^{\infty}$ converges to a unique stationary point $(\pi^\star, q_T^\star) \in \Pi \times \mathcal{Q}_T$, where $Q^{\pi^\star}(\mathbf{s}_t, \mathbf{a}_t; q_T^\star) > Q^{\pi}(\mathbf{s}_t, \mathbf{a}_t; q_T^i)$ for all $\pi \in \Pi$ and any $q_T^i \in \mathcal{Q}_T$ and $(\mathbf{s}, \mathbf{a}) \in \mathcal{S} \times \mathcal{A}$.

□

Corollary 3 (Optimality of Variational Outcome Driven Policy Iteration)**.** *Variational Dynamic-Discount Behavior-Driven Policy Iteration on $(\pi, q_T) \in \Pi \times \mathcal{Q}_T$ results in an optimal policy at least as good or better than any optimal policy attainable from policy iteration on $\pi \in \Pi$ alone.*

Remark 2. *The convergence proof of ODPE assumes a transition-dependent discount factor [12], because the variational distribution used in Equation (B.67) depends on the next state and action as well as on the desired outcome.*

B.5 Lemmas

Lemma 1. *Let $q(T = t) \doteq q(T = t|T \geq t) \prod_{i=1}^{t} q(T \neq i-1|T \geq i-1)$ be a discrete probability distribution with support $\mathbb{N}_0$. Then for any $t \in \mathbb{N}_0$, we have that*

$$q(T \geq t) = \sum_{i=t}^{\infty} q(T = i|T \geq i) \prod_{j=1}^{i} q(T \neq j-1|T \geq j-1) = \prod_{i=1}^{t} q(T \neq i-1|T \geq i-1). \tag{B.90}$$

Proof. We proof the statement by induction on t.

Base case: For $t = 0$, $q(T \geq 0) = 1$ by definition of the empty product.

Inductive case: Note that $q(T \leq t) = \prod_{i=1}^{t} q(T = i-1|T \geq i-1)$. Show that

$$q(T \geq t) = \prod_{i=1}^{t} q(T \neq i-1|T \geq i-1) \Longrightarrow q(T \geq t+1) = \prod_{i=1}^{t+1} q(T \neq i-1|T \geq i-1). \tag{B.91}$$

Consider $q(T \geq t+1) = \sum_{i=t+1}^{\infty} q(T = i|T \geq i) \prod_{j=1}^{i} q(T \neq j-1|T \geq j-1)$. To proof the inductive hypothesis, we need to show that the following equality is true:

$$\sum_{i=t+1}^{\infty} q(T = i|T \geq i) \prod_{j=1}^{i} q(T \neq j-1|T \geq j-1) = \prod_{i=1}^{t+1} q(T \neq i-1|T \geq i-1) \tag{B.92}$$

$$\begin{aligned} \Longleftrightarrow \sum_{i=t}^{\infty} q(T = i|T \geq i) \prod_{j=1}^{i} q(T \neq j-1|T \geq j-1) - q(T = t|T \geq t) \prod_{j=1}^{t} q(T \neq j-1|T \geq j-1) \\ = q(T \neq t|T \geq t) \prod_{i=1}^{t} q(T \neq i-1|T \geq i-1). \end{aligned} \tag{B.93}$$

By the inductive hypothesis,

$$q(T \geq t) = \sum_{i=t}^{\infty} q(T = i|T \geq i) \prod_{j=1}^{i} q(T \neq j-1|T \geq j-1) = \prod_{i=1}^{t} q(T \neq i-1|T \geq i-1), \tag{B.94}$$

and so

$$\text{Equation (B.93)} \Longleftrightarrow \prod_{j=1}^{t} q(T \neq j|T \geq j) - q(T \neq t+1|T \geq t+1 \tag{B.95}$$

$$\cdot \prod_{j=1}^{t} q(T = j|T \geq j) = q(T \neq t|T \geq t) \prod_{i=1}^{t} q(T \neq i-1|T \geq i-1). \tag{B.96}$$

Factoring out $\prod_{i=1}^{t} q(T \neq i-1 | T \geq i-1)$, we get

$$\iff \prod_{j=1}^{t} q(T \neq j-1 | T \geq j-1) \underbrace{(1 - q(T = t | T \geq t))}_{= q(T \neq t | T \geq t)} = q(T \neq t | T \geq t) \prod_{j=1}^{t} q(T = j-1 | T \geq j-1) \tag{B.97}$$

$$\iff q(T \neq t | T \geq t) \prod_{j=1}^{t} q(T \neq j-1 | T \geq j-1) = q(T \neq t | T \geq t) \prod_{j=1}^{t} q(T \neq j-1 | T \geq j-1), \tag{B.98}$$

which proves the inductive hypothesis. □

Lemma 2. *Let $q_T(t)$ and $p_T(t)$ be discrete probability distributions with support $\mathbb{N}_0$, let Δ_t be a Bernoulli random variable, with success defined as $T = t+1$ given that $T \geq t$, and let q_{Δ_t} be a discrete probability distribution over Δ_t for $t \in \mathbb{N} \backslash \{0\}$, so that*

$$\begin{aligned} q_{\Delta_{t+1}}(\Delta_{t+1} = 0) &\doteq q(T \neq t \,|\, T \geq t) \\ q_{\Delta_{t+1}}(\Delta_{t+1} = 1) &\doteq q(T = t \,|\, T \geq t). \end{aligned} \tag{B.99}$$

Then we can write $q(T = t) = q_{\Delta_{t+1}}(\Delta_{t+1} = 1) \prod_{i=1}^{t} q_{\Delta_i}(\Delta_i = 0)$ for any $t \in \mathbb{N}_0$ and have that

$$q(T \geq t) = \sum_{i=t}^{\infty} q_{\Delta_{i+1}}(\Delta_{i+1} = 1) \prod_{j=1}^{i} q_{\Delta_j}(\Delta_j = 0) = \prod_{i=1}^{t} q_{\Delta_i}(\Delta_i = 0). \tag{B.100}$$

Proof. By Lemma 1, we have that for any $t \in \mathbb{N}_0$

$$q(T \geq t) = \sum_{i=t}^{\infty} q(T = i | T \geq i) \prod_{j=1}^{i} q(T \neq j-1 | T \geq j-1) = \prod_{i=1}^{t} q(T \neq i-1 | T \geq i-1). \tag{B.101}$$

The result follows by replacing $q(T = i | T \geq i)$ by $q_{\Delta_{i+1}}(\Delta_{i+1} = 1)$, $q(T \neq j-1 | T \geq j-1)$ by $q_{\Delta_j}(\Delta_j = 0)$, and $q(T \neq i-1 | T \geq i-1)$ by $q_{\Delta_i}(\Delta_i = 0)$. □

Lemma 3. *Let $q_T(t)$ and $p_T(t)$ be discrete probability distributions with support $\mathbb{N}_0$. Then for any $k \in \mathbb{N}_0$,*

$$\begin{aligned} &\mathbb{E}_{t \sim q(T \,|\, T \geq k)} \left[\log \frac{q(T = t \,|\, T \geq k)}{p(T = t \,|\, T \geq k)} \right] \\ &\quad = f(q, p, k) + q(T \neq k \,|\, T \geq k)\, \mathbb{E}_{t \sim q(T \,|\, T \geq k+1)} \left[\log \frac{q(T = t \,|\, T \geq k+1)}{p(T = t \,|\, T \geq k+1)} \right]. \end{aligned} \tag{B.102}$$

Proof. Consider $\mathbb{E}_{t \sim q(T \,|\, T \geq k)} \left[\log \frac{q(T=t \,|\, T \geq k)}{p(T=t \,|\, T \geq k)} \right]$ and note that by the law of total expectation we can rewrite it as

$$\begin{aligned} &\mathbb{E}_{t \sim q(T \,|\, T \geq k)} \left[\log \frac{q(T = t \,|\, T \geq k)}{p(T = t \,|\, T \geq k)} \right] \\ &= q(T = k \,|\, T \geq k)\, \mathbb{E}_{t \sim q(T \,|\, T = k)} \left[\log \frac{q(T = t \,|\, T \geq k)}{p(T = t \,|\, T \geq k)} \right] \\ &\quad + q(T \neq k \,|\, T \geq k)\, \mathbb{E}_{t \sim q(T \,|\, T \geq k+1)} \left[\log \frac{q(T = t \,|\, T \geq k)}{p(T = t \,|\, T \geq k)} \right] \end{aligned} \tag{B.103}$$

$$= q(T = k \,|\, T \geq k) \log \frac{q(T = k \,|\, T \geq k)}{p(T = k \,|\, T \geq k)} + q(T \neq k \,|\, T \geq k)\, \mathbb{E}_{t \sim q(T | T \geq k+1)} \left[\log \frac{q(T = t \,|\, T \geq k)}{p(T = t \,|\, T \geq k)} \right]. \tag{B.104}$$

For all values of $T \geq k+1$, we have that

$$q(T = t \,|\, T \geq k) = q(T = t \,|\, T \geq k+1) q(T \neq k \,|\, T \geq k) \tag{B.105}$$

$$p(T = t \,|\, T \geq k) = p(T = t \,|\, T \geq k+1) p(T \neq k \,|\, T \geq k) \tag{B.106}$$

and so we can rewrite the expectation in Equation (B.104) as

$$\mathbb{E}_{t \sim q(T \,|\, T \geq k+1)} \left[\log \frac{q(T = t \,|\, T \geq k)}{p(T = t \,|\, T \geq k)} \right] = \mathbb{E}_{t \sim q(T \,|\, T \geq k+1)} \left[\log \frac{q(T = t \,|\, T \geq k)}{p(T = t \,|\, T \geq k)} + \log \frac{q(T \neq k \,|\, T \geq k)}{p(T \neq k \,|\, T \geq k)} \right] \tag{B.107}$$

$$= \mathbb{E}_{t \sim q(T \,|\, T \geq k+1)} \left[\log \frac{q(T = t \,|\, T \geq k)}{p(T = t \,|\, T \geq k)} \right] + \log \frac{q(T \neq k \,|\, T \geq k)}{p(T \neq k \,|\, T \geq k)} \tag{B.108}$$

Combining Equation (B.108) with Equation (B.104), we have

$$\begin{aligned}
&\mathbb{E}_{t\sim q(T\,|\,T\geq k)}\left[\log\frac{q(T=t\,|\,T\geq k)}{p(T=t\,|\,T\geq k)}\right] \\
&\quad=\underbrace{q(T=k\,|\,T\geq k)\log\frac{q(T=k\,|\,T\geq k)}{p(T=k\,|\,T\geq k)}+q(T\neq k\,|\,T\geq k)\log\frac{q(T\neq k\,|\,T\geq k)}{p(T\neq k\,|\,T\geq k)}}_{\doteq\, f(q,p,k)} \\
&\qquad+q(T\neq k\,|\,T\geq k)\,\mathbb{E}_{t\sim q(T\,|\,T\geq k+1)}\left[\log\frac{q(T=t\,|\,T\geq k+1)}{p(T=t\,|\,T\geq k+1)}\right],
\end{aligned} \tag{B.109}$$

which concludes the proof. □

Lemma 4. *Let $q_T(t)$ and $p_T(t)$ be discrete probability distributions with support $\mathbb{N}_0$. Then the KL divergence from q_T to p_T can be written as*

$$\mathbb{D}_{\mathrm{KL}}(q_T\,||\,p_T)=\sum_{t=0}^{\infty}q(T\geq t)f(q_T,p_T,t) \tag{B.110}$$

where $f(q_T,p_T,t)$ is shorthand for

$$f(q_T,p_T,t)=q(T=t\,|\,T\geq t)\log\frac{q(T=t\,|\,T\geq t)}{p(T=t\,|\,T\geq t)}+q(T\neq t\,|\,T\geq t)\log\frac{q(T\neq t\,|\,T\geq t)}{p(T\neq t\,|\,T\geq t)}. \tag{B.111}$$

Proof. Note that $q(T=k)$ denotes the probability that the distribution q assigns to the event $T=k$ and $q(T\geq m)$ denotes the tail probability, that is, $q(T\geq m)=\sum_{t=m}^{\infty}q(T=t)$. We will write $q(T|T\geq m)$ to denote the conditional distribution of q given $T\geq m$, that is, $q(T=k|T\geq m)=\mathbb{1}[k\geq m]q(T=k)/q(T\geq m)$. We will use analogous notation for p.

By the definition of the KL divergence and using the fact that, since the support is lowerbounded by $T=0$, $q(T=0)=q(T=0\,|\,T\geq 0)$, we have

$$\mathbb{D}_{\mathrm{KL}}(q_T\parallel p_T)=\mathbb{E}_{t\sim q(T)}\left[\log\frac{q(T=t)}{p(T=t)}\right]=\mathbb{E}_{t\sim q(T\,|\,T\geq 0)}\left[\log\frac{q(T=t\,|\,T\geq 0)}{p(T=t\,|\,T\geq 0)}\right]. \tag{B.112}$$

Using Lemma 3 with $k=0,1,2,3,\ldots$, we can expand the above expression to get

$$\mathbb{D}_{\mathrm{KL}}(q_T\parallel p_T)=f(q_T,p_T,0)+q(T\neq 0\,|\,T\geq 0)\,\mathbb{E}_{t\sim q(T\,|\,T\geq 1)}\left[\log\frac{q(T=t\,|\,T\geq 1)}{p(T=t\,|\,T\geq 1)}\right] \tag{B.113}$$

$$\begin{aligned}
&=f(q,p,0)+q(T\neq 0\,|\,T\geq 1)f(q_T,p_T,1) \\
&\qquad+q(T\neq 0\,|\,T\geq 0)q(T\neq 1\,|\,T\geq 1)\,\mathbb{E}_{t\sim q(T\,|\,T\geq 2)}\left[\log\frac{q(T=t\,|\,T\geq 2)}{p(T=t\,|\,T\geq 2)}\right]
\end{aligned} \tag{B.114}$$

$$\begin{aligned}
&=\underbrace{1}_{=q(T\geq 0)}\cdot f(q,p,0) \\
&\qquad+\underbrace{q(T\neq 0\,|\,T\geq 0)}_{=q(T\geq 1)}f(q,p,1) \\
&\qquad+\underbrace{q(T\neq 0\,|\,T\geq 0)q(T\neq 1\,|\,T\geq 1)}_{=q(T\geq 2)}f(q_T,p_T,2) \\
&\qquad+\underbrace{q(T\neq 0\,|\,T\geq 0)q(T\neq 1\,|\,T\geq 1)q(T\neq 2\,|\,T\geq 2)}_{=q(T\geq 3)}\,\mathbb{E}_{t\sim q(T\,|\,T\geq 3)}\left[\log\frac{q(T=t\,|\,T\geq 3)}{p(T=t\,|\,T\geq 3)}\right]
\end{aligned} \tag{B.115}$$

$$=\sum_{t=0}^{\infty}q(T\geq t)f(q_T,p_T,t), \tag{B.116}$$

where $f(q_T,p_T,t)$ is shorthand for

$$f(q_T,p_T,t)=q(T=t\,|\,T\geq t)\log\frac{q(T=t\,|\,T\geq t)}{p(T=t\,|\,T\geq t)}+q(T\neq t\,|\,T\geq t)\log\frac{q(T\neq t\,|\,T\geq t)}{p(T\neq t\,|\,T\geq t)}. \tag{B.117}$$

and we used the fact that, by Lemma 1,

$$q(T\geq t)=\prod_{k=1}^{t}q(T\neq k-1\,|\,T\geq k-1). \tag{B.118}$$

This completes the proof. □

Lemma 5. *Let $q_T(t)$ and $p_T(t)$ be discrete probability distributions with support $\mathbb{N}_0$, let Δ_t be a Bernoulli random variable, with success defined as $T = t$ given that $T \geq t$, and let q_{Δ_t} and p_{Δ_t} be discrete probability distributions over Δ_t for $t \in \mathbb{N}_0 \backslash \{0\}$, so that*

$$q_{\Delta_{t+1}}(\Delta_{t+1} = 0) \doteq q(T \neq t \,|\, T \geq t) \qquad q_{\Delta_{t+1}}(\Delta_{t+1} = 1) \doteq q(T = t \,|\, T \geq t) \tag{B.119}$$

$$p_{\Delta_{t+1}}(\Delta_{t+1} = 0) \doteq p(T \neq t \,|\, T \geq t) \qquad p_{\Delta_{t+1}}(\Delta_{t+1} = 1) \doteq p(T = t \,|\, T \geq t). \tag{B.120}$$

Then the KL divergence from q_T to p_T can be written as

$$\mathbb{D}_{\mathrm{KL}}(q_T \,||\, p_T) = \sum_{t=0}^{\infty} \Big(\prod_{k=1}^{t} q_{\Delta_t}(\Delta_t = 0) \Big) \mathbb{D}_{\mathrm{KL}}(q_{\Delta_{t+1}} \,||\, p_{\Delta_{t+1}}) \tag{B.121}$$

Proof. The result follows from Lemma 4, Equation (B.118), Equation (B.119), and the definition of f. In detail, from Lemma 1, and Equation (B.119) we have that

$$q(T \geq t) = \prod_{k=1}^{t} q(T \neq k-1 \,|\, T \geq k-1) = \prod_{k=1}^{t} q_{\Delta_k}(\Delta_k = 0). \tag{B.122}$$

From the definition of $f(q_T, p_T, t)$, we have

$$f(q_T, p_T, t) = q(T = t \,|\, T \geq t) \log \frac{q(T = t \,|\, T \geq t)}{p(T = t \,|\, T \geq t)} + q(T \neq t \,|\, T \geq t) \log \frac{q(T \neq t \,|\, T \geq t)}{p(T \neq t \,|\, T \geq t)} \tag{B.123}$$

$$= q_{\Delta_{t+1}}(\Delta_{t+1} = 0) \log \frac{q_{\Delta_{t+1}}(\Delta_{t+1} = 0)}{p_{\Delta_{t+1}}(\Delta_{t+1} = 0)} + q(\Delta_{t+1} = 1) \log \frac{q_{\Delta_{t+1}}(\Delta_{t+1} = 1)}{p_{\Delta_{t+1}}(\Delta_{t+1} = 1)} \tag{B.124}$$

$$= \mathbb{D}_{\mathrm{KL}}(q_{\Delta_{t+1}} \parallel p_{\Delta_{t+1}}). \tag{B.125}$$

Combining Equation (B.122), Equation (B.125), and Equation (B.110) completes the proof. □