

Greedification Operators for Policy Optimization: Investigating Forward and Reverse KL Divergences

Alan Chan

Hugo Silva

Sungsu Lim

Tadashi Kozuno

A. Rupam Mahmood

Martha White

Department of Computing Science, Alberta Machine Intelligence Institute (Amii)

University of Alberta

Edmonton, Alberta, Canada

ACHAN4@UALBERTA.CA

HUGOLUIS@UALBERTA.CA

SUNGSU@UALBERTA.CA

KOZUNO@UALBERTA.CA

ARMAHMOOD@UALBERTA.CA

WHITEM@UALBERTA.CA

Editor: George Konidaris

Abstract

Approximate Policy Iteration (API) algorithms alternate between (approximate) policy evaluation and (approximate) greedification. Many different approaches have been explored for approximate policy evaluation, but less is understood about approximate greedification and what choices guarantee policy improvement. In this work, we investigate approximate greedification when reducing the KL divergence between the parameterized policy and the Boltzmann distribution over action values. In particular, we investigate the difference between the forward and reverse KL divergences, with varying degrees of entropy regularization; these are chosen because they underlie many existing policy optimization approaches, as we highlight in this work. We show that the reverse KL has stronger policy improvement guarantees, and that reducing the forward KL can result in a worse policy. We also demonstrate, however, that a large enough reduction of the forward KL can induce improvement under additional assumptions. Empirically, we show on simple continuous-action environments that the forward KL can induce more exploration, but at the cost of a more suboptimal policy. No significant differences were observed in the discrete-action setting or on a suite of benchmark problems. This work provides novel theoretical and empirical insights about the forward KL and reverse KL for greedification, and clear next steps for understanding and improving our policy optimization algorithms.

Keywords: reinforcement learning, policy gradient, policy iteration, kl divergence

1. Introduction

A canonical approach to learn policies in reinforcement learning (RL) is Policy Iteration (PI). PI interleaves policy evaluation—understanding how a policy is currently performing by computing a value function—and policy improvement—making the current policy better based on the value function. The policy improvement step is sometimes called the *greedification* step, because typically the policy is set to a greedy policy. That is, the policy is set to take the action that maximizes the current action-value function, in each state. In the tabular setting, this procedure is guaranteed to result in iteratively better policies and

converge to the optimal policy (Bertsekas, 2019). The greedification step can also be soft, in that some probability is placed on all other actions. In certain such cases, like with entropy regularization, PI converges to the optimal soft policy (Geist et al., 2019).

Practically, however, it is not always feasible to perform each step to completion. Approximate PI (API) (Bertsekas, 2011; Scherrer, 2014) allows for each step to be done incompletely, and still maintain convergence guarantees. The agent can perform an approximate policy evaluation step, where it obtains an improved estimate of the values without achieving the true values. The agent can also only perform approximate greedification by updating the policy to be closer to the (soft) greedy policy under the current values. The first approximation underlies algorithms like Sarsa, where the action-value estimates are updated with one new sample, upon which the new policy is immediately set to the soft greedy policy (approximate evaluation, exact greedification).

It is not as common to consider approximate greedification. One of the reasons is that obtaining the (soft) greedy policy is straightforward for discrete actions.¹ For continuous actions, however, obtaining the greedy action for given action-values is non-trivial, requiring the computation of the maximum value (or supremum) over the continuous domain. Some methods have considered optimization approaches to compute it, to get continuous-action Q-learning methods (Amos et al., 2017; Gu et al., 2016; Kalashnikov et al., 2018; Ryu et al., 2020; Gu, 2019). It is more common, though, to instead turn to policy gradient methods and learn a parameterized policy.

This switch to parameterized policies, however, does not evade the question of how to perform approximate greedification. Indeed, many policy gradient (PG) methods—those approximating a gradient of the policy objective—can actually be seen as instances of API. The connection between PG and API arises because efficient implementation of PG methods requires the estimation of a value function. Actor-critic methods estimate value functions through temporal-difference methods (Sutton and Barto, 2018). We explicitly show in this work that the basic actor-critic method can be seen as API with a particular approximate greedification step. In general, numerous papers have already linked PG methods to policy iteration (Sutton et al., 1999; Kakade and Langford, 2002; Perkins and Pendrith, 2002; Perkins and Precup, 2003; Wagner, 2011, 2013; Scherrer and Geist, 2014; Bhandari and Russo, 2019; Ghosh et al., 2020; Vieillard et al., 2020a), including recent work connecting maximum-entropy PG and value-based methods (O’Donoghue et al., 2017; Nachum et al., 2017b; Schulman et al., 2017a; Nachum et al., 2019).

Moreover, most so-called PG methods used in practice are better thought of as API methods, rather than as PG methods. Many PG methods use a biased estimate of the policy gradient. The correct state weighting is not used in either the on-policy setting (Thomas, 2014; Nota and Thomas, 2020) or the off-policy setting (Imani et al., 2018). Additionally, the use of function approximators to estimate action-values generally results in biased gradient estimates without any further guarantees, such as a compatibility condition (Sutton et al.,

1. Even under discrete actions, there is a reasonable argument that approximate greedification may be preferable, even if exact greedification is possible. We typically only have estimates of the value function, and exact greedification on the estimates can potentially harm the agent’s performance (Kakade and Langford, 2002). Further, having an explicit parameterized policy, even under discrete actions, can be beneficial to avoid an effect known as delusional bias (Lu et al., 2018), where directly computing the greedy value in action-value updates can result in inconsistent action choices.

1999). This bias can be reduced by using n -step return estimates for the policy update, but is not completely removed. Understanding approximate greedification within API, therefore, is one direction for better understanding the PG methods actually used in practice.

The question is what approximate greedification approach should be used. One answer is to define a target policy, that would provide policy improvement if we could represent it, and learn a policy to approximate that target. The classical policy improvement theorem (Sutton and Barto, 2018) guarantees that if a new policy is greedy with respect to the action-value function of an old policy, then the new policy is at least as good as the old policy. For parameterized policies (e.g., neural-network policies), exact greedification in each state is rarely possible as not all policies will be representable by a given function class. Instead of the greedy policy, we can use the Boltzmann distribution over the action values as the target policy, which is known to provide policy improvement (Haarnoja et al., 2018). Of particular importance for us and as we show in this work, stepping towards this target policy—approximate greedification—on average across the state space—rather than for each state—is also guaranteed to provide policy improvement. As such, it is a reasonable target policy to explore for approximate greedification under function approximation.

We explore minimizing the Kullback-Leibler (KL) divergence to this target policy.² Other options are possible, such as total variation or Wasserstein distance. We focus on the KL because it underlies many existing methods—as has been previously shown Vieillard et al. (2020a) and as we more comprehensively summarize in Section 3.2. Further, the KL divergence is a convenient choice because stochastic estimation of this objective only requires the ability to sample from the distributions and evaluate them at single points.

Even though the KL has been used, it is as yet unclear whether to use the reverse or the forward KL divergence, here also called RKL and FKL respectively. That is, should the first argument of the KL divergence be policy π , or should it be the Boltzmann distribution over the action values? Neumann (2011) argues in favour of the reverse KL divergence to obtain a most cost-average policy, whereas Norouzi et al. (2016) uses the forward KL divergence to induce a more exploratory policy (i.e., more diverse state visitation distribution).

The typical default is the reverse KL. The reverse KL without entropy regularization corresponds to a standard Actor-Critic update and is easy to compute, as we show in Section 3.2. More recently, it was shown that the reverse KL guarantees policy improvement when the KL can be minimized separately for each state (Haarnoja et al., 2018, p. 4); this finding motivated the development of Soft-Actor Critic. Regret analyses involving Bregman divergences, like for mirror descent (Orabona, 2019; Shani et al., 2020), also tend to imply results for the reverse KL, but not for the forward KL.

Some work, though, has used the forward KL (Norouzi et al., 2016; Nachum et al., 2017a; Agarwal et al., 2019; Vieillard et al., 2020b), including implicitly some work in classification for RL (Lagoudakis and Parr, 2003; Lazaric et al., 2010; Farahmand et al., 2015). For contextual bandits, Chen et al. (2019) showed improved performance when using a surrogate,

2. We use the KL divergence to project the target policy to the space of parameterized policies. This differs from using the KL (or Bregman) divergence to regularize policy updates to force a new policy be close to a previous one (Peters et al., 2010; Schulman et al., 2015; Abdolmaleki et al., 2018; Geist et al., 2019; Vieillard et al., 2020a). While such regularization changes the target policy and may confer benefits to algorithms (Schulman et al., 2015; Vieillard et al., 2020a), there still remains a question of how to project the target policy, which is the main focus of the present paper.

forward KL objective for the smoothed risk. Others used the forward KL to prevent mode collapse, given that the forward KL is mode-covering (Agarwal et al., 2019; Mei et al., 2019).

Though both have been used and advocated for, there is no comprehensive investigation into their differences for approximate greedification. The closest work is Neumann (2011), but they do KL divergence reduction in the context of EM-based policy search using the variational inference framework, whereas we frame the problem as approximate policy iteration, which leads to different optimization processes and cost functions. Their reverse KL target, for example, is a reward weighted trajectory distribution, which is different from the Boltzmann distribution we use here and they minimize the KL divergence with respect to the variational distribution, while we minimize it directly with respect to the policy. Their work does not provide any theoretical results and their experimental settings are limited to single step decision making, whereas we experiment on sequential decision making.

The goal of this work is to investigate the differences between using a forward or reverse KL divergence, under entropy regularization, for approximate greedification. We ask, given that we optimize a policy to reduce either the forward or the reverse KL divergence to a Boltzmann distribution over the action values, what is the quality of the resulting policy? We provide some clarity on this question with the following contributions.

1. We highlight four choices for greedification: forward KL (FKL) or reverse KL (RKL) to a Boltzmann distribution on the action-values, with or without entropy regularization. We show that many existing methods can be categorized into one of these four quadrants, and particularly show that the standard Actor-Critic update corresponds to using the RKL.
2. We extend the policy improvement result for the RKL (Haarnoja et al., 2018) in two ways. (a) Instead of reducing the RKL for all states, we only need to reduce it on average under a certain state-weighting. (b) We characterize improvement under approximate action-values, rather than only exact action-values.
3. We further extend our theoretical results under a condition where the policy does not change too much; these results provide an extension of the seminal improvement results for conservative policy iteration (Kakade and Langford, 2002, Theorem 4.1), to parameterized policies using gradient descent on the RKL for greedification.
4. We show via a counterexample that merely reducing the FKL is not sufficient to guarantee improvement and discuss additional sufficient conditions to guarantee improvement.
5. We investigate optimization differences in small MDPs, and find that, particularly under continuous actions, (a) the RKL can converge faster, but sometimes to suboptimal local minima solutions, but (b) the optimal solution of the FKL can be worse than the corresponding RKL, particularly under higher entropy regularization.
6. In a maze environment, with neural network function approximation, we show that the FKL promotes more exploration under continuous actions, by maintaining a higher variance in the learned policy, but for discrete actions, exploration is very similar for both.

In addition to these carefully controlled experiments, we tested the approaches in benchmark environments. We found that performance between the two was similar. We hypothesize that the reason for this outcome is that the action-values and the corresponding Boltzmann policy are largely unimodal for the benchmark problems; bigger differences

should arise for the multi-modal setting. We conclude the work with a discussion about open questions and key next steps, including how to leverage these insights about FKL and RKL to potentially obtain improved policy optimization algorithms and theory.

2. Problem Formulation

We formalize the reinforcement learning problem (Sutton and Barto, 2018) as a Markov Decision Process (MDP): a tuple $(\mathcal{S}, \mathcal{A}, \gamma, r, p)$ where \mathcal{S} is the state space; \mathcal{A} is the action space; $\gamma \in [0, 1]$ is the discount factor; $r : \mathcal{S} \times \mathcal{A} \rightarrow \mathbb{R}$ is the reward function; and, for every $(s, a) \in \mathcal{S} \times \mathcal{A}$, $p(\cdot | s, a)$ gives the conditional transition probabilities over \mathcal{S} . A *policy* is a mapping $\pi : \mathcal{S} \rightarrow \Delta_{\mathcal{A}}$, where $\Delta_{\mathcal{A}}$ is the space of probability distributions over \mathcal{A} . At every discrete time step t , an agent observes a state S_t , from which it draws an action from its policy: $A_t \sim \pi(\cdot | S_t)$. The agent sends the action A_t to the environment, from which it receives the reward signal $r(S_t, A_t)$ and the next state S_{t+1} .

In this work we focus on the episodic problem setting, where the goal of the RL agent is to maximize the *expected return*—the expectation of a discounted sum of rewards—from the set of start states. To formalize this goal, we define the *value function* V^π for policy π as

$$V^\pi(s) := \mathbb{E}_\pi \left[\sum_{k=0}^{\infty} \gamma^k r(S_k, A_k) \mid S_0 = s \right].$$

The expectation above is over the trajectory $(S_0, A_0, S_1, A_1, \dots)$ induced by π and the transition kernel p . For simplicity, we omit p in the subscript, because the expectation is always according to p . We similarly define the *action-value function*:

$$Q^\pi(s, a) := \mathbb{E}_\pi \left[\sum_{k=0}^{\infty} \gamma^k r(S_k, A_k) \mid S_0 = s, A_0 = a \right] = r(s, a) + \gamma \mathbb{E}[V^\pi(S') \mid S = a, A = a].$$

A common objective in policy optimization is the value of the policy π_θ averaged over the start state distribution ρ_0 :

$$\eta(\pi_\theta) := \int_{\mathcal{S}} \rho_0(s) \int_{\mathcal{A}} \pi_\theta(a | s) Q^{\pi_\theta}(s, a) da ds. \tag{1}$$

The policy gradient theorem gives us the gradient of $J(\theta)$ (Sutton et al., 1999),

$$\nabla_\theta \eta(\pi_\theta) = \frac{1}{1 - \gamma} \int_{\mathcal{S}} d^{\pi_\theta}(s) \int_{\mathcal{A}} Q^{\pi_\theta}(s, a) \nabla_\theta \pi_\theta(a | s) da ds, \tag{2}$$

where $d^{\pi_\theta}(s) := (1 - \gamma) \sum_{t=0}^{\infty} \gamma^t p(S_t = s | S_0 \sim \rho_0)$ is the normalized discounted state visitation distribution.

Because we do not have access to Q^{π_θ} , we instead approximate. For example, in REINFORCE (Williams, 1992), a sampled return from (s, a) is used as an unbiased estimate of $Q^{\pi_\theta}(s, a)$. This method, however, assumes on-policy returns and tends to be sample inefficient. Commonly, a biased but lower-variance choice is to use a learned estimate Q of Q^{π_θ} , obtained through policy evaluation algorithms like SARSA (Sutton and Barto, 2018). In these Actor-Critic algorithms, the actor—the policy—updates with a (biased) estimate of the above gradient, given by this Q —the critic.

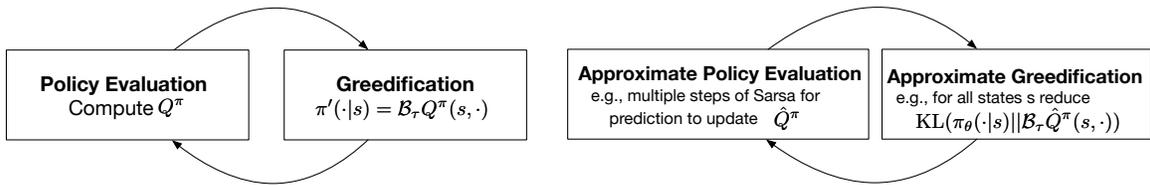


Figure 1: **Contrasting Policy Iteration (PI) and Approximation PI (API)**, left and right respectively. **[Left-hand Side]** In PI, the policy evaluation step and the greedification step are done exactly. In the policy evaluation step, the action values for the current policy π are computed. In the greedification step—also called the policy improvement step—the policy π' is set to the (soft) greedy policy. Then this π' is handed to the policy evaluation step, as the new π , the new action values are computed, and this process repeats. In this work, the soft greedy policy that we consider is the Boltzmann policy, $\mathcal{B}_\tau Q^\pi$, defined in Equation (3). We discuss why this choice for greedification provides policy improvement in Section 3.1. **[Right-hand Side]** In API, we use the same target policy for greedification—namely the Boltzmann policy—but we only approximate this target policy. We investigate minimizing—or at least reducing—a KL divergence between a parameterized π_θ to the target policy for this approximate greedification. This learned π_θ corresponds to the new policy π' that is handed back to the approximate policy evaluation step. The approximate greedification step can fully minimize the KL or only reduce it. There is a similar choice for the approximate policy evaluation step: we can either obtain the best approximate action-values with a batch algorithm like least-squares TD, or simply improve the estimate from the existing estimate using multiple stochastic updates to the action-values under the new policy.

This Actor-Critic procedure with learned Q can be interpreted as Approximate Policy Iteration (API). API methods alternate between approximate policy evaluation to obtain a new Q and approximate greedification to get a policy π that is more greedy with respect to Q . We depict this approach in Figure 1, and contrast it to PI. As we show in the next section, the gradient in Equation (2) can be recast as the gradient of a KL divergence to a policy peaked at maximal actions under Q ; reducing this KL divergence updates the policy to increase the probability of these maximal actions, and so become more greedy with respect to Q . Under this API view, we obtain a clear separation between estimating Q and greedifying π . We can be agnostic to the strategy for updating Q —we can even use soft action values (Ziebart, 2010) or Q-learning (Watkins and Dayan, 1992)—and focus on answering: for a given Q , how can we perform an approximate greedification step?

This work focuses on understanding the differences between using the forward and reverse KL divergences towards the Boltzmann policy for approximate greedification. We explicitly define each of these approaches in the next section. Other choices for approximate greedification are possible—other target policies and other divergences or metrics—but we constrain our investigation to a feasible scope. In the next section, we further motivate why we investigate these approaches for approximate greedification, and later summarize how these variants underlie a variety of policy optimization methods.

3. Approximate Greedification

In this section, we formalize how to do approximate greedification. First, we discuss an appropriate choice for the (soft) greedy policy π_{target} . Given access to such a policy π_{target} , we can update our existing policy π to be closer to π_{target} , using the KL divergence. The KL divergence, however, is not symmetric and has an entropy parameter τ , resulting in four variants. We present these four variants that we use throughout the paper, and derive the updates under each choice. Finally, we discuss the importance of the state weighting in the final objective, which weights the divergence to the target policy in each state.

3.1 Defining a Target Policy

A reasonable choice for the target policy is the Boltzmann distribution, as we motivate in this section. The Boltzmann distribution we use here is also common in pseudo-likelihood methods (Kober and Peters, 2008; Neumann, 2011; Levine, 2018), ensuring that one has a target distribution based on the action-values.

Let Q be an action-value function estimate. For a given $\tau > 0$, the Boltzmann distribution $\mathcal{B}_\tau Q(s, \cdot)$ for a state s is defined as

$$\mathcal{B}_\tau Q(s, a) := \frac{\exp(Q(s, a)\tau^{-1})}{Z(s)} \quad \text{for } Z(s) := \int_{\mathcal{A}} \exp(Q(s, b)\tau^{-1}) db \quad (3)$$

where $Z(s)$ is known as the partition function. The definition in Equation (3) does not depend upon a particular policy: we can input any Q that is a function of states and actions. For larger τ , the Boltzmann distribution is more stochastic: it has higher entropy.

This distribution can be derived by solving for the entropy-regularized greedy policy on Q . To see why, recall the definition of the entropy of a distribution, which captures how spread out the distribution is:

$$\mathcal{H}(\pi(\cdot | s)) := - \int_{\mathcal{A}} \pi(a | s) \log \pi(a | s) da.$$

The higher the entropy, the less the probability mass of $\pi(\cdot | s)$ is concentrated in any particular area. Let \mathcal{F} be the set of all nonnegative functions on \mathcal{A} that integrate to 1. At a given state, the entropy-regularized greedy policy is given by

$$\pi_{\text{target}}(\cdot | s) := \arg \max_{p \in \mathcal{F}} \int p(a)(Q(s, a) - \tau \log p(a)) da. \quad (4)$$

The integrand can be rewritten as follows:

$$\begin{aligned} p(a) (Q(s, a) - \tau \log p(a)) &= \tau p(a) \log \frac{\exp(Q(s, a)/\tau)}{p(a)} \\ &= \tau p(a) \log \frac{\mathcal{B}_\tau Q(s, a)}{p(a)} + \tau p(a) \log \int \exp\left(\frac{Q(s, a)}{\tau}\right) da. \end{aligned}$$

The right summand becomes a constant when integrated, so $\pi_{\text{target}}(\cdot | s)$ can be rewritten as

$$\pi_{\text{target}}(\cdot | s) = \arg \min_{p \in \mathcal{F}} \int p(a) \log \frac{p(a)}{\mathcal{B}_\tau Q(s, a)} da = \mathcal{B}_\tau Q(s, \cdot).$$

where the first term is actually the KL divergence between p and $\mathcal{B}_\tau Q(s, a)$, which is minimized by setting p to $\mathcal{B}_\tau Q(s, a)$.

The use of entropy-regularization avoids obtaining deterministic, greedy policies that can be problematic in policy gradient methods. Instead, this approach allows for soft greedification, giving the most greedy policy under the constraint that the entropy of the policy remains non-negligible. This policy can be shown to provide guaranteed policy improvement, but under a different criteria: according to soft value functions (Ziebart, 2010).

Soft value functions are value functions where an entropy term is added to the reward.

$$V_\tau^\pi(s) := \mathbb{E}_\pi \left[\sum_{k=0}^{\infty} \gamma^k [r(S_k, A_k) + \tau \mathcal{H}(\pi(\cdot | S_k))] \mid S_0 = s \right]$$

We can define the soft action-value function in terms of the soft value function.

$$Q_\tau^\pi(s, a) := r(s, a) + \gamma \mathbb{E}[V_\tau^\pi(S') \mid S = s, A = a]$$

We can also write the state-value function in terms of the action-value function.

$$V_\tau^\pi(s) = \mathbb{E}_\pi [Q_\tau^\pi(s, A) - \tau \log \pi(A \mid s)].$$

These soft value functions corresponds to a slightly different RL problem described as entropy-regularized MDPs (Geist et al., 2019).

For these soft value functions, we can guarantee policy improvement under greedification with the Boltzmann distribution. If we set $\pi'(\cdot | s) = \mathcal{B}_\tau Q^\pi(s, \cdot)$ for all $s \in \mathcal{S}$, then $Q_\tau^{\pi'}(s, a) \geq Q_\tau^\pi(s, a)$ for all (s, a) (Haarnoja et al., 2017, Theorem 4). This parallels the classical policy improvement result in policy iteration (Sutton and Barto, 2018). This guaranteed policy improvement is a motivation for using $\mathcal{B}_\tau Q$ as a target policy for greedification. We extend this policy improvement result to hold under weaker conditions in Section 5.

3.2 Approximate Greedification with the KL

In this section we discuss how to use a KL divergence to bring π_θ closer to $\mathcal{B}_\tau Q$. One might wonder why we do not just set $\pi(\cdot | s) = \mathcal{B}_\tau Q(s, \cdot)$. Indeed, for discrete action spaces, we can draw actions from $\mathcal{B}_\tau Q(s, \cdot)$ easily at each time step. However, for continuous actions, even calculating $\mathcal{B}_\tau Q(s, \cdot)$ requires approximating a generally intractable integral. Furthermore, even in the discrete-action regime, using $\mathcal{B}_\tau Q$ might not be desirable as Q is usually just an action-value *estimate*. Greedifying with respect to an action-value estimate does not guarantee greedification with respect to the action-value.

In this work we focus on the KL divergence to measure the difference between π and π_{target} . Given two probability distributions p, q on \mathcal{A} , the KL divergence between p and q is

$$\text{KL}_q^p := \int_{\mathcal{A}} p(a) \log \frac{p(a)}{q(a)} da, \tag{5}$$

where p is assumed to be absolutely continuous (Billingsley, 2008) with respect to q (i.e. p is never nonzero where q is zero), to ensure that the KL divergence exists. The KL divergence is zero iff $p = q$ almost everywhere, and is always non-negative. Stochastic estimation of

the KL divergence has the advantage of requiring just the ability to sample from p and to calculate p and q . This feature is in contrast to the Wasserstein metric for example, which generally requires solving an optimization problem just to compute it.

The KL divergence is not symmetric. For example, KL_q^p may be defined while KL_p^q may not even exist if q is not absolutely continuous with respect to p . This asymmetry leads to the two possible choices for measuring differences between distributions: the reverse KL and the forward KL. Assume that q is a true distribution that we would like to match with our learned distribution p_θ , where p_θ is smooth with respect to $\theta \in \mathbb{R}^k$. The *forward* KL divergence is $\text{KL}_{p_\theta}^q$ and the *reverse* KL divergence is $\text{KL}_q^{p_\theta}$.

We define the **Reverse KL** (RKL) for greedification on Q at a given state s as

$$\text{RKL}_{\mathcal{B}_\tau Q}^{\pi_\theta}(s) := \text{KL}_{\mathcal{B}_\tau Q}^{\pi_\theta}(s),$$

where we additionally define for any two policies π_1, π_2 ,

$$\text{KL}_{\pi_2}^{\pi_1}(s) := \int_{\mathcal{A}} \pi_1(a | s) \log \frac{\pi_1(a | s)}{\pi_2(a | s)} da.$$

This Q is any action-value on which we perform approximate greedification; it can be a soft action value or not. We can rewrite the RKL as follows:

$$\begin{aligned} \text{RKL}_{\mathcal{B}_\tau Q}^{\pi_\theta}(s) &= \int_{\mathcal{A}} \pi_\theta(a | s) \log \frac{\pi_\theta(a | s)}{\mathcal{B}_\tau Q(s, a)} da \\ &= \int_{\mathcal{A}} \pi_\theta(a | s) \left(\log \pi_\theta(a | s) - \frac{Q(s, a)}{\tau} + \log Z(s) \right) da \\ &= -\mathcal{H}(\pi_\theta(\cdot | s)) - \int_{\mathcal{A}} \pi_\theta(a | s) \frac{Q(s, a)}{\tau} da + \log Z(s), \end{aligned}$$

with gradient

$$\nabla_\theta \text{RKL}_{\mathcal{B}_\tau Q}^{\pi_\theta}(s) = -\nabla_\theta \mathcal{H}(\pi_\theta(\cdot | s)) - \int_{\mathcal{A}} \nabla_\theta \pi_\theta(a | s) \frac{Q(s, a)}{\tau} da.$$

If we scale by τ to get $\tau \text{RKL}(\theta; s, Q)$, we can see that τ plays the role of an entropy regularization parameter:³ a larger τ results in more entropy regularization on $\pi_\theta(\cdot | s)$.

For a finite action space, we can take the limit to get the **Hard Reverse KL**.⁴

$$\begin{aligned} \lim_{\tau \rightarrow 0} \tau \text{RKL}_{\mathcal{B}_\tau Q}^{\pi_\theta}(s) &= \lim_{\tau \rightarrow 0} \tau \sum_a \pi_\theta(a | s) \log \pi_\theta(a | s) - \tau \sum_{\mathcal{A}} \pi_\theta(a | s) \log \mathcal{B}_\tau Q(s, a) \\ &= 0 - \lim_{\tau \rightarrow 0} \sum_a \pi_\theta(a | s) \left(Q(s, a) - \tau \log \sum_b \exp(Q(s, b) \tau^{-1}) \right) \\ &= \left(-\sum_a \pi_\theta(a | s) Q(s, a) \right) + \lim_{\tau \rightarrow 0} \tau \log \sum_b \exp(Q(s, b) \tau^{-1}) \\ &= \left(-\sum_a \pi_\theta(a | s) Q(s, a) \right) + (n_s) \max_a Q(s, a), \end{aligned}$$

3. For a fixed $\mathcal{B}_\tau Q$, the policy that minimizes the RKL is the same regardless of the scaling by a constant in front, so we use the more standard unscaled KL to define the RKL.

4. When investigating $\tau \rightarrow 0$, it is not straightforward to extend the calculations we do here to integrals, so we derive them for the discrete case and extrapolate to the continuous case.

where n_s is the number of maximizing actions in s . Since the last term of the RHS does not depend on π_θ , we are motivated to define the Hard Reverse KL as follows, for both finite and infinite action spaces.

$$\text{Hard RKL}_{\mathcal{B}_\tau Q}^{\pi_\theta}(s) := - \int_{\mathcal{A}} \pi_\theta(a | s) Q(s, a) da,$$

with gradient

$$\nabla_\theta \text{Hard RKL}_{\mathcal{B}_\tau Q}^{\pi_\theta}(s) = - \int_{\mathcal{A}} \nabla_\theta \pi_\theta(a | s) Q(s, a) da.$$

If Q is equal to Q^{π_θ} , then this gradient is exactly the negative of the inner term of the policy gradient in Equation (2).⁵ This similarity in form means that the typical policy gradient update in actor-critic can be thought of as a greedification step with the Hard RKL.

Similarly, we can define the **Forward KL** (FKL) for greedification:

$$\text{FKL}_{\mathcal{B}_\tau Q}^{\pi_\theta}(s) := \text{KL}_{\pi_\theta}^{\mathcal{B}_\tau Q}(s).$$

We can rewrite the FKL as

$$\begin{aligned} \text{FKL}_{\mathcal{B}_\tau Q}^{\pi_\theta}(s) &= \int_{\mathcal{A}} \mathcal{B}_\tau Q(s, a) \log \frac{\mathcal{B}_\tau Q(s, a)}{\pi_\theta(a | s)} da \\ &= \int_{\mathcal{A}} \mathcal{B}_\tau Q(s, a) \log \mathcal{B}_\tau Q(s, a) da - \int_{\mathcal{A}} \mathcal{B}_\tau Q(s, a) \log \pi_\theta(a | s) da \\ &= -\mathcal{H}(\mathcal{B}_\tau Q(s, \cdot)) - \int_{\mathcal{A}} \mathcal{B}_\tau Q(s, a) \log \pi_\theta(a | s) da \end{aligned}$$

with gradient

$$\nabla_\theta \text{FKL}_{\mathcal{B}_\tau Q}^{\pi_\theta}(s) = - \int_{\mathcal{A}} \mathcal{B}_\tau Q(s, a) \nabla_\theta \log \pi_\theta(a | s) da.$$

We can again consider the limit $\tau \rightarrow 0$ in the case of a finite action space (in this case there is no need to multiply the KL divergence by τ). Assume that there are n_s maximizing actions of $Q(s, \cdot)$, indexed by a_i^* .

$$\begin{aligned} \lim_{\tau \rightarrow 0} \text{FKL}_{\mathcal{B}_\tau Q}^{\pi_\theta}(s) &= - \lim_{\tau \rightarrow 0} \mathcal{H}(\mathcal{B}_\tau Q(s, \cdot)) - \lim_{\tau \rightarrow 0} \sum_a \frac{\exp(Q(s, a)\tau^{-1})}{\sum_b \exp(Q(s, b)\tau^{-1})} \log \pi_\theta(a | s) \\ &= - \lim_{\tau \rightarrow 0} \mathcal{H}(\mathcal{B}_\tau Q(s, \cdot)) - \sum_a \lim_{\tau \rightarrow 0} \frac{\exp(Q(s, a)\tau^{-1})}{\sum_b \exp(Q(s, b)\tau^{-1})} \log \pi_\theta(a | s) \\ &= - \lim_{\tau \rightarrow 0} \mathcal{H}(\mathcal{B}_\tau Q(s, \cdot)) - \frac{1}{n_s} \sum_{i=1}^{n_s} \log \pi_\theta(a_i^* | s). \end{aligned}$$

5. We are unaware of a previous statement of this result in the literature, though similar results have been reported. For example, Kober and Peters (2008) derive the policy gradient update from a pseudo-likelihood method. Belousov and Peters (2019) also derive it as a special case of f-divergence constrained relative entropy policy search (Peters et al., 2010). Some references to a connection between value-based methods with entropy regularization and policy gradient can be found in (Nachum et al., 2017b).

KL	Formula	Gradient	Comment
RKL	$\text{KL}_{\mathcal{B}_\tau Q}^{\pi_\theta}(s)$	$-\int_{\mathcal{A}} \nabla_\theta \pi_\theta(a s) \tau^{-1} Q(s, a) da - \nabla_\theta \mathcal{H}(\pi_\theta(\cdot s))$	A likelihood-based Soft Actor-Critic. ⁶
Hard RKL	$-\int_{\mathcal{A}} \pi_\theta(a s) Q(s, a) da$	$-\int_{\mathcal{A}} \nabla_\theta \pi_\theta(a s) Q(s, a) da$	Equivalent to vanilla actor-critic if action value is unregularized.
FKL	$\text{KL}_{\pi_\theta}^{\mathcal{B}_\tau Q}(s)$	$-\int_{\mathcal{A}} \mathcal{B}_\tau Q(s, a) \nabla_\theta \log \pi_\theta(a s) da$	Like classification with cross-entropy loss and $\mathcal{B}_\tau Q$ the distribution over the correct label.
Hard FKL	$-\frac{1}{n_s} \sum_{i=1}^{n_s} \log \pi_\theta(a_i^* s)$	$-\frac{1}{n_s} \sum_{i=1}^{n_s} \nabla_\theta \log \pi_\theta(a_i^* s)$	Like classification with cross-entropy loss and a_i^* the correct labels.

Table 1: This table summarizes the four KL variants for greedification, including the objective, gradient, and a descriptive comment. We fix a state s and define $a_i^* \in \arg \max_a Q(s, a)$ with n_s maximal actions. To reduce clutter, we have assumed that there can only be finitely many maximizing actions for the Hard FKL.

As the first term does not depend upon the policy parameters, we ignore it to define the **Hard Forward KL** as

$$\text{Hard FKL}_{\mathcal{B}_\tau Q}^{\pi_\theta}(s) := -\frac{1}{n_s} \sum_{i=1}^{n_s} \log \pi_\theta(a_i^* | s)$$

For continuous actions, if we have a compact action space, then the arg max exists over a and provides valid actions. We can use a similar definition to the discrete case, assuming there are a finite number of maximal actions. If there are flat regions, with intervals of maximal actions, then the sum is replaced with an integral; for simplicity we assume a finite set and provide the definition under this assumption. The gradient for the Hard FKL is

$$\nabla_\theta \text{Hard FKL}_{\mathcal{B}_\tau Q}^{\pi_\theta}(s) = -\frac{1}{n_s} \sum_{i=1}^{n_s} \nabla_\theta \log \pi_\theta(a_i^* | s)$$

The Hard FKL expression looks quite similar to the cross-entropy loss in supervised classification, if one views the maximum action of $Q(s, \cdot)$ as the correct class of state s . We are unaware of any literature that analyzes the Hard FKL for approximate greedification.

We summarize the main expressions, gradients and results in Table 1. Interestingly, many existing policy gradient methods have policy updates that fit into one of these four quadrants. We also summarize this categorization in Table 2, with more justification for the categorization in Appendix A. For example, TRPO uses the Hard RKL, but with an

	Reverse KL	Forward KL
Hard No entropy	Actor-Critic (Sutton and Barto, 2018); Trust Region Policy Optimization (TRPO) (Schulman et al., 2015) and variants including MPO (Abdolmaleki et al., 2018), Mirror descent policy iteration (Geist et al., 2019), REPS (Peters et al., 2010); PPO (Schulman et al., 2017b); Deep Deterministic Policy Gradient (DDPG) (Silver et al., 2014; Lillicrap et al., 2016); Politex (Abbasi-Yadkori et al., 2019)	Deep Conservative Policy Iteration (DCPI) (Vieillard et al., 2020b); Policy Greedification as Classification (Lagoudakis and Parr, 2003; Lazaric et al., 2010; Farahmand et al., 2015)
Soft With entropy	Soft Q-learning (SQL) (Haarnoja et al., 2017); Soft Actor-Critic (SAC) (Haarnoja et al., 2018); A3C (Mnih et al., 2016); Conservative value iteration (Kozuno et al., 2019)	UREX (Nachum et al., 2017a; Agarwal et al., 2019); Exploratory Conservative Policy Optimization (ECPO) (Mei et al., 2019)

Table 2: A categorization of existing policy optimization methods, in terms of which underlying approach is used to update towards the Boltzmann policy. This list is not exhaustive, because (a) we have likely missed a few methods, given so many variants are based on an Actor-Critic update and (b) some policy optimization methods cannot be seen as updating with a KL divergence to the Boltzmann policy. See Appendix A for justification for this categorization.

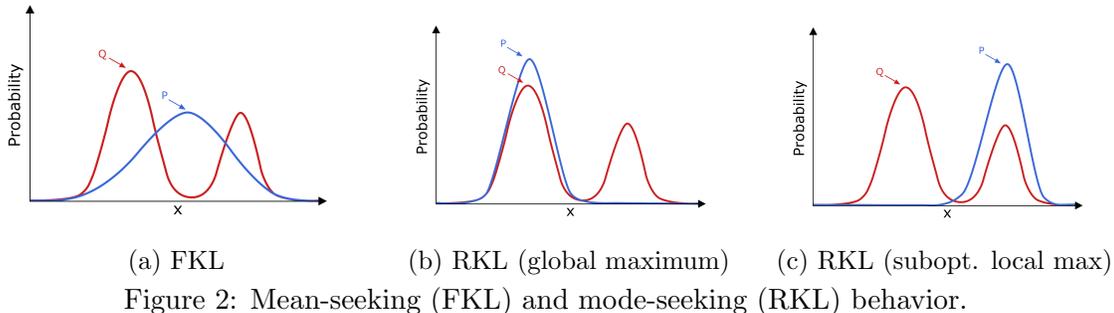
additional constraint that the policy should not change too much after an update. This is encoded as a KL divergence to the previous policy, which is a difference use of the KL than described above. The algorithms in this summary table have many properties beyond the underlying update to the Boltzmann target policy; we do not intend to imply they are solely defined by the use of the RKL or FKL, with or without entropy regularization. Nonetheless, it provides another lens on understanding similarities and differences between these algorithms, due to the choice of this underlying update.

3.3 Differences between the FKL and RKL

Although switching π_θ and $\mathcal{B}_\tau Q$ might seem like a small change, there are several differences in terms of the optimization and the solution itself. In terms of the optimization, it can be simpler to sample the gradient of the RKL, because actions do not have to be sampled according to $\mathcal{B}_\tau Q$. The FKL, on the other hand, requires actions to be sampled from the $\mathcal{B}_\tau Q$, which can be expensive. But, favourably for the FKL, if $\pi_\theta(a | s) \propto \exp(\theta_a)$, then the FKL is convex with respect to θ because $\log \sum_i \exp(\theta_i)$ is convex.⁷ The RKL, on the other hand, is generally not convex with respect to θ , even if π_θ is parameterized with a Boltzmann distribution.

The other critical difference is in terms of the solution itself. If $\mathcal{B}_\tau Q$ is representable by π_θ for some θ , then the FKL and RKL both have the same solution: $\mathcal{B}_\tau Q$. Otherwise, they

7. This fact can be seen by showing that the Hessian is positive semi-definite.



make different trade-offs. Of particular note is the well-known fact that the forward KL causes mean-seeking behavior and the reverse KL causes mode-seeking behavior (Bishop, 2006). To understand the reason, we look at the expressions for each divergence. For a target distribution q , the forward KL is $\int_x q(x) \log \left(\frac{q(x)}{p(x)} \right) dx$. If there is an x where $q(x) \gg 0$ (i.e. $q(x)$ is significantly greater than zero) and $p(x)$ is very close to zero, then $q(x) \log \left(\frac{q(x)}{p(x)} \right)$ is large. In fact, as $p(x) \rightarrow 0$, we have that $q(x) \log \left(\frac{q(x)}{p(x)} \right) \rightarrow \infty$. Therefore, to keep the forward KL small, whenever $q(x) \gg 0$ then we also need $p(x) \gg 0$.

This can result in p that is quite different from q , if the parameterized p cannot represent q . Particularly, if q is multimodal and p is unimodal, then the p^* that minimizes the forward KL will try to cover all of the modes simultaneously, even if the cost for that is placing high mass in regions where $q(x) \approx 0$. For us, this corresponds to regions where the action-values are low. This forward KL solution, which is also called the M-projection (the M stands for moment), is known to be moment-matching. In the case that the family of distributions parameterizing p is exponential and has some element whose moments match those of q , then the moments of p^* and q will match (Koller and Friedman, 2009).

The reverse KL, on the other hand, has the expression $\int_x p(x) \log \left(\frac{p(x)}{q(x)} \right) dx$. Even if $q(x) \gg 0$, we can choose $p(x) = 0$ without causing $p(x) \log \left(\frac{p(x)}{q(x)} \right)$ to get big. This means that p can select one mode, if q is multi-modal. This can be desirable in RL, because it can concentrate the action probability on the region with highest value, even if there is another region with somewhat high action values. This distinction provides another helpful mnemonic: the **forward** KL cares about the **full** support of the target distribution q and the **reverse** KL can **restrict** the support.

However, the reverse KL can get stuck in sub-optimal solutions. If $q(x)$ is near zero for some x , and we pick $p(x) \gg 0$, then $p(x) \log \left(\frac{p(x)}{q(x)} \right)$ can be large. As $q(x)$ gets closer to zero, this number goes to infinity. Therefore, reducing the reverse KL will lead to p such that when $q(x) \approx 0$ then $p(x) \approx 0$. Because of this, the reverse KL is sometimes called zero-forcing or cost-averse. Similarly to the forward KL, this has certain consequences if p is parameterized in a way that cannot represent q . In the multimodal target example we used for the forward KL, p^* that minimizes the reverse KL, often called the I-projection (the I stands for information), will try to avoid placing mass in regions where $q(x) \approx 0$. This

means that p may end up in a sub-optimal mode when reducing the reverse KL via gradient descent. Both behaviors are illustrated in Figure 2.

When approximating some distributions, reducing the FKL can cause overestimation of the tail of the target because of the mean-seeking behavior, whereas reducing the RKL underestimates it because of the mode-seeking behavior. In variational inference, the posterior distribution is approximated using a variational distribution. Importance sampling (IS) can be used to debias estimates obtained from some Bayesian inference procedures. The IS proposal distribution is commonly obtained by minimizing the RKL. However, because of this underestimation of the tail of the target, the quality of the IS estimates sometimes suffer, and a proposal distribution that overestimates the tail of the target is more desirable. Recent work (Jerfel et al., 2021) has shown that the FKL can be superior in this case. Better understanding the implications of reducing the FKL in the context of approximate policy improvement will facilitate incorporate such advantages into future RL algorithms.

3.4 The Weighting over States

The above greedification objectives, and corresponding gradients, are defined per state. To specify the full greedification objective across states, we need a weighting $d : \mathcal{S} \rightarrow \mathbb{R}^+$. Under function approximation, the agent requires this distribution to trade-off accuracy of greedification across states. For example, the full objective for the RKL is

$$\int_{\mathcal{S}} d(s) \text{RKL}_{B^*, Q}^{\pi_\theta}(s) ds.$$

The other objectives are specified similarly.

The state weighting specifies how function approximation resources should be allocated for greedification. If there are no trade-offs, such as if the Boltzmann policy can be perfectly represented in each state, then the state weighting plays almost no role. It simply needs to be positive in a state to ensure the KL is minimized for that state. Otherwise, it may be that to make the policy closer to the Boltzmann policy in one state, it has to make it further in another state. The state weighting specifies which states to prioritize in this trade-off.

Algorithms in practice use a replay buffer, where, without reweighting, d implicitly corresponds to the state frequency in the replay buffer. We might expect early on that the implicit weighting is similar to the state visitation distribution under a random policy, and later more similar to the state visitation under a near-optimal policy—if learning is effective. The ramifications of allowing d to be chosen implicitly by the replay buffer are as yet not well understood. In practice, algorithms seem to perform reasonably well, even without carefully controlling this weighting, possibly in part due to the fact that large neural networks are used to parameterize the policy that are capable of representing the target policy.

There is, however, some evidence that the weighting can matter, particularly from theoretical work on policy gradient methods. This role of the weighting might seem quite different from the typical role in the policy gradient, but there are some clear connections. When averaging the gradient of the Hard RKL with weighting d , we have $-\int_{\mathcal{S}} d(s) \int_{\mathcal{A}} Q(s, a) \nabla \pi_\theta(a | s) da ds$. If $d = d^{\pi_\theta}$ and $Q = Q^{\pi_\theta}$, then we have the true policy gradient; otherwise, for different weightings d , it may not correspond to the gradient of any function (Nota and Thomas, 2020). A similar issue has been highlighted for the off-policy policy gradient setting (Imani et al., 2018), where using the wrong weighting results in

convergence to a poor stationary point. These counterexamples have implications for API, as they suggest that with accurate policy evaluation steps, the iteration between evaluation and greedification—the policy gradient step—may converge to poor solutions, without carefully selecting the state weighting.

At the same time, this does not mean that the weighting must correspond to the policy state visitation distribution. Outside these counterexamples, many other choices could result in good policies. In fact, the work on CPI indicates that the weighting with d^π can require a large number of samples to get accurate gradient estimates, and moving to a more uniform weighting over states can be significantly better (Kakade and Langford, 2002). The choice of weighting remains an important open question. For this work, we do not investigate this question, and simply opt for the typical choice in practice—using replay.

4. An API Algorithm with FKL or RKL

In this section, we provide a concrete algorithm that uses either the FKL or RKL for greedification within an API framework. The algorithm resembles Soft Actor-Critic (SAC) (Haarnoja et al., 2018), which was originally described as a policy iteration algorithm. The key choices in the algorithm include (1) how to learn the (soft) action-values, (2) how to obtain an estimate of the RKL or FKL for a given state, and (3) how to sample states.

To sample states, we use the standard strategy of maintaining a buffer of the most recent experience. We sample states uniformly from this buffer. To obtain an estimate of the FKL or RKL for a given state, we need to estimate the gradient that has a sum or integral over actions. For the discrete action setting, we can simply sum over all actions. The **All-Actions** updates from a state s correspond to

$$\begin{aligned} \text{RKL} : \nabla_\theta \sum_{a \in \mathcal{A}} \pi_\theta(a | s) \log \pi_\theta(a | s) - \sum_{a \in \mathcal{A}} \nabla_\theta \pi_\theta(a | s) \log \mathcal{B}_\tau Q(a | s) \\ = - \sum_{a \in \mathcal{A}} \nabla_\theta \pi_\theta(a | s) \left[\frac{Q(s, a)}{\tau} - \log \pi_\theta(a | s) \right] \end{aligned} \quad (6)$$

$$\text{FKL} : - \sum_{a \in \mathcal{A}} \mathcal{B}_\tau Q(a | s) \nabla_\theta \log \pi_\theta(a | s) \quad (7)$$

$$\text{Hard RKL} : - \sum_{a \in \mathcal{A}} \nabla_\theta \pi_\theta(a | s) Q(s, a) \quad (8)$$

$$\text{Hard FKL} : - \nabla_\theta \log \pi_\theta \left(\arg \max_b Q(s, b) | s \right). \quad (9)$$

For the Hard FKL, when there is more than one maximal action, we assume that ties are broken randomly. For the continuous action setting, we can try to estimate the All-Actions update with numerical integration. More practically, we can simply sample actions.

Sampling actions is more straightforward for the RKL than the FKL. For the RKL and Hard RKL, we simply need to sample actions from the policy. In this case, we assume we sample n actions a_1, \dots, a_n from $\pi_\theta(\cdot | s)$ and, using $\nabla_\theta \pi_\theta(a | s) = \pi_\theta(a | s) \nabla_\theta \log \pi_\theta(a | s)$ and $\sum_{a \in \mathcal{A}} \pi_\theta(a | s) \nabla_\theta \log \pi_\theta(a | s) \approx \sum_{i=1}^n \frac{\nabla_\theta \log \pi_\theta(a_i | s)}{n}$, compute a **Sampled-Action** update,

where we also change Q to $Q - V$

$$\mathbf{RKL} : -\frac{1}{n} \sum_{i=1}^n \nabla_{\theta} \log \pi_{\theta}(a_i | s) \left(\frac{Q(s, a_i) - V(s)}{\tau} - \log \pi_{\theta}(a_i | s) \right) \quad (10)$$

$$\mathbf{Hard RKL} : -\frac{1}{n} \sum_{i=1}^n \nabla_{\theta} \log \pi_{\theta}(a_i | s) (Q(s, a_i) - V(s)) \quad (11)$$

The inclusion of a baseline V reduces variance due to sampling actions and does not introduce bias. Alternatively, for certain distributions, we can use the reparametrization trick and compute alternative sampled-action updates. For the case where the policy is parametrized as a multivariate normal, with $\pi_{\theta}(\cdot | s) \sim \mathcal{N}(\mu_{\theta}(s), \text{Diag}(\sigma_{\theta}(s)))$, where $\text{Diag}(\cdot)$ converts a vector to a diagonal matrix, an action sampled $a \sim \pi_{\theta}(\cdot | s)$ can be written as $a = a_{\theta}(s, a') = \mu_{\theta}(s) + \text{Diag}(\sigma_{\theta}(s))a'$ for $a' \sim \mathcal{N}(0, I)$. This reparameterization allows gradients to flow through sampled actions by using the chain rule. We can write:

$$\begin{aligned} \nabla_{\theta} \mathbb{E}_{\pi_{\theta}}[f(\pi_{\theta}(\cdot | s))] &= \nabla_{\theta} \int_a \pi_{\theta}(a | s) f(\pi_{\theta}(a | s)) da = \nabla_{\theta} \int_{a'} p(a') f(\pi_{\theta}(a_{\theta}(s, a') | s)) da' \\ &= \int_{a'} p(a') \nabla_{\theta} f(\pi_{\theta}(a_{\theta}(s, a') | s)) da' \approx \frac{1}{N} \sum_{i=1}^N \nabla_{\theta} f(\pi_{\theta}(a_{\theta}(s, a'_i) | s)) \end{aligned}$$

Applying this to the formulas in Table 1, the updates are

$$\mathbf{RKL} : \frac{1}{n} \sum_{i=1}^n \nabla_{\theta} \log(\pi_{\theta}(a_{\theta}(s, a'_i) | s)) - \left(\frac{\nabla_{\theta} Q(s, a_{\theta}(s, a'_i))}{\tau} \right) \quad (12)$$

$$\mathbf{Hard RKL} : -\frac{1}{n} \sum_{i=1}^n \nabla_{\theta} Q(s, a_{\theta}(s, a'_i)) \quad (13)$$

For the FKL, we need to sample according to $\mathcal{B}_{\tau} Q(\cdot | s)$, which can be expensive. Instead, we will use weighted importance sampling, similarly to a previous method that minimizes FKL (Nachum et al., 2017a). We can sample actions from π_{θ} , and compute importance sampling ratios $\rho_i := \frac{\mathcal{B}_{\tau} Q(a_i | s)}{\pi_{\theta}(a_i | s)} \propto \frac{\exp(Q(s, a_i) \tau^{-1})}{\pi(a_i | s)}$. To reduce variance, we use weighted importance sampling with $\tilde{\rho}_i := \frac{\rho_i}{\sum_{j=1}^n \rho_j}$, to get the update

$$\mathbf{FKL} : -\sum_{i=1}^n \tilde{\rho}_i \nabla_{\theta} \log \pi_{\theta}(a_i | s) \quad (14)$$

The Hard FKL update is the same as in the discrete action setting, with the additional complication that computing the argmax action is more difficult for continuous actions. The simplest strategy is to do gradient ascent on $Q(s, \cdot)$, to find a maximal action. There are, however, smarter strategies that have been explored for continuous action Q-learning (Amos et al., 2017; Gu et al., 2016; Kalashnikov et al., 2018; Ryu et al., 2020; Gu, 2019). For example, input convex neural networks (ICNNs) ensure the learned neural network is convex in the input action, so that a gradient descent search to find the minimal action of the negative of the action-values is guaranteed to find a maximal action. We could use these

approaches to find maximal actions, and then also learn an explicit policy with the hard FKL update that increases the likelihood of these maximizing actions.

Finally, we use a standard bootstrapping approach to learn the soft action-values. We perform bootstrapping as per the recommendations in Pardo et al. (2018). For a non-terminal transition (s, a, r, s') , the action values $Q(s, a)$ are updated with the bootstrap target $r + \gamma V(s')$. For a terminal transition, the target is simply r . Note that an episode cut-off—where the agent is teleported to a start state if it reaches a maximum number of steps in the episode—is not a terminal transition and is updated with the usual $r + \gamma V(s')$.

To compute this bootstrap target, we learn a separate V . It is possible to instead simply use $Q(s', a') - \tau \log(\pi(a'|s'))$ for the bootstrap target, but this has higher variance. Instead, a lower variance approach is to use the idea behind Expected Sarsa, which is to compute the expected value for the given policy in the next state. $V(s')$ is a direct estimate of this expected value, rather than computing it from $Q(s', \cdot)$. To update V , we can use the same bootstrap target for Q , but need to incorporate an importance sampling ratio to correct the distribution over actions. To avoid using importance sampling, another option is to use the approach in SAC, where the target for $V(s)$ is $Q(s, a) - \tau \log(\pi(a|s))$. The complete algorithm, putting this all together, is in Algorithm 1.

Algorithm 1 Approximate Policy Iteration (API) with KL Greedification

Input: choice of KL divergence; temperature $\tau \geq 0$; learning rates $\alpha_\theta, \alpha_v, \alpha_w$
Initialize: policy π_θ (parameters θ); action-value estimate Q_β (parameters β); state-value estimate V_w (parameters w); experience replay buffer \mathcal{B}
 Get initial state s_0
for $t = 0, \dots$ **do**
 Apply action $a_t \sim \pi_\theta(\cdot | s_t)$ and observe r, s_{t+1}, done
 $\mathcal{B} = \mathcal{B} \cup (s_t, a_t, s_{t+1}, r, \text{done})$
 if $|\mathcal{B}| \geq$ batch size **then**
 Draw minibatch $\mathcal{D} \sim \mathcal{B}$
 Calculate $g_\theta \approx \mathbb{E}_{\mathcal{D}}[\nabla_\theta KL]$ using one of Equations 6 - 14
 Calculate g_w, g_v using Algorithm 2
 $\theta = \theta - \alpha_\theta g_\theta$
 $w = w - \alpha_w g_w$
 $v = v - \alpha_v g_v$

Algorithm 2 GetValueUpdates

Given: policy π_θ ; Q_β ; V_w ; $\tau \geq 0$; batch of data \mathcal{D}
 $g_w \leftarrow 0, g_v \leftarrow 0$
for (s, a, r, s') in \mathcal{D} **do**
 Draw $\tilde{a} \sim \pi_\theta(\cdot | s)$.
 $g_w \leftarrow g_w - (Q_\beta(s, \tilde{a}) - \tau \log \pi_\theta(\tilde{a} | s) - V_w(s)) \nabla_w V_w(s)$
 $g_v \leftarrow g_v - (r + \gamma \cdot (1 - \text{done}) \cdot V_w(s') - Q_\beta(s, a)) \nabla_v Q_\beta(s, a)$
 $g_w \leftarrow \frac{g_w}{|\mathcal{D}|}, g_v \leftarrow \frac{g_v}{|\mathcal{D}|}$

5. Theoretical Results on Policy Improvement Guarantees

We study the theoretical policy improvement guarantees under the RKL and FKL. We start with definitions and by motivating the choice of the entropy regularized setting. We then consider the guarantees, or lack thereof, for the RKL and FKL. We first provide an extension of Lemma 2 in (Haarnoja et al., 2018) to rely only upon RKL minimization *on average* across states, and then further extend the result to approximate action values. We then show we can obtain a more practical result, under an additional condition that the policy update does not take too big of a step away from the current policy.

Geist et al. (2019) performed error propagation analysis of entropy-regularized approximate dynamic programming algorithms, but they did not provide a monotonic policy improvement guarantee in the non-tabular setting, we do. Shani et al. (2020) analyzed TRPO and provide monotonic policy improvement guarantee (Lemma 15). However, their result relies on the tabular representation of the policy, and it does not necessarily apply to RKL reduction for general, non-tabular policies. Lemma 2 of Lan (2021) is the same result as our Theorem 6. Zhu and Matsubara (2020) provided monotonic policy improvement guarantee similar to Theorem 6. Compared to their result, we take a further step to show that RKL reduction on average across states suffices for monotonic policy improvement. Moreover, we provide additional results that do not depend on having the true action values or improbable state distributions.

Then we investigate the FKL, for which there are currently no existing policy improvements results. We show that the FKL does not have as strong of policy improvement guarantees as the RKL. We provide a counterexample where optimizing the FKL *does not* induce policy improvement. But, this counterexample does not imply that FKL reduction cannot provide policy improvement. We discuss further assumptions that can be made to ensure that FKL does induce policy improvement. All proofs are contained in Appendix B.

5.1 Definitions and Assumptions

We characterize performance of the policy in the entropy regularized setting. First, it will be useful to introduce some concepts for unregularized MDPs and then present their counterparts for entropy regularized MDPs. Throughout, we assume that the class of policies Π consists of policies whose entropies are finite. This assumption is not restrictive for finite action-spaces, as entropy is always finite in that setting; for continuous action-spaces, for most distributions used in practice like Gaussians with finite variance, the entropy will also be finite. The assumption of finite entropies is necessary to ensure that the soft value functions are well-defined.

For some of the theoretical results for the FKL, we will restrict our attention further to finite action-spaces, to ensure we have non-negative entropies and to use the total variation distance for discrete sets. We use sums instead of integrals throughout our proofs to enhance clarity; unless we explicitly assume a finite action space, all of our results hold as well for general action spaces given standard measure-theoretic assumptions.

Assumption 1. *Every $\pi \in \Pi$ has finite entropy: $\mathcal{H}(\pi(\cdot|s)) < \infty$ for all $s \in \mathcal{S}$.*

Definition 1 (Unregularized Performance Criterion). *For a start state distribution ρ_0 , the performance criterion is defined as*

$$\eta(\pi) := \mathbb{E}_{\rho_0}[V^\pi(S)].$$

Definition 2 (Unregularized Advantage). *For any policy π , the advantage is*

$$A^\pi(s, a) := Q^\pi(s, a) - V^\pi(s).$$

The advantage asks: what is the average benefit if I take action a in state s , as opposed to drawing an action from π ? The soft extensions of these quantities are as follows.

Definition 3 (Soft Performance Criterion). *For a start state distribution ρ_0 and temperature $\tau > 0$, the soft performance criterion is defined as*

$$\eta_\tau(\pi) := \mathbb{E}_{\rho_0}[V_\tau^\pi(S)].$$

It will also be helpful to have a soft version of the advantage. An intuition for the advantage in the non-soft setting is that it should be zero when averaged over π . To enforce this requirement in the soft setting, we require a small modification.

Definition 4 (Soft Advantage). *For a policy π and temperature $\tau > 0$, the soft advantage is*

$$A_\tau^\pi(s, a) := Q_\tau^\pi(s, a) - \tau \log \pi(a | s) - V_\tau^\pi(s).$$

If $\tau = 0$, we recover the usual definition of the advantage function. Like unregularized advantage functions, this definition also ensures $\mathbb{E}_\pi[A_\tau^\pi(s, A)] = 0$.

5.2 Why Use the Entropy Regularized Framework?

Since the actual goal of RL is to optimize the unregularized objective, it might sound unnatural to instead study guarantees in its regularized counterpart. We can view the entropy regularized setting as a surrogate for the unregularized setting, or simply of alternative interest. In the first case, it may be too difficult to optimize $\eta(\pi)$; entropy regularization can improve the optimization landscape and potentially promote exploration. Optimizing $\eta_\tau(\pi)$ is more feasible and can still get us close enough to a good solution of $\eta(\pi)$. In the second case, we may in fact want to reason about optimal stochastic policies, obtained through entropy regularization. In either setting, it is sensible to understand if we can obtain policy improvement guarantees under entropy regularization.

There have been several recent papers highlighting that entropy regularization can improve the optimization behavior of policy gradient algorithms. Mei et al. (2020b) studied how entropy regularization affects convergence rates in the tabular case, considering policies parametrized by a softmax. By using a proof technique based on Lojasiewicz inequalities, they were able to show that policy gradients without entropy regularization converge to the optimal policy at a $O(1/t)$ rate. Furthermore, they also showed a $\Omega(1/t)$ bound for this same method, concluding that the bound is unimprovable for vanilla policy gradients. By adding entropy regularization, the convergence rate can be improved to $O(e^{-t})$.

Ahmed et al. (2019) empirically studied how adding entropy regularization changes the optimization landscape for policy gradient methods. By sampling multiple directions in

parameter space for some suboptimal policy and visualizing scatter plots of curvature and gradient values around that policy, combined with visualization techniques that linearly interpolate policies, they concluded that adding entropy regularization likely connects local optima. The optimization landscape can be made smoother, while also allowing the use of higher learning rates.

Finally, Ghosh et al. (2020) provided theoretical justification that (nearly) deterministic policies can stall learning progress. They first provide an operator view of policy gradient methods, particularly showing that REINFORCE can be seen as a repeated application of an improvement operator and a projection operator (Ghosh et al., 2020, Proposition 1). They then showed (Ghosh et al., 2020, Proposition 5) that the performance of the (non-projected) improved policy π' , $\eta(\pi')$, is equal to $\eta(\pi_{\text{old}})$ times a term including the variance: $\eta(\pi') = \eta(\pi_{\text{old}})(1 + \frac{\text{Variance of Return under } \pi_{\text{old}}}{\text{Expected Return under } \pi_{\text{old}}}) \geq \eta(\pi_{\text{old}})$. This means that if the variance under π_{old} is near zero, then $\eta(\pi') \approx \eta(\pi_{\text{old}})$. In that sense, having higher variance can help the algorithm make consistent progress. A common way of achieving higher variance is by adding entropy regularization.

Finally, there is some theoretical work relating the solutions under the unregularized and regularized objectives. From (Geist et al., 2019, Proposition 3), if the entropy is bounded for all policies with constants L_τ, U_τ giving $L_\tau \leq -\tau\mathcal{H}(\pi) \leq U_\tau$, then we know that $V^\pi(s) - \frac{U_\tau}{1-\gamma} \leq V_\tau^\pi(s) \leq V^\pi(s) - \frac{L_\tau}{1-\gamma}$. Using this result, we can take expectations across the state space with respect to the starting state distribution, to get

$$\eta(\pi) - \frac{U_\tau}{1-\gamma} \leq \eta_\tau(\pi) \leq \eta(\pi) - \frac{L_\tau}{1-\gamma}.$$

Hence, if the upper bound is tight, increasing $\eta_\tau(\pi)$ will increase $\eta(\pi)$. A similar result exists for single-step decision making with discrete actions (Chen et al., 2019, Proposition 2).

5.3 Policy Improvement with the RKL

First, we note a strengthening of the original result for policy improvement under RKL reduction (Haarnoja et al., 2018). Particularly, they take π_{new} to be the policy that *minimizes* the RKL to $\mathcal{B}_\tau Q_\tau^{\pi_{\text{old}}}(s, \cdot)$ at every state. Examining their proof reveals that their new policy π_{new} does not have to be the minimizer; rather, it suffices that π_{new} is smaller in RKL than π_{old} at every state s . We therefore restate their lemma with this slight modification.

Lemma 5 (Restatement of Lemma 2 (Haarnoja et al., 2018)). *For $\pi_{\text{old}}, \pi_{\text{new}} \in \Pi$, if for all s*

$$\text{RKL}_{\mathcal{B}_\tau Q_\tau^{\pi_{\text{old}}}}^{\pi_{\text{new}}}(s) \leq \text{RKL}_{\mathcal{B}_\tau Q_\tau^{\pi_{\text{old}}}}^{\pi_{\text{old}}}(s),$$

then $Q_\tau^{\pi_{\text{new}}}(s, a) \geq Q_\tau^{\pi_{\text{old}}}(s, a)$ for all (s, a) and $\tau > 0$.

Proof Same proof as in Haarnoja et al. (2018). ■

We extend this result by considering an RKL reduction in average across states, rather than requiring RKL reduction in every state. To prove our result, it will be useful to prove a soft counterpart to the classical performance difference lemma (Kakade and Langford, 2002).

Lemma 6. *[Soft Performance Difference] For any policies $\pi_{\text{old}}, \pi_{\text{new}}$, any $\tau \geq 0$, we have*

$$\eta_\tau(\pi_{\text{new}}) - \eta_\tau(\pi_{\text{old}}) = \frac{1}{1-\gamma} \mathbb{E}_{d^{\pi_{\text{new}}}} \left[\mathbb{E}_{\pi_{\text{new}}} [A_\tau^{\pi_{\text{old}}}(S, A)] - \tau \text{KL}_{\pi_{\text{old}}}^{\pi_{\text{new}}}(S) \right].$$

If we set $\tau = 0$, we recover the classical performance difference lemma. Now, we can show that reducing the RKL on average is sufficient and necessary for policy improvement.

Proposition 7. *[Improvement Under Average RKL Reduction] For $\pi_{\text{old}}, \pi_{\text{new}} \in \Pi$, define*

$$\Delta\text{RKL}_{\pi_{\text{new}}}^{\pi_{\text{old}}}(S) := \text{RKL}_{\mathcal{B}_\tau Q_\tau^{\pi_{\text{old}}}}^{\pi_{\text{old}}}(S) - \text{RKL}_{\mathcal{B}_\tau Q_\tau^{\pi_{\text{old}}}}^{\pi_{\text{new}}}(S).$$

$$\text{For } \tau > 0: \quad \eta_\tau(\pi_{\text{new}}) - \eta_\tau(\pi_{\text{old}}) = \frac{\tau}{1 - \gamma} \mathbb{E}_{d^{\pi_{\text{new}}}} [\Delta\text{RKL}_{\pi_{\text{new}}}^{\pi_{\text{old}}}(S)]. \quad (15)$$

Furthermore, $\eta_\tau(\pi_{\text{new}}) \geq \eta_\tau(\pi_{\text{old}})$ if and only if $\mathbb{E}_{d^{\pi_{\text{new}}}} [\Delta\text{RKL}_{\pi_{\text{new}}}^{\pi_{\text{old}}}(S)] \geq 0$.

This result shows that reducing the RKL on average, under weighting $d^{\pi_{\text{new}}}$, guarantees improvement. Notice that the more stringent condition of RKL reduction in every state from Theorem 5 ensures reduction under the weighting $d^{\pi_{\text{new}}}$; this new result is therefore more general. Ensuring reduction under $d^{\pi_{\text{new}}}$, however, may be difficult in practice, as we do not have access to data under π_{new} ; rather, we have data from π_{old} . We extend this result in Section 5.3.2, to a weighting under π_{old} , by adding a condition on how far π_{new} moves from π_{old} . This more practical result relies on the above result, and so we present the above result first as a standalone to highlight the key reason for the policy improvement.

This result provides some theoretical support for using stochastic gradient descent for the RKL, as is done in practice. It is unlikely that we will completely minimize the RKL on every step, nor reduce it in every state. With sufficient reduction of the average RKL on each step, this iterative procedure between approximate greedification and exact policy evaluation should converge to an optimal policy. In fact, by inspecting Equation (15), we can see that any optimal policy satisfies, for any fixed π_0 ,

$$\pi^* \in \arg \max_{\pi \in \Pi} \eta_\tau(\pi) = \arg \max_{\pi \in \Pi} \eta_\tau(\pi) - \eta_\tau(\pi_0) = \arg \max_{\pi \in \Pi} \frac{\tau}{1 - \gamma} \mathbb{E}_{d^\pi} [\Delta\text{RKL}_\pi^{\pi_0}(S)]$$

Note, however, that we cannot generally guarantee that this procedure will converge to the optimal policy. This is because the average RKL reduction may decrease to zero prematurely, in the sense that $\lim_{i \rightarrow \infty} \eta_\tau(\pi_i)$ may be less than $\sup_{\pi \in \Pi} \eta_\tau(\pi)$.

5.3.1 EXTENSION TO ACTION-VALUE ESTIMATES

In the previous section we focused on approximate greedification with exact action-values. The theoretical results allowed for improvements on average across the state space, better reflecting what is done in practice. However, algorithms in practice are also not likely to have exact action-values. In this section, we further extend the theoretical results to allow for both approximate greedification and approximation policy evaluation.

First, we prove an analogue of the soft performance difference lemma for approximate action-values.

Lemma 8. *[Approximate Soft Performance Difference] Let $\pi_{\text{old}}, \pi_{\text{new}}$ be any policies and let $\tau \geq 0$. Let $\hat{Q} : \mathcal{S} \times \mathcal{A} \rightarrow \mathbb{R}$ represent an action-value estimate and let $\epsilon(s, a) := \hat{Q}(s, a) - Q_\tau^{\pi_{\text{old}}}(s, a)$ be the per-state approximation error. As well, define*

$$\begin{aligned} \hat{\Delta}\text{RKL}_{\pi_{\text{new}}}^{\pi_{\text{old}}}(S) &:= \text{RKL}_{\mathcal{B}_\tau \hat{Q}}^{\pi_{\text{old}}}(S) - \text{RKL}_{\mathcal{B}_\tau \hat{Q}}^{\pi_{\text{new}}}(S), \\ \bar{\epsilon} &:= \mathbb{E}_{d^{\pi_{\text{new}}}} [\mathbb{E}_{\pi_{\text{new}}} [\epsilon(S, A)]] - \mathbb{E}_{\pi_{\text{old}}} [\epsilon(S, A)]. \end{aligned}$$

$$\text{We have: } \eta_\tau(\pi_{\text{new}}) - \eta_\tau(\pi_{\text{old}}) + \bar{\epsilon} = \frac{\tau}{1-\gamma} \mathbb{E}_{d^{\pi_{\text{new}}}} [\hat{\Delta} \text{RKL}_{\pi_{\text{new}}}^{\pi_{\text{old}}}(S)]. \quad (16)$$

As a corollary, we have a corresponding policy improvement result.

Corollary 9. *[Approximate RKL Reduction] Under the assumptions of Theorem 8, $\eta_\tau(\pi_{\text{new}}) \geq \eta_\tau(\pi_{\text{old}})$ iff $\frac{\tau}{1-\gamma} \mathbb{E}_{d^{\pi_{\text{new}}}} \Delta \text{RKL}_{\pi_{\text{new}}}^{\pi_{\text{old}}}(S) \geq \bar{\epsilon}$.*

If we only have access to an estimate \hat{Q} of $Q_\tau^{\pi_{\text{old}}}$, reducing the RKL is not enough to guarantee policy improvement. When one reduces the RKL, one must also take care that $\bar{\epsilon}$ not be larger than the RKL amount reduced. The quantity $\bar{\epsilon}$ represents the difference in average approximation error over the action distributions of π_{new} and π_{old} . For $\bar{\epsilon}$ to be small, the approximation error averaged over the π_{new} distribution should not be much larger than the approximation error averaged over the π_{old} distribution. For example, $\bar{\epsilon}$ could be small if π_{new} is similar to π_{old} , or if \hat{Q} already approximates $Q_\tau^{\pi_{\text{old}}}$ well under the state-action distribution induced by π_{new} .

5.3.2 EXTENSIONS TO WEIGHTING UNDER $d^{\pi_{\text{old}}}$

CPI (Kakade and Langford, 2002) derives a lower bound of the performance difference under which one is able to guarantee policy improvement using $d^{\pi_{\text{old}}}$, assuming π_{new} is sufficiently close to π_{old} . Similar considerations allow us to derive a corresponding bound in our setting.

Proposition 10. *If $\text{KL}_{\pi_{\text{old}}}^{\pi_{\text{new}}}(s) \leq \alpha$ for all $s \in \mathcal{S}$,*

$$\begin{aligned} \eta_\tau(\pi_{\text{new}}) - \eta_\tau(\pi_{\text{old}}) &\geq \frac{1}{1-\gamma} \mathbb{E}_{d^{\pi_{\text{old}}}} \left[\sum_a \pi_{\text{new}}(a|S) A_\tau^{\pi_{\text{old}}}(S, a) - \tau \text{KL}_{\pi_{\text{old}}}^{\pi_{\text{new}}}(S) - 4\sqrt{2}V_{\tau, \max} \sqrt{\alpha} \right] \\ &= \frac{1}{1-\gamma} \mathbb{E}_{d^{\pi_{\text{old}}}} \left[\Delta \text{RKL}_{\pi_{\text{new}}}^{\pi_{\text{old}}} - 4\sqrt{2}V_{\tau, \max} \sqrt{\alpha} \right] \end{aligned} \quad (17)$$

With knowledge of $V_{\tau, \max}$ and the exact value functions, one could optimize the lower bound in Theorem 10 as a function of both α and π_{new} , without knowledge of $d^{\pi_{\text{old}}}$. If $V_{\tau, \max}$ is large, then α must be rather small to ensure that the RHS of Theorem 10 is non-negative. Intuitively, the larger the maximum return, the greater the possible error in using $d^{\pi_{\text{old}}}$ rather than $d^{\pi_{\text{new}}}$.

We can also combine Theorem 10 and Theorem 8.

Proposition 11. *If $\text{KL}_{\pi_{\text{old}}}^{\pi_{\text{new}}}(s) \leq \alpha$ for all $s \in \mathcal{S}$,*

$$\begin{aligned} \eta_\tau(\pi_{\text{new}}) - \eta_\tau(\pi_{\text{old}}) &\geq \frac{1}{1-\gamma} \mathbb{E}_{d^{\pi_{\text{old}}}} \left[\tau \hat{\Delta} \text{RKL}_{\pi_{\text{new}}}^{\pi_{\text{old}}}(S) \right. \\ &\quad \left. + \sum_a \epsilon(s, a) (\pi_{\text{old}}(a) - \pi_{\text{new}}(a)) - 4V_{\tau, \max} \sqrt{2\alpha} \right]. \end{aligned} \quad (18)$$

5.4 Policy Improvement with the FKL

In this section, we study the policy improvement properties of reducing the FKL. First, in Section 5.4.1 we provide a counterexample showing that reducing the FKL leads to a

strictly worse policy. Second, in Section 5.4.2 we provide a sufficient condition on the FKL reduction to ensure policy improvement. The plots in that section show that this bound is non-trivial, but that unfortunately the required reduction is close to the maximum possible reduction. Third, in Section 5.4.3, we discuss when reducing the FKL may be used as a surrogate for reducing the RKL, in particular by providing an upper bound for the RKL in terms of the FKL. It will turn out that reducing the FKL alone is insufficient for reducing the RKL because this bound involves not only the FKL, but another term that depends upon π_{new} . We conclude with a discussion about the implications for the use of FKL for approximate greedification.

5.4.1 COUNTEREXAMPLE FOR POLICY IMPROVEMENT UNDER FKL REDUCTION

Unfortunately, the FKL does not enjoy the same policy improvement guarantees as the RKL. In the next proposition, we provide a counterexample where reducing the FKL makes the policy worse. The intuition behind this example is that π_{old} almost always chooses the good action, but is made close to deterministic and thus arbitrarily large in FKL to $\mathcal{B}_\tau Q_\tau^{\pi_{\text{old}}}$, while π_{new} , by being less deterministic, reduces the FKL to $\mathcal{B}_\tau Q_\tau^{\pi_{\text{old}}}$ but it almost always chooses the bad action, thus being worse in the soft-objective.

Again we use the notation

$$\Delta\text{FKL}_{\pi_{\text{new}}}^{\pi_{\text{old}}}(s) := \text{FKL}_{\mathcal{B}_\tau Q_\tau^{\pi_{\text{old}}}}^{\pi_{\text{old}}}(s) - \text{FKL}_{\mathcal{B}_\tau Q_\tau^{\pi_{\text{old}}}}^{\pi_{\text{new}}}(s)$$

where $\Delta\text{FKL}_{\pi_{\text{new}}}^{\pi_{\text{old}}}(s) > 0$ means we obtained FKL reduction: the new policy has lower FKL than the old policy.

Proposition 12. *[FKL Counterexample] There exists an MDP such that, for any $\tau \geq 0$, there exists a pair of policies $(\pi_{\text{old}}, \pi_{\text{new}})$ where $\Delta\text{FKL}_{\pi_{\text{new}}}^{\pi_{\text{old}}}(s) > 0$ at every state but the new policy has lower value: $Q_\tau^{\pi_{\text{new}}}(s, a) < Q_\tau^{\pi_{\text{old}}}(s, a)$ at every state-action pair $(s, a) \in \mathcal{S} \times \mathcal{A}$, $V_\tau^{\pi_{\text{new}}}(s) < V_\tau^{\pi_{\text{old}}}(s)$ at every state $s \in \mathcal{S}$, and $\eta_\tau(\pi_{\text{new}}) < \eta_\tau(\pi_{\text{old}})$.*

5.4.2 POLICY IMPROVEMENT UNDER SUFFICIENT FKL REDUCTION

We know that completely reducing the FKL, under the assumption that we can represent all policies, will guarantee improvement, since then we would have $\pi_{\text{new}} = \mathcal{B}_\tau Q_\tau^{\pi_{\text{old}}}$. One might hope that additional conditions, that ensure *sufficient* reduction in the FKL, might imply policy improvement. Because policy improvement is obtained if and only if the RKL is reduced, as per Theorem 7, we can equivalently ask what conditions on the FKL ensure we obtain RKL reduction. In this section, we provide a lower bound on the FKL reduction, that guarantees the RKL is reduced and so the policy is improved. We numerically investigate the magnitude of required reduction under this condition, to see how much lower it is than completely reducing the FKL.

As before, we could first prove this result under reduction per-state; this is a more restricted setting that implies reduction in average across the states. We therefore provide only the more general result, which averages across states, as the connection between per-state and across state has already been clearly shown above and is not useful to repeat.

Proposition 13 (Improvement Under Average Sufficient FKL Reduction). *Assume the action set is finite. If*

$$\mathbb{E}_{d^{\pi_{\text{new}}}} \left[\text{RKL}_{\mathcal{B}_\tau Q_\tau^{\pi_{\text{old}}}}^{\pi_{\text{old}}}(S) + \mathbb{E}_{\mathcal{B}_\tau Q_\tau^{\pi_{\text{old}}}} [\log \pi_{\text{old}}(\cdot|S)] \right] \geq 0, \quad (19)$$

$$\begin{aligned} \mathbb{E}_{d^{\pi_{\text{new}}}} [\Delta \text{FKL}_{\pi_{\text{new}}}^{\pi_{\text{old}}}(S)] &\geq \mathbb{E}_{d^{\pi_{\text{new}}}} [\text{FKL}_{\mathcal{B}_\tau Q_\tau^{\pi_{\text{old}}}}^{\pi_{\text{old}}}(S)] - \frac{1}{2} \left(\frac{\tau}{\|Q_\tau^{\pi_{\text{old}}}\|_\infty} \left(\mathbb{E}_{d^{\pi_{\text{new}}}} \left[\text{RKL}_{\mathcal{B}_\tau Q_\tau^{\pi_{\text{old}}}}^{\pi_{\text{old}}}(S) \right] \right. \right. \\ &\quad \left. \left. + \mathbb{E}_{d^{\pi_{\text{new}}}} \left[\mathbb{E}_{\mathcal{B}_\tau Q_\tau^{\pi_{\text{old}}}} [\log(\pi_{\text{old}}(\cdot|S))] \right] \right) \right)^2, \end{aligned} \quad (20)$$

$$\text{and } \forall s \in \mathcal{S}, \Delta \text{FKL}_{\pi_{\text{new}}}^{\pi_{\text{old}}}(S) \geq 0, \quad (21)$$

then $\eta_\tau(\pi_{\text{new}}) \geq \eta_\tau(\pi_{\text{old}})$.

Equation (19) essentially says that $\mathbb{E}_{d^{\pi_{\text{new}}}} \left[\text{RKL}_{\mathcal{B}_\tau Q_\tau^{\pi_{\text{old}}}}^{\pi_{\text{old}}}(S) \right]$ is greater than or equal to $\mathbb{E}_{d^{\pi_{\text{new}}}} \left[\text{FKL}_{\mathcal{B}_\tau Q_\tau^{\pi_{\text{old}}}}^{\pi_{\text{old}}}(S) + \mathcal{H}(\mathcal{B}_\tau Q_\tau^{\pi_{\text{old}}}(S, \cdot)) \right]$. When might the RKL be larger than the FKL and the entropy? If $\mathcal{B}_\tau Q_\tau^{\pi_{\text{old}}}$ has low entropy across states, then $\mathcal{B}_\tau Q_\tau^{\pi_{\text{old}}}$ will have low probability mass placed on certain actions. If π_{old} places probability mass on these actions, the RKL will likely be high because the RKL incentivizes mode-matching. Unfortunately, this result also assumes that FKL reduction is non-negative in all states.

Just as with the RKL, we can extend these results to use an action-value estimate \hat{Q} . We provide these extensions and their proofs in Appendix B.2.1.

We know that fully reducing the FKL, so that π_{new} equals $\mathcal{B}_\tau Q_\tau^{\pi_{\text{old}}}$, guarantees improvement, which is a strong requirement; we can ask how much less strict the above condition is in comparison. We can check this numerically, in a simple bandit setting with $|\mathcal{A}| = 5$. We test different π_{old} calculated as:

$$\pi_{\text{old}} = (1 - \lambda)\pi_{\text{rand}} + (\lambda)\mathcal{B}_\tau Q$$

for $\lambda \in \{0, 1/3, 2/3, 0.99\}$, π_{rand} the random policy and each of these policies corresponding to probability vectors over the 5 actions. Varying λ allows us to see the impact on the bound for policies far from the target $\mathcal{B}_\tau Q$ ($\lambda = 0$) to very close to the target ($\lambda = 0.99$). Additionally, the temperature plays an important role in the bound. We therefore measure the bound for a variety of τ , for each λ . We include the results for 30 seeds in Figure 3.

We can see that the bound is very conservative and a near-maximum FKL reduction is necessary for many temperatures. These plots suggest that additional conditions are needed for FKL reduction to guarantee improvement, as we discuss further at the end of this section.

5.4.3 UPPER BOUNDING THE RKL IN TERMS OF THE FKL

The primary goal for this section is to highlight a potential direction towards connecting minimizing FKL to the RKL. We show that the FKL times a term that depends on the new and old policies gives an upper bound on the RKL. We discuss how this connection provides insight into why FKL reduction may not result in improvement. We omit dependence on

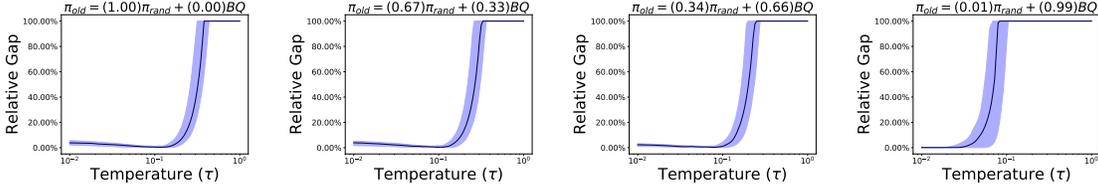


Figure 3: The Relative Gap between the maximal ΔFKL and our bound. We measure the ratio between our bound on the required level of reduction in Equation (20) and the maximum possible FKL reduction (i.e. $\Delta\text{FKL} = \text{FKL}_{\mathcal{B}_\tau Q}^{\pi_{\text{old}}}$). The Relative Gap is 1-ratio, where values close to zero indicate near-maximal reduction is necessary for improvement in the soft objective. We show the result for policies far from the target (leftmost) to policies very close to the target (rightmost). The solid line is the median over 30 runs, with the shaded region showing the 25% percentile and the 75% percentile.

the state in the following result, but it holds per state. This result is a straightforward application of a result from Sason and Verdú (2016). The result uses the Rényi divergence of order ∞ :

$$D_\infty(P \parallel Q) := \log \left(\max_i \frac{p_i}{q_i} \right)$$

where P and Q are two discrete probability distributions, with elements p_i and q_i respectively, that are absolutely continuous with respect to each other (i.e. one is never nonzero where the other one is zero).

Lemma 14 (An Upper Bound on RKL in Terms of the FKL). *Assume the action set is finite. For $\kappa(t) := \frac{t \log t + 1 - t}{t - 1 - \log t}$ where κ is defined on $(0, 1) \cup (1, \infty)$ and all $s \in \mathcal{S}$,*

$$\text{RKL}_{\mathcal{B}_\tau Q_\tau}^{\pi_{\text{new}} \pi_{\text{old}}}(s) \leq \kappa(\exp(D_\infty(\pi_{\text{new}} \parallel \mathcal{B}_\tau Q_\tau^{\pi_{\text{old}}})) \text{FKL}_{\mathcal{B}_\tau Q_\tau}^{\pi_{\text{new}} \pi_{\text{old}}}(s) \quad (22)$$

Proof To obtain this result, we bound the difference between the two choices of KL divergence. Define

$$\beta_1 := \exp(-D_\infty(P \parallel Q)) \quad \text{and} \quad \beta_2 := \exp(-D_\infty(Q \parallel P)),$$

Then, from Equation (161) in Sason and Verdú (2016), as long as $P \neq Q$, we have

$$\kappa(\beta_2) \leq \frac{\text{KL}_Q^P}{\text{KL}_P^Q} \leq \kappa(\beta_1^{-1}).$$

Setting $P = \pi_{\text{new}}$ and $Q = \mathcal{B}_\tau Q_\tau^{\pi_{\text{old}}}$ at a particular state s , where we omit the dependence on s , we have

$$\begin{aligned} \text{RKL}_{\mathcal{B}_\tau Q_\tau}^{\pi_{\text{new}} \pi_{\text{old}}}(s) &\leq \kappa(\beta_1^{-1}) \text{FKL}_{\mathcal{B}_\tau Q_\tau}^{\pi_{\text{new}} \pi_{\text{old}}}(s) \\ &= \kappa(\exp(D_\infty(\pi_{\text{new}} \parallel \mathcal{B}_\tau Q_\tau^{\pi_{\text{old}}})) \text{FKL}_{\mathcal{B}_\tau Q_\tau}^{\pi_{\text{new}} \pi_{\text{old}}}(s) \end{aligned} \quad (23)$$

■

To reduce the RKL as a function of π_{new} , it thus suffices to reduce the right-hand side of the inequality. There are, however, problems with this approach. First, the bound itself may not be tight; even if we could reduce FKL and the multiplicand $\kappa(\beta_1^{-1})$, we still may not obtain a reduction in RKL. Second, we have only developed a mechanism to reduce the FKL, rather than the FKL and the multiplicand. A simple proxy could be to just focus on reducing the FKL.

The bound given above includes $\kappa(\beta_1^{-1})$, which also depends on π_{new} . It is possible that in reducing the FKL, we actually also increase $\kappa(\beta_1^{-1})$, possibly offsetting our reduction of the FKL. For example, because of limited function approximation capacity, reducing the FKL might result in π_{new} covering a low-probability region of $\mathcal{B}_\tau Q_\tau^{\pi_{\text{old}}}$ in order to place some mass at multiple high-probability regions of $\mathcal{B}_\tau Q_\tau^{\pi_{\text{old}}}$. While such a π_{new} might have a moderate value of FKL, the resulting $D_\infty(\pi_{\text{new}} \parallel \mathcal{B}_\tau Q_\tau^{\pi_{\text{old}}})$ would be large, making β_1^{-1} large. Correspondingly, because κ is a monotone increasing function (Sason and Verdú, 2016), $\kappa(\beta_1^{-1})$ would also be large. Consequently, $\kappa(\beta_1^{-1})\text{FKL}_{\mathcal{B}_\tau Q_\tau^{\pi_{\text{old}}}}^{\pi_{\text{new}}}$ may not be small enough to enforce a reduction in RKL.

On a more positive note, however, we know that the κ term in Equation (23),

$$\kappa(\beta_1^{-1}) = \kappa(\exp(D_\infty(\pi_{\text{new}} \parallel \mathcal{B}_\tau Q_\tau^{\pi_{\text{old}}})) = \kappa\left(\max_i \frac{(\pi_{\text{new}})_i}{(\mathcal{B}_\tau Q_\tau^{\pi_{\text{old}}})_i}\right),$$

only grows logarithmically with $\max_i \frac{(\pi_{\text{new}})_i}{(\mathcal{B}_\tau Q_\tau^{\pi_{\text{old}}})_i}$. Particularly,

$$\lim_{x \rightarrow \infty} \frac{\kappa(x)}{\log(x)} = 1 \quad \text{and so} \quad \kappa(x) = \Theta(\log(x)).$$

Therefore, β_1^{-1} has to increase by orders of magnitude to significantly increase $\kappa(\beta_1^{-1})$.

A modification to the FKL reduction strategy could be to use $\kappa(\beta_1^{-1})\text{FKL}_{\mathcal{B}_\tau Q_\tau^{\pi_{\text{old}}}}^{\pi_{\text{new}}}$ as an objective. The main difficulty with this approach is that β_1 is not differentiable because of the max operation in the calculation of $D_\infty(\pi_{\text{new}} \parallel \mathcal{B}_\tau Q)$. It might be possible to approximate this maximum with smooth operations like LogSumExp, but we leave exploration of this avenue for future work.

5.5 Summary and Discussion

There are two key takeaways from the above results. First, the RKL has a stronger policy improvement result than the FKL as it requires only that the RKL of π_{new} be no greater than the RKL of π_{old} . In fact, RKL reduction under a certain state-distribution is a necessary and sufficient condition for improvement to occur. Second, the FKL can fail to induce policy improvement, but sufficient reduction guarantees such improvement. The current bounds, although sufficient, are not a necessary condition for improvement to occur.

The theoretical results suggest that the FKL is inferior to the RKL for improving the policy, and that the FKL requires additional conditions. We hypothesize that the nature of these conditions has to do with the mean-seeking and mode-seeking behavior. Approximating a target distribution via RKL reduction is very sensitive to placing non-negligible probabilities in regions where the target distribution is close to zero. The FKL, on the other hand, focuses on placing high probabilities in the regions where the target probability is high. It is not

hard to see that reducing FKL can increase RKL, if for example we approximate a bimodal distribution with a unimodal one, or if we use a Gaussian parameterization and the target distribution is highly skewed. As we discussed, to obtain improvement, we need a sufficient reduction in the FKL to ensure RKL reduction. Our bound provided one such condition, but as we found with numerical experiments, this bound was relatively loose. It remains an open question to understand the conditions that guarantee improvement, and when FKL reduction does not give RKL reduction.

The settings where the RKL and FKL are significantly different—meaning that FKL reduction can actually cause the RKL to increase—may not be as prevalent in practice. For example, if the target distributions are unimodal and symmetric, we may find that RKL and FKL have similar empirical performance. We will see in our experiments that the FKL is often able to induce policy improvement in practice, suggesting a gap between the theory developed and the practical performance.

An important next step is to leverage these policy improvement results to prove convergence to an optimal policy under approximate greedification. When completely reducing the RKL per state, it is known that the iterative procedure between policy evaluation and greedification with RKL minimization converges to the optimal policy in the policy set (Haarnoja et al., 2018, Theorem 1). This result should similarly hold, under only RKL reduction, as long as that reduction is sufficient on each step. A next step is to understand the conditions on how much reduction is needed per step, for both the RKL and FKL, to obtain this result.

6. Empirical Results Comparing the FKL and RKL

In this section, we complement the theoretical results with an investigation of the other practical properties of the FKL and RKL. The theory focused on their differences in terms of inducing policy improvement. We also care about (1) the optimization behavior of these greedification operators, when using (stochastic) gradient descent and (2) the nature of the policies induced during learning, in particular whether stochasticity collapses quickly and differences in encourage exploratory behavior. We may also want to understand generally how these two approaches perform in the wild, on a suite of problems. We investigate these three questions empirically in this section.

6.1 Optimization Behavior in Microworlds

The goal in this section is to understand differences between FKL and RKL in terms of (1) the loss surface and (2) the behavior of iterates optimized under the losses. By behavior, we mean whether the iterates reach multiple local optima, how stable iterates under that loss are, and how often iterates reach the global optimum (or optima). Given the fine-grained nature of our questions, we focus upon small-scale environments, which we call *microworlds*. Doing so allows us to avoid any possible confounding factors associated with larger, more complicated environments, and furthermore allows us to more fully separate any issues to do with stochasticity.

We use two continuous action low-dimensional microworlds to allow us to visualize and thoroughly investigate behavior. Our first microworld is a **Bimodal Bandit** in Figure 4a. For continuous actions, we designed a continuous bandit with action space $[-1, 1]$ and

reward function $Q(a) := \exp(-\frac{1}{2}(\frac{2a+1}{0.2})^2) + \frac{3}{2} \exp(-\frac{1}{2}(\frac{2a-1}{0.2})^2)$. The two unequal modes at -0.5 and 0.5 enable us to test the mean-seeking and mode-seeking behavior as well as simulate a realistic scenario where the agent’s policy parameterization (here, unimodal) cannot represent the true distribution (bimodal).

Our second microworld is the **Switch-Stay** domain in Figure 4b. From s_0 , action 0 (stay) gives a reward of 1 and transitions to state 0. From s_1 , action 0 gives a reward of 2 and transitions to s_1 . From s_0 , action 1 (switch) gives a reward of -1 and transitions to s_1 , while action 1 from s_1 gives a reward of 0 and transitions to s_0 . To adapt this environment to the continuous action setting, we treat actions > 0 as switch and actions ≤ 0 as stay.⁸ We set $\gamma = 0.9$ to ensure that the optimal action from s_0 is to switch, which ensures the existence of a short-term/long-term trade-off inherent to realistic RL environments.

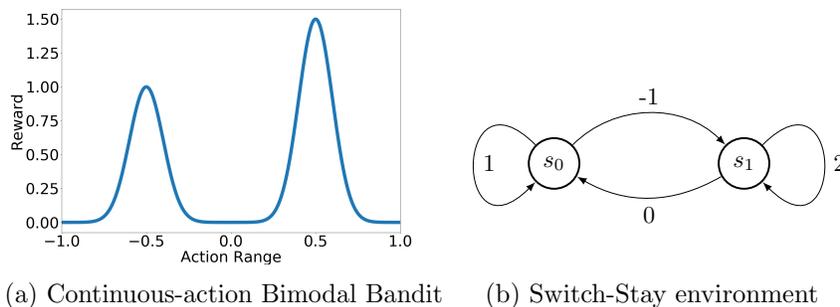


Figure 4: The Microworld environments used to investigate and visualize the optimization behavior of FKL and RKL.

6.1.1 IMPLEMENTATION DETAILS

All policies are tabular in the state. To calculate the FKL and RKL under continuous actions, we use the Clenshaw-Curtis (Clenshaw and Curtis, 1960) numerical integration scheme with 1024 points from the package quadpy,⁹ excluding the first and the last points at -1 and 1 because of numerical stability. We use the true action-values when calculating the KL losses. In the Bimodal Bandit, the action-value is given by the reward function, while in Switch-Stay it is calculated (i.e., not learned). To calculate the Hard FKL, we use the true maximum action as determined by the environment. For Switch-Stay, we calculate and optimize the mean KL across the two states.

For policy parameterizations, in continuous action settings we use a Gaussian policy with mean and standard deviation learned as $(\hat{\mu}, \log(1 + \exp(\hat{\sigma})))$. The action sampled from the learned Gaussian is passed through tanh to ensure that the action is in the feasible range $[-1, 1]$ and to avoid the bias induced in the policy gradient when action ranges are not enforced (Chou et al., 2017).

8. Note that we also compared the RKL and FKL for the discrete action variant of Switch-Stay. Under a softmax parameterization, we found no significant differences between the RKL and FKL.

9. <https://pypi.org/project/quadpy/>

Finally, we use the RMSprop optimizer (Tieleman and Hinton, 2012). Overall trends for Adam (Kingma and Ba, 2015) were similar to those for RMSprop, while results for SGD resulted in slower learning for both FKL and RKL and a wider range of limit points, most likely due to oscillation from the constant step-size. We focus on RMSprop here to avoid any confounding factors associated with momentum.

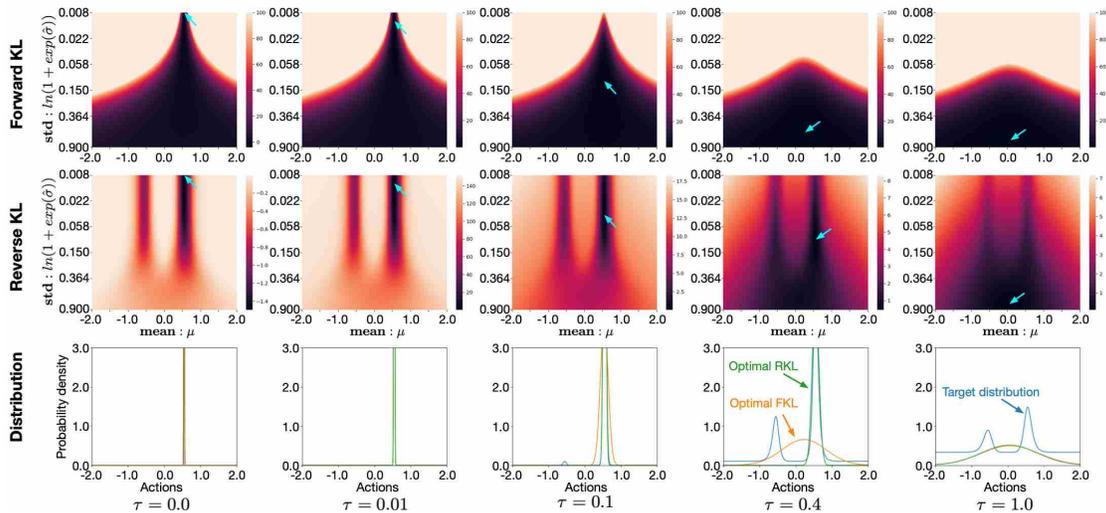


Figure 5: KL loss over mean and standard deviation across temperature. The heatmaps depict the loss for each mean and standard deviation pair. The last row depicts the target distribution over which the KL loss is optimized. Note that the actual action taken applies \tanh to the samples of the resulting distribution (i.e., the optimal mean is at $\tanh^{-1}(0.5) \approx 0.55$). FKL loss has been upper-bounded for better visualization of minima. Arrows indicate the global minimum.

6.1.2 LOSS SURFACE IN THE BIMODAL BANDIT

We might expect the FKL to have a smoother loss surface. Given that policies often are part of an exponential family (e.g., softmax policy), having the policy π be the second argument of KL_q^p removes the exponential of π , resulting in an objective that is an affine function of the features. For example, if $\pi(a | s) \propto \exp(\phi(s, a))$ for features (s, a) , the resulting FKL becomes a sum of a term that is linear in $\phi(s, a)$ and a term involving $\text{LogSumExp}(\phi)$, which is convex.

We visualize the KL loss surfaces in Figure 5 with five different temperatures. The surfaces suggest the following.

- 1) The FKL surface has a single valley, while the RKL surface has two valleys that are separated from one another. In this sense, the FKL surface seems much smoother than the RKL surface, suggesting that iterates under the FKL will more likely reach the global optimum than iterates under the RKL, which seem likely to fall into either of the valleys.

2) The smoothness of the RKL landscape increases with temperature as the gap between the peaks becomes less steep. A higher temperature also causes the valley in the FKL map to become less sharply peaked, and for the optimal μ to move closer to 0.

3) The optimal μ for the FKL seems to move more quickly to zero, as τ increases, than the optimal μ for the RKL, although both eventually reach 0. It is possible that the FKL may become suboptimal sooner than the RKL as τ increases, likely because it is mean-seeking. Interestingly, even the RKL appears to be mean-seeking for high τ , because selecting one mode would have lower entropy.

As a note, it may seem strange that two valleys exist for the RKL at $\tau = 0$ given that the target distribution is unimodal. When $\tau = 0$, however, the loss function is no longer a distributional loss; that is, we are no longer minimizing any pseudo-distance between the policy and a distribution.

6.1.3 SOLUTION QUALITY IN SWITCH-STAY

In this section, we investigate the properties of the solutions under the FKL and RKL for an environment with more than one state. Given our previous results, we might expect the FKL to result in better solutions, because FKL iterates can reach the global optimum more easily. But this depends on the quality of this solution. The global minimum of the FKL objective may not correspond well with the optimal solution of the original, unregularized objective, as we investigate below.

The Switch-Stay environment is appropriate to investigate the quality of the stationary points of the RKL and FKL for two reasons. First, it is a simple instantiation of the full RL problem, we are interested in understanding any possible differences between FKL and RKL in the presence of short-term/long-term trade-offs. On Switch-Stay, the agent incur a short-term penalty by switching from state 0 to state 1, but longer term this maximizes return. Second, the Switch-Stay environment facilitates visualization. Since the MDP has only two states, we can plot any value function as a point on a 2-dimensional plane. In particular, one can view the entire space of value functions, shown recently to be a polytope in the discrete-action setting (Dadashi et al., 2019).

We can similarly visualize the value function polytope for continuous actions in Switch-Stay. Recall that we treat any action ≤ 0 as stay, and any action > 0 as switch. To calculate the value function corresponding to a continuous policy π , we convert π to an equivalent discrete policy π_{discrete} of the underlying discrete MDP. The conversion requires the calculation of the probability that π outputs an action ≤ 0 in each state, which we do with numerical integration of the policy PDF. We then calculate the value function of π as $(I - \gamma P_{\pi_{\text{discrete}}})^{-1} r_{\pi_{\text{discrete}}}$, where P_{π} and r_{π} are respectively the transition matrix and the reward function induced by π .

For the hard FKL, we require access to the greedy action of the action-value function. In the continuous-action setting, this greedy action is usually infeasible to obtain. For the purposes of this experiment, if the greedy action is stay, we represent it in $[-1, 1]$ by drawing a uniform random number from $[-1, 0]$. If the greedy action is switch, we represent it as a uniform random number in $[0, 1]$. This design choice is meant to simulate noisy access to the greedy action in practice.

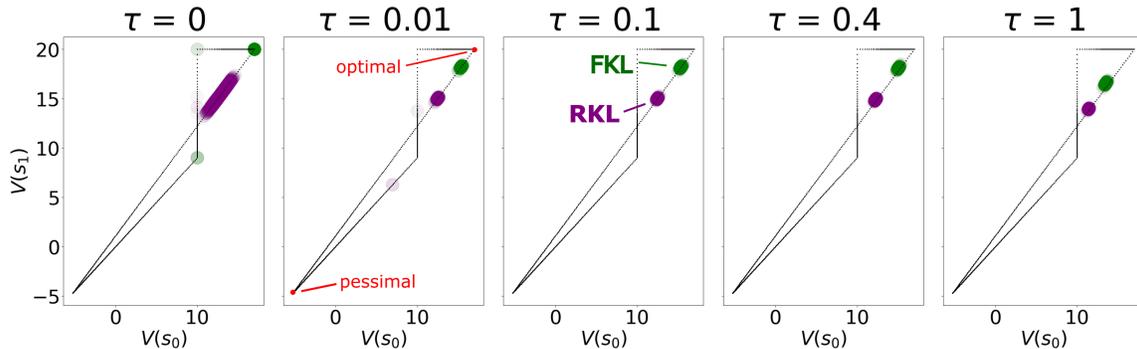


Figure 6: We plot the final value functions on the continuous-action version of Switch-Stay after 500 gradient steps for 1000 iterates, for a variety of temperatures. For clarity, we plot in black the boundary of the value function polytope. All points on the boundary, and all points in the interior of this shape, correspond to value functions of some policy. The optimal and pessimal value functions are fixed for all τ , as we plot the value functions corresponding to the unregularized MDP. Each iterate is represented by a translucent dot. RMSprop with an initial learning rate of 0.01 was used.

For all of these experiments, we initialized means in the range $(-0.95, 0.95)$. All experiments are run for 500 gradient steps and each experiment has 1000 iterates. We plot the value function of the final policy for each iterate and experiment in Figure 6 by visualizing the value function polytope (Dadashi et al., 2019). That is, for finite state and action spaces, the set of all value functions is a polytope (a union of convex polytopes). By plotting the value functions of our policies on the value function polytope, we are able to concisely gauge the performance of an algorithm relative to other algorithms.

1) FKL with $\tau = 0$ converged noticeably slower than the other temperatures, which seems to be an artifact of our encoding of continuous actions to the underlying discrete dynamics of Switch-Stay, and the fact that we used random tie-breaking when computing the arg max for hard FKL.

2) RKL iterates converge slightly faster than FKL iterates across all temperature settings. RKL iterates with $\tau = 0$ sometimes converged to non-optimal value functions on the corners.

3) The limiting value functions of the FKL iterates seem more suboptimal than the limiting value functions of the RKL iterates. The latter are closer to the optimal value function of the original MDP. This result is consistent with our observations in the continuous bandit. Although the FKL optimum may be more easily reached, that optimal point may be suboptimal with respect to the unregularized objective.

6.1.4 THE IMPACT OF STOCHASTICITY IN THE UPDATE

Although with discrete actions it is practical to sum across all actions when calculating the KL losses, difficulty emerges with high-dimensional continuous action spaces. Quadrature methods scale exponentially with the dimension of the action-space, leaving methods like

Clenshaw-Curtis impractical. Monte-Carlo integration—in this case sampling actions from the current policy to estimate the update—seems the only feasible answer in this setting. An important distinction between FKL and RKL, therefore, is how they perform when using a noisier estimate of their updates.

We repeated the experiment in Switch-Stay, now using Monte-Carlo integration instead of Clenshaw-Curtis quadrature to estimate the update for a state, averaged across the sampled actions. As discussed in Section 4, we can estimate the gradients of the RKL and FKL using sampled actions rather than full integration. The hard and soft RKL gradient updates are estimated using sampled actions from the current policy π , and the soft FKL gradient update is estimated using weighted importance sampling. Note that since Hard FKL only depends upon the maximum action, we do not modify the algorithm in this experiment.

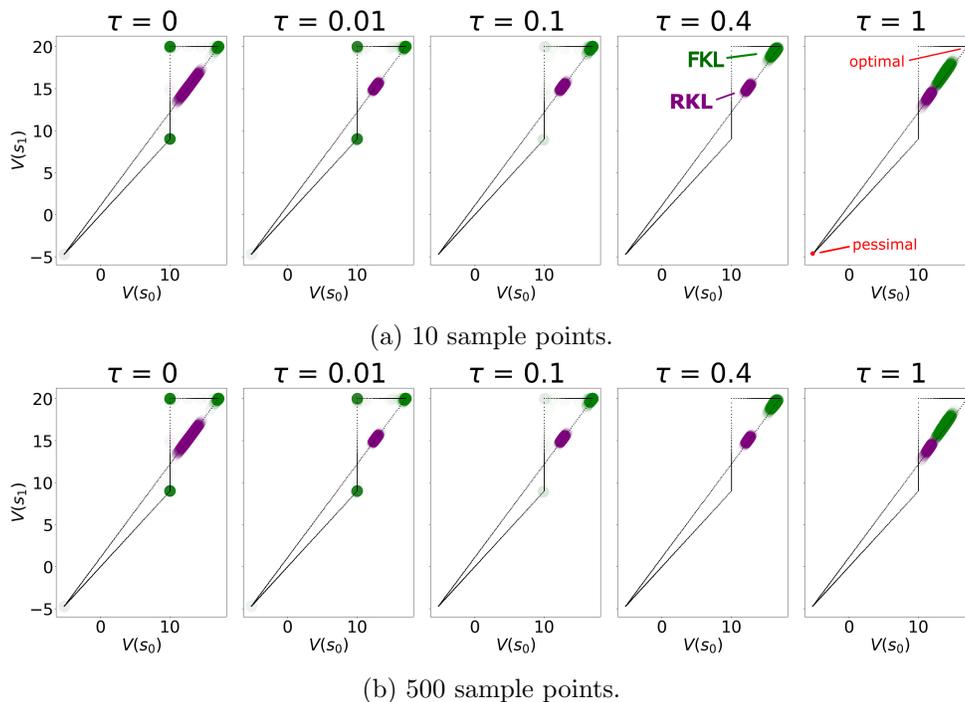


Figure 7: The final value functions in Switch-Stay with stochastic estimation of the gradient using sampled actions, and otherwise the same settings as in Figure 6.

We can see in Figure 7 that the RKL is much more variable than the FKL with a smaller number of sampled actions (10 vs 500). RKL iterates converged to minima to which they did not converge in the Clenshaw-Curtis regime, even for 500 sampled actions. In Figure 7b, there is an interesting trend across temperatures. Temperatures below 0.4 induced many suboptima far from the optimal value function, while temperatures 0.4 and 1 seemed better at clustering RKL iterates near the optimal value function. On the other hand, FKL seemed relatively insensitive both to the temperature and the number of sample points. This relative insensitivity could be due to having a smoother loss landscape to begin with, which tends to direct iterates to a single global optimum. As noted before, though, this global optimum is

quite suboptimal with respect to the unregularized MDP; nonetheless, the FKL can reach its optimum more robustly under noise.

6.2 Exploration Differences between the FKL and RKL

The focus of this section is to study whether there are any significant differences in exploration when using the FKL and RKL. To obtain sufficient exploration, the approach should induce a state visitation distribution whose support is larger, namely that covers more of the state space. Accumulating more transitions from more diverse parts of the state space presumably allows for more accurate estimates of the action value function, and hence more reliable policy improvement. Entropy-regularized RL, as it is currently formulated, only benefits exploration by proxy, through penalizing the negative entropy of the policy. In the context of reward maximization, entropy is only a means to an end; at times, the means may conflict with the end. A policy with higher entropy may have a more diverse state visitation distribution, but it may be prevented from exploiting that information to the fullest capacity because of the penalty to negative entropy.

There has been some work discussing the potential differences between the FKL and RKL for exploration. Neumann (2011) argues in favour of the reverse KL divergence as such a resulting policy would be cost-averse, but also mentions that the forward KL averages over all modes of the target distribution, which may cause it to include regions of low reward in its policy. While in principle it may seem like a bad idea to include those, we note that the value function estimates can be highly inaccurate (Ilyas et al., 2020), causing this inclusion to possibly be beneficial for exploration. Indeed, Norouzi et al. (2016) use the forward KL divergence to induce a policy that is more exploratory.

We hypothesize that the FKL benefits exploration by causing the agent’s policy to commit more slowly to actions that apparently have high value under the current value function estimate. This could benefit exploration both because it causes the agent to explore more and avoids incorrectly committing too quickly to value function estimates that are inaccurate. This non-committal behavior may help the policy avoid converging quickly to a suboptimal policy. Conversely, we hypothesize that the RKL will more quickly reduce the probability of actions that seem low-valued under our current (potentially inaccurate) value estimates. We investigate the differences first in continuous-action Switch-Stay, and then in a Maze environment with a misleading, suboptimal goal. Once again, we observed few differences for the discrete action experiments, even in the Maze environment; we include these results in Appendix D.1.

6.2.1 EXPLORATION UNDER CONTINUOUS ACTIONS IN SWITCH-STAY

We first revisit the Switch-Stay environment, and examine if the FKL and RKL exhibited differences in the variance of their policies. Recall that we examined the value functions for the final policies under the FKL and RKL, in Figure 6. We noted that the FKL converged to more suboptimal policies than the RKL with the same τ , when evaluated under the unregularized objective. A natural hypothesis is that the FKL policy is a more stochastic policy, which is further from the optimal deterministic policy.

To see if this is the case, we plot the final standard deviations of the learned policies for the learning rate of 0.01. In Figure 8, we see that the final FKL iterates have higher

standard deviation for each τ , meaning that the final policies are further from the optimal deterministic policy of the unregularized MDP. Put informally, the FKL tends to commit less than the RKL. This means that even when using a target Boltzmann policy with the same level of entropy-regularization, that level of entropy induces a more stochastic policy under the FKL.

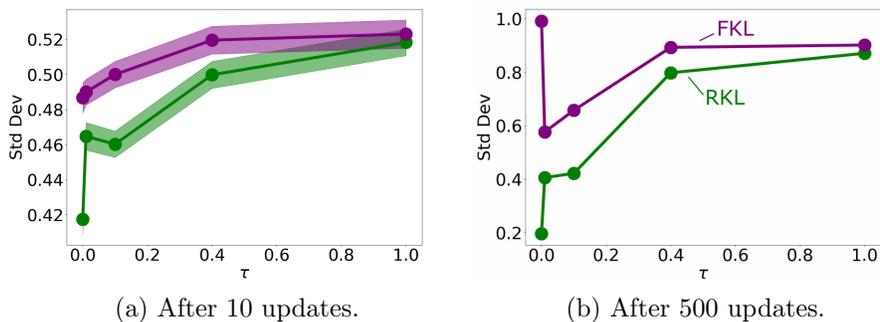


Figure 8: We plot the standard deviation (y-axis) on the continuous-action version of Switch-Stay, with the temperature varied on the x -axis. RMSprop is used with learning rate 0.01. Each dot is the mean of 1000 iterates, with shaded areas corresponding to the standard error. This plot is for state 0, and recall that every action ≤ 0 is treated as “stay” and every action > 0 is treated as “switch”. The large final standard deviation for $\tau = 0$ with the FKL is an artifact of our encoding of the maximum action as a uniform random point in either $[0, 1]$ or $[-1, 0]$, depending on if the maximum action is respectively stay or switch.

6.2.2 EXPLORATION UNDER CONTINUOUS ACTIONS IN MISLEADING MAZE

We next investigate exploration behavior in a more difficult exploration problem: a maze with a misleading goal. We also include an experiment in a maze with discrete actions, but find behavior between the FKL and RKL is very similar (see Appendix D.1).

Figure 9a illustrates the maze we use in this section. The agent starts in the center of the green block and has to get close to the center of the blue block. The red blocks are the obstacles and the yellow block corresponds to the misleading exit, which terminates the episode but gives a reward much lower than the real exit. Since the misleading exit is closer to the starting point than the actual exit, the agent can only find out about the higher reward after exploring the maze. The coordinates the agent sees are normalized to the range $[-1, 1]$, but its actions are given as a tuple (dx, dy) , with $dx, dy \in [-1, 1]$ corresponding to the direction it will try to move (for the actions, 1 corresponds to the length of one block, as opposed to one unit in state-space). The reward is -1 if the agent lands in a normal block, -10 if it hits an obstacle or a wall, 1000 if it lands close to the center of the misleading exit and $100,000$ if it lands close to the center of the actual exit. In case the agent hits a wall or obstacle, any attempt to move in a direction that does not point to the opposite direction will result in no movement and a reward of -10 . Additionally, there is a timeout of

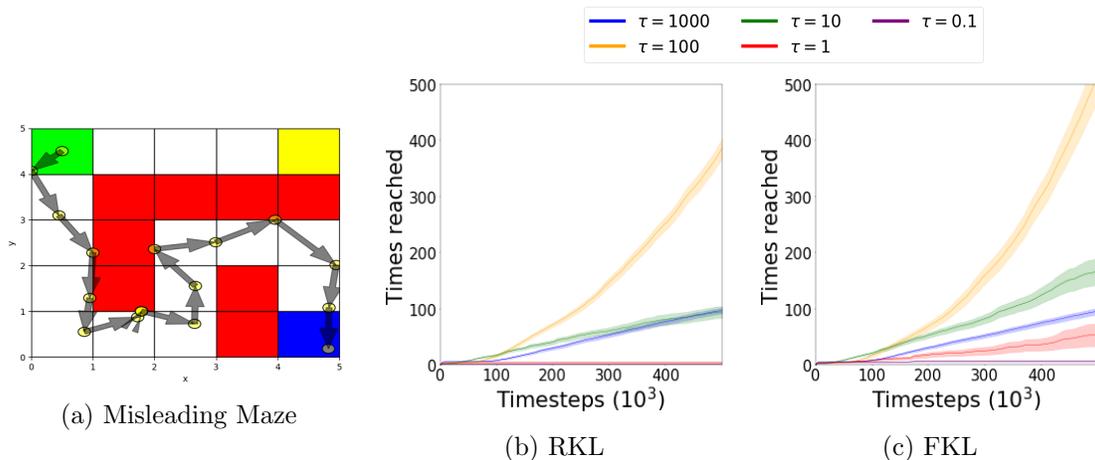


Figure 9: (a) The optimal policy is to go to the blue goal. The yellow goal is misleading, in that it has a positive reward and is easier to get to. Plots in (b) and (c) show the cumulative number of times the correct exit is reached throughout training for multiple temperatures, averaged over 30 runs with the standard errors reported as the shaded regions.

10,000 timesteps, after which the agent has its position reset to the starting position without episode termination. To implement this Misleading Maze, we adapt code from GridMap.¹⁰

The FKL and RKL agents are the same as those used in the following benchmark problems in Section 6.3, with pseudocode in Section 4. The FKL is used with weighted importance sampling and RKL is used with the reparametrization trick. The actor and critic are both parametrized as two-layer neural networks of size 128 with ReLU activations, with the actor corresponding to an unimodal Gaussian. The critic learning rate is set to $1e - 4$ and the actor learning rate is set to $1e - 5$, the optimizer is RMSProp, the batch size is 32 and we sample 128 actions to estimate the gradients. Experiments were done using 30 seeds.

In Figure 9 we show the mean and standard error of the cumulative number of times the correct exit is reached throughout training for multiple temperatures, where FKL and RKL are plotted separately. Appendix D.2 gives more detailed plots of these experiments, where we also show the cumulative number of times the misleading exit is reached and plots for 2M timesteps, instead of the 500k steps we show here.

We conclude that (1) the FKL seems to be more exploratory than the RKL in this environment and (2) the performance when using FKL seems to be a little more robust to the temperature hyperparameter. We also note that this is the only experiment in the paper where the higher temperatures such as 10 and 100 performed better than the lower ones. Figure 9a shows a sampled trajectory from one of the resulting FKL policies with temperature 100, trained for 2M timesteps. Although these policies often reach the goal, they are still highly stochastic. A natural next step is to make them more deterministic over time, by using temperature annealing.

10. <https://github.com/huyaoyu/GridMap>

6.3 Performance in Benchmark Environments

Finally, we compared the FKL and RKL on benchmark continuous and discrete-action environments, using non-linear function approximation. For continuous actions, we experimented on Pendulum, Reacher, Swimmer and HalfCheetah (Todorov et al., 2012). For discrete actions, we experimented on three OpenAI Gym environments (Brockman et al., 2016) and five MinAtar environments (Young and Tian, 2019). We selected policy parameterizations (unimodal Gaussians) and other hyperparameters as is typically done for these problems, as a first investigation into differences under typical conditions.

We found almost no difference between the FKL and RKL for discrete actions and only minimal differences for continuous actions. For completeness and posterity, we still include these systematic results in Appendix C. This contrasts the several differences we found above, in both theory and in more controlled experiments. It is possible that in these environments, the action-values are largely unimodal and so the resulting Boltzmann policy representable by our parameterized policies. The differences between the FKL and RKL are less likely to be pronounced in such settings. A natural next step is to investigate the differences between the two in benchmark problems where it is common to use or need multimodal policies.

7. Discussion and Conclusion

In this work, we investigated the forward KL (FKL) and reverse KL (RKL) divergences to the Boltzmann policy for approximate greedification. To the best of our knowledge, this is the first systematic study into these two natural alternatives, in terms of understanding 1) their theoretical policy improvement properties, 2) their differences in optimization behavior and 3) differences in terms of promoting exploration. Our goal was to highlight that this choice for approximate greedification can have important theoretical and empirical ramifications; we did not advocate for either the FKL or RKL, and found each had useful properties that suit different situations.

We motivated the importance of understanding this question by explaining that many policy optimization algorithms can actually be well-thought of as approximate policy iteration algorithms, with alternating approximate policy evaluation and greedification steps. We categorized many existing algorithms based on whether they (implicitly) use the FKL and RKL, providing another lens to understand these algorithms. Given this approximate policy iteration perspective, understanding improvements to either step can help us improve our policy optimization algorithms. This work focused on understanding the greedification step, which has been much less investigated than policy evaluation.

Based on our theoretical and empirical results, we can summarize our findings as follows.

Theoretically, we found reduction in the RKL is both sufficient and necessary to obtain policy improvement (Proposition 7). The FKL, on the other hand, is not guaranteed to induce policy improvement as reliably as the RKL, with a counterexample is a simple one-state MDP (Proposition 5.4.1). Policy improvement can still occur if a sufficiently high reduction in FKL occurs (Proposition 13). We further found that we could upper bound the RKL with the FKL times another term based on the Renyi divergence (Lemma 14), both suggesting that reducing the FKL on its own may be insufficient without additional conditions that control this term as well.

Empirically, on the microworld and maze experiments, there were more differences between FKL and RKL in the continuous-action setting, whereas no significant differences were observed in the discrete-action setting. In the continuous-action setting, the FKL tended to have a smoother loss landscape that directed iterates to a global optimum of the entropy-regularized objective. This global optimum, however, was sometimes less optimal with respect to the unregularized objective, especially with higher temperatures, than the optima of the RKL. Moreover, the greater suboptimality of the FKL limit points was correlated with the final FKL policies having higher action variance than the corresponding final policies for the RKL. Further experiments supported the claim that the FKL induces more stochastic policies than the RKL, which is consistent with previously described mean-seeking behaviour of the FKL. The fact that significant differences were only observed in the continuous-action regime suggests the important role of policy parameterization.

Finally, on our benchmark experiments, there did not seem to be much difference between choosing one divergence or the other for most of the environments, with performance being more heavily dictated by the choice of temperature and learning rate. In fact, plots of both performance and sensitivity to learning rates were highly similar. The algorithms had some sensitivity to hyperparameters and the best learning rate-temperature combination was also highly environment dependent.

This work highlights a variety of potential next steps. One conclusion from this work is that, though it has been rarely used, the FKL is a promising direction for policy greedification and warrants further investigation. For $\tau \neq 0$, weighted importance sampling allows us to estimate the forward KL objective stochastically, making it practically feasible. Although the FKL performed similarly to the RKL in the benchmark problems, it has properties which can be useful, such as committing less quickly to actions, being more robust to stochastic samples of the gradient of the FKL for a state—which is pertinent as mini-batch estimates will be stochastic—and having a smoother loss surface.

A natural question from this study is why the differences were the largest for continuous actions. One potential reason is the policy parameterization: the Gaussian policy is likely more restrictive than the softmax as it cannot capture multimodal structure. Learning the standard deviation of a Gaussian policy may be another source of instability. In contrast, a softmax policy can represent multiple modes, and does not separate the parameterization of the measure of central tendency (e.g., mean) and variation (e.g., standard deviation).

A promising next step is to compare the FKL and RKL with different policy parameterizations for continuous actions. Recent work into alternative policy parameterizations has explored the Beta distribution (Chou et al., 2017), quantile regression (Richter and Wattenhofer, 2019), and normalizing flows (Ward et al., 2019). While the latter two works in particular have focused on the motivation of multimodality for domains that have multiple goals, we believe that the relevance of multimodality for optimization is as important.

The choice of target distribution in the greedification objective is another interesting question. The Boltzmann distribution over action values is a natural choice for entropy-regularized RL, but one might not want to be tied this framework, especially given sensitivity to the temperature parameter and exploration that is undirected. As yet, there are not too many alternatives. The escort transform has been defined for discrete actions (Mei et al., 2020a), to avoid the overcommittal behavior of the softmax (Boltzmann) policy. For settings where action-values are guaranteed to be positive, the normalized action-values have been

used to create a target distribution (Ghosh et al., 2020). An interesting question is what other alternatives there are to the Boltzmann distribution, especially for continuous actions.

In addition to the choice of target distribution, there are many other possible choices for a greedification objective. Besides the KL divergences, one may consider the Wasserstein distance, Cramér distance, the JS divergence, and many more. Some of our analysis for the KL could actually be used to obtain some of these extensions. Our derivations for the FKL were based on inequalities connecting f-divergences. Sason and Verdú (2016) compiled a list of inequalities connecting f-divergences, so it is possible to follow similar steps to the ones we followed here to derive bounds for other divergences as well. Since each divergence has situations where it works best, having theoretical guarantees for all of them will make it easier to design algorithms that work in each case.

Though we focused on episodic problems in this work, some of the approaches could be used for the continuing setting with average reward. Recent work has analyzed the regret in continuing RL problems for a policy iteration algorithm, called Politex (Abbasi-Yadkori et al., 2019). The policy is a Boltzmann distribution over the sum of all previous action-value function estimates. For continuous actions, a natural alternative is to consider approximating this distribution with a parameterized policy. Work using differential action-values for the average reward setting could benefit from explicit policy parameterizations that use approximate greedification with respect to those differential action-values.

Finally, the connection between policy gradients approaches and API could provide new directions to theoretically analyze policy gradient algorithms. A recent global convergence result for policy gradient methods relies on a connection to policy iteration and showing guaranteed policy improvement (Bhandari and Russo, 2019). These results are similar to earlier results in Scherrer and Geist (2014), and in particular, both works rely on convex policy classes and a closure of the policy class under greedification.

Many other works have exploited linear action-value parameterizations. Perkins and Precup (2003) proved that API converges with linear function approximation for the action-values and soft policy improvement. This work, though, requires the best linear action-value approximation in the function class. Other works have assumed instead exact greedification, with approximate action-values, such as Politex (Abbasi-Yadkori et al., 2019) which bounded regret for an API algorithm, with linear action-values learned online. Scherrer (2014) provides error propagation analyses of many variants of API; these analyses could be extended to finite-sample guarantees by following (Scherrer et al., 2015). Finite-sample analyses for LSPI (Lazaric et al., 2012) and classification-based PI (Lazaric et al., 2016) also exist, but again were designed for approximate policy evaluation, and exact greedification.

It is clear there is a rich theoretical literature for API from which to draw, to fill in gaps understanding existing policy optimization algorithms. An important next step is to investigate if this theory can be extended to approximate greedification, and so make steps towards better characterizing many existing policy optimization algorithms.

Acknowledgements

We gratefully acknowledge funding from NSERC, the Canada CIFAR AI Chair program, and the Alberta Machine Intelligence Institute (Amii). Special thanks as well to Nicolas Le Roux for comments on an earlier version of this work.

References

- Yasin Abbasi-Yadkori, Peter Bartlett, Kush Bhatia, Nevena Lazic, Csaba Szepesvari, and Gellért Weisz. Politex: Regret bounds for policy iteration using expert prediction. In *International Conference on Machine Learning*, 2019.
- Abbas Abdolmaleki, Jost Tobias Springenberg, Yuval Tassa, Remi Munos, Nicolas Heess, and Martin Riedmiller. Maximum a posteriori policy optimisation. In *International Conference on Learning Representations*, 2018.
- Rishabh Agarwal, Chen Liang, Dale Schuurmans, and Mohammad Norouzi. Learning to generalize from sparse and underspecified rewards. In *International Conference on Machine Learning*, 2019.
- Zafarali Ahmed, Nicolas Le Roux, Mohammad Norouzi, and Dale Schuurmans. Understanding the impact of entropy on policy optimization. In *International Conference on Machine Learning*, 2019.
- Brandon Amos, Lei Xu, and J. Zico Kolter. Input convex neural networks. In *International Conference on Machine Learning*, 2017.
- Kavosh Asadi and Michael L. Littman. An alternative softmax operator for reinforcement learning. In *International Conference on Machine Learning*, 2017.
- Boris Belousov and Jan Peters. Entropic regularization of markov decision processes. *Entropy*, 21(7), 2019.
- Dimitri P. Bertsekas. Approximate policy iteration: A survey and some new methods. *Journal of Control Theory and Applications*, 9(3), 2011.
- Dimitri P Bertsekas. *Reinforcement learning and optimal control*. Athena Scientific Belmont, MA, 2019.
- Jalaj Bhandari and Daniel Russo. Global optimality guarantees for policy gradient methods. *arXiv:1906.01786*, 2019.
- Patrick Billingsley. *Probability and Measure*. John Wiley & Sons, 2008.
- Christopher M. Bishop. *Pattern Recognition and Machine Learning*. Springer-Verlag, Berlin, Heidelberg, 2006.
- Greg Brockman, Vicki Cheung, Ludwig Pettersson, Jonas Schneider, John Schulman, Jie Tang, and Wojciech Zaremba. Openai gym. *arXiv:1606.01540*, 2016.
- Minmin Chen, Ramki Gummadi, Chris Harris, and Dale Schuurmans. Surrogate objectives for batch policy optimization in one-step decision making. In *Advances in Neural Information Processing Systems*, 2019.
- Po-Wei Chou, Daniel Maturana, and Sebastian Scherer. Improving stochastic policy gradients in continuous control with deep reinforcement learning using the beta distribution. In *International Conference on Machine Learning*, 2017.

- Charles W Clenshaw and Alan R Curtis. A method for numerical integration on an automatic computer. *Numerische Mathematik*, 2(1), 1960.
- Robert Dadashi, Adrien Ali Taiga, Nicolas Le Roux, Dale Schuurmans, and Marc G. Bellemare. The value function polytope in reinforcement learning. In *International Conference on Machine Learning*, 2019.
- Amir-massoud Farahmand, Doina Precup, André MS Barreto, and Mohammad Ghavamzadeh. Classification-based approximate policy iteration. *IEEE Transactions on Automatic Control*, 60(11), 2015.
- Justin Fu, Katie Luo, and Sergey Levine. Learning robust rewards with adversarial inverse reinforcement learning. In *International Conference on Learning Representations*, 2018.
- Matthieu Geist, Bruno Scherrer, and Olivier Pietquin. A theory of regularized markov decision processes. In *International Conference on Machine Learning*, 2019.
- Seyed Kamyar Seyed Ghasemipour, Richard Zemel, and Shixiang Gu. A divergence minimization perspective on imitation learning methods. In *Conference on Robot Learning*, 2020.
- Dibya Ghosh, Marlos C. Machado, and Nicolas Le Roux. An operator view of policy gradient methods. In *Advances in Neural Information Processing Systems*, 2020.
- Shixiang Gu. *Sample-efficient deep reinforcement learning for continuous control*. PhD thesis, University of Cambridge, 2019.
- Shixiang Gu, Timothy P. Lillicrap, Ilya Sutskever, and Sergey Levine. Continuous deep q-learning with model-based acceleration. In *International Conference on Machine Learning*, 2016.
- Tuomas Haarnoja, Haoran Tang, Pieter Abbeel, and Sergey Levine. Reinforcement learning with deep energy-based policies. In *International Conference on Machine Learning*, 2017.
- 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.
- Jonathan Ho and Stefano Ermon. Generative adversarial imitation learning. In *Advances in Neural Information Processing Systems*, 2016.
- Andrew Ilyas, Logan Engstrom, Shibani Santurkar, Dimitris Tsipras, Firdaus Janoos, Larry Rudolph, and Aleksander Madry. A closer look at deep policy gradients. In *International Conference on Learning Representations*, 2020.
- Ehsan Imani, Eric Graves, and Martha White. An off-policy policy gradient theorem using emphatic weightings. In *Advances in Neural Information Processing Systems*, 2018.
- Ghassen Jerfel, Serena Lutong Wang, Clara Fannjiang, Katherine A Heller, Yian Ma, and Michael Jordan. Variational refinement for importance sampling using the forward kullback-leibler divergence. In *Advances in Approximate Bayesian Inference*, 2021.

- Sham Kakade and John Langford. Approximately optimal approximate reinforcement learning. In *International Conference on Machine Learning*, 2002.
- Dmitry Kalashnikov, Alex Irpan, Peter Pastor, Julian Ibarz, Alexander Herzog, Eric Jang, Deirdre Quillen, Ethan Holly, Mrinal Kalakrishnan, Vincent Vanhoucke, and Sergey Levine. Qt-opt: Scalable deep reinforcement learning for vision-based robotic manipulation. In *Conference on Robot Learning*, 2018.
- Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *International Conference on Learning Representations*, 2015.
- Jens Kober and Jan R Peters. Policy search for motor primitives in robotics. In *Advances in Neural Information Processing Systems*, 2008.
- Daphne Koller and Nir Friedman. *Probabilistic graphical models: principles and techniques*. MIT press, 2009.
- Tadashi Kozuno, Eiji Uchibe, and Kenji Doya. Theoretical analysis of efficiency and robustness of softmax and gap-increasing operators in reinforcement learning. In *International Conference on Artificial Intelligence and Statistics*, 2019.
- Michail G Lagoudakis and Ronald Parr. Reinforcement learning as classification: Leveraging modern classifiers. In *International Conference on Machine Learning*, 2003.
- Guanghui Lan. Policy mirror descent for reinforcement learning: Linear convergence, new sampling complexity, and generalized problem classes. *arXiv:2102.00135*, 2021.
- Alessandro Lazaric, Mohammad Ghavamzadeh, and Rémi Munos. Analysis of a classification-based policy iteration algorithm. In *International Conference on Machine Learning*, 2010.
- Alessandro Lazaric, Mohammad Ghavamzadeh, and Rémi Munos. Finite-sample analysis of least-squares policy iteration. *Journal of Machine Learning Research*, 13(98), 2012.
- Alessandro Lazaric, Mohammad Ghavamzadeh, and Rémi Munos. Analysis of classification-based policy iteration algorithms. *Journal of Machine Learning Research*, 17(19), 2016.
- Sergey Levine. Reinforcement learning and control as probabilistic inference: Tutorial and review. *arXiv:1805.00909*, 2018.
- Timothy P Lillicrap, Jonathan J. Hunt, Alexander Pritzel, Nicolas Heess, Tom Erez, Yuval Tassa, David Silver, and Daan Wierstra. Continuous control with deep reinforcement learning. In *International Conference on Learning Representations*, 2016.
- Boyi Liu, Qi Cai, Zhuoran Yang, and Zhaoran Wang. Neural trust region/proximal policy optimization attains globally optimal policy. In *Advances in Neural Information Processing Systems*, 2019.
- Tyler Lu, Dale Schuurmans, and Craig Boutilier. Non-delusional q-learning and value-iteration. In *Advances in Neural Information Processing Systems*, 2018.

- Jincheng Mei, Chenjun Xiao, Ruitong Huang, Dale Schuurmans, and Martin Müller. On principled entropy exploration in policy optimization. In *International Joint Conference on Artificial Intelligence*, 2019.
- Jincheng Mei, Chenjun Xiao, Bo Dai, Lihong Li, Csaba Szepesvári, and Dale Schuurmans. Escaping the gravitational pull of softmax. *Advances in Neural Information Processing Systems*, 33:21130–21140, 2020a.
- Jincheng Mei, Chenjun Xiao, Csaba Szepesvari, and Dale Schuurmans. On the global convergence rates of softmax policy gradient methods. In *International Conference on Machine Learning*, 2020b.
- Volodymyr Mnih, Adria Puigdomenech Badia, Mehdi Mirza, Alex Graves, Timothy Lillicrap, Tim Harley, David Silver, and Koray Kavukcuoglu. Asynchronous methods for deep reinforcement learning. In *International Conference on Machine Learning*, 2016.
- Ofir Nachum, Mohammad Norouzi, and Dale Schuurmans. Improving policy gradient by exploring under-appreciated rewards. In *International Conference on Learning Representations*, 2017a.
- Ofir Nachum, Mohammad Norouzi, Kelvin Xu, and Dale Schuurmans. Bridging the gap between value and policy based reinforcement learning. In *Advances in Neural Information Processing Systems*, 2017b.
- Ofir Nachum, Bo Dai, Ilya Kostrikov, Yinlam Chow, Lihong Li, and Dale Schuurmans. Algaedice: Policy gradient from arbitrary experience. *arXiv:1912.02074*, 2019.
- Gergely Neu, Anders Jonsson, and Vicenç Gómez. A unified view of entropy-regularized markov decision processes. *arXiv:1705.07798*, 2017.
- Gerhard Neumann. Variational inference for policy search in changing situations. In *International Conference on Machine Learning*, 2011.
- Mohammad Norouzi, Samy Bengio, zhifeng Chen, Navdeep Jaitly, Mike Schuster, Yonghui Wu, and Dale Schuurmans. Reward augmented maximum likelihood for neural structured prediction. In *Advances in Neural Information Processing Systems*, 2016.
- Chris Nota and Philip S. Thomas. Is the policy gradient a gradient? In *International Conference on Autonomous Agents and Multiagent Systems*, 2020.
- Brendan O’Donoghue, Remi Munos, Koray Kavukcuoglu, and Volodymyr Mnih. Combining policy gradient and q-learning. *International Conference on Learning Representations*, 2017.
- Francesco Orabona. A modern introduction to online learning. *arXiv:1912.13213*, 2019.
- Fabio Pardo, Arash Tavakoli, Vitaly Levnik, and Petar Kormushev. Time limits in reinforcement learning. In *International Conference on Machine Learning*, 2018.

- Theodore J. Perkins and Mark D. Pendrith. On the existence of fixed points for q-learning and sarsa in partially observable domains. In *International Conference on Machine Learning*, 2002.
- Theodore J Perkins and Doina Precup. A convergent form of approximate policy iteration. In *Advances in Neural Information Processing Systems*, 2003.
- Jan Peters, Katharina Mülling, and Yasemin Altün. Relative entropy policy search. In *AAAI Conference on Artificial Intelligence*, 2010.
- Mark S Pinsker. *Information and information stability of random variables and processes*. Holden-Day, 1964.
- Oliver Richter and Roger Wattenhofer. Learning policies through quantile regression. *arXiv:1906.11941*, 2019.
- Moonkyung Ryu, Yinlam Chow, Ross Anderson, Christian Tjandraatmadja, and Craig Boutilier. Caql: Continuous action q-learning. In *International Conference on Learning Representations*, 2020.
- Igal Sason and Sergio Verdú. f -divergence inequalities. *IEEE Transactions on Information Theory*, 62(11), 2016.
- Bruno Scherrer. Approximate policy iteration schemes: A comparison. In *International Conference on Machine Learning*, 2014.
- Bruno Scherrer and Matthieu Geist. Local policy search in a convex space and conservative policy iteration as boosted policy search. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, 2014.
- Bruno Scherrer, Mohammad Ghavamzadeh, Victor Gabillon, Boris Lesner, and Matthieu Geist. Approximate modified policy iteration and its application to the game of tetris. *Journal of Machine Learning Research*, 16(49), 2015.
- John Schulman, Sergey Levine, Pieter Abbeel, Michael Jordan, and Philipp Moritz. Trust region policy optimization. In *International Conference on Machine Learning*, 2015.
- John Schulman, Xi Chen, and Pieter Abbeel. Equivalence between policy gradients and soft q-learning. *arXiv:1704.06440*, 2017a.
- John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv:1707.06347*, 2017b.
- Lior Shani, Yonathan Efroni, and Shie Mannor. Adaptive trust region policy optimization: Global convergence and faster rates for regularized mdps. In *AAAI Conference on Artificial Intelligence*, 2020.
- David Silver, Guy Lever, Nicolas Heess, Thomas Degris, Daan Wierstra, and Martin Riedmiller. Deterministic policy gradient algorithms. In *International Conference on Machine Learning*, 2014.

- Richard S. Sutton and Andrew G. Barto. *Reinforcement Learning: An Introduction*. MIT press, 2018.
- Richard S. Sutton, David A. McAllester, Satinder P. Singh, and Yishay Mansour. Policy gradient methods for reinforcement learning with function approximation. In *Advances in Neural Information Processing Systems*, 1999.
- Philip Thomas. Bias in natural actor-critic algorithms. In *International Conference on Machine Learning*, 2014.
- Tijmen Tieleman and Geoffrey Hinton. Lecture 6.5-rmsprop: Divide the gradient by a running average of its recent magnitude. *Coursera: Neural Networks for Machine Learning*, 2012.
- Emanuel Todorov, Tom Erez, and Yuval Tassa. Mujoco: A physics engine for model-based control. In *International Conference on Intelligent Robots and Systems*, 2012.
- Nino Vieillard, Tadashi Kozuno, Bruno Scherrer, Olivier Pietquin, Remi Munos, and Matthieu Geist. Leverage the average: an analysis of kl regularization in reinforcement learning. In *Advances in Neural Information Processing Systems*, 2020a.
- Nino Vieillard, Olivier Pietquin, and Matthieu Geist. Deep conservative policy iteration. In *AAAI Conference on Artificial Intelligence*, 2020b.
- Paul Wagner. A reinterpretation of the policy oscillation phenomenon in approximate policy iteration. In *Advances in Neural Information Processing Systems*, 2011.
- Paul Wagner. Optimistic policy iteration and natural actor-critic: A unifying view and a non-optimality result. In *Advances in Neural Information Processing Systems*, 2013.
- Patrick Nadeem Ward, Ariella Smofsky, and Avishek Joey Bose. Improving exploration in soft-actor-critic with normalizing flows policies. *Workshop on Invertible Neural Networks and Normalizing Flows at International Conference for Machine Learning*, 2019.
- Christopher J.C.H. Watkins and Peter Dayan. Q-learning. *Machine Learning*, 8(3-4), 1992.
- Ronald J Williams. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine Learning*, 8(3-4), 1992.
- Kenny Young and Tian Tian. Minatar: An atari-inspired testbed for more efficient reinforcement learning experiments. *arXiv:1903.03176*, 2019.
- Lingwei Zhu and Takamitsu Matsubara. Ensuring monotonic policy improvement in entropy-regularized value-based reinforcement learning. *arXiv:2008.10806*, 2020.
- B. Ziebart. Modeling purposeful adaptive behavior with the principle of maximum causal entropy. *PhD thesis, Carnegie Mellon University*, 2010.

Appendix A. Categorizing Existing Algorithms by their Greedification Operators

There are many existing policy optimization algorithms. As we motivated in the introduction, many of these can actually be seen as doing API, though they are typically described as policy gradient methods. We categorize these methods based on which of these four KLs underlie their policy update.

It is important to note that these methods are not only characterized by which KL variant they use for updating towards the Boltzmann policy. In fact, in many cases other properties are the critical novelties of the algorithms. For example, as noted below, the Hard RKL underlies TRPO. However, a defining characteristic of TRPO is preventing the policy from changing too much using a KL divergence to the previous policy. Therefore, in addition to using a Hard RKL towards the Boltzmann policy, it also uses a KL divergence to the old policy; these two KL divergences play different roles.

A.1 RKL without Entropy Regularization

Many actor-critic approaches that do not have entropy regularization are implicitly optimizing a hard RKL when performing a policy update.

Vanilla Actor-Critic (Sutton and Barto, 2018) uses the gradient of the Hard RKL for its policy update, in a given state. For the episodic objective $\eta(\pi_\theta)$ given in Equation (1), the policy gradient theorem shows that gradient is

$$\nabla_\theta \eta(\pi_\theta) = \int_{\mathcal{S}} d^{\pi_\theta}(s) \int_{\mathcal{A}} Q^{\pi_\theta}(s, a) \nabla_\theta \pi_\theta(a | s) da ds.$$

For each state, the inner update is exactly the Hard RKL. By selecting $d = d^{\pi_\theta}$, the Hard RKL averaged across all states exactly equals the policy gradient underlying actor-critic. This weighting is obtained by simply acting on-policy and weighting the update by the discount raised to the power of the step in the episode (Thomas, 2014). In practice, this weighting by the discount is often omitted and the Hard RKL update performed in each state visited under the policy.

Trust Region Policy Optimization (TRPO) (Schulman et al., 2015) has the same hard RKL objective as vanilla actor-critic, but with an *additional constraint* in the projection step that the policy should not change too much after an update. This strategy builds on the earlier Conservative Policy Iteration (CPI) algorithm (Kakade and Langford, 2002), which motivates that the old policy and old action-values can be used in the objective. This objective is sometimes called the linearized objective; with the addition of the constraint, the objective corresponds to

$$J(\theta) = \mathbb{E}_{S \sim d^{\pi_{\theta_{old}}}, A \sim \pi_{\theta_{old}}} [Q^{\pi_{\theta_{old}}}(S, A)] = \mathbb{E}_{S \sim d^{\pi_{\theta_{old}}}, A \sim \pi_{\theta_{old}}} \left[\frac{\pi_\theta(A|S)}{\pi_{\theta_{old}}(A|S)} Q^{\pi_{\theta_{old}}}(S, A) \right]$$

subject to $\delta \geq \mathbb{E}_{S \sim d^{\pi_{\theta_{old}}}} [\text{KL}_{\pi_\theta}^{\pi_{\theta_{old}}}(S)]$

For a given state sampled from $d^{\pi_{\theta_{old}}}$, the inner optimization is precisely a hard RKL to the action-values for the old policy. The objective can be written with actions sampled either according to π_θ or according to $\pi_{\theta_{old}}$ with an importance sampling ratio.

Note, though, that the actual TRPO algorithm uses an approximation that results in a natural policy gradient update, rather than a gradient of the above objective. So, though it is motivated by optimizing the above (Equation (14) in their work), it actually solves (Schulman et al., 2015, Equation 17). This has further been re-interpreted as a mirror descent update and used to provide convergence of TRPO (Neu et al., 2017; Liu et al., 2019; Shani et al., 2020).

Proximal Policy Optimization (PPO) (Schulman et al., 2017b) the baseline they compare against uses the same objective as TRPO, but uses a KL penalty instead of a constraint. The policy objective with the KL penalty is

$$J(\theta) = \mathbb{E}_{S \sim d^{\pi_{\theta_{old}}}, A \sim \pi_{\theta_{old}}} \left[\frac{\pi_{\theta}(A|S)}{\pi_{\theta_{old}}(A|S)} Q^{\pi_{old}}(S, A) \right] - \beta \mathbb{E}_{S \sim d^{\pi_{\theta_{old}}}} \left[\text{KL}_{\pi_{\theta}}^{\pi_{\theta_{old}}}(S) \right]$$

The greedification component of this objective—the first term—can again be seen as using $d = d^{\pi_{old}}$ with a Hard RKL to action-values $Q^{\pi_{old}}$. The second term simply modifies how the Hard RKL is optimized, because the convergence point is unchanged: when θ_{old} is optimal, setting $\theta = \theta^*$ maximizes the first term while keeping the second term 0.

The actual objective used in PPO clips hyperparameter $\epsilon \geq 0$ and is written as

$$J(\theta) = \mathbb{E}_{\pi_{\theta_{old}}} \left[\min \left(\frac{\pi_{\theta}(a|s)}{\pi_{\theta_{old}}(a|s)} \hat{A}^{\pi}(s, a), \text{clip} \left(\frac{\pi_{\theta}(a|s)}{\pi_{\theta_{old}}(a|s)}, 1 - \epsilon, 1 + \epsilon \right) \hat{A}^{\pi}(s, a) \right) \right]$$

The clipped objective forces the probability ratio $\frac{\pi_{\theta}(a|s)}{\pi_{\theta_{old}}(a|s)}$ to stay within $[1 - \epsilon, 1 + \epsilon]$.

Other trust region policy optimization methods can be also understood as minimizing the Hard RKL and a KL constraint, as noted in the beginning of the section. Maximum a posteriori policy optimization (Abdolmaleki et al., 2018) resembles TRPO, but uses a target policy obtained by replacing the entropy term in Equation (4) with an FKL between p and an old policy. Politex (Abbasi-Yadkori et al., 2019) can be understood as MPO without a KL constraint. Mirror descent policy iteration, which is mirror descent modified policy iteration (Geist et al., 2019) with $m = \infty$, similarly uses a target policy obtained by replacing the entropy term with a Bregman divergence from an old policy to p . Another policy optimization algorithm based on the trust region method is relative entropy policy search (Peters et al., 2010).

Deep Deterministic Policy Gradient (DDPG) (Silver et al., 2014; Lillicrap et al., 2016) can be viewed as a “degenerate” Hard RKL with a deterministic policy, taking care to differentiate through the action.

A.2 RKL with Entropy Regularization

The RKL has been used for an actor-critic algorithm and a continuous action Q-learning algorithm.

Soft Actor-Critic (SAC) (Haarnoja et al., 2018) minimizes the RKL, but is written slightly differently because it subsumes the temperature inside the reward and scales the reward to control the temperature. The policy objective is

$$J(\theta) = \mathbb{E}_{S \sim \mathcal{D}} \left[\text{KL} \left(\pi_{\theta}(\cdot|S) \parallel \frac{\exp(Q^{\pi_{old}}(S, \cdot))}{Z(S)} \right) \right]$$

where the states s are sampled from a dataset given by \mathcal{D} . We note that A3C (Mnih et al., 2016) updates the policy with the same objective, but does not learn the soft value functions.

Soft Q-learning (SQL) (Haarnoja et al., 2017) introduces a soft Bellman optimality update, that iterates towards a soft optimal action-value Q^* using a soft maximization. To avoid sampling the Boltzmann policy, which is expensive, they proposed using Stein variational gradient descent. This approach requires introducing an approximate sampling network, which can be alternatively seen as a policy. The resulting objective corresponds exactly to an RKL, with the Boltzmann policy defined on Q^*

$$J(\theta) = \mathbb{E}_{S \sim \mathcal{D}} \left[\text{KL} \left(\pi_{\theta}(\cdot | S) \parallel \frac{\exp(Q^*(S, \cdot)\tau^{-1})}{Z(S)} \right) \right]$$

Two other approaches are similar to SQL, but for discrete actions. Asadi and Littman (2017) propose generalized value iteration with the mellomax operator, where the maximum-entropy mellomax policy is a Boltzmann policy with state-dependent temperature. Conservative value iteration (Kozuno et al., 2019) can be seen as a variant of SQL in which a trust region method is used, formally proven in (Vieillard et al., 2020a).

A.3 FKL without Entropy Regularization

The FKL without entropy regularization and Boltzmann target policy has not previously been explored. However, the FKL to other target distributions has been considered.

Deep Conservative Policy Iteration (DCPI) (Vieillard et al., 2020b) learns both an action value function Q and a policy π_{θ} , with target networks Q^- and π^- for each. To update the policy, DCPI minimizes an FKL loss between π_{θ} and a regularized greedification of Q . In other words, DCPI uses a forward KL, but to different target distribution than the Boltzmann distribution.

$$J(\theta) := \mathbb{E}_{S \sim \mathcal{D}} [\text{KL}_{\pi_{\theta}}^{((1-\alpha)\pi^- + \alpha\mathcal{G}(Q))}(S)]$$

for $\alpha \in [0, 1]$ and $\mathcal{G}(Q(S, \cdot))$ the greedy policy w.r.t. Q for a given state S .

Policy Greedification as Classification uses multi-class classification, where actions that maximize $Q(s, a)$ at each state are labeled positive, and the policy is updated to predict that greedy action for every state (Lagoudakis and Parr, 2003; Lazaric et al., 2010; Farahmand et al., 2015). This approach is related to using the Hard FKL, because the FKL corresponds to the cross-entropy loss which can be used for classification. Other classification methods, however, like SVMs, are not directly related.

A.4 FKL with Entropy Regularization

Under-appreciated Reward EXploration (UREX) (Nachum et al., 2017a) optimizes a mixture of forward and reverse KLs. Their reverse KL is the usual vanilla actor-critic objective, while the forward KL is given by $\text{KL}_{\pi_{\theta}}^{\pi^*}$, where they approximate $\pi^* \propto \exp(\tau^{-1}G - \log \pi_{\theta})$ with G being the return received at the end of the episode. Subsequent work in Agarwal et al. (2019) also employs both a forward and reverse KL, where the forward KL is initially used to collect diverse trajectories, and the reverse KL is used to learn a robust policy, which performs well in sparse and under-specified reward settings.

Exploratory Conservative Policy Optimization (ECPO) (Mei et al., 2019) splits policy optimization into a *project* and a *lift* step. The project step minimizes the forward KL divergence to a target policy $\bar{\pi}_{\tau, \tau'}$ that is the optimal policy under the entropy-regularized objective, with a KL penalty to the old policy.

$$\begin{aligned} \text{Project} &: \arg \min_{\pi_\theta} \text{KL}_{\pi_\theta}^{\bar{\pi}_{\tau, \tau'}} \\ \text{Lift} &: \bar{\pi}_{\tau, \tau'} = \arg \max_{\pi} \mathbb{E}_{S_0} [V^{\pi_{old}}(S_0)] - \tau \text{KL}_{\pi_{old}}^{\pi} + \tau' \mathcal{H}(\pi) \end{aligned}$$

There have also been relevant works in supervised learning using entropy regularization and forward KL. In **Reward Augmented Maximum Likelihood (RAML)** (Norouzi et al., 2016), instead of optimizing over the traditional maximum likelihood framework (Hard FKL), a target reward distribution is defined, and the model distribution minimizes the forward KL to that target distribution.

It is interesting to note that the FKL has been used in other area of reinforcement learning. In Inverse Reinforcement Learning and Imitation Learning, Ghasemipour et al. (2020) have shown that, under the cost regularized framework proposed by Ho and Ermon (2016), most methods can be categorized as reducing some divergence. Notably among them is behavior cloning, which corresponds to a FKL reduction between policies, whereas AIRL (Fu et al., 2018), one of the state-of-the-art methods, reduces the RKL between occupancy measures. The paper then proposes a variant of AIRL based on reducing the FKL, obtaining competitive results.

Appendix B. Proofs

In this section we provide all the proofs. We give results organized according to the same subsections as in the main body.

B.1 Proofs for Section 5.3: Policy Improvement with RKL

Lemma 15. *[Soft Performance Difference] For any policies π_{old}, π_{new} , any $\tau \geq 0$, we have*

$$\eta_\tau(\pi_{new}) - \eta_\tau(\pi_{old}) = \frac{1}{1 - \gamma} \mathbb{E}_{d^{\pi_{new}}} \left[\mathbb{E}_{\pi_{new}} [A_\tau^{\pi_{old}}(S, A)] - \tau \text{KL}_{\pi_{old}}^{\pi_{new}}(S) \right].$$

Proof When we write $\mathbb{E}_{\rho_0}^{\pi_{new}}$, we mean the expectation over the trajectory distribution induced by ρ_0 and π_{new} .

$$\begin{aligned} \frac{1}{1 - \gamma} \mathbb{E}_{d^{\pi_{new}, \pi_{new}}} [A_\tau^{\pi_{old}}(S, A)] &= \mathbb{E}_{\rho_0}^{\pi_{new}} \left[\sum_{t=0}^{\infty} \gamma^t A_\tau^{\pi_{old}}(S_t, A_t) \right] \\ &= \mathbb{E}_{\rho_0}^{\pi_{new}} \left[\sum_{t=0}^{\infty} \gamma^t (Q_\tau^{\pi_{old}}(S_t, A_t) - \tau \log \pi_{old}(\cdot | S_t) - V_\tau^{\pi_{old}}(S_t)) \right] \end{aligned}$$

where the first equality follows from the definition of the visitation distribution and the second from the definition of the soft advantage. We can simplify the term inside the

expectation as follows.

$$\begin{aligned}
 & \sum_{t=0}^{\infty} \gamma^t (Q_{\tau}^{\pi_{\text{old}}}(S_t, A_t) - \tau \log \pi_{\text{old}}(\cdot | S_t) - V_{\tau}^{\pi_{\text{old}}}(S_t)) \\
 &= \sum_{t=0}^{\infty} \gamma^t (r(S_t, A_t) + \gamma V_{\tau}^{\pi_{\text{old}}}(S_{t+1}) - \tau \log \pi_{\text{old}}(\cdot | S_t) - V_{\tau}^{\pi_{\text{old}}}(S_t)) \\
 &= \left(\sum_{t=0}^{\infty} \gamma^t (r(S_t, A_t) - \tau \log \pi_{\text{old}}(\cdot | S_t)) \right) - V_{\tau}^{\pi_{\text{old}}}(S_0),
 \end{aligned}$$

where the second line follows from expanding $Q_{\tau}^{\pi_{\text{old}}}$ and the second from the telescoping series $\gamma V_{\tau}^{\pi_{\text{old}}}(S_{t+1}) - V_{\tau}^{\pi_{\text{old}}}(S_t)$. Plugging this back into the expectation, and using $\mathbb{E}_{\rho_0}^{\pi_{\text{new}}}[V_{\tau}^{\pi_{\text{old}}}(S_0)] = \eta_{\tau}(\pi_{\text{old}})$ we get

$$\begin{aligned}
 \mathbb{E}_{\rho_0}^{\pi_{\text{new}}} \left[\sum_{t=0}^{\infty} \gamma^t A_{\tau}^{\pi_{\text{old}}}(S_t, A_t) \right] &= -\eta_{\tau}(\pi_{\text{old}}) + \mathbb{E}_{\rho_0}^{\pi_{\text{new}}} \left[\sum_{t=0}^{\infty} \gamma^t (r(S_t, A_t) - \tau \log \pi_{\text{old}}(\cdot | S_t)) \right] \\
 &= -\eta_{\tau}(\pi_{\text{old}}) + \eta_{\tau}(\pi_{\text{new}}) + \mathbb{E}_{\rho_0}^{\pi_{\text{new}}} \left[\sum_{t=0}^{\infty} \gamma^t \tau (\log \pi_{\text{new}}(\cdot | S_t) - \log \pi_{\text{old}}(\cdot | S_t)) \right] \\
 &= -\eta_{\tau}(\pi_{\text{old}}) + \eta_{\tau}(\pi_{\text{new}}) + \frac{\tau}{1-\gamma} \mathbb{E}_{d^{\pi_{\text{new}}}} [\text{KL}_{\pi_{\text{old}}}^{\pi_{\text{new}}}(S)],
 \end{aligned}$$

where the second equality is obtained by adding and subtracting $\tau \log \pi_{\text{new}}(\cdot | S_t)$. \blacksquare

Proposition 16. *[Improvement Under Average RKL Reduction] For $\pi_{\text{old}}, \pi_{\text{new}} \in \Pi$, define*

$$\Delta \text{RKL}_{\pi_{\text{new}}}^{\pi_{\text{old}}}(S) := \text{RKL}_{\mathcal{B}_{\tau} Q_{\tau}^{\pi_{\text{old}}}}^{\pi_{\text{old}}}(S) - \text{RKL}_{\mathcal{B}_{\tau} Q_{\tau}^{\pi_{\text{old}}}}^{\pi_{\text{new}}}(S).$$

$$\text{For } \tau > 0: \quad \eta_{\tau}(\pi_{\text{new}}) - \eta_{\tau}(\pi_{\text{old}}) = \frac{\tau}{1-\gamma} \mathbb{E}_{d^{\pi_{\text{new}}}} [\Delta \text{RKL}_{\pi_{\text{new}}}^{\pi_{\text{old}}}(S)]. \quad (15)$$

Furthermore, $\eta_{\tau}(\pi_{\text{new}}) \geq \eta_{\tau}(\pi_{\text{old}})$ if and only if $\mathbb{E}_{d^{\pi_{\text{new}}}} [\Delta \text{RKL}_{\pi_{\text{new}}}^{\pi_{\text{old}}}(S)] \geq 0$.

Proof We start by writing the RHS of Theorem 6.

$$\begin{aligned}
 & \frac{1}{1-\gamma} \mathbb{E}_{d^{\pi_{\text{new}}, \pi_{\text{new}}}} [A_{\tau}^{\pi_{\text{old}}}(S, A)] - \frac{\tau}{1-\gamma} \mathbb{E}_{d^{\pi_{\text{new}}}} [\text{KI}_{\pi_{\text{old}}}^{\pi_{\text{new}}}(S)] = \\
 & \frac{\tau}{1-\gamma} \mathbb{E}_{d^{\pi_{\text{new}}, \pi_{\text{new}}}} \left[\frac{Q_{\tau}^{\pi_{\text{old}}}(S, A)}{\tau} - \log(\pi_{\text{old}}(A|S)) - \frac{V_{\tau}^{\pi_{\text{old}}}(S)}{\tau} \right] \\
 & \quad - \frac{\tau}{1-\gamma} \mathbb{E}_{d^{\pi_{\text{new}}, \pi_{\text{new}}}} [\log(\pi_{\text{new}}(A|S)) - \log(\pi_{\text{old}}(A|S))] = \\
 & \frac{\tau}{1-\gamma} \left(\mathbb{E}_{d^{\pi_{\text{new}}, \pi_{\text{new}}}} \left[\log \left(e^{\frac{Q_{\tau}^{\pi_{\text{old}}}(S, A)}{\tau}} \right) \right] - \mathbb{E}_{d^{\pi_{\text{new}}, \pi_{\text{old}}}} \left[\frac{Q_{\tau}^{\pi_{\text{old}}}(S, A)}{\tau} - \log(\pi_{\text{old}}(A|S)) \right] \right) \\
 & \quad - \frac{\tau}{1-\gamma} \mathbb{E}_{d^{\pi_{\text{new}}, \pi_{\text{new}}}} [\log(\pi_{\text{new}}(A|S))] = \\
 & \frac{\tau}{1-\gamma} \mathbb{E}_{d^{\pi_{\text{new}}, \pi_{\text{new}}}} \left[\log \left(e^{\frac{Q_{\tau}^{\pi_{\text{old}}}(S, A)}{\tau}} \right) - \log(\pi_{\text{new}}(A|S)) \right] \\
 & \quad - \frac{\tau}{1-\gamma} \mathbb{E}_{d^{\pi_{\text{new}}, \pi_{\text{old}}}} \left[\log \left(e^{\frac{Q_{\tau}^{\pi_{\text{old}}}(S, A)}{\tau}} \right) - \log(\pi_{\text{old}}(A|S)) \right] = \\
 & \frac{\tau}{1-\gamma} \mathbb{E}_{d^{\pi_{\text{new}}}} [\Delta \text{RKL}_{\pi_{\text{new}}}^{\pi_{\text{old}}}(S)].
 \end{aligned}$$

The last equality follows by adding and subtracting $\mathbb{E}_{d^{\pi_{\text{new}}}}(\log(Z(S)))$ and rearranging. Plugging that in the equation from Theorem 6 we get:

$$\eta_{\tau}(\pi_{\text{new}}) - \eta_{\tau}(\pi_{\text{old}}) = \frac{\tau}{1-\gamma} \mathbb{E}_{d^{\pi_{\text{new}}}} [\Delta \text{RKL}_{\pi_{\text{new}}}^{\pi_{\text{old}}}(S)],$$

which can only be nonnegative if $\mathbb{E}_{d^{\pi_{\text{new}}}} [\Delta \text{RKL}_{\pi_{\text{new}}}^{\pi_{\text{old}}}(S)] \geq 0$. ■

B.1.1 PROOFS FOR SECTION 5.3.1: EXTENSIONS TO APPROXIMATE ACTION-VALUES

Lemma 17. *[Approximate Soft Performance Difference] Let $\pi_{\text{old}}, \pi_{\text{new}}$ be any policies and let $\tau \geq 0$. Let $\hat{Q} : \mathcal{S} \times \mathcal{A} \rightarrow \mathbb{R}$ represent an action-value estimate and let $\epsilon(s, a) := \hat{Q}(s, a) - Q_{\tau}^{\pi_{\text{old}}}(s, a)$ be the per-state approximation error. As well, define*

$$\begin{aligned}
 \hat{\Delta} \text{RKL}_{\pi_{\text{new}}}^{\pi_{\text{old}}}(S) & := \text{RKL}_{\mathcal{B}_{\tau} \hat{Q}}^{\pi_{\text{old}}}(S) - \text{RKL}_{\mathcal{B}_{\tau} \hat{Q}}^{\pi_{\text{new}}}(S), \\
 \bar{\epsilon} & := \mathbb{E}_{d^{\pi_{\text{new}}}} [\mathbb{E}_{\pi_{\text{new}}}[\epsilon(S, A)]] - \mathbb{E}_{\pi_{\text{old}}}[\epsilon(S, A)].
 \end{aligned}$$

$$\text{We have: } \quad \eta_{\tau}(\pi_{\text{new}}) - \eta_{\tau}(\pi_{\text{old}}) + \bar{\epsilon} = \frac{\tau}{1-\gamma} \mathbb{E}_{d^{\pi_{\text{new}}}} [\hat{\Delta} \text{RKL}_{\pi_{\text{new}}}^{\pi_{\text{old}}}(S)]. \quad (16)$$

Proof We focus on a particular state s and omit function arguments when they are clear from context.

$$\begin{aligned}
 & \tau(\text{KL}_{\mathcal{B}_\tau \hat{Q}}^{\pi_{\text{old}}} - \text{KL}_{\mathcal{B}_\tau \hat{Q}}^{\pi_{\text{new}}}) \\
 &= \tau[-H(\pi_{\text{old}}) + H(\pi_{\text{new}}) - \sum_a \pi_{\text{old}}(a)(\hat{Q}(s, a)/\tau - \log Z(s)) + \sum_a \pi_{\text{new}}(a)(\hat{Q}(s, a)/\tau - \log Z(s))] \\
 &= -\tau H(\pi_{\text{old}}) + \tau H(\pi_{\text{new}}) - \sum_a \pi_{\text{old}}(a)\hat{Q}(s, a) + \sum_a \pi_{\text{new}}(a)\hat{Q}(s, a) \\
 &\quad \triangleright \text{policies sum to one and } Z(s) \text{ does not depend on } a \\
 &= -\tau H(\pi_{\text{old}}) + \tau H(\pi_{\text{new}}) - \sum_a \pi_{\text{old}}(a)(Q_\tau^{\pi_{\text{old}}}(s, a) + \epsilon(s, a)) \\
 &\quad + \sum_a \pi_{\text{new}}(a)(Q_\tau^{\pi_{\text{old}}}(s, a) + \epsilon(s, a) + \tau \log \pi_{\text{old}}(a) - \tau \log \pi_{\text{old}}(a)) \\
 &\quad \triangleright \text{expanding } \hat{Q} \\
 &= -\tau H(\pi_{\text{old}}) - \tau \text{KL}_{\pi_{\text{old}}}^{\pi_{\text{new}}} - \sum_a \pi_{\text{old}}(a)(Q_\tau^{\pi_{\text{old}}}(s, a) + \epsilon(s, a)) \\
 &\quad + \sum_a \pi_{\text{new}}(a)(Q_\tau^{\pi_{\text{old}}}(s, a) + \epsilon(s, a) - \tau \log \pi_{\text{old}}(a)) \\
 &\quad \triangleright \text{folding term into KL} \\
 &= -\tau H(\pi_{\text{old}}) - \tau \text{KL}_{\pi_{\text{old}}}^{\pi_{\text{new}}} - V_\tau(s) - \sum_a \pi_{\text{old}}(a)(\tau \log \pi_{\text{old}}(a) + \epsilon(s, a)) \\
 &\quad + \sum_a \pi_{\text{new}}(a)(Q_\tau^{\pi_{\text{old}}}(s, a) + \epsilon(s, a) - \tau \log \pi_{\text{old}}(a)) \\
 &\quad \triangleright \text{using average of action-value} \\
 &= -\tau \text{KL}_{\pi_{\text{old}}}^{\pi_{\text{new}}} - \sum_a \pi_{\text{old}}(a)\epsilon(s, a) + \sum_a \pi_{\text{new}}(a)(A_\tau^{\pi_{\text{old}}}(s, a) + \epsilon(s, a)) \\
 &\quad \triangleright \text{cancelling entropy and using definition of soft advantage}
 \end{aligned}$$

Upon averaging both sides over $d^{\pi_{\text{new}}}$, dividing by $1 - \gamma$, and using the performance difference lemma, we obtain the result.

$$\begin{aligned}
 & \frac{\tau}{1 - \gamma} \mathbb{E}_{\pi_{\text{new}}} [\text{KL}_{\mathcal{B}_\tau \hat{Q}}^{\pi_{\text{old}}} - \text{KL}_{\mathcal{B}_\tau \hat{Q}}^{\pi_{\text{new}}}] \\
 &= \eta(\pi_{\text{new}}) - \eta(\pi_{\text{old}}) + \mathbb{E}_{d^{\pi_{\text{new}}}} [\mathbb{E}_{\pi_{\text{new}}} \epsilon(s, a) - \mathbb{E}_{\pi_{\text{old}}} \epsilon(s, a)].
 \end{aligned}$$

■

Corollary 18. *[Approximate RKL Reduction] Under the assumptions of Theorem 8, $\eta_\tau(\pi_{\text{new}}) \geq \eta_\tau(\pi_{\text{old}})$ iff $\frac{\tau}{1 - \gamma} \mathbb{E}_{d^{\pi_{\text{new}}}} \Delta \text{RKL}_{\pi_{\text{new}}}^{\pi_{\text{old}}}(S) \geq \bar{\epsilon}$.*

Proof The proof is simply rearrangement of Theorem 8. ■

B.1.2 PROOFS FOR SECTION 5.3.2: EXTENSIONS TO WEIGHTING UNDER $d^{\pi_{\text{old}}}$

Proposition 19. *If $\text{KL}_{\pi_{\text{old}}}^{\pi_{\text{new}}}(s) \leq \alpha$ for all $s \in \mathcal{S}$,*

$$\begin{aligned} \eta_{\tau}(\pi_{\text{new}}) - \eta_{\tau}(\pi_{\text{old}}) &\geq \frac{1}{1-\gamma} \mathbb{E}_{d^{\pi_{\text{old}}}} \left[\sum_a \pi_{\text{new}}(a|S) A_{\tau}^{\pi_{\text{old}}}(S, a) - \tau \text{KL}_{\pi_{\text{old}}}^{\pi_{\text{new}}}(S) - 4\sqrt{2}V_{\tau, \max} \sqrt{\alpha} \right] \\ &= \frac{1}{1-\gamma} \mathbb{E}_{d^{\pi_{\text{old}}}} \left[\Delta \text{RKL}_{\pi_{\text{new}}}^{\pi_{\text{old}}} - 4\sqrt{2}V_{\tau, \max} \sqrt{\alpha} \right] \end{aligned} \quad (17)$$

Proof

Reversing the role of π_{new} and π_{old} in Theorem 6,

$$\eta_{\tau}(\pi_{\text{new}}) - \eta_{\tau}(\pi_{\text{old}}) = \frac{1}{1-\gamma} \mathbb{E}_{d^{\pi_{\text{old}}, \pi_{\text{old}}}} \left[\tau \text{KL}_{\pi_{\text{new}}}^{\pi_{\text{old}}}(S) - A_{\tau}^{\pi_{\text{new}}}(S, A) \right].$$

Let Δ be $A_{\tau}^{\pi_{\text{new}}} - A_{\tau}^{\pi_{\text{old}}}$. Since $\sum_{a \in \mathcal{A}} \pi(a|s) A_{\tau}^{\pi}(s, a) = 0$, we have that

$$\begin{aligned} \sum_{a \in \mathcal{A}} \pi_{\text{old}}(a|s) A_{\tau}^{\pi_{\text{new}}}(s, a) &= \sum_{a \in \mathcal{A}} \pi_{\text{old}}(a|s) \Delta(s, a) \\ &= \sum_{a \in \mathcal{A}} (\pi_{\text{old}}(a|s) - \pi_{\text{new}}(a|s)) \Delta(s, a) + \sum_{a \in \mathcal{A}} \pi_{\text{new}}(a|s) \Delta(s, a) \\ &= \sum_{a \in \mathcal{A}} (\pi_{\text{old}}(a|s) - \pi_{\text{new}}(a|s)) \Delta(s, a) - \sum_{a \in \mathcal{A}} \pi_{\text{new}}(a|s) A_{\tau}^{\pi_{\text{old}}}(s, a). \end{aligned}$$

Furthermore, note that

$$\Delta(s, a) = \underbrace{Q_{\tau}^{\pi_{\text{new}}}(s, a) - V_{\tau}^{\pi_{\text{new}}}(s) - Q_{\tau}^{\pi_{\text{old}}}(s, a) + V_{\tau}^{\pi_{\text{old}}}(s)}_{:= \Delta'(s, a)} + \tau \log \frac{\pi_{\text{old}}(a|s)}{\pi_{\text{new}}(a|s)}.$$

Therefore, letting $V_{\tau, \max} := \frac{r_{\max} + \tau \log A}{1-\gamma}$

$$\begin{aligned} &\sum_{a \in \mathcal{A}} \pi_{\text{old}}(a|s) A_{\tau}^{\pi_{\text{new}}}(s, a) \\ &= \sum_{a \in \mathcal{A}} (\pi_{\text{old}}(a|s) - \pi_{\text{new}}(a|s)) \Delta'(s, a) + \tau [\text{KL}_{\pi_{\text{new}}}^{\pi_{\text{old}}}(s) + \text{KL}_{\pi_{\text{old}}}^{\pi_{\text{new}}}(s)] - \sum_{a \in \mathcal{A}} \pi_{\text{new}}(a|s) A_{\tau}^{\pi_{\text{old}}}(s, a) \\ &\leq 4V_{\tau, \max} \sum_{a \in \mathcal{A}} |\pi_{\text{old}}(a|s) - \pi_{\text{new}}(a|s)| + \tau [\text{KL}_{\pi_{\text{new}}}^{\pi_{\text{old}}}(s) + \text{KL}_{\pi_{\text{old}}}^{\pi_{\text{new}}}(s)] - \sum_{a \in \mathcal{A}} \pi_{\text{new}}(a|s) A_{\tau}^{\pi_{\text{old}}}(s, a) \\ &\leq 4V_{\tau, \max} \sqrt{2\text{KL}_{\pi_{\text{old}}}^{\pi_{\text{new}}}(s)} + \tau [\text{KL}_{\pi_{\text{new}}}^{\pi_{\text{old}}}(s) + \text{KL}_{\pi_{\text{old}}}^{\pi_{\text{new}}}(s)] - \sum_{a \in \mathcal{A}} \pi_{\text{new}}(a|s) A_{\tau}^{\pi_{\text{old}}}(s, a). \end{aligned}$$

Using this result,

$$\begin{aligned} &\eta_{\tau}(\pi_{\text{new}}) - \eta_{\tau}(\pi_{\text{old}}) \\ &\geq \frac{1}{1-\gamma} \mathbb{E}_{d^{\pi_{\text{old}}}} \left[\sum_{a \in \mathcal{A}} \pi_{\text{new}}(a|S) A_{\tau}^{\pi_{\text{old}}}(S, a) - \tau \text{KL}_{\pi_{\text{old}}}^{\pi_{\text{new}}}(S) - 4V_{\tau, \max} \sqrt{2\text{KL}_{\pi_{\text{old}}}^{\pi_{\text{new}}}(S)} \right] \\ &\geq \frac{1}{1-\gamma} \mathbb{E}_{d^{\pi_{\text{old}}}} \left[\sum_{a \in \mathcal{A}} \pi_{\text{new}}(a|S) A_{\tau}^{\pi_{\text{old}}}(S, a) - \tau \text{KL}_{\pi_{\text{old}}}^{\pi_{\text{new}}}(S) - 4V_{\tau, \max} \sqrt{2\alpha} \right]. \end{aligned}$$

The final result follows by noting that $\Delta\text{RKL}_{\pi_{\text{new}}}^{\pi_{\text{old}}} = \sum_{a \in \mathcal{A}} \pi_{\text{new}}(a|S) A_{\tau}^{\pi_{\text{old}}}(S, a) - \tau \text{KL}_{\pi_{\text{old}}}^{\pi_{\text{new}}}(S)$, as in the proof of Theorem 7. \blacksquare

Proposition 20. *If $\text{KL}_{\pi_{\text{old}}}^{\pi_{\text{new}}}(s) \leq \alpha$ for all $s \in \mathcal{S}$,*

$$\begin{aligned} \eta_{\tau}(\pi_{\text{new}}) - \eta_{\tau}(\pi_{\text{old}}) &\geq \frac{1}{1-\gamma} \mathbb{E}_{d^{\pi_{\text{old}}}} \left[\tau \hat{\Delta}\text{RKL}_{\pi_{\text{new}}}^{\pi_{\text{old}}}(S) \right. \\ &\quad \left. + \sum_a \epsilon(s, a) (\pi_{\text{old}}(a) - \pi_{\text{new}}(a)) - 4V_{\tau, \max} \sqrt{2\alpha} \right]. \end{aligned} \quad (18)$$

Proof From the proof of Theorem 8, we have

$$\sum_{a \in \mathcal{A}} \pi_{\text{new}}(a|S) A_{\tau}^{\pi_{\text{old}}}(S, a) - \tau \text{KL}_{\pi_{\text{old}}}^{\pi_{\text{new}}}(S) = \tau (\text{KL}_{\mathcal{B}_{\tau}\hat{Q}}^{\pi_{\text{old}}} - \text{KL}_{\mathcal{B}_{\tau}\hat{Q}}^{\pi_{\text{new}}}) + \sum_a \epsilon(s, a) (\pi_{\text{old}}(a) - \pi_{\text{new}}(a)).$$

Using this result in Theorem 10,

$$\begin{aligned} &\eta_{\tau}(\pi_{\text{new}}) - \eta_{\tau}(\pi_{\text{old}}) \\ &\geq \frac{1}{1-\gamma} \mathbb{E}_{d^{\pi_{\text{old}}}} \left[\sum_{a \in \mathcal{A}} \pi_{\text{new}}(a|S) A_{\tau}^{\pi_{\text{old}}}(S, a) - \tau \text{KL}_{\pi_{\text{old}}}^{\pi_{\text{new}}}(S) - 4V_{\tau, \max} \sqrt{2\alpha} \right] \\ &= \frac{1}{1-\gamma} \mathbb{E}_{d^{\pi_{\text{old}}}} \left[\tau (\text{KL}_{\mathcal{B}_{\tau}\hat{Q}}^{\pi_{\text{old}}} - \text{KL}_{\mathcal{B}_{\tau}\hat{Q}}^{\pi_{\text{new}}}) + \sum_a \epsilon(s, a) (\pi_{\text{old}}(a) - \pi_{\text{new}}(a)) - 4V_{\tau, \max} \sqrt{2\alpha} \right] \\ &= \frac{1}{1-\gamma} \mathbb{E}_{d^{\pi_{\text{old}}}} \left[\tau \hat{\Delta}\text{RKL}_{\pi_{\text{new}}}^{\pi_{\text{old}}}(S) + \sum_a \epsilon(s, a) (\pi_{\text{old}}(a) - \pi_{\text{new}}(a)) - 4V_{\tau, \max} \sqrt{2\alpha} \right]. \end{aligned}$$

\blacksquare

B.2 Proofs for Section 5.4: Policy Improvement with FKL

Proposition 21. *[FKL Counterexample] There exists an MDP such that, for any $\tau \geq 0$, there exists a pair of policies $(\pi_{\text{old}}, \pi_{\text{new}})$ where $\Delta\text{FKL}_{\pi_{\text{new}}}^{\pi_{\text{old}}}(s) > 0$ at every state but the new policy has lower value: $Q_{\tau}^{\pi_{\text{new}}}(s, a) < Q_{\tau}^{\pi_{\text{old}}}(s, a)$ at every state-action pair $(s, a) \in \mathcal{S} \times \mathcal{A}$, $V_{\tau}^{\pi_{\text{new}}}(s) < V_{\tau}^{\pi_{\text{old}}}(s)$ at every state $s \in \mathcal{S}$, and $\eta_{\tau}(\pi_{\text{new}}) < \eta_{\tau}(\pi_{\text{old}})$.*

Proof Consider the environment with a single state s and two actions: a_1 and a_2 . Regardless of the action chosen, the agent always transition to s . We will omit dependency on the state in the following notation. The rewards are defined as

$$r(a_1) = -1 \quad , \quad r(a_2) = 1.$$

Take π_{old} and π_{new} as follows.

$$\begin{aligned} \pi_{\text{old}}(a_1) &= \epsilon_1 \quad , \quad \pi_{\text{new}}(a_1) = 1 - \epsilon_2, \\ \pi_{\text{old}}(a_2) &= 1 - \epsilon_1 \quad , \quad \pi_{\text{new}}(a_2) = \epsilon_2. \end{aligned}$$

We will prove the result by making ϵ_1 arbitrarily small, which we will show forces $\text{FKL}_{\mathcal{B}_\tau Q_\tau^{\pi_{\text{old}}}}^{\pi_{\text{old}}}$ to ∞ while keeping ϵ_2 fixed, causing $\text{FKL}_{\mathcal{B}_\tau Q_\tau^{\pi_{\text{old}}}}^{\pi_{\text{new}}}$ to be finite. Note that

$$\begin{aligned} \text{FKL}_{\mathcal{B}_\tau Q_\tau^{\pi_{\text{old}}}}^{\pi_{\text{old}}} &= \sum_{i=1}^2 \mathcal{B}_\tau Q_\tau^{\pi_{\text{old}}}(a_i) \log\left(\frac{\mathcal{B}_\tau Q_\tau^{\pi_{\text{old}}}(a_i)}{\pi_{\text{old}}(a_i)}\right) \\ &= -\mathcal{H}(\mathcal{B}_\tau Q_\tau^{\pi_{\text{old}}}) - \sum_{i=1}^2 \mathcal{B}_\tau Q_\tau^{\pi_{\text{old}}}(a_i) \log(\pi_{\text{old}}(a_i)), \\ \lim_{\epsilon_1 \rightarrow 0} \text{FKL}_{\mathcal{B}_\tau Q_\tau^{\pi_{\text{old}}}}^{\pi_{\text{old}}} &= \underbrace{-\lim_{\epsilon_1 \rightarrow 0} \mathcal{H}(\mathcal{B}_\tau Q_\tau^{\pi_{\text{old}}})}_{\geq -\log 2 \text{ and } \leq 0} - \lim_{\epsilon_1 \rightarrow 0} \mathcal{B}_\tau Q_\tau^{\pi_{\text{old}}}(a_1) \log(\epsilon_1) + 0. \end{aligned}$$

To calculate the limit of the middle summand, we note that if $\lim_{\epsilon_1 \rightarrow 0} \mathcal{B}_\tau Q_\tau^{\pi_{\text{old}}}(a_1) > 0$, the middle summand will go to infinity, since $\lim_{\epsilon_1 \rightarrow 0} -\log(\epsilon_1) = \infty$. We can verify that this is indeed the case.

$$\begin{aligned} \lim_{\epsilon_1 \rightarrow 0} Q_\tau^{\pi_{\text{old}}}(a_1) &= -1 + \frac{\gamma}{1-\gamma} & ; & \quad \lim_{\epsilon_1 \rightarrow 0} Q_\tau^{\pi_{\text{old}}}(a_2) = \frac{1}{1-\gamma} \\ \lim_{\epsilon_1 \rightarrow 0} \mathcal{B}_\tau Q_\tau^{\pi_{\text{old}}}(a_1) &= \frac{e^{(-1+\frac{\gamma}{1-\gamma})\frac{1}{\tau}}}{Z} & ; & \quad \lim_{\epsilon_1 \rightarrow 0} \mathcal{B}_\tau Q_\tau^{\pi_{\text{old}}}(a_2) = \frac{e^{(\frac{1}{1-\gamma})\frac{1}{\tau}}}{Z}, \end{aligned} \quad (24)$$

where $Z := e^{(-1+\frac{\gamma}{1-\gamma})\frac{1}{\tau}} + e^{(\frac{1}{1-\gamma})\frac{1}{\tau}}$. Since, for fixed γ and τ , the quantities in Equation (24) are fixed, we have that $\lim_{\epsilon_1 \rightarrow 0} -\mathcal{B}_\tau Q_\tau^{\pi_{\text{old}}}(a_1) \log(\epsilon_1) = \infty$, causing $\lim_{\epsilon_1 \rightarrow 0} \text{FKL}_{\mathcal{B}_\tau Q_\tau^{\pi_{\text{old}}}}^{\pi_{\text{old}}} = \infty$. Moreover, $\text{FKL}_{\mathcal{B}_\tau Q_\tau^{\pi_{\text{old}}}}^{\pi_{\text{new}}}$ has a similar form of the above FKL, but, since ϵ_2 is assumed to be fixed, this quantity will be finite. The point is that for any ϵ_2 , we can find ϵ_1 such that $\text{FKL}_{\mathcal{B}_\tau Q_\tau^{\pi_{\text{old}}}}^{\pi_{\text{old}}} > \text{FKL}_{\mathcal{B}_\tau Q_\tau^{\pi_{\text{old}}}}^{\pi_{\text{new}}}$.

It remains to be seen that we can have ϵ_2 that guarantees $V_\tau^{\pi_{\text{new}}} < V_\tau^{\pi_{\text{old}}}$. We write

$$\begin{aligned} V_\tau^{\pi_{\text{new}}} &= \mathbb{E} \left[\sum_{t=0}^{\infty} \gamma^t (R + \tau \mathcal{H}(\pi_{\text{new}})) \right] = \mathbb{E} \left[\sum_{t=0}^{\infty} \gamma^t R \right] + \frac{\tau \mathcal{H}(\pi_{\text{new}})}{1-\gamma} \\ &= \left(\sum_{t=0}^{\infty} \gamma^t \mathbb{E}[R] \right) + \frac{\tau \mathcal{H}(\pi_{\text{new}})}{1-\gamma} \\ &= \frac{2\epsilon_2 - 1}{1-\gamma} + \frac{\tau \mathcal{H}(\pi_{\text{new}})}{1-\gamma}. \end{aligned}$$

Additionally, we know that $\lim_{\epsilon_1 \rightarrow 0} V_\tau^{\pi_{\text{old}}} = \frac{1}{1-\gamma}$. If we can find ϵ_2 such that $V_\tau^{\pi_{\text{new}}} < \frac{1}{1-\gamma}$, then we can find ϵ_1 such that simultaneously $V_\tau^{\pi_{\text{new}}} < V_\tau^{\pi_{\text{old}}}$ and $\text{FKL}_{\mathcal{B}_\tau Q_\tau^{\pi_{\text{old}}}}^{\pi_{\text{old}}} > \text{FKL}_{\mathcal{B}_\tau Q_\tau^{\pi_{\text{old}}}}^{\pi_{\text{new}}}$. $V_\tau^{\pi_{\text{new}}} < \frac{1}{1-\gamma}$ will hold if

$$\begin{aligned} \frac{1}{1-\gamma} (2\epsilon_2 - 1 + \tau \mathcal{H}(\pi_{\text{new}})) &< \frac{1}{1-\gamma} \\ 2\epsilon_2 - 1 + \tau(-(1-\epsilon_2) \log(1-\epsilon_2) - \epsilon_2 \log(\epsilon_2)) &< 1. \end{aligned} \quad (25)$$

We use $f(\epsilon_2)$ to denote the LHS of eq. (25). We have

$$\lim_{\epsilon_2 \rightarrow 0} f(\epsilon_2) = -1,$$

which is less than the RHS of Equation (25). Formally, for any $\epsilon > 0$, we can find $\delta > 0$ such that if $|\epsilon_2| < \delta$, $|f(\epsilon_2) - 1| < \epsilon$. Any $\epsilon < 2$ will suffice: we can conclude that, for any fixed $\tau \geq 0$, there is some $\epsilon_2 \in (0, 1)$ satisfying eq. (25).

It is possible to extend this to an MDP with multiple states. Consider a case where $\mathcal{S} = \{s_1, s_2, s_3, \dots, s_N\}$, and at each $s_n \in \mathcal{S}$ except s_N , a state transition to s_{n+1} occurs. At state s_N , the transition is back to itself. Setting the reward, π_{old} , and π_{new} similarly to the above environment, we can show the same result with multiple states. \blacksquare

B.2.1 PROOFS FOR SECTION 5.4.2: POLICY IMPROVEMENT UNDER SUFFICIENT FKL REDUCTION

Lemma 22. *Let the action space \mathcal{A} be finite. For any state s and any vector $\hat{Q} : \mathcal{S} \times \mathcal{A} \rightarrow \mathbb{R}$, if $\text{FKL}_{\mathcal{B}_\tau \hat{Q}}^{\pi_{\text{new}}}(s) \leq \text{FKL}_{\mathcal{B}_\tau \hat{Q}}^{\pi_{\text{old}}}(s)$ then*

$$\text{RKL}_{\mathcal{B}_\tau \hat{Q}}^{\pi_{\text{new}}}(s) \leq \frac{\|\hat{Q}\|_\infty}{\tau} \sqrt{2\text{FKL}_{\mathcal{B}_\tau \hat{Q}}^{\pi_{\text{new}}}(s) - \mathcal{B}_\tau \hat{Q}(s, \cdot)^\top \log(\boldsymbol{\pi}_{\text{old}}(\cdot | s))}$$

Proof We will suppress the argument s . Start with the following.

$$\begin{aligned} 0 &\leq \text{FKL}_{\mathcal{B}_\tau Q}^{\pi_{\text{new}}} \\ &\leq \mathcal{B}_\tau Q^\top \log\left(\frac{\mathcal{B}_\tau Q}{\boldsymbol{\pi}_{\text{new}}}\right) \\ &= \mathcal{B}_\tau Q^\top \left(\frac{\mathbf{q}}{\tau} - \log(\boldsymbol{\pi}_{\text{new}})\right) - \log(Z) \\ &\quad \triangleright \text{Expanding the inner } \mathcal{B}_\tau Q \\ 0 &\leq \mathcal{B}_\tau Q^\top (\mathbf{q} - \tau \log(\boldsymbol{\pi}_{\text{new}})) - \tau \log(Z) \\ &\quad \triangleright \text{Multiplying both sides by } \tau \\ &= \boldsymbol{\pi}_{\text{new}}^\top (\mathbf{q} - \tau \log \boldsymbol{\pi}_{\text{new}}) + \mathbf{q}^\top (\mathcal{B}_\tau Q - \boldsymbol{\pi}_{\text{new}}) + \\ &\quad \tau \log \boldsymbol{\pi}_{\text{new}}^\top (\boldsymbol{\pi}_{\text{new}} - \mathcal{B}_\tau Q) - \tau \log(Z) \\ &\quad \triangleright \text{Adding and subtracting } \boldsymbol{\pi}_{\text{new}}^\top \mathbf{q} \text{ and } \tau \boldsymbol{\pi}_{\text{new}}^\top \log \boldsymbol{\pi}_{\text{new}} \\ &= -\tau \text{RKL}_{\mathcal{B}_\tau Q}^{\pi_{\text{new}}} + \mathbf{q}^\top (\mathcal{B}_\tau Q - \boldsymbol{\pi}_{\text{new}}) + \\ &\quad \tau (\boldsymbol{\pi}_{\text{new}}^\top \log(\boldsymbol{\pi}_{\text{new}}) - \mathcal{B}_\tau Q^\top \log(\boldsymbol{\pi}_{\text{new}})) \\ &\quad \triangleright \text{Absorbing the partition function.} \end{aligned} \tag{26}$$

We analyze some of the summands from Equation (26) in turn. By Hölder's inequality,

$$\mathbf{q}^\top (\mathcal{B}_\tau Q - \boldsymbol{\pi}_{\text{new}}) \leq \|\mathbf{q}\|_\infty \sum_i |(\mathcal{B}_\tau Q_i - \pi_{\text{new}i})|.$$

By Pinsker's inequality (Pinsker, 1964),

$$\sum_i |(\mathcal{B}_\tau Q)_i - \pi_{\text{new}i}| \leq \sqrt{2\text{FKL}_{\mathcal{B}_\tau Q}^{\pi_{\text{new}}}}.$$

Therefore,

$$\mathbf{q}^\top (\mathcal{B}_\tau Q - \boldsymbol{\pi}_{\text{new}}) \leq \|\mathbf{q}\|_\infty \sqrt{2\text{FKL}_{\mathcal{B}_\tau Q}^{\pi_{\text{new}}}}.$$

The other summand from Equation (26) can be written as

$$\tau(\boldsymbol{\pi}_{\text{new}}^\top \log(\boldsymbol{\pi}_{\text{new}}) - \mathcal{B}_\tau Q^\top \log(\boldsymbol{\pi}_{\text{new}})) \leq \tau(0 - \mathcal{B}_\tau Q^\top \log(\boldsymbol{\pi}_{\text{old}})),$$

where we used the fact that the negative entropy of $\boldsymbol{\pi}_{\text{new}}$ is less than or equal to zero and that, since the underlying assumption is that we have non-negative FKL reduction, we also have $-\mathcal{B}_\tau Q^\top \log(\boldsymbol{\pi}_{\text{new}}) \leq -\mathcal{B}_\tau Q^\top \log(\boldsymbol{\pi}_{\text{old}})$ (by the FKL definition).

We substitute these upper bounds into Equation (26).

$$\begin{aligned} 0 &\leq -\tau \text{RKL}_{\mathcal{B}_\tau Q}^{\pi_{\text{new}}} + \mathbf{q}^\top (\mathcal{B}_\tau Q - \boldsymbol{\pi}_{\text{new}}) + \\ &\quad \tau(\boldsymbol{\pi}_{\text{new}}^\top \log(\boldsymbol{\pi}_{\text{new}}) - \mathcal{B}_\tau Q^\top \log(\boldsymbol{\pi}_{\text{new}})) \\ &\leq -\tau \text{RKL}_{\mathcal{B}_\tau Q}^{\pi_{\text{new}}} + \|\mathbf{q}\|_\infty \sqrt{2\text{FKL}_{\mathcal{B}_\tau Q}^{\pi_{\text{new}}}} + \\ &\quad \tau(-\mathcal{B}_\tau Q^\top \log(\boldsymbol{\pi}_{\text{old}})), \end{aligned}$$

which then implies

$$\text{RKL}_{\mathcal{B}_\tau Q}^{\pi_{\text{new}}} \leq \frac{\|\mathbf{q}\|_\infty}{\tau} \sqrt{2\text{FKL}_{\mathcal{B}_\tau Q}^{\pi_{\text{new}}}} - \mathcal{B}_\tau Q^\top \log(\boldsymbol{\pi}_{\text{old}}). \quad (27)$$

■

We start by showing this result in bandits, to make it simpler to show across states.

Definition 23 (Entropy-Regularized Bandits). *We denote $\boldsymbol{\pi}$ as a vector in $\mathbb{R}^{|\mathcal{A}|}$ that satisfies $\boldsymbol{\pi} \geq \mathbf{0}$ and $\boldsymbol{\pi}^\top \mathbf{1} = 1$, with $\mathbf{1}$ and $\mathbf{0}$ being vectors containing respectively only entries equal to 1 and 0. Further, we consider a single state and denote the corresponding action-values as a vector $\mathbf{q} \in \mathbb{R}^{|\mathcal{A}|}$. The objective is then defined as $\eta_\tau(\boldsymbol{\pi}) = \boldsymbol{\pi}^\top (\mathbf{q} - \tau \log(\boldsymbol{\pi}))$. Moreover, $\mathcal{B}_\tau Q = \mathcal{B}_\tau \mathbf{q} = \frac{\exp(\frac{\mathbf{q}}{\tau})}{Z}$ with $Z = \exp(\frac{\mathbf{q}}{\tau})^\top \mathbf{1}$ being the normalizing constant. We also have $\text{FKL}_{\mathcal{B}_\tau Q}^{\boldsymbol{\pi}} = \mathcal{B}_\tau Q^\top \log\left(\frac{\mathcal{B}_\tau Q}{\boldsymbol{\pi}}\right)$ and $\text{RKL}_{\mathcal{B}_\tau Q}^{\boldsymbol{\pi}} = \boldsymbol{\pi}^\top \log\left(\frac{\boldsymbol{\pi}}{\mathcal{B}_\tau Q}\right)$.*

The maximal possible FKL reduction is obtained by moving π_{new} all the way to $\mathcal{B}_\tau Q_{\tau}^{\pi_{\text{old}}}$, to give $\Delta\text{FKL} = \text{FKL}_{\mathcal{B}_\tau Q_{\tau}^{\pi_{\text{old}}}}^{\pi_{\text{old}}}$; for this π_{new} , we can guarantee RKL reduction. The question is if we can still obtain RKL reduction, even without stepping all the way to this maximal possible FKL reduction. We provide a condition on how much FKL reduction is sufficient to ensure that we obtain policy improvement, first in the bandit setting and then generalized to the MDP setting.

Proposition 24 (Sufficient FKL Reduction in Bandits). *For two policies $\boldsymbol{\pi}_{old}, \boldsymbol{\pi}_{new} \in \mathbb{R}^{|\mathcal{A}|}$ in the bandit setting, if*

$$\begin{aligned} & \Delta \text{FKL}_{\boldsymbol{\pi}_{new}}^{\boldsymbol{\pi}_{old}} \\ & \geq \max\{0, \text{FKL}_{\mathcal{B}_\tau Q}^{\boldsymbol{\pi}_{old}} - \frac{1}{2} \left(\frac{\tau}{\|\mathbf{q}\|_\infty} \left(\text{RKL}_{\mathcal{B}_\tau Q}^{\boldsymbol{\pi}_{old}} + \mathcal{B}_\tau Q^\top \log(\boldsymbol{\pi}_{old}) \right) \right)^2 \} \end{aligned} \quad (28)$$

and

$$\text{RKL}_{\mathcal{B}_\tau Q}^{\boldsymbol{\pi}_{old}} + \mathcal{B}_\tau Q^\top \log(\boldsymbol{\pi}_{old}) \geq 0,$$

then $\Delta \text{RKL}_{\boldsymbol{\pi}_{new}}^{\boldsymbol{\pi}_{old}} \geq 0$ and $\eta_\tau(\boldsymbol{\pi}_{new}) \geq \eta_\tau(\boldsymbol{\pi}_{old})$.

It is straightforward to extend this result to MDPs when we have reduction in every state.

Corollary 25 (All-state Sufficient FKL reduction). *Assume the action set is finite. If the assumptions in Theorem 24 are satisfied for all $s \in \mathcal{S}$, then $Q_\tau^{\boldsymbol{\pi}_{new}}(s, a) \geq Q_\tau^{\boldsymbol{\pi}_{old}}(s, a)$ for all states and actions and $\eta_\tau(\boldsymbol{\pi}_{new}) \geq \eta_\tau(\boldsymbol{\pi}_{old})$.*

Proof We know that reducing the RKL in all states (i.e., the RKL of $\boldsymbol{\pi}_{new}$ to $\mathcal{B}_\tau Q_\tau^{\boldsymbol{\pi}_{old}}$ is smaller than the RKL of $\boldsymbol{\pi}_{old}$ to $\mathcal{B}_\tau Q_\tau^{\boldsymbol{\pi}_{old}}$) will lead to $Q_\tau^{\boldsymbol{\pi}_{new}}(s, a) \geq Q_\tau^{\boldsymbol{\pi}_{old}}(s, a)$ for all (s, a) . From our RKL results, we also know that it implies $\eta_\tau(\boldsymbol{\pi}_{new}) \geq \eta_\tau(\boldsymbol{\pi}_{old})$. Therefore, FKL reduction following Equation (28) in all states will also lead to these improvements, since, by the same argument as Theorem 24, it leads to RKL reduction in all states. \blacksquare

Proposition 26 (Sufficient FKL Reduction in Bandits). *For two policies $\boldsymbol{\pi}_{old}, \boldsymbol{\pi}_{new} \in \mathbb{R}^{|\mathcal{A}|}$ in the bandit setting, if*

$$\begin{aligned} & \Delta \text{FKL}_{\boldsymbol{\pi}_{new}}^{\boldsymbol{\pi}_{old}} \\ & \geq \max\{0, \text{FKL}_{\mathcal{B}_\tau Q}^{\boldsymbol{\pi}_{old}} - \frac{1}{2} \left(\frac{\tau}{\|\mathbf{q}\|_\infty} \left(\text{RKL}_{\mathcal{B}_\tau Q}^{\boldsymbol{\pi}_{old}} + \mathcal{B}_\tau Q^\top \log(\boldsymbol{\pi}_{old}) \right) \right)^2 \} \end{aligned} \quad (28)$$

and

$$\text{RKL}_{\mathcal{B}_\tau Q}^{\boldsymbol{\pi}_{old}} + \mathcal{B}_\tau Q^\top \log(\boldsymbol{\pi}_{old}) \geq 0,$$

then $\Delta \text{RKL}_{\boldsymbol{\pi}_{new}}^{\boldsymbol{\pi}_{old}} \geq 0$ and $\eta_\tau(\boldsymbol{\pi}_{new}) \geq \eta_\tau(\boldsymbol{\pi}_{old})$.

Proof We start with the result of Theorem 22. If the RHS of Equation (27) is less than or equal to $\text{RKL}_{\mathcal{B}_\tau Q}^{\boldsymbol{\pi}_{old}}$, we will have $\text{RKL}_{\mathcal{B}_\tau Q}^{\boldsymbol{\pi}_{new}} \leq \text{RKL}_{\mathcal{B}_\tau Q}^{\boldsymbol{\pi}_{old}}$, which in turn implies improvement. The assumption that the RHS of Equation (27) is $\leq \text{RKL}_{\mathcal{B}_\tau Q}^{\boldsymbol{\pi}_{old}}$ can be written as

$$\frac{\|\mathbf{q}\|_\infty}{\tau} \sqrt{2\text{FKL}_{\mathcal{B}_\tau Q}^{\boldsymbol{\pi}_{new}} - \mathcal{B}_\tau Q^\top \log(\boldsymbol{\pi}_{old})} \leq \text{RKL}_{\mathcal{B}_\tau Q}^{\boldsymbol{\pi}_{old}}.$$

With some algebraic manipulation, we get that this assumption is equivalent to

$$\frac{\tau}{\|\mathbf{q}\|_\infty} (\text{RKL}_{\mathcal{B}_\tau Q}^{\boldsymbol{\pi}_{old}} + \mathcal{B}_\tau Q^\top \log(\boldsymbol{\pi}_{old})) \geq \sqrt{2\text{FKL}_{\mathcal{B}_\tau Q}^{\boldsymbol{\pi}_{new}}}.$$

Assuming that $\text{RKL}_{\mathcal{B}_\tau Q}^{\pi_{old}} + \mathcal{B}_\tau Q^\top \log(\boldsymbol{\pi}_{old}) \geq 0$, the above is equivalent to

$$\left(\frac{\tau}{\|\mathbf{q}\|_\infty} (\text{RKL}_{\mathcal{B}_\tau Q}^{\pi_{old}} + \mathcal{B}_\tau Q^\top \log(\boldsymbol{\pi}_{old})) \right)^2 \geq 2\text{FKL}_{\mathcal{B}_\tau Q}^{\pi_{new}}.$$

Dividing by 2, adding $\text{FKL}_{\mathcal{B}_\tau Q}^{\pi_{old}}$, and rearranging yields that the assumption is equivalent to the following.

$$\Delta\text{FKL}_{\pi_{new}}^{\pi_{old}} \geq \text{FKL}_{\mathcal{B}_\tau Q}^{\pi_{old}} - \frac{1}{2} \left(\frac{\tau}{\|\mathbf{q}\|_\infty} (\text{RKL}_{\mathcal{B}_\tau Q}^{\pi_{old}} + \mathcal{B}_\tau Q^\top \log(\boldsymbol{\pi}_{old})) \right)^2. \quad (29)$$

The claim follows. ■

Proposition 27 (Improvement Under Average Sufficient FKL Reduction). *Assume the action set is finite. If*

$$\mathbb{E}_{d^{\pi_{new}}} \left[\text{RKL}_{\mathcal{B}_\tau Q_\tau^{\pi_{old}}}^{\pi_{old}}(S) + \mathbb{E}_{\mathcal{B}_\tau Q_\tau^{\pi_{old}}} [\log \pi_{old}(\cdot|S)] \right] \geq 0, \quad (19)$$

$$\begin{aligned} \mathbb{E}_{d^{\pi_{new}}} [\Delta\text{FKL}_{\pi_{new}}^{\pi_{old}}(S)] &\geq \mathbb{E}_{d^{\pi_{new}}} [\text{FKL}_{\mathcal{B}_\tau Q_\tau^{\pi_{old}}}^{\pi_{old}}(S)] - \frac{1}{2} \left(\frac{\tau}{\|Q_\tau^{\pi_{old}}\|_\infty} \left(\mathbb{E}_{d^{\pi_{new}}} \left[\text{RKL}_{\mathcal{B}_\tau Q_\tau^{\pi_{old}}}^{\pi_{old}}(S) \right] \right. \right. \\ &\quad \left. \left. + \mathbb{E}_{\mathcal{B}_\tau Q_\tau^{\pi_{old}}} \left[\mathbb{E}_{\mathcal{B}_\tau Q_\tau^{\pi_{old}}} [\log(\pi_{old}(\cdot|S))] \right] \right) \right)^2, \end{aligned} \quad (20)$$

$$\text{and } \forall s \in \mathcal{S}, \Delta\text{FKL}_{\pi_{new}}^{\pi_{old}}(S) \geq 0, \quad (21)$$

then $\eta_\tau(\pi_{new}) \geq \eta_\tau(\pi_{old})$.

Proof The strategy is the same as the one in Theorem 24. We know from Theorem 7 that if we have

$$\mathbb{E}_{d^{\pi_{new}}} [\Delta\text{RKL}_{\pi_{new}}^{\pi_{old}}(S)] \geq 0,$$

then $\eta_\tau(\pi_{new}) \geq \eta_\tau(\pi_{old})$. Applying expectations to both sides of the result of Theorem 22, we obtain

$$\begin{aligned} &\mathbb{E}_{d^{\pi_{new}}} \left[\frac{\|Q_\tau^{\pi_{old}}(S, \cdot)\|_\infty}{\tau} \sqrt{2\text{FKL}_{\mathcal{B}_\tau Q_\tau^{\pi_{old}}}^{\pi_{new}}(S)} - \mathbb{E}_{\mathcal{B}_\tau Q_\tau^{\pi_{old}}} [\log(\pi_{old}(\cdot|S))] \right] \\ &\geq \mathbb{E}_{d^{\pi_{new}}} \left[\text{RKL}_{\mathcal{B}_\tau Q_\tau^{\pi_{old}}}^{\pi_{new}}(S) \right]. \end{aligned}$$

If we can show that the LHS is smaller than

$$\mathbb{E}_{d^{\pi_{new}}} \left[\text{RKL}_{\mathcal{B}_\tau Q_\tau^{\pi_{old}}}^{\pi_{old}}(S) \right],$$

then, by Theorem 7, we will be guaranteed improvement. The condition can be written as

$$\begin{aligned} \mathbb{E}_{d^{\pi_{\text{new}}}} \left[\text{RKL}_{\mathcal{B}_\tau Q_\tau^{\pi_{\text{old}}}}^{\pi_{\text{old}}}(S) \right] &\geq \\ \mathbb{E}_{d^{\pi_{\text{new}}}} \left[\frac{\|Q_\tau^{\pi_{\text{old}}}(S, \cdot)\|_\infty}{\tau} \sqrt{2\text{FKL}_{\mathcal{B}_\tau Q_\tau^{\pi_{\text{old}}}}^{\pi_{\text{new}}}(S)} - \mathbb{E}_{\mathcal{B}_\tau Q_\tau^{\pi_{\text{old}}}} [\log(\pi_{\text{old}}(\cdot|S))] \right]. \end{aligned} \quad (30)$$

Our goal is to derive the assumption on $\text{FKL}_{\mathcal{B}_\tau Q_\tau^{\pi_{\text{old}}}}^{\pi_{\text{new}}}(S)$ that must be made for Equation (30) to hold.

From rearranging to have the FKL term on one side, the following condition is equivalent to Equation (30).

$$\begin{aligned} \frac{\tau}{\|Q_\tau^{\pi_{\text{old}}}\|_\infty} \left(\mathbb{E}_{d^{\pi_{\text{new}}}} \left[\text{RKL}_{\mathcal{B}_\tau Q_\tau^{\pi_{\text{old}}}}^{\pi_{\text{old}}}(S) + \mathbb{E}_{\mathcal{B}_\tau Q_\tau^{\pi_{\text{old}}}} [\log(\pi_{\text{old}}(\cdot|S))] \right] \right) \\ \geq \mathbb{E}_{d^{\pi_{\text{new}}}} \left[\sqrt{2\text{FKL}_{\mathcal{B}_\tau Q_\tau^{\pi_{\text{old}}}}^{\pi_{\text{new}}}(S)} \right]. \end{aligned}$$

We now square both sides. The following is also equivalent to Equation (30) by assumption in Equation (19).

$$\begin{aligned} \left(\frac{\tau}{\|Q_\tau^{\pi_{\text{old}}}\|_\infty} \left(\mathbb{E}_{d^{\pi_{\text{new}}}} \left[\text{RKL}_{\mathcal{B}_\tau Q_\tau^{\pi_{\text{old}}}}^{\pi_{\text{old}}}(S) + \mathbb{E}_{\mathcal{B}_\tau Q_\tau^{\pi_{\text{old}}}} [\log(\pi_{\text{old}}(\cdot|S))] \right] \right) \right)^2 \\ \geq \left(\mathbb{E}_{d^{\pi_{\text{new}}}} \left[\sqrt{2\text{FKL}_{\mathcal{B}_\tau Q_\tau^{\pi_{\text{old}}}}^{\pi_{\text{new}}}(S)} \right] \right)^2. \end{aligned}$$

Jensen's inequality applied to the RHS shows that the following *implies* Equation (30).

$$\begin{aligned} \left(\frac{\tau}{\|Q_\tau^{\pi_{\text{old}}}\|_\infty} \left(\mathbb{E}_{d^{\pi_{\text{new}}}} \left[\text{RKL}_{\mathcal{B}_\tau Q_\tau^{\pi_{\text{old}}}}^{\pi_{\text{old}}}(S) + \mathbb{E}_{\mathcal{B}_\tau Q_\tau^{\pi_{\text{old}}}} [\log(\pi_{\text{old}}(\cdot|S))] \right] \right) \right)^2 \\ \geq \mathbb{E}_{d^{\pi_{\text{new}}}} \left[2\text{FKL}_{\mathcal{B}_\tau Q_\tau^{\pi_{\text{old}}}}^{\pi_{\text{new}}}(S) \right]. \end{aligned}$$

Some straightforward rearrangement yields the assumption in the statement of the Proposition.

$$\begin{aligned} -\frac{1}{2} \left(\frac{\tau}{\|Q_\tau^{\pi_{\text{old}}}\|_\infty} \left(\mathbb{E}_{d^{\pi_{\text{new}}}} \left[\text{RKL}_{\mathcal{B}_\tau Q_\tau^{\pi_{\text{old}}}}^{\pi_{\text{old}}}(S) + \mathbb{E}_{\mathcal{B}_\tau Q_\tau^{\pi_{\text{old}}}} [\log(\pi_{\text{old}}(\cdot|S))] \right] \right) \right)^2 \\ + \mathbb{E}_{d^{\pi_{\text{new}}}} \left[\text{FKL}_{\mathcal{B}_\tau Q_\tau^{\pi_{\text{old}}}}^{\pi_{\text{old}}}(S) \right] \\ \leq \mathbb{E}_{d^{\pi_{\text{new}}}} \left[\Delta\text{FKL}_{\pi_{\text{new}}}^{\pi_{\text{old}}}(S) \right]. \end{aligned}$$

■

B.2.2 PROOFS FOR FKL WITH APPROXIMATE ACTION-VALUES

We can also extend the above results to approximate action-values. We first state all the results upfront, and then provide their proofs.

We first reformulate Theorem 24 with action-value estimates.

Proposition 28 (Sufficient Approximate FKL Reduction in Bandits). *Let \hat{Q} be an action-value estimate of $Q_\tau^{\pi_{old}}$, let $\epsilon := \hat{Q} - Q$ be the approximation error, and let $\bar{\epsilon} := \mathbb{E}_{\pi_{new}}[\epsilon(A)] - \mathbb{E}_{\pi_{old}}[\epsilon(A)]$. As well, define*

$$\hat{\Delta}\text{FKL}_{\pi_{new}}^{\pi_{old}} := \text{FKL}_{\mathcal{B}_\tau \hat{Q}}^{\pi_{old}} - \text{FKL}_{\mathcal{B}_\tau \hat{Q}}^{\pi_{new}}.$$

For two policies $\pi_{old}, \pi_{new} \in \mathbb{R}^{|A|}$ in the bandit setting, if

$$\begin{aligned} & \hat{\Delta}\text{FKL}_{\pi_{new}}^{\pi_{old}} \\ & \geq \max \left\{ 0, \text{FKL}_{\mathcal{B}_\tau \hat{Q}}^{\pi_{old}} - \frac{1}{2} \left(\frac{\tau}{\|\hat{Q}\|_\infty} \left(\text{RKL}_{\mathcal{B}_\tau \hat{Q}}^{\pi_{old}} + \mathcal{B}_\tau \hat{Q}^\top \log(\pi_{old}) \right) \right)^2 \right\}, \end{aligned}$$

$$\text{RKL}_{\mathcal{B}_\tau \hat{Q}}^{\pi_{old}} - \bar{\epsilon} + \mathcal{B}_\tau \hat{Q}^\top \log(\pi_{old}) \geq 0,$$

and $\bar{\epsilon} \leq \hat{\Delta}\text{RKL}_{\pi_{new}}^{\pi_{old}}(s)$, then $\Delta\text{RKL}_{\pi_{new}}^{\pi_{old}} \geq 0$ and $\eta_\tau(\pi_{new}) \geq \eta_\tau(\pi_{old})$.

An analogous result to Theorem 25 applies, with the same proof.

Corollary 29 (All-state Sufficient Approximate FKL reduction). *Assume the action set is finite. If the assumptions in Theorem 28 are satisfied for all $s \in \mathcal{S}$, then $Q_\tau^{\pi_{new}}(s, a) \geq Q_\tau^{\pi_{old}}(s, a)$ for all states and actions and $\eta_\tau(\pi_{new}) \geq \eta_\tau(\pi_{old})$.*

Finally, we note an analogous result to Theorem 13.

Proposition 30 (Approximate Average FKL Reduction). *Let \hat{Q} be an action-value estimate of $Q_\tau^{\pi_{old}}$, let $\epsilon := \hat{Q} - Q$ be the approximation error, and let $\bar{\epsilon} := \mathbb{E}_{d^{\pi_{new}}}[\mathbb{E}_{\pi_{new}}[\epsilon(S, A)]] - \mathbb{E}_{\pi_{old}}[\epsilon(S, A)]$.*

For two policies $\pi_{old}, \pi_{new} \in \mathbb{R}^{|A|}$ in the bandit setting, if

$$\begin{aligned} & \hat{\mathbb{E}}_{d^{\pi_{new}}}[\Delta\text{FKL}_{\pi_{new}}^{\pi_{old}}] \\ & \geq \mathbb{E}_{d^{\pi_{new}}}[\text{FKL}_{\mathcal{B}_\tau \hat{Q}}^{\pi_{old}}] - \frac{1}{2} \left(\frac{\tau}{\|\hat{Q}\|_\infty} \left(\mathbb{E}_{d^{\pi_{new}}}[\text{RKL}_{\mathcal{B}_\tau \hat{Q}}^{\pi_{old}} + \mathcal{B}_\tau \hat{Q}^\top \log(\pi_{old})] \right) \right)^2, \end{aligned}$$

$$\mathbb{E}_{d^{\pi_{new}}}[\text{RKL}_{\mathcal{B}_\tau \hat{Q}}^{\pi_{old}} + \mathcal{B}_\tau \hat{Q}^\top \log(\pi_{old})] \geq \bar{\epsilon},$$

and $\bar{\epsilon} \leq \mathbb{E}_{d^{\pi_{new}}}[\hat{\Delta}\text{RKL}_{\pi_{new}}^{\pi_{old}}(s)]$, then $\eta_\tau(\pi_{new}) \geq \eta_\tau(\pi_{old})$.

Proposition 31 (Sufficient Approximate FKL Reduction in Bandits). *Let \hat{Q} be an action-value estimate of $Q_\tau^{\pi_{old}}$, let $\epsilon := \hat{Q} - Q$ be the approximation error, and let $\bar{\epsilon} := \mathbb{E}_{\pi_{new}}[\epsilon(A)] - \mathbb{E}_{\pi_{old}}[\epsilon(A)]$. As well, define*

$$\hat{\Delta}\text{FKL}_{\pi_{new}}^{\pi_{old}} := \text{FKL}_{\mathcal{B}_\tau \hat{Q}}^{\pi_{old}} - \text{FKL}_{\mathcal{B}_\tau \hat{Q}}^{\pi_{new}}.$$

For two policies $\pi_{old}, \pi_{new} \in \mathbb{R}^{|A|}$ in the bandit setting, if

$$\begin{aligned} & \hat{\Delta}\text{FKL}_{\pi_{new}}^{\pi_{old}} \\ & \geq \max \left\{ 0, \text{FKL}_{\mathcal{B}_\tau \hat{Q}}^{\pi_{old}} - \frac{1}{2} \left(\frac{\tau}{\|\hat{Q}\|_\infty} \left(\text{RKL}_{\mathcal{B}_\tau \hat{Q}}^{\pi_{old}} + \mathcal{B}_\tau \hat{Q}^\top \log(\pi_{old}) \right) \right)^2 \right\}, \end{aligned}$$

$$\text{RKL}_{\mathcal{B}_\tau \hat{Q}}^{\pi_{old}} - \bar{\epsilon} + \mathcal{B}_\tau \hat{Q}^\top \log(\pi_{old}) \geq 0,$$

and $\bar{\epsilon} \leq \hat{\Delta}\text{RKL}_{\pi_{new}}^{\pi_{old}}(s)$, then $\Delta\text{RKL}_{\pi_{new}}^{\pi_{old}} \geq 0$ and $\eta_\tau(\pi_{new}) \geq \eta_\tau(\pi_{old})$.

Proof We will again use Theorem 22. To obtain improvement by Theorem 9, we need that $\text{RKL}_{\mathcal{B}_\tau \hat{Q}}^{\pi_{new}} \leq \text{RKL}_{\mathcal{B}_\tau \hat{Q}}^{\pi_{old}} - \bar{\epsilon}$. Using Theorem 22, we will require that

$$\frac{\|\hat{Q}\|_\infty}{\tau} \sqrt{2\text{FKL}_{\mathcal{B}_\tau \hat{Q}}^{\pi_{new}}} - \mathcal{B}_\tau \hat{Q}^\top \log(\pi_{old}) \leq \text{RKL}_{\mathcal{B}_\tau \hat{Q}}^{\pi_{old}} - \bar{\epsilon}. \quad (31)$$

With some algebraic manipulation and assuming that $\text{RKL}_{\mathcal{B}_\tau \hat{Q}}^{\pi_{old}} - \bar{\epsilon} + \mathcal{B}_\tau \hat{Q}^\top \log(\pi_{old}) \geq 0$, the above is equivalent to

$$\left(\frac{\tau}{\|\hat{Q}\|_\infty} \left(\text{RKL}_{\mathcal{B}_\tau \hat{Q}}^{\pi_{old}} - \bar{\epsilon} + \mathcal{B}_\tau \hat{Q}^\top \log(\pi_{old}) \right) \right)^2 \geq 2\text{FKL}_{\mathcal{B}_\tau \hat{Q}}^{\pi_{new}}.$$

Dividing by 2, adding $\text{FKL}_{\mathcal{B}_\tau \hat{Q}}^{\pi_{old}}$, and rearranging yields that the assumption is equivalent to the following.

$$\hat{\Delta}\text{FKL}_{\pi_{new}}^{\pi_{old}} \geq \text{FKL}_{\mathcal{B}_\tau \hat{Q}}^{\pi_{old}} - \frac{1}{2} \left(\frac{\tau}{\|\hat{Q}\|_\infty} \left(\text{RKL}_{\mathcal{B}_\tau \hat{Q}}^{\pi_{old}} + \mathcal{B}_\tau \hat{Q}^\top \log(\pi_{old}) \right) \right)^2. \quad (32)$$

■

Proposition 32 (Approximate Average FKL Reduction). *Let \hat{Q} be an action-value estimate of $Q_\tau^{\pi_{old}}$, let $\epsilon := \hat{Q} - Q$ be the approximation error, and let $\bar{\epsilon} := \mathbb{E}_{d^{\pi_{new}}}[\mathbb{E}_{\pi_{new}}[\epsilon(S, A)]] - \mathbb{E}_{\pi_{old}}[\epsilon(S, A)]$.*

For two policies $\boldsymbol{\pi}_{old}, \boldsymbol{\pi}_{new} \in \mathbb{R}^{|\mathcal{A}|}$ in the bandit setting, if

$$\begin{aligned} & \hat{\mathbb{E}}_{d^{\pi_{new}}} [\Delta \text{FKL}_{\boldsymbol{\pi}_{new}}^{\boldsymbol{\pi}_{old}}] \\ & \geq \mathbb{E}_{d^{\pi_{new}}} [\text{FKL}_{\mathcal{B}_\tau \hat{Q}}^{\boldsymbol{\pi}_{old}}] - \frac{1}{2} \left(\frac{\tau}{\|\hat{Q}\|_\infty} \left(\mathbb{E}_{d^{\pi_{new}}} [\text{RKL}_{\mathcal{B}_\tau \hat{Q}}^{\boldsymbol{\pi}_{old}} + \mathcal{B}_\tau \hat{Q}^\top \log(\boldsymbol{\pi}_{old})] \right) \right)^2, \\ & \mathbb{E}_{d^{\pi_{new}}} [\text{RKL}_{\mathcal{B}_\tau \hat{Q}}^{\boldsymbol{\pi}_{old}} + \mathcal{B}_\tau \hat{Q}^\top \log(\boldsymbol{\pi}_{old})] \geq \bar{\epsilon}, \end{aligned}$$

and $\bar{\epsilon} \leq \mathbb{E}_{d^{\pi_{new}}} [\hat{\Delta} \text{RKL}_{\boldsymbol{\pi}_{new}}^{\boldsymbol{\pi}_{old}}(s)]$, then $\eta_\tau(\boldsymbol{\pi}_{new}) \geq \eta_\tau(\boldsymbol{\pi}_{old})$.

Proof The proof closely follows that of Theorem 13. Applying expectations to both sides of Equation (31), we obtain

$$\begin{aligned} & \mathbb{E}_{d^{\pi_{new}}} \left[\frac{\|\hat{Q}\|_\infty}{\tau} \sqrt{2\text{FKL}_{\mathcal{B}_\tau \hat{Q}}^{\boldsymbol{\pi}_{new}}(S)} - \mathcal{B}_\tau \hat{Q}(\cdot | S)^\top \log(\boldsymbol{\pi}_{old}(\cdot | S)) \right] \\ & \leq \mathbb{E}_{d^{\pi_{new}}} [\text{RKL}_{\mathcal{B}_\tau \hat{Q}}^{\boldsymbol{\pi}_{old}}(s)] - \bar{\epsilon}. \end{aligned} \quad (33)$$

As per the discussion in Theorem 28, satisfying this assumption will result in improvement. We hence derive an assumption on the FKL that implies Equation (36).

From rearranging to have the FKL term on one side, the following condition is equivalent to Equation (36).

$$\begin{aligned} & \frac{\tau}{\|\hat{Q}\|_\infty} \left(\mathbb{E}_{d^{\pi_{new}}} \left[\text{RKL}_{\mathcal{B}_\tau \hat{Q}}^{\boldsymbol{\pi}_{old}}(S) - \bar{\epsilon} + \mathbb{E}_{\mathcal{B}_\tau \hat{Q}} [\log(\boldsymbol{\pi}_{old}(\cdot | S))] \right] \right) \\ & \geq \mathbb{E}_{d^{\pi_{new}}} \left[\sqrt{2\text{FKL}_{\mathcal{B}_\tau \hat{Q}}^{\boldsymbol{\pi}_{new}}(S)} \right]. \end{aligned}$$

We now square both sides. The following is also equivalent to Equation (36) by assumption that $\mathbb{E}_{d^{\pi_{new}}} [\text{RKL}_{\mathcal{B}_\tau \hat{Q}}^{\boldsymbol{\pi}_{old}} + \mathcal{B}_\tau \hat{Q}^\top \log(\boldsymbol{\pi}_{old})] \geq \bar{\epsilon}$.

$$\begin{aligned} & \left(\frac{\tau}{\|\hat{Q}\|_\infty} \left(\mathbb{E}_{d^{\pi_{new}}} \left[\text{RKL}_{\mathcal{B}_\tau \hat{Q}}^{\boldsymbol{\pi}_{old}}(S) - \bar{\epsilon} + \mathbb{E}_{\mathcal{B}_\tau \hat{Q}} [\log(\boldsymbol{\pi}_{old}(\cdot | S))] \right] \right) \right)^2 \\ & \geq \left(\mathbb{E}_{d^{\pi_{new}}} \left[\sqrt{2\text{FKL}_{\mathcal{B}_\tau \hat{Q}}^{\boldsymbol{\pi}_{new}}(S)} \right] \right)^2. \end{aligned}$$

Jensen's inequality applied to the RHS shows that the following *implies* Equation (36).

$$\begin{aligned} & \left(\frac{\tau}{\|\hat{Q}\|_\infty} \left(\mathbb{E}_{d^{\pi_{new}}} \left[\text{RKL}_{\mathcal{B}_\tau \hat{Q}}^{\boldsymbol{\pi}_{old}}(S) - \bar{\epsilon} + \mathbb{E}_{\mathcal{B}_\tau \hat{Q}} [\log(\boldsymbol{\pi}_{old}(\cdot | S))] \right] \right) \right)^2 \\ & \geq \mathbb{E}_{d^{\pi_{new}}} \left[2\text{FKL}_{\mathcal{B}_\tau \hat{Q}}^{\boldsymbol{\pi}_{new}}(S) \right]. \end{aligned}$$

Some straightforward rearrangement yields the assumption in the statement of the Proposition.

$$\begin{aligned}
 & -\frac{1}{2} \left(\frac{\tau}{\|\hat{Q}\|_\infty} \left(\mathbb{E}_{d^{\pi_{\text{new}}}} \left[\text{RKL}_{\mathcal{B}_\tau \hat{Q}}^{\pi_{\text{old}}}(S) - \bar{\epsilon} + \mathbb{E}_{\mathcal{B}_\tau \hat{Q}}[\log(\pi_{\text{old}}(\cdot|S))] \right] \right) \right)^2 \\
 & \quad + \mathbb{E}_{d^{\pi_{\text{new}}}} \left[\text{FKL}_{\mathcal{B}_\tau \hat{Q}}^{\pi_{\text{old}}}(S) \right] \\
 & \leq \mathbb{E}_{d^{\pi_{\text{new}}}} \left[\Delta \text{FKL}_{\pi_{\text{new}}}^{\pi_{\text{old}}}(S) \right].
 \end{aligned}$$

■

We now turn to proving the above results.

Proposition 33 (Sufficient Approximate FKL Reduction in Bandits). *Let \hat{Q} be an action-value estimate of $Q_\tau^{\pi_{\text{old}}}$, let $\epsilon := \hat{Q} - Q$ be the approximation error, and let $\bar{\epsilon} := \mathbb{E}_{\pi_{\text{new}}}[\epsilon(A)] - \mathbb{E}_{\pi_{\text{old}}}[\epsilon(A)]$. As well, define*

$$\hat{\Delta} \text{FKL}_{\pi_{\text{new}}}^{\pi_{\text{old}}} := \text{FKL}_{\mathcal{B}_\tau \hat{Q}}^{\pi_{\text{old}}} - \text{FKL}_{\mathcal{B}_\tau \hat{Q}}^{\pi_{\text{new}}}.$$

For two policies $\pi_{\text{old}}, \pi_{\text{new}} \in \mathbb{R}^{|A|}$ in the bandit setting, if

$$\begin{aligned}
 & \hat{\Delta} \text{FKL}_{\pi_{\text{new}}}^{\pi_{\text{old}}} \\
 & \geq \max \left\{ 0, \text{FKL}_{\mathcal{B}_\tau \hat{Q}}^{\pi_{\text{old}}} - \frac{1}{2} \left(\frac{\tau}{\|\hat{Q}\|_\infty} \left(\text{RKL}_{\mathcal{B}_\tau \hat{Q}}^{\pi_{\text{old}}} + \mathcal{B}_\tau \hat{Q}^\top \log(\pi_{\text{old}}) \right) \right)^2 \right\},
 \end{aligned}$$

$$\text{RKL}_{\mathcal{B}_\tau \hat{Q}}^{\pi_{\text{old}}} - \bar{\epsilon} + \mathcal{B}_\tau \hat{Q}^\top \log(\pi_{\text{old}}) \geq 0,$$

and $\bar{\epsilon} \leq \hat{\Delta} \text{RKL}_{\pi_{\text{new}}}^{\pi_{\text{old}}}(s)$, then $\Delta \text{RKL}_{\pi_{\text{new}}}^{\pi_{\text{old}}} \geq 0$ and $\eta_\tau(\pi_{\text{new}}) \geq \eta_\tau(\pi_{\text{old}})$.

Proof We will again use Theorem 22. To obtain improvement by Theorem 9, we need that $\text{RKL}_{\mathcal{B}_\tau \hat{Q}}^{\pi_{\text{new}}} \leq \text{RKL}_{\mathcal{B}_\tau \hat{Q}}^{\pi_{\text{old}}} - \bar{\epsilon}$. Using Theorem 22, we will require that

$$\frac{\|\hat{Q}\|_\infty}{\tau} \sqrt{2 \text{FKL}_{\mathcal{B}_\tau \hat{Q}}^{\pi_{\text{new}}} - \mathcal{B}_\tau \hat{Q}^\top \log(\pi_{\text{old}})} \leq \text{RKL}_{\mathcal{B}_\tau \hat{Q}}^{\pi_{\text{old}}} - \bar{\epsilon}. \quad (34)$$

With some algebraic manipulation and assuming that $\text{RKL}_{\mathcal{B}_\tau \hat{Q}}^{\pi_{\text{old}}} - \bar{\epsilon} + \mathcal{B}_\tau \hat{Q}^\top \log(\pi_{\text{old}}) \geq 0$, the above is equivalent to

$$\left(\frac{\tau}{\|\hat{Q}\|_\infty} \left(\text{RKL}_{\mathcal{B}_\tau \hat{Q}}^{\pi_{\text{old}}} - \bar{\epsilon} + \mathcal{B}_\tau \hat{Q}^\top \log(\pi_{\text{old}}) \right) \right)^2 \geq 2 \text{FKL}_{\mathcal{B}_\tau \hat{Q}}^{\pi_{\text{new}}}.$$

Dividing by 2, adding $\text{FKL}_{\mathcal{B}_\tau \hat{Q}}^{\pi_{old}}$, and rearranging yields that the assumption is equivalent to the following.

$$\hat{\Delta} \text{FKL}_{\pi_{new}}^{\pi_{old}} \geq \text{FKL}_{\mathcal{B}_\tau \hat{Q}}^{\pi_{old}} - \frac{1}{2} \left(\frac{\tau}{\|\hat{Q}\|_\infty} \left(\text{RKL}_{\mathcal{B}_\tau \hat{Q}}^{\pi_{old}} + \mathcal{B}_\tau \hat{Q}^\top \log(\boldsymbol{\pi}_{old}) \right) \right)^2. \quad (35)$$

■

Proposition 34 (Approximate Average FKL Reduction). *Let \hat{Q} be an action-value estimate of $Q_\tau^{\pi_{old}}$, let $\epsilon := \hat{Q} - Q$ be the approximation error, and let $\bar{\epsilon} := \mathbb{E}_{d^{\pi_{new}}} [\mathbb{E}_{\pi_{new}} [\epsilon(S, A)]] - \mathbb{E}_{\pi_{old}} [\epsilon(S, A)]$.*

For two policies $\boldsymbol{\pi}_{old}, \boldsymbol{\pi}_{new} \in \mathbb{R}^{|A|}$ in the bandit setting, if

$$\begin{aligned} & \hat{\mathbb{E}}_{d^{\pi_{new}}} [\Delta \text{FKL}_{\pi_{new}}^{\pi_{old}}] \\ & \geq \mathbb{E}_{d^{\pi_{new}}} [\text{FKL}_{\mathcal{B}_\tau \hat{Q}}^{\pi_{old}}] - \frac{1}{2} \left(\frac{\tau}{\|\hat{Q}\|_\infty} \left(\mathbb{E}_{d^{\pi_{new}}} [\text{RKL}_{\mathcal{B}_\tau \hat{Q}}^{\pi_{old}} + \mathcal{B}_\tau \hat{Q}^\top \log(\boldsymbol{\pi}_{old})] \right) \right)^2, \end{aligned}$$

$$\mathbb{E}_{d^{\pi_{new}}} [\text{RKL}_{\mathcal{B}_\tau \hat{Q}}^{\pi_{old}} + \mathcal{B}_\tau \hat{Q}^\top \log(\boldsymbol{\pi}_{old})] \geq \bar{\epsilon},$$

and $\bar{\epsilon} \leq \mathbb{E}_{d^{\pi_{new}}} [\hat{\Delta} \text{RKL}_{\pi_{new}}^{\pi_{old}}(s)]$, then $\eta_\tau(\boldsymbol{\pi}_{new}) \geq \eta_\tau(\boldsymbol{\pi}_{old})$.

Proof The proof closely follows that of Theorem 13. Applying expectations to both sides of Equation (31), we obtain

$$\begin{aligned} & \mathbb{E}_{d^{\pi_{new}}} \left[\frac{\|\hat{Q}\|_\infty}{\tau} \sqrt{2 \text{FKL}_{\mathcal{B}_\tau \hat{Q}}^{\pi_{new}}(S)} - \mathcal{B}_\tau \hat{Q}(\cdot | S)^\top \log(\boldsymbol{\pi}_{old}(\cdot | S)) \right] \\ & \leq \mathbb{E}_{d^{\pi_{new}}} [\text{RKL}_{\mathcal{B}_\tau \hat{Q}}^{\pi_{old}}(S)] - \bar{\epsilon}. \end{aligned} \quad (36)$$

As per the discussion in Theorem 28, satisfying this assumption will result in improvement. We hence derive an assumption on the FKL that implies Equation (36).

From rearranging to have the FKL term on one side, the following condition is equivalent to Equation (36).

$$\begin{aligned} & \frac{\tau}{\|\hat{Q}\|_\infty} \left(\mathbb{E}_{d^{\pi_{new}}} \left[\text{RKL}_{\mathcal{B}_\tau \hat{Q}}^{\pi_{old}}(S) - \bar{\epsilon} + \mathbb{E}_{\mathcal{B}_\tau \hat{Q}} [\log(\boldsymbol{\pi}_{old}(\cdot | S))] \right] \right) \\ & \geq \mathbb{E}_{d^{\pi_{new}}} \left[\sqrt{2 \text{FKL}_{\mathcal{B}_\tau \hat{Q}}^{\pi_{new}}(S)} \right]. \end{aligned}$$

We now square both sides. The following is also equivalent to Equation (36) by assumption that $\mathbb{E}_{d^{\pi_{\text{new}}}} \left[\text{RKL}_{\mathcal{B}_\tau \hat{Q}}^{\pi_{\text{old}}} + \mathcal{B}_\tau \hat{Q}^\top \log(\pi_{\text{old}}) \right] \geq \bar{\epsilon}$.

$$\begin{aligned} & \left(\frac{\tau}{\|\hat{Q}\|_\infty} \left(\mathbb{E}_{d^{\pi_{\text{new}}}} \left[\text{RKL}_{\mathcal{B}_\tau \hat{Q}}^{\pi_{\text{old}}}(S) - \bar{\epsilon} + \mathbb{E}_{\mathcal{B}_\tau Q_\tau^{\pi_{\text{old}}}} [\log(\pi_{\text{old}}(\cdot|S))] \right] \right) \right)^2 \\ & \geq \left(\mathbb{E}_{d^{\pi_{\text{new}}}} \left[\sqrt{2\text{FKL}_{\mathcal{B}_\tau \hat{Q}}^{\pi_{\text{new}}}(S)} \right] \right)^2. \end{aligned}$$

Jensen’s inequality applied to the RHS shows that the following *implies* Equation (36).

$$\begin{aligned} & \left(\frac{\tau}{\|\hat{Q}\|_\infty} \left(\mathbb{E}_{d^{\pi_{\text{new}}}} \left[\text{RKL}_{\mathcal{B}_\tau \hat{Q}}^{\pi_{\text{old}}}(S) - \bar{\epsilon} + \mathbb{E}_{\mathcal{B}_\tau Q_\tau^{\pi_{\text{old}}}} [\log(\pi_{\text{old}}(\cdot|S))] \right] \right) \right)^2 \\ & \geq \mathbb{E}_{d^{\pi_{\text{new}}}} \left[2\text{FKL}_{\mathcal{B}_\tau \hat{Q}}^{\pi_{\text{new}}}(S) \right]. \end{aligned}$$

Some straightforward rearrangement yields the assumption in the statement of the Proposition.

$$\begin{aligned} & -\frac{1}{2} \left(\frac{\tau}{\|\hat{Q}\|_\infty} \left(\mathbb{E}_{d^{\pi_{\text{new}}}} \left[\text{RKL}_{\mathcal{B}_\tau \hat{Q}}^{\pi_{\text{old}}}(S) - \bar{\epsilon} + \mathbb{E}_{\mathcal{B}_\tau \hat{Q}} [\log(\pi_{\text{old}}(\cdot|S))] \right] \right) \right)^2 \\ & \quad + \mathbb{E}_{d^{\pi_{\text{new}}}} \left[\text{FKL}_{\mathcal{B}_\tau \hat{Q}}^{\pi_{\text{old}}}(S) \right] \\ & \leq \mathbb{E}_{d^{\pi_{\text{new}}}} \left[\Delta\text{FKL}_{\pi_{\text{new}}}^{\pi_{\text{old}}}(S) \right]. \end{aligned}$$

■

Appendix C. Complete Results for Performance in Benchmark Environments

In this section, we compare the KL methods on benchmark continuous and discrete-action environments, using non-linear function approximation. Here, we wish to understand (1) if our observations from the microworld experiments apply to more complicated environments, (2) if there are any new differences as a result of function approximation or increased environment complexity and (3) if any of the KL divergences is more robust to hyperparameter choices than the other.

C.1 Implementation Details

The agents use the API Algorithm with KL Greedification, in Algorithm 1. For the discrete action environments, we use the All-Actions updates, and for the continuous action environments we use the Sampled-Actions update, for both RKL and FKL, with 128 sampled actions. When evaluating the integral of the gradient for the RKL, we tested using the log-likelihood trick as well as the reparametrization trick. Since the last outperformed the

first, we report results using reparametrization. All agents use experience replay with a buffer size of 10^6 and use batch sizes of 32.

Hyperparameter sweeps are performed separately for each domain. We use RMSprop for both the actor and critic. In the continuous action-setting, we sweep over the actor learning rates $\{10^{-5}, 10^{-4}, 10^{-3}, 10^{-2}\}$ and critic learning rates $\{10^{-5}, 10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}\}$. In the discrete-action setting we have a shared learning rate because of a shared architecture and the sweep is done over the learning rates $\{10^{-5}, 10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}\}$. We sweep temperatures in $\{10^{-3}, 5 \times 10^{-3}, 10^{-2}, 5 \times 10^{-2}, 10^{-1}, 5 \times 10^{-1}, 1\}$ for the soft action-value methods and additionally include runs with the hard action-value methods. The temperature in $\mathcal{B}_\tau Q$ and the temperature in the soft action-value function are set to be the same value. For example, if $\tau = 0.01$, then we learn a soft action-value function with $\tau = 0.01$ and use a KL target distribution proportional to $\exp(Q(s, a)\tau^{-1})$.

On our continuous-action domains, all policy and value function networks are implemented as two-layer neural networks of size 128, with ReLU activations. On our discrete-action domains, we employ the following architectures. In the OpenAI Gym environments, the architecture is a two-layer neural network of size 128 with ReLU activations, with the policy and value functions as separate heads off of the main two-layer body. In MinAtar, the architecture is a convolutional network into one fully-connected layer for each of the policy, action value function, and state value function. The convolutional layer has 16 3x3 convolutions with stride 1, the same as in Young and Tian (2019). The size of the fully-connected layer is 128, with ReLU activations used between layers.

C.2 Performance

For continuous actions, we experiment on Pendulum, Reacher, Swimmer and HalfCheetah (Todorov et al., 2012). For discrete actions, we experiment on OpenAI Gym environments (Brockman et al., 2016) and MinAtar environments (Young and Tian, 2019). In this section, we plot only a summary of the performance, detailed plots showing how it varies throughout training can be found in Appendix D.3. Temperatures [1.0, 0.5, 0.1] were grouped together and labeled as “High RKL/FK”, whereas [0.05, 0.01, 0.005, 0.001, 0.0]¹¹ were grouped together and named “Low RKL/FKL”. For each temperature, 30 seeds were run per hyperparameter setting and the best performing 20% settings were selected, which were then grouped together based on the temperature as “High/Low”. Figures 10 and 11 report the average of the last half of the area under the curve for each group, as well as standard errors between all runs in the group. Returns are normalized between 0 and 1, with 0 corresponding to the lower limit of the returns from the curves in Appendix D.3 and 1 to the highest. There was no striking pattern regarding which temperature is best overall, the choice seems to be highly environment dependent. Furthermore, FKL and RKL seem to perform comparably overall, with no clear dominance of one over the other, although the FKL performed slightly better in the few cases they were different.

11. Zero was excluded for FKL with continuous actions, see Appendix D.3

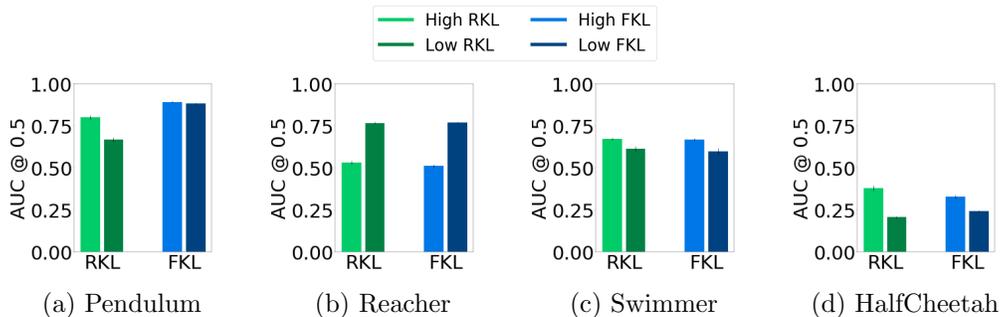


Figure 10: **Average return on the continuous-action environments.** The reported performance is the average return over the last half of learning, normalized between 0 and 1 and averaged over 30 runs with standard errors shown.

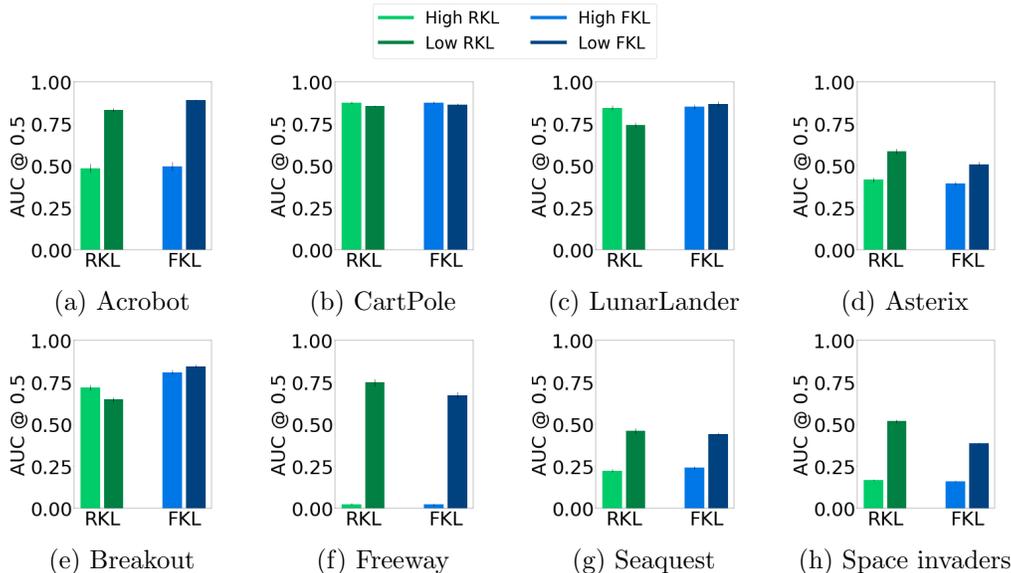


Figure 11: **Average return on the discrete-action environments.** The reported performance is the average return over the last half of learning, normalized between 0 and 1 and averaged over 30 runs with standard errors shown.

C.3 Hyperparameter sensitivity

We wrap up the experiments on benchmark problems by investigating the sensitivity of each divergence to hyperparameters. We focus on studying the sensitivity to the ones that seem to influence performance the most: learning rate and temperature. Particularly, we vary learning rates for each different temperature. For the continuous environments, where the actor and critic were separate networks, we have both the actor learning rate and the critic learning rate, whereas for the discrete environments we only have one learning rate.

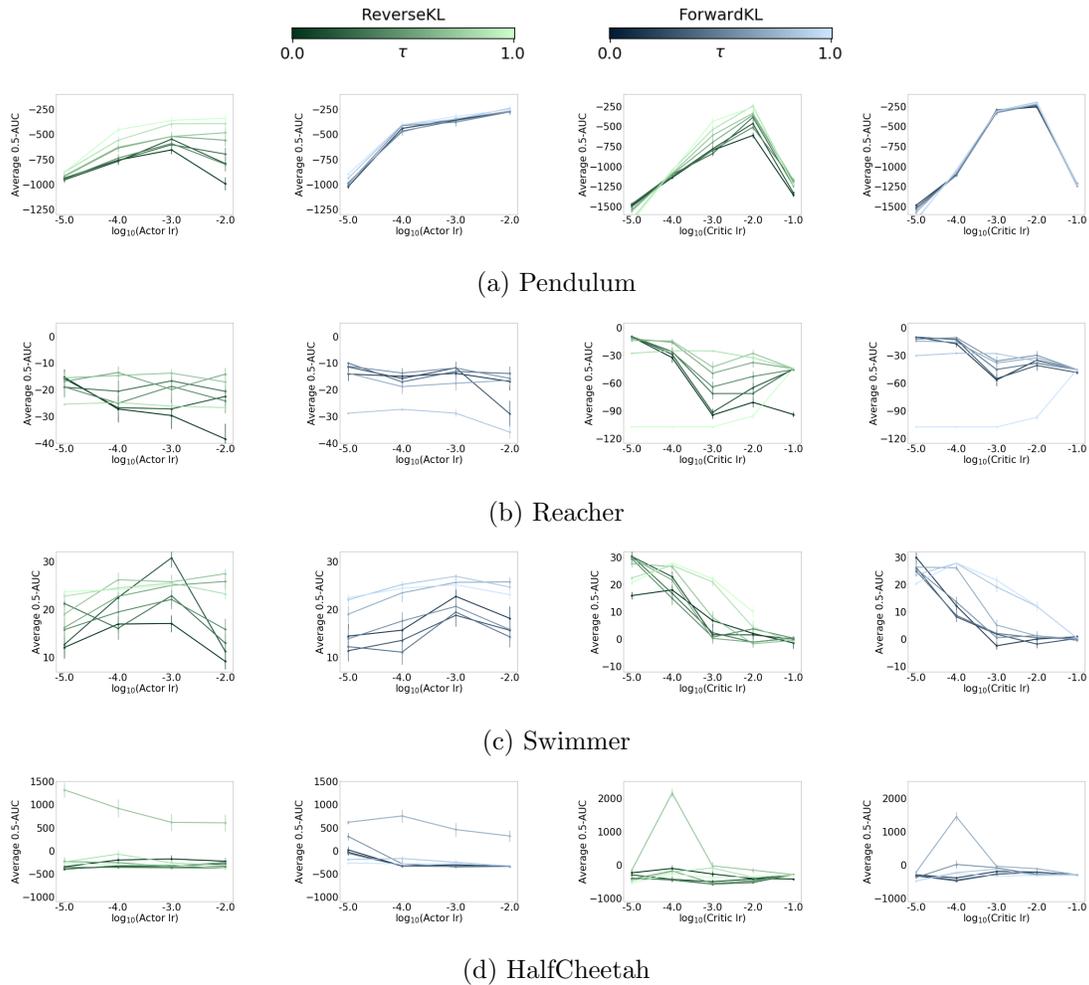


Figure 12: Sensitivity plots for continuous-action environments. The x axis represent the hyperparameter values and the y axis represent the averaged area under the last half of the learning curve. Each point corresponds to the average of the top 20% performing settings that had the corresponding learning rate and temperature

From looking at Figures 12 and 13, we see that both methods are sensitive to the hyperparameters, with the best parameters being highly environment dependent. For a given temperature and environment, RKL and FKL have very similar behavior: if the performance goes up for a certain learning rate for RKL, it also goes up for FKL and the same applies to decreases in performance, with very few exceptions and, even in those cases, the overall tendency is still the same for the two divergences. On Pendulum, represented in Figure 12a, for example, the worst learning rates perform better on FKL than in RKL, but the overall tendency is still for performance to go up as actor learning rates increase to 0.01 and as critic learning rates increase to 0.01, followed by a decrease when critic learning rate further increases to 0.1. The main takeaway is that, for a given choice of environment

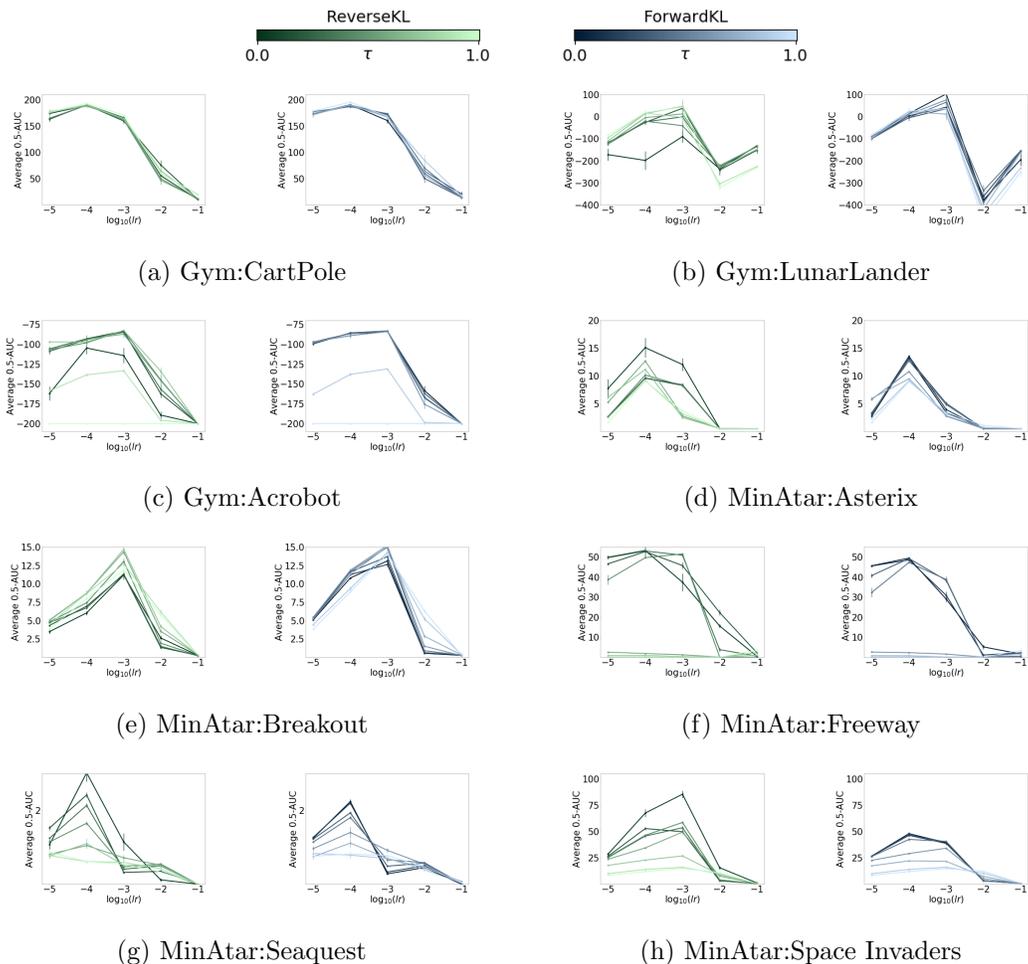


Figure 13: Sensitivity plots for discrete-action environments. Plot settings are identical to Figure 12

and temperature, learning rate sensitivity is not significantly influenced by the choice of divergence.

Appendix D. Additional Experimental Results

D.1 Exploration in a Discrete Maze

For these experiments, we use Gym-maze¹², choosing their fixed 10×10 maze to plot the changes in state-visitation distribution throughout training. This way, some of the differences we may be able to inspect are: if any one of the divergences becomes deterministic quicker than the other; if any of the divergences get stuck in local optima while the other succeeds in finding the optimal policy; how spread-out the state-visitation distribution is in the

12. <https://github.com/MattChanTK/Gym-maze>

early phases of training, when the agent is more likely to be exploring; if there are different priorities between regions of the maze when the agents are exploring.

For the Gym-maze, the dynamics $\Pr(s_{t+1} | s_t, a_t)$ are deterministic and the actions are the four directions. The agent remains in place if it tries to move to a position where there is a wall. The reward is -0.1 divided by the total number of cells if s_{t+1} is not the goal state and is 1.0 for the goal state. There is a timeout if the agent does not reach the goal after 10,000 steps. The agent is given a tabular representation: a one-hot encoding of the (x,y) position. This means that the agent has to create some sort of mental map of the maze from the positions it visited, causing the environment to be a harder exploratory problem than it might seem at first glance. We use the same agent that will be used in Section 6.3, based on Algorithm 1, which was introduced in Section 4.

On each iteration, one gradient descent step is performed to update the policy, for a given value function.

D.1.1 EXPLORATION WITH TRUE VALUES

We want to understand differences in exploration when minimizing the KL divergences with respect to the estimated value functions. Even so, it is important to also study how the state-visitation distribution changes when using accurate value functions, which we will henceforth refer to as true values, noting that they are still only approximations of the actual values that are calculated more precisely. In this setting, the agent does not need to explore or even interact with the environment directly; the dynamics are given. These studies make clear not only how entropy affects the convergence of policies but also make it easier to disentangle which behaviors and results are due to the use of estimated values and which happen even with true values.

By hand annotating the dynamics, we can use dynamic programming to compute more accurate estimates of the value function, as opposed to the common approach of using gradient descent methods based on the Bellman equation and least squares. The gradient for the greedification step is computed over all states, as opposed to using a buffer, and the learning rate is set to 0.1 with RMSprop. The total number of iterations is 100.

The stopping condition for dynamic programming is when the relative difference between successive Q's is less than 0.01% for 10 consecutive iterations. The policy update, for either FKL or RKL, requires both Q and V. We compute V directly from Q, by summing over all four actions weighted by the current policy. The most representative timesteps are illustrated in Figure 14.

D.1.2 EXPLORATION WITH ESTIMATED VALUES

In this section, we use the more practical approach of updating value function estimates via expected semi-gradient updates. The optimizer used was RMSprop with learning rate 0.001 and the total number of iterations is 20000. We use a mini-batch of 32 states sampled from a buffer. The buffer size is 10000.

Results are illustrated in Figure 15, where we show only the subset of timesteps most representative of changes and only the plots corresponding to $\tau = 0$. The start state is in the top left position and the goal state is in the bottom right. To obtain these images, 30 seeds were used for each temperature-divergence combination and trained for the full

number of iterations. The image at timestep t corresponds to 100 trajectories generated from each of the 30 policies at that timestep. Particularly, we take the visitation counts of each of these 30×100 trajectories, normalize them and average them, producing an image representative of the overall exploratory behavior of that divergence-temperature combination. The figure shows that, given a certain temperature, both RKL and FKL have state-visitation distributions that evolve very similarly, as can be seen by comparing the pairwise green and blue images for each timestep.

Figure 16 makes it clear that, for this setting, entropy regularization seems to do more harm than good: the agents converge to policies that tend to go to the correct trajectory, but waste time exploring when they already have all information they need. This is one example where the maximization of entropy conflicts with the true objective, ideally, one would prefer to have a deterministic policy for this environment at the end of training.

D.2 Exploration in a Continuous Maze

This section gives a more in-depth view of the experiments from Section 6.2.2. Figure 17 plots the cumulative number of times both the misleading and the correct exits are reached throughout training using 30 seeds and 2M steps, instead of 500k. For $\tau = 1000$, RKL and FKL have very similar curves; for $\tau = 100$ and $\tau = 10$ the RKL visits the misleading exit more and the correct exit less. For the remaining temperatures, FKL visits both exits more than the RKL. For lower temperatures, the misleading exit is visited orders of magnitude more than the regular one. These all corroborate with the conclusion that the FKL is more exploratory in this setting.

D.3 Performance

Implementation details are the same as in Section 6.3. We perform 30 runs for all hyperparameter settings and plot the mean return averaged over the past 20 episodes. Shaded areas represent standard errors.

D.3.1 CONTINUOUS-ACTIONS RESULTS

We compare agents on Pendulum (Brockman et al., 2016), Reacher, Swimmer and HalfCheetah (Todorov et al., 2012), with results shown in Figure 18. We exclude Hard FKL in our comparison since it requires access to $\max_a Q(s, a)$, which is difficult to obtain with continuous actions. The leftmost plot shows all temperatures from RKL, the middle plot shows all temperatures for FKL and the rightmost plot averages all high temperatures and all low temperatures for each divergence. Temperatures [1.0, 0.5, 0.1] were considered high and [0.05, 0.01, 0.005, 0.001, 0.0] were considered low.

Except for Pendulum, which is the simplest environment and where FKL seems to perform more consistently across temperatures, the overall behavior of both divergences is very similar, with performance being much more dependent on the choice of temperature than on the choice of divergence. On Reacher, Swimmer and Pendulum, FKL and RKL with high temperatures have the worst performance, but on HalfCheetah this pattern is reversed, meaning the benefits of entropy regularization are environment dependent.

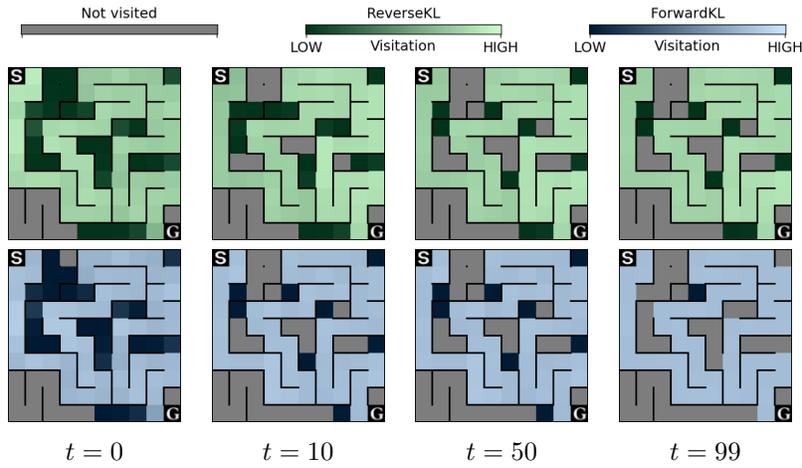
It is difficult to comment on the importance of the policy parameterization for these experiments relative to our microworld experiments. Any influence from the Gaussian policy

parameterization is conflated with function approximation. Moreover, as we will see below, no stark pattern seems to divide continuous and discrete action settings, as one did in our microworld experiments.

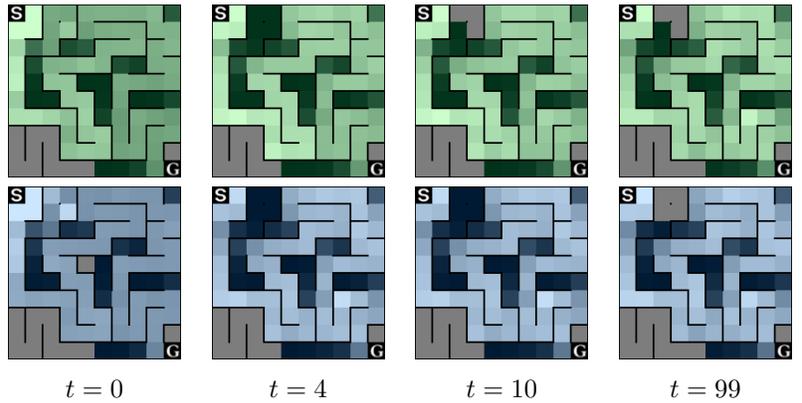
D.3.2 DISCRETE-ACTIONS RESULTS

We report results for environments from the OpenAI Gym (Brockman et al., 2016) and MinAtar (Young and Tian, 2019). Analogously to the continuous action setting, the OpenAI Gym results reported in Figure 19 show that both FKL and RKL behave similarly for any given choice of temperature, with the left and middle plots being similar to one another. The main difference is that, for some of the non-optimal temperatures for each problem, FKL seemed to learn faster than RKL, this becomes specially clear on Acrobot. The higher temperatures performed better on CartPole and LunarLander, but worse on Acrobot, confirming that the influence of this hyperparameter is highly environment dependent.

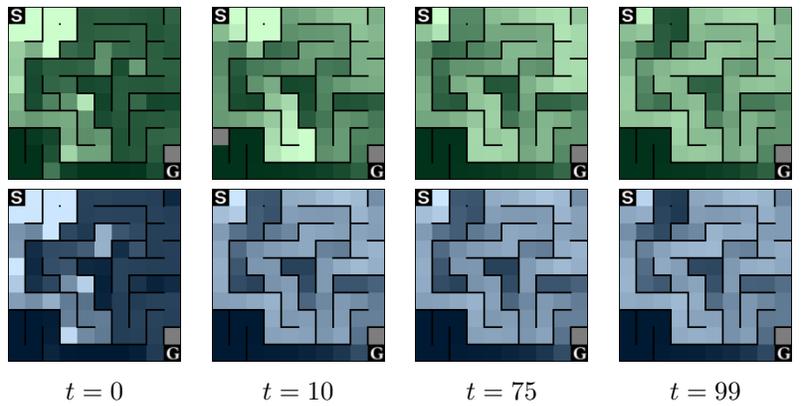
Finally, for the MinAtar results represented in Figure 20, there is once more no consistent dominance of either KL over the other: the plots of both FKL and RKL are again highly similar to each other. The slight superiority of FKL for non-optimal temperatures is present only on Breakout and Seaquest. On Asterix, Freeway, Seaquest and Space Invaders highest temperatures performed the worse, but they performed the best on Breakout, showing a pattern opposite of the one seen in the continuous environments of Figure 18 and confirming that the optimal temperature is going to vary according to the environment.



(a) $\tau = 0.0$



(b) $\tau = 0.01$



(c) $\tau = 0.1$

Figure 14: Evolution of state-visitation distributions throughout training for each temperature with true values.

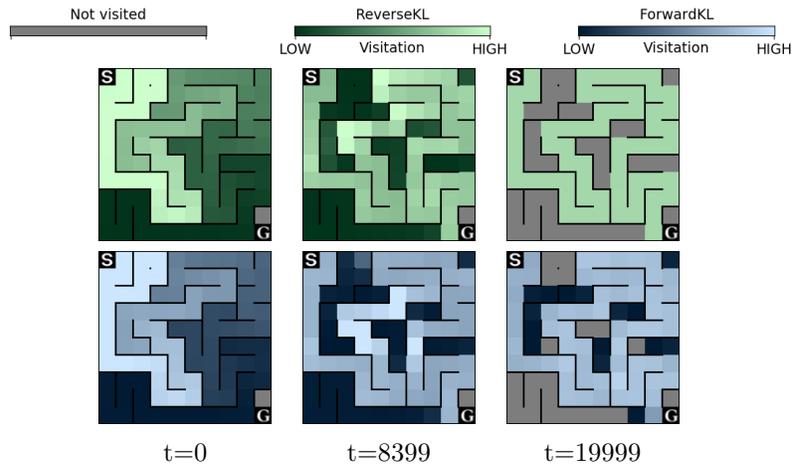


Figure 15: Evolution of state-visitation distributions throughout training for $\tau = 0$

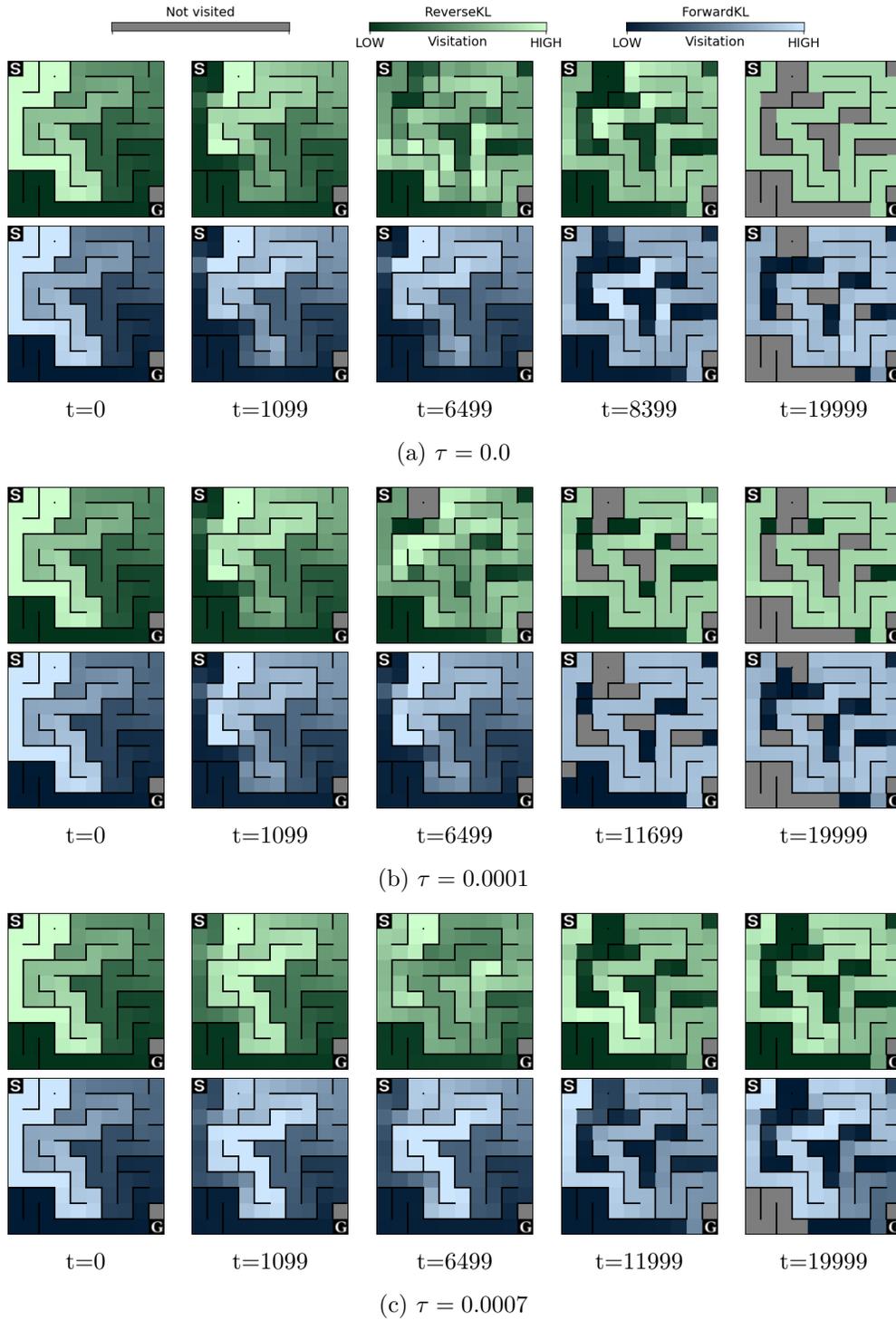


Figure 16: Evolution of state-visitation distributions throughout training for each temperature with estimated values.

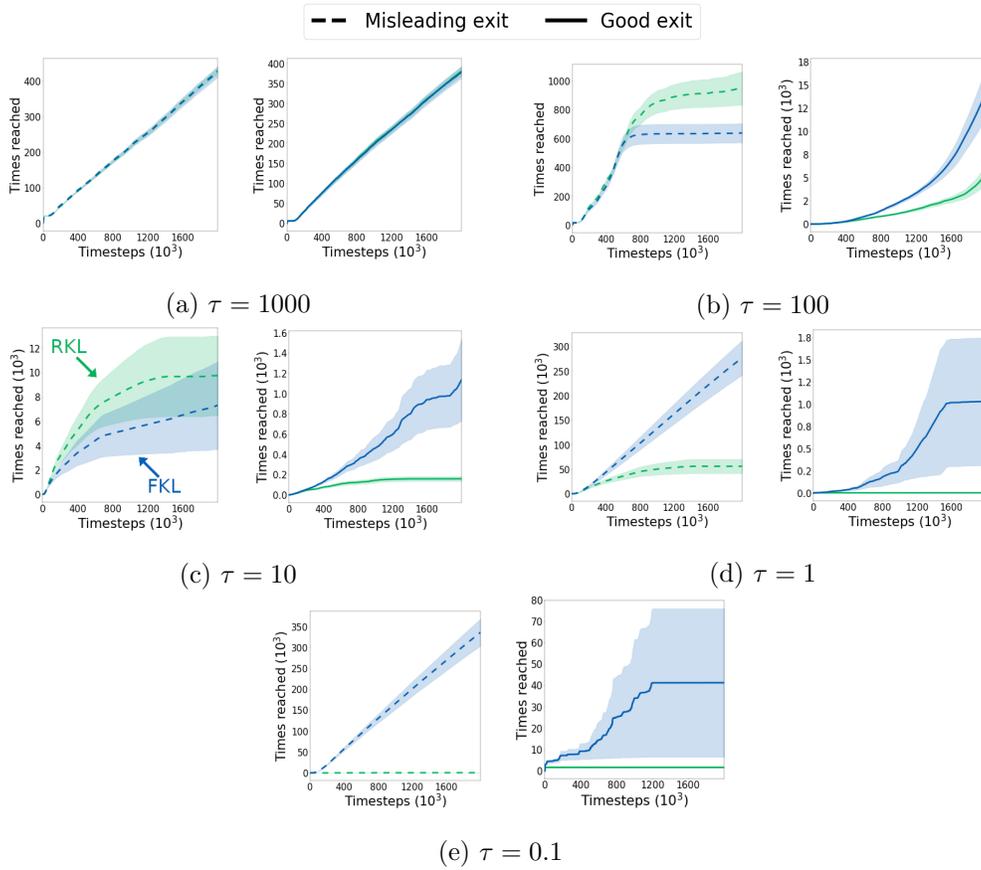


Figure 17: Cumulative number of times the each exit is reached throughout training for multiple temperatures plotted separately. RKL corresponds to green curves and FKL to blue curves.

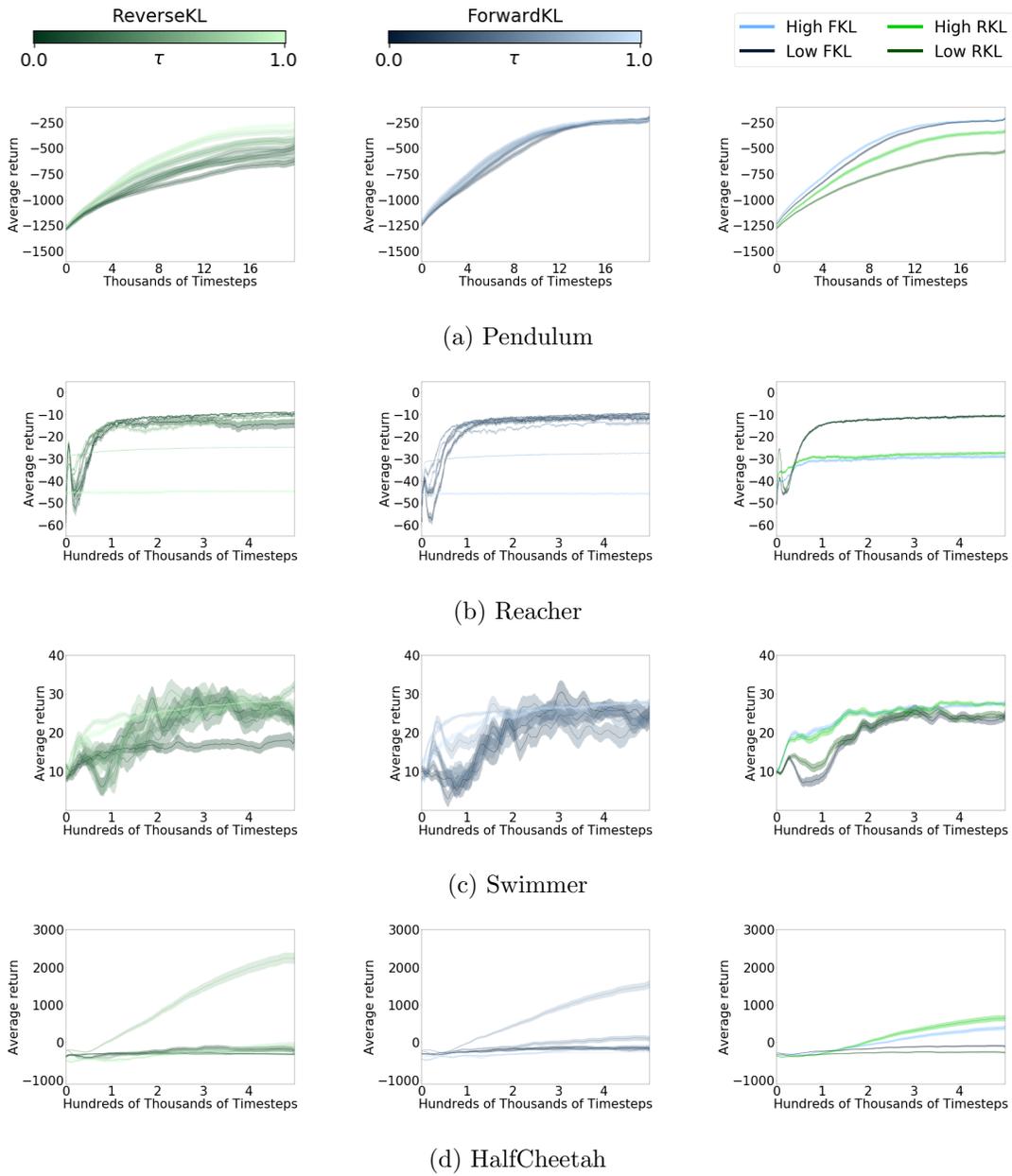


Figure 18: Continuous-action environments. Shown are the averaged top 20% performing hyperparameter settings for each algorithm, which are selected by largest area under the last half of the learning curve. In the left we have RKL, in the middle we have FKL and in the right we grouped the high and low temperatures for both FKL and RKL. Temperatures for the curves in the left and middle plots correspond to the respective colormaps and in the rightmost plot the curves correspond to the legend

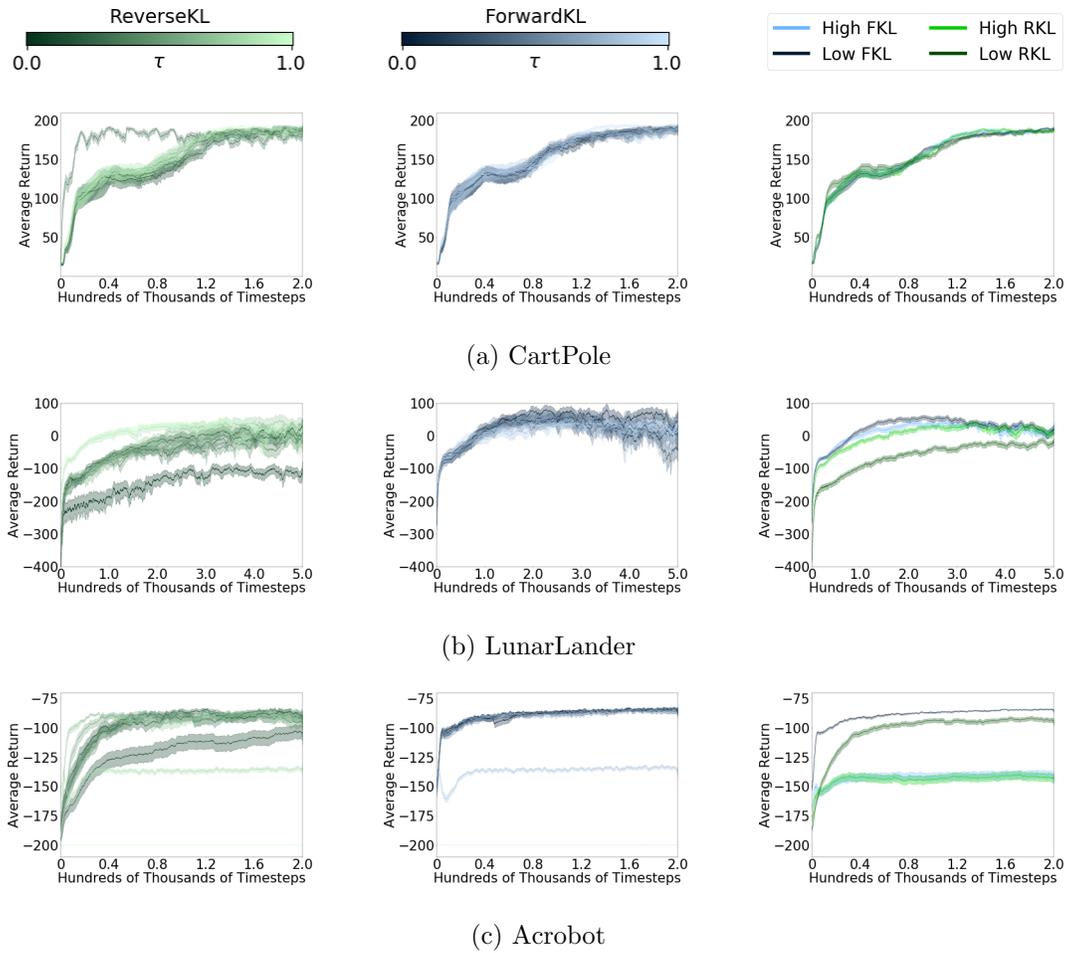


Figure 19: OpenAI Gym discrete-action environments. Plot settings are identical to Figure 18.

GREEDIFICATION OPERATORS FOR POLICY OPTIMIZATION

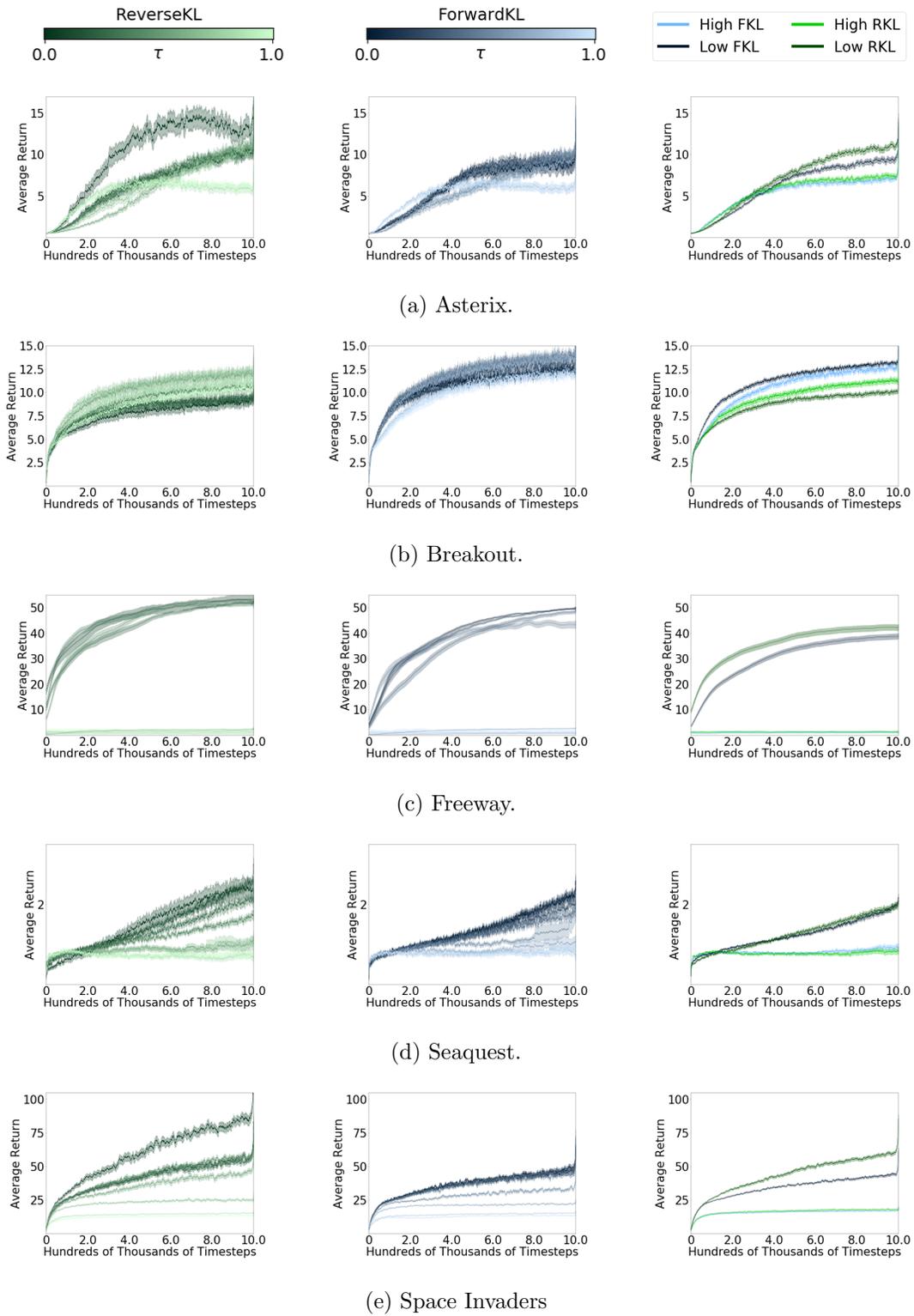


Figure 20: MinAtar discrete-action environments. Plot settings are identical to those in Figure 18.