SOFT Q-LEARNING WITH MUTUAL-INFORMATION REGULARIZATION

Anonymous authors
Paper under double-blind review

ABSTRACT

We propose a reinforcement learning (RL) algorithm that uses mutual-information regularization to optimize the prior action distribution for better performance and exploration. Entropy-based regularization has previously been shown to improve both exploration and robustness in challenging sequential decision-making tasks. It does so by encouraging policies to put probability mass on all actions. However, entropy regularization might be undesirable when actions have significantly different importance. In this paper, we propose a theoretically motivated framework that dynamically weights the importance of actions by using the mutual-information. In particular, we express the RL problem as an inference problem where the prior probability distribution over actions is subject to optimization. We show that the prior optimization introduces a mutual-information regularizer in the RL objective. This regularizer encourages the policy to be close to a non-uniform distribution that assigns higher probability mass to more important actions. We empirically demonstrate that our method significantly improves over entropy regularization methods, attaining state-of-the-art performance.

1 INTRODUCTION

Reinforcement Learning (RL) (Sutton & Barto, 1998) is a framework for solving sequential decision-making problems under uncertainty. Contemporary state-of-the-art RL methods often use an objective that includes an entropy regularization term (Haarnoja et al., 2018; 2017; Teh et al., 2017). Entropy regularized RL has been shown to capture multi-modal behaviour, as well as exhibiting superior exploration (Haarnoja et al., 2017). Additionally, the learned policies are more robust, as the entropy bonus accounts for future action stochasticity (Grau-Moya et al., 2016) and reduces value overestimation (Fox et al., 2016).

While encouraging high-entropy policies can provide several benefits, it is possible to devise examples where entropy regularization actually impedes exploration. Since high-entropy policies tend to spread the probability mass across all actions equally, they can perform poorly when the RL problem contains actions that are rarely useful. In this paper, we design a reinforcement learning algorithm that dynamically adjusts the importance of actions while learning. We motivate our algorithm by phrasing RL as an inference problem with an adaptive prior action distribution. Where previous work assumes a uniform prior distribution over the actions (Rawlik et al., 2012; Levine, 2018), we generalize the formulation by optimizing the prior. We show that this optimization process leads to an RL objective function with a regularizer based on the mutual-information between states and actions.

We develop a novel algorithm that uses mutual-information regularization to obtain an optimal action prior for better performance and exploration. This novel regularizer for RL encourages policies to be close to the marginal distribution over actions. This results in assigning higher probability to actions frequently used by the optimal policy, while actions that are used infrequently have lower probability under the prior. We demonstrate significant improvement on 19 Atari games over a Deep Q-Network (Mnih et al., 2015) (DQN) baseline without regularization and over Soft Q-Learning (Schulman et al., 2017) (SQL) that employs standard entropy regularization.

2 BACKGROUND

2.1 REINFORCEMENT LEARNING

We consider the standard Markov decision process (MDP) setting. Formally, an MDP is defined as the tuple $\langle \mathcal{S}, \mathcal{A}, P, R, \gamma \rangle$ where $\mathcal{S}$ is the state space, $\mathcal{A}$ the action space, and $P : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to [0, 1]$ denotes the state transition distribution. Upon taking action $a_t \in \mathcal{A}$ in state $s_t \in \mathcal{S}$, the agent transitions to s_{t+1} with probability $P(s_{t+1}|s_t, a_t)$. The reward function $\mathcal{R} : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ quantifies the agent's performance. The goal is to find a policy that maximizes the value function, i.e. $\pi^*(a|s) = \arg\max_\pi V^\pi(s)$, where $V^\pi(s) = \mathbb{E}\Big[\sum_{t=0}^{T} \gamma^t r(s_t, a_t)|s_0 = s\Big]$. Here γ is a discount factor ($0 < \gamma < 1$) that allows to account for the future in different ways.

The policy-dependent state transition probabilities are defined as $P_\pi(s'|s) := \sum_a P(s'|a, s)\pi(a|s)$ which can be written in matrix notation as $\boldsymbol{P}_\pi \in \mathbb{R}^{|\mathcal{S}|} \times \mathbb{R}^{|\mathcal{S}|}$ where the rows are indexed by s and the columns by s'. This allows to us conveniently define the agent's stationary distribution over states $\mu_\pi(s)$ and actions $\rho_\pi(a)$ as follows:

Definition 1 (Stationary distribution over states)**.** *The stationary distribution over states is defined in vector form as $\boldsymbol{\mu}_\pi^\top := \lim_{t\to\infty} \boldsymbol{\nu}_0^\top \boldsymbol{P}_\pi^t$ with $\boldsymbol{\nu}_0$ being an arbitrary vector of probabilities over states at time $t = 0$. The stationary distribution satisfies $\mu_\pi(s') = \sum_s P_\pi(s'|s)\mu_\pi(s)$ and therefore is a fixed point under the state transition probabilities $\boldsymbol{\mu}_\pi^\top = \boldsymbol{\mu}_\pi^\top \boldsymbol{P}_\pi$.*

Definition 2 (Stationary distribution over actions)**.** *Let $\mu_\pi(s)$ be the stationary distribution over states induced by the policy π. Then the stationary distribution over actions under the policy π is defined as $\rho_\pi(a) := \sum_{s\in\mathcal{S}} \mu_\pi(s)\pi(a|s)$.*

2.2 MAXIMUM ENTROPY REINFORCEMENT LEARNING

Maximum Entropy Reinforcement Learning augments the standard RL reward objective with an additional policy entropy term. The optimal value function under entropy regularization (Haarnoja et al., 2017) is defined as:

$$\mathcal{V}^*(s) = \max_\pi \mathbb{E}\Bigg[\sum_{t=0}^{T} \Bigg(r(s_t, a_t) - \frac{1}{\beta} \log \pi(a_t|s_t)\Bigg)\Bigg], \tag{1}$$

where $s_0 = s$, $\frac{1}{\beta}$ trades off between reward and entropy maximization, and the expectation operation is over state-action trajectories. Here we focus on the undiscounted case for ease of exposition. We will introduce the discounted case below. The optimal policy that solves (1) can be written in closed form as:

$$\pi^*(a|s) = \frac{e^{\beta \mathcal{Q}^*(s,a)}}{\sum_{a\in\mathcal{A}} e^{\beta \mathcal{Q}^*(s,a)}},$$

where $\mathcal{Q}^*(s, a) := r(s, a) + \sum_{s'\in\mathcal{S}} P(s'|s, a)\mathcal{V}^*(s')$. Note that the above represents a generalization of standard RL settings, where $\beta \to \infty$ corresponds to a standard RL valuation ($\lim_{\beta\to\infty} \mathcal{V}^*(s) = \max_\pi V^\pi(s)$), while for $\beta \to 0$ we recover the valuation under a random uniform policy. For intermediate values of β we can trade off between reward maximization and entropy maximization.

Interestingly, one can formulate the maximum entropy RL optimization as an inference problem (Levine, 2018) by specifying a prior distribution over trajectories that assumes a fixed uniform distribution over actions. Precisely this assumption is what encourages the policies to maximize entropy. As outlined in the introduction, encouraging policies to be close to a uniform distribution might be undesirable when some actions are simply non-useful or not frequently used. In the following section, we show that relaxing this assumption by allowing for prior optimization we obtain a novel variational inference formulation of the RL problem that constrains the policies' mutual-information between states and actions. We show that such policies must be close to the marginal distribution over actions which automatically assign high probability mass to overall useful actions and low probability to non-important or infrequently used actions.

3 MUTUAL-INFORMATION REINFORCEMENT LEARNING

The mutual information, core to our approach, is a basic quantity in information theory to measure the statistical dependence between two random variables. Machine learning applications that use the mutual information are numerous and include the information-bottleneck method (Tishby et al., 1999), rate-distortion theory (Cover & Thomas, 2006), clustering (Still & Bialek, 2004) and curiosity driven exploration (Still & Precup, 2012). Typically, the mutual information is defined as as $I(X;Y) := \sum_{x,y} p(x,y) \log \frac{p(x,y)}{p_Y(y)p_X(x)} = \sum_x p_X(x)\mathrm{KL}(p_{Y|X}(\cdot|x)||p_Y(\cdot))$ where $\mathrm{KL}(p(\cdot)||q(\cdot)) = \sum_x p(x) \log \frac{p(x)}{q(x)}$ is the Kullback-Leibler (KL) divergence, p_X, p_Y are the marginal distributions, and $p_{Y|X}$ the conditional distribution with respect to the joint $p(x,y)$. However, in this paper, we use an alternative formulation of the mutual information, that will be useful later on in order to phrase our optimization objective for RL.

Proposition 1 (Mutual Information)**.** *Let I_f be a functional, in particular:*

$$I_f(p_X, p_{Y|X}, q_Y) := \sum_x p_X(x)\mathrm{KL}(p_{Y|X}(\cdot|x)||q_Y(\cdot)),$$

where $p_X(x)$ is the distribution of the input, $p_{Y|X}(y|x)$ is the channel, and $q_Y(y)$ a variational distribution. Then, the mutual information can be recovered with

$$I[X,Y] = \min_{q_Y} I_f(p_X, p_{Y|X}, q_Y),$$

where the optimal variational distribution is $q_Y^\star(y) = \sum_x p_X(x)p_{Y|X}(y|x)$. See e.g. (Cover & Thomas, 2006, Lemma 10.8.1) for details.

As expected, the original definition of the mutual information is recovered since the optimal variational distribution q that minimizes the information functional is the marginal distribution p_Y.

In the following section, we propose an inference based formulation where the mutual information arises as a consequence of allowing for optimizing the action-prior distributions.

3.1 VARIATIONAL INFERENCE RL FORMULATION WITH OPTIMAL ACTION PRIORS

The RL problem can be expressed as an inference problem by introducing a binary random variable R that denotes whether the trajectory $\tau := (s_0, a_0, \dots s_T, a_T)$ is optimal ($R = 1$) or not ($R = 0$). The likelihood of an optimal trajectory can then be expressed as $p(R = 1|\tau) \propto \exp(\sum_{t=0}^T r(s_t, a_t))$ (Levine, 2018). We additionally introduce a scaling factor $\beta > 0$ into the exponential, i.e. $p(R = 1|\tau) \propto \exp(\beta \sum_{t=0}^T r(s_t, a_t))$. This will allow us to trade off reward and entropy maximization [1]. Next, we can define the posterior trajectory probability assuming optimality, i.e. $p(\tau|R = 1)$. Here we treat τ as a latent variable with prior probability $p(\tau)$, and we specify the log-evidence as $\log p(R = 1) = \log \int p(R = 1|\tau)p(\tau)d\tau$. We now introduce a variational distribution $q(\tau)$ to approximate the posterior $p(\tau|R = 1)$. This leads to an Evidence Lower BOund (ELBO) of the previous expression (scaled by $\frac{1}{\beta}$) [2]:

$$\begin{aligned}\frac{1}{\beta}\log p(R=1) &= \frac{1}{\beta}\log \int p(R=1|\tau)p(\tau)d\tau \\ &\geq \frac{1}{\beta}\mathbb{E}_{\tau\sim q(\tau)}\left[\log \frac{p(R=1|\tau)p(\tau)}{q(\tau)}\right] \end{aligned} \tag{2}$$

The generative model is written as $p(\tau) = p(s_0)\prod_{t=0}^{T-1} p(a_t)P(s_{t+1}|s_t, a_t)$ and the variational distribution as $q(\tau) = p(s_0)\prod_{t=0}^{T-1} \pi(a_t|s_t)P(s_{t+1}|s_t, a_t)$. The RL problem can now be stated as a maximization of the ELBO w.r.t π. The maximum entropy RL objective is recovered when assuming a fixed uniform prior distribution over actions, i.e. $p(a_t) = \frac{1}{|\mathcal{A}|}$ for all t.

[1] Other authors absorb this scaling factor into the reward, but we keep it as an explicit hyperparameter.

[2] Obtained by multiplying and dividing by $q(\tau)$ inside the integral and applying Jensen's inequality $f(\mathbb{E}_{x\sim q}[x]) \geq \mathbb{E}_{x\sim q}[f(x)]$ (assuming concave f).

We obtain a novel variational RL formulation by introducing an adaptive prior over actions ρ. Contrary to maximum entropy RL, where the priors of the generative model are fixed and uniform, here the prior over actions is subject to optimization. Therefore, the generative model is written as $p(\tau) = p(s_0)\prod_{t=0}^{T-1}\rho(a_t)P(s_{t+1}|s_t,a_t)$ whereas the variational distribution $q(\tau)$ remains unchanged. Starting from Equation (2) and substituting $p(\tau)$ and $q(\tau)$ we obtain the following ELBO:

$$\max_{\pi,\rho}\mathbb{E}_q\left[\sum_{t=0}^{T}\left(r(s_t,a_t)-\frac{1}{\beta}\log\frac{\pi(a_t|s_t)}{\rho(a_t)}\right)\right].$$

Since we are interested in infinite horizon problems, we introduce a discount factor and take the limit $\lim_{T\to\infty}$ (see (Haarnoja et al., 2017)). This leads to the optimization objective that we use in our experiments:

$$\max_{\pi,\rho}\mathbb{E}_q\left[\sum_{t=0}^{\infty}\gamma^t\left(r(s_t,a_t)-\frac{1}{\beta}\log\frac{\pi(a_t|s_t)}{\rho(a_t)}\right)\right], \tag{3}$$

where $0<\gamma<1$ is the discount factor. In the following, we show that the the solution for the prior and the policy can be expressed in closed form giving rise to a novel RL regularization scheme.

3.2 Recursion, Optimal Policies and Priors

Crucial for the construction of a practical algorithm are closed form expressions for the optimal policy and prior. More concretely, the optimal policy takes the form of a Boltzmann distribution weighted by the prior ρ. When fixing the policy, the optimal prior is the marginal distribution over actions under the discounted stationary distribution over states. This finding is important to obtain a practical algorithm for learning an optimal prior.

Optimal policy for a fixed prior ρ: We start by defining the value function with the information cost as $\mathcal{V}_{\pi,\rho}(s) := \mathbb{E}\left[\sum_{t=0}^{\infty}\gamma^t\left(r(s_t,a_t)-\frac{1}{\beta}\log\frac{\pi(a_t|s_t)}{\rho(a_t)}\right)|s_0=s\right]$ where one can show that $\mathcal{V}_{\pi,\rho}$ satisfies a recursion similar to the Bellman equations where:

$$\mathcal{V}_{\pi,\rho}(s)=\mathbb{E}_\pi\left[r(s,a)-\frac{1}{\beta}\log\frac{\pi(a|s)}{\rho(a)}+\gamma\mathbb{E}_{s'}[\mathcal{V}_{\pi,\rho}(s')]\right]. \tag{4}$$

Additionally, when considering a fixed ρ, the problem of maximizing Equation (4) over the policy can be solved analytically by standard variational calculus (Rubin et al., 2012; Genewein et al., 2015). The optimal policy then is

$$\pi^*(a|s) := \frac{1}{Z}\rho(a)\exp(\beta Q_{\pi^*,\rho}(s,a)) \tag{5}$$

with $Z=\sum_a\rho(a)\exp(\beta Q_{\pi^*,\rho}(s,a))$ and the following definition of the soft Q-function as

$$Q_{\pi,\rho}(s,a) := r(s,a)+\gamma\mathbb{E}_{s'}[\mathcal{V}_{\pi,\rho}(s')]. \tag{6}$$

Being able to write the optimal policy in this way for a given Q-function is needed in order to estimate the optimal prior as we show next.

Optimal prior for a fixed policy: In order to solve for the optimal prior, we rewrite the problem in Equation (3) as

$$\max_{\pi,\rho}\sum_{t=0}^{\infty}\sum_s\gamma^t\nu_t(s)\sum_a\pi(a|s)\left(r(s_t,a_t)-\frac{1}{\beta}\log\frac{\pi(a|s)}{\rho(a)}\right),$$

where we have defined the marginal distribution over states at time t as

$$\nu_t(s_t) := \sum_{s_0,a_0,\ldots,s_{t-1},a_{t-1}} p(s_0)\prod_{i=0}^{t-1}\pi(a_i|s_i)T(s_{i+1}|s_i,a_i). \tag{7}$$

For a fixed π, we eliminate the max operator for π and all components that do not depend on ρ:

$$\max_\rho -\frac{1}{\beta}\sum_{t=0}^{\infty}\gamma^t\sum_s\nu_t(s)\mathrm{KL}(\pi(\cdot|s)||\rho(\cdot)).$$

Swapping the sums and letting $p(s) := \sum_{t=0}^{\infty} \gamma^t \nu_t(s)$ be the unnormalized discounted marginal distribution over states, we obtain $\max_\rho -\frac{1}{\beta} \sum_s p(s) \mathrm{KL}(\pi(\cdot|s)||\rho(\cdot))$. The solution to the latter, $\rho^\star(a) = \frac{\sum_s p(s)\pi(a|s)}{\sum_{a,s} p(s)\pi(a|s)}$, can easily obtained by adding the constraint that the action prior is a valid distribution (i.e., $\sum_a \rho(a) = 1$ and $\rho(a) > 0,\ \forall a$) and using the method of Lagrange multipliers and standard variational calculus.

With the form of the optimal prior for a fixed policy at hand, one can easily devise a stochastic approximation method (e.g. $\rho_{k+1}(a) = (1-\alpha)\rho_k(a) + \alpha\pi(a|s)$ with $\alpha \in [0,1]$ and $s \sim p(\cdot)$) to estimate the optimal ρ using the current estimate of the optimal policy from Equation (5).

Before proposing a practical algorithm (see Section 4), we note that up to this stage there is still no clear connection between the proposed approach and the mutual information. To provide further insight, it will be instructive to show next that, under some limit conditions, the optimization problem proposed by the ELBO can be mapped to an optimization problem with a mutual-information constraint between the stationary distributions over states and actions.

3.3 Equivalence between Optimizing Priors and Mutual Information

The goal of this section is to show that when $\gamma \to 1$, Equation (3) can be expressed as the following average-reward formulation (Puterman, 1994) with a constraint on the stationary mutual information

$$\max_\pi \mathbb{E}_{s\sim\mu_\pi}\left[\sum_a \pi(a|s) r(s,a)\right] \quad \text{s.t. } I_f(\mu_\pi, \pi, \rho_\pi) \leq C, \tag{8}$$

where μ_π and ρ_π are the stationary distributions induced by policy π over states and actions, respectively, and thus $I_f(\mu_\pi, \pi, \rho_\pi)$ is defined as the stationary mutual-information. Note that for a fixed stationary distribution over states, this problem coincides exactly with the well-known rate-distortion problem (Cover & Thomas, 2006).

We start by expressing (3) as a constrained problem

$$\max_\pi \mathbb{E}_q\left[\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t)\right] \text{ s.t. } \min_\rho \sum_{t=0}^{\infty} \gamma^t I_f(\nu_t, \pi, \rho) \leq K(\gamma), \tag{9}$$

where $K(\gamma) := \frac{C}{1-\gamma}$ (although we have set $K(\cdot)$ as a function of γ, it is without loss of generality since we can always obtain a desired $K(\cdot)$ by choosing an appropriate C for a given γ) and the marginal probability of state s_t at time t following the Markovian dynamics is written as in Equation (7).

A standard result in the MDP literature (Bertsekas, 1995) is that

$$\arg\max_\pi \lim_{\gamma\to\infty} (1-\gamma)\mathbb{E}_q\left[\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t)\right] \iff \arg\max_\pi \mathbb{E}_{s\sim\mu_\pi}\left[\sum_a \pi(a|s) r(s,a)\right]$$

which basically says that the optimal policy for the limit $\gamma \to 1$ of an infinite horizon problem is equivalent to the average reward formulation. Now it only remains to make explicit a similar equivalence on the constraint.

We rewrite the constraint by multiplying on both sides by $(1-\gamma)$ assuming $\gamma \in (0,1)$

$$\min_\rho \sum_{t=0}^{\infty} \gamma^t I_f(\nu_t, \pi, \rho) \leq K(\gamma) \iff \min_\rho (1-\gamma) \sum_{t=0}^{\infty} \gamma^t I_f(\nu_t, \pi, \rho) \leq C.$$

Taking the limit $\gamma \to 1$ in the last inequality and interchanging the limit and the min operators[3], we obtain the constraint $\min_\rho \lim_{\gamma\to\infty} (1-\gamma) \sum_{t=0}^{\infty} \gamma^t I_f(\nu_t, \pi, \rho) \leq C$. Then, we see the connection between the last inequality and the constraint on Equation (8) using the following Proposition 2.

[3]We can interchange the operators because the l.h.s. of the inequality is finite for any $0 < \gamma < 1$.

Algorithm 1 MIRL

Input: the learning rates α_ρ, α_Q and α_β, a Q-network $Q_\theta(s,a)$, a target network $Q_{\bar{\theta}}(s,a)$, a behavioural policy π_b, an initial prior ρ_0 and parameters θ_0 at $t=0$.
for $i = 1$ to N iterations **do**
  Get environment state s_i and apply action $a_i \sim \pi_b(\cdot|s_i)$
  Get r_i, s_{i+1} and store (s_i, a_i, r_i, s_{i+1}) in replay memory $\mathcal{M}$
  Update prior $\rho_{i+1}(\cdot) = \rho_i(\cdot)(1-\alpha_\rho) + \alpha_\rho \pi_i(\cdot|s_i)$
  if $i \mod$ update frequency **then**
    Update Q-function $\theta_{i+1} = \theta_i - \alpha_Q \nabla_\theta L(\theta, \rho_{i+1})$ (Equation (12))
    Update parameter $\beta_{i+1} = (1-\alpha_\beta)\beta_i + \alpha_\beta \left(\frac{1}{L(\theta_i, \rho_{i+1})}\right)$
  end if
end for

Proposition 2. *Let $\mu_\pi(s)$ and $\rho_\pi(a)$ be the stationary distribution over states and actions under policy π according to Definitions* (1) *and* (2)*. Then the stationary mutual information defined as $I_f(\mu_\pi, \pi, \rho_\pi) := \min_\rho I_f(\mu_\pi \pi, \rho)$ can also be written as*

$$I_f(\mu_\pi, \pi, \rho_\pi) = \min_\rho \lim_{\gamma \to 1} (1-\gamma) \sum_{t=0}^{\infty} \gamma^t I_f(\nu_t, \pi, \rho) \tag{10}$$

where $\nu_t(s)$ is defined as in (7)*.*

The proof can be found in the Appendix. Since in practice we use a discount factor $\gamma \lessapprox 1$, our original problem formulation in (3) can be seen as an approximation to the problem with stationary mutual-information constraints in (8).

The conclusion of this section is that we have established a clear link between the ELBO with optimizable priors and the average reward formulation with stationary mutual-information constraints. Having the basic theory in place, in the next section we propose a practical algorithm to optimize the problem in (3) using the analytical solutions for the policy and the priors from Section 3.2.

4 MIRL: A Mutual Information RL Algorithm

In this section, we present the MIRL agent, our proposed algorithm (see Algorithm 1) for mutual-information RL. MIRL optimizes Equation (3) using a parametric function Q-function approximator Q_θ with parameters θ, and approximates the optimal prior over actions ρ during learning. The data used to learn those functions is collected with a behavioural policy that leverages the existence of the adjusted prior over actions. More concretely, at iteration i, MIRL holds an estimate of the Q-function denoted by Q_{θ_i} and an estimate of the optimal prior ρ_i. The optimal policy for a given estimate of Q_{θ_i} and ρ_i at iteration i is written as $\pi_i(a|s) = \frac{1}{Z}\rho_i(a)\exp(\beta_i Q_{\theta_i}(s,a))$ (see Equation (5)). Note that we have introduced an iteration-dependent β_i which will be useful to slowly shrink the effect of the constraint over time for improved performance. Now it only remains to show the form of the behavioural policy and how MIRL updates ρ_i, θ_i and β_i at each iteration.

Prior Updates: We approximate the optimal prior by employing the following update equation,

$$\rho_{i+1}(a) = (1-\alpha)\rho_i(a) + \alpha \pi_i(a|s_i) \tag{11}$$

where $s_i \sim \nu_i(\cdot)$. Note that assuming a fixed policy it is easy to show that this iterative update converges to $\rho_\pi(a) = \sum_s \mu_\pi(s)\pi(a|s)$, thus estimating correctly the optimal prior.

Q-function Updates: Concurrently to learning the prior, MIRL updates the parameters of the Q-function by minimizing the following loss,

$$L(\theta, \rho) := \mathbb{E}_{s,a,r,s' \sim \mathcal{M}} \left[\left(Q_\theta(s,a) - (T^\rho_{\text{soft}} Q_{\bar{\theta}})(s,a) \right)^2 \right] \tag{12}$$

where $\mathcal{M}$ is a replay memory (Mnih et al., 2015), $Q_{\bar{\theta}}$ is a target network that is updated after a certain number of training iterations, ρ is the estimate of the optimal prior and $T^\rho_{\text{soft}} Q$ is the empirical soft-operator defined as $(T^\rho_{\text{soft}} Q)(s,a) := r(s,a) + \gamma \frac{1}{\beta} \log \sum_{a'} \rho(a') \exp\left(\beta Q(s',a')\right)$. Importantly, this

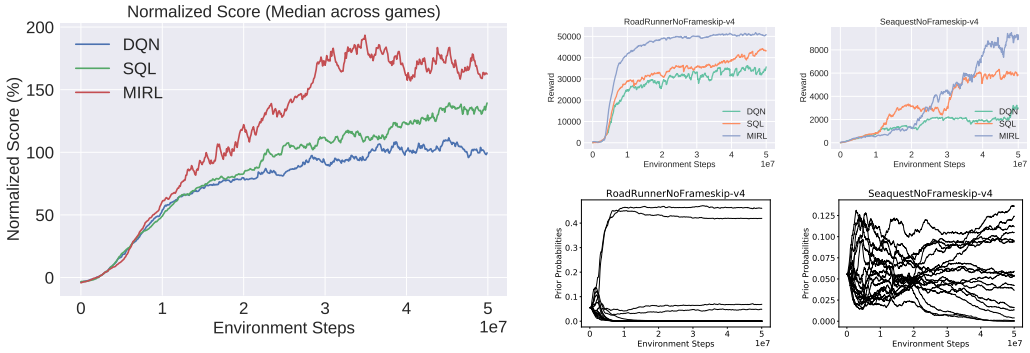

Figure 1: **Left panel:** Median Normalized score across 19 Atari games. Comparison between mutual information RL (MIR)L, Soft Q-learning (SQL) and DQN. **Right panels:** Top figures show the raw score for exemplary games i.e., RoadRunner and Seaquest. Bottom shows the evolution of the estimated prior over actions. We observe that for RoadRunner the prior converges to stable values, whereas for Seaquest it does not. See Appendix for details and plots for all environments. The curves are smoothed with an exponential moving average with effective window size of 10^6 environment steps.

operator differs from other soft operators arising when employing entropy regularization. The main difference lies in that previous approaches consider a fixed uniform prior, whereas in our case the optimal prior is estimated in the course of learning.

Behavioural policy: Since the Q-function can be learned off-policy, the samples in the replay memory can conveniently be sampled from a behavioural policy π_b different from the current estimate of the optimal policy. As such, the behavioural policy used in our experiments is similar in spirit to an epsilon greedy policy but it better exploits the existence of the estimated optimal prior when both exploring and exploiting. When exploring, MIRL's behavioural policy samples from the current estimate of the optimal prior ρ_i which has adjusted probabilities, in contrast to vanilla ϵ-greedy that samples all actions with equal frequency. Additionally, when exploiting, MIRL selects the maximum probability action that depends not only on the Q-values but also the current estimate of the optimal action prior, instead of selecting the action with highest Q-value as in traditional ϵ-greedy. More formally, given a random sample $u \sim$ Uniform[0,1] and epsilon ϵ, the action a_t is obtained by

$$a_t = \begin{cases} \arg\max_a \pi_i(a|s_t^{(i)}), & \text{if } u > \epsilon \\ a^\star \quad \text{where} \quad a^\star \sim \rho_i(\cdot) & \text{if } u \leq \epsilon. \end{cases}$$

Parameter β Updates: The parameter β can be seen as the Lagrange multiplier of Equation (9) that is directly controlled by the constant K in the constraint. As such, a small fixed value of β would restrict the class of available policies and evidently constrain the asymptotic performance of MIRL. In order to remedy this problem and obtain better asymptotic performance, we use the same adaptive β-scheduling from (Leibfried et al., 2017) in which β_i is updated according to the inverse of the empirical loss of the Q-function i.e, $\beta_{i+1} = (1-\alpha_\beta)\beta_i + \alpha_\beta(\frac{1}{L(\theta_i,\rho_{i+1})})$. This update favours small values of β at the beginning of training and large values towards the end of training when the error over Q-values is small. Therefore, towards the end of training when β is large, MIRL is solving the problem in (9) but without the constraint. This ensures that the asymptotic performance of MIRL is not hindered.

5 Experiments

We conducted experiments on 19 Atari games (Brockman et al., 2016) with Algorithm 1 (MIRL) and two baselines, Deep Q-Networks (DQN) (Mnih et al., 2015) that does not use any regularization and soft Q-learning (SQL) (Leibfried et al., 2017; Haarnoja et al., 2018) which employs entropy regularization. The three algorithms use a neural network for the estimation of Q-values as in (Mnih

et al., 2015). The network receives as an input the state s which is composed of the last four frames of the game with some extra pre-processing (see Appendix), and it outputs a vector of Q-values, i.e. one value for each valid action. We train the network for $5 \cdot 10^7$ environment steps, where a training iteration is performed every four steps. The target network $Q_{\bar{\theta}}$ is updated every 10^4 training iterations. Both SQL and MIRL update the Lagrange multiplier β over time by using an exponential moving average of the inverse loss (Leibfried et al., 2017) with $\alpha_\beta = 3 \cdot 10^5$. MIRL additionally updates the estimate of the optimal prior by using a learning rate $\alpha_\rho = 5 \cdot 10^{-5}$. Additional experimental details can be found in the Appendix.

During the training of each algorithm, snapshots of the network parameters (and the estimate of prior in the case of MIRL) are stored every 10^5 environment steps for evaluation. The evaluation for a single snapshot is conducted by running the policy for 30 episodes that last a maximum of $4.5 \cdot 10^3$ environment steps. The epsilon value when in evaluation mode is set to $\epsilon = 0.05$.

Game	DQN (%)	SQL (%)	MIRL (%)
Alien	**101.58**	51.02	40.23
Assault	250.61	283.62	**357.40**
Asterix	166.32	242.73	**330.19**
Asteroids	**9.74**	8.57	7.80
BankHeist	97.12	94.62	**166.26**
BeamRider	99.16	113.64	**117.21**
Boxing	2178.57	2283.33	**2338.89**
ChopperCommand	**72.71**	26.37	65.03
DemonAttack	350.95	451.78	**469.30**
Gopher	474.18	**538.87**	429.44
Kangaroo	351.48	393.16	**405.9**
Krull	843.16	886.68	**1036.04**
KungFuMaster	122.14	**142.04**	121.41
Riverraid	77.21	**109.37**	76.02
RoadRunner	548.90	613.62	**695.88**
Seaquest	21.95	36.00	**64.86**
SpaceInvaders	166.62	**200.38**	164.79
StarGunner	653.44	**681.12**	574.89
UpNDown	183.19	230.82	**394.21**
Mean	356.26	388.83	**413.46**

Table 1: Median Normalized score in 19 Atari games for Deep Q-Networks (DQN), Soft Q-Learning (SQL) and our approach (MIRL).

To summarize the results across all games, we normalized the episodic reward obtained in the evaluation. The normalized episodic rewards are computed as follows $z_{\text{normalized}} = \frac{z - z_{\text{random}}}{z_{\text{human}} - z_{\text{random}}} \cdot 100\%$, where z stands for the score obtained from our agent at test time, z_{random} stands for the score that a random agent would obtain and z_{human} for the score a human would obtain. Random and human scores are taken from (Mnih et al., 2015; Van Hasselt et al., 2016). As seen in Figure 1, our algorithm significantly outperforms the baselines in terms of the median normalized score. In particular, after $50M$ interactions we obtain about 30% higher median normalized score compared to Soft-Q-Learning and 50% higher compared to DQN. MIRL reaches the SQL final performance in about half the amount of interactions with the environment and, similarly, it reaches DQN final performance in about five times less interactions.

In Table 1 we show the comparison between best-performing agents for all the environments, where a best-performing agent is the agent that achieves the best score in evaluation mode considering all evaluation snapshots. Although this measure is not very robust, we include it since it is a commonly reported measure of performance in the field. MIRL outperforms the other baselines in 11 out of 19 games compared to SQL and DQN that are best on 5 and 3 games respectively.

6 Conclusion

Using a variational inference perspective, we derived a novel RL objective that allows optimization of the prior over actions. This generalizes previous methods in the literature that assume fixed uniform priors. We show that our formulation is equivalent to applying a mutual information regularization and derive a novel algorithm (MIRL) that learns the prior over actions. We demonstrate that MIRL significantly improves performance over Soft Q-Learning and Deep Q-Networks.

An interesting direction for future work would be to investigate the convergence properties of the alternating optimization problem presented here. We believe that, at least in the tabular case, the framework of stochastic approximation for two timescales (Borkar, 2009) is sufficient to prove convergence. On the experimental side, one could also investigate how our approach can be combined with the Rainbow framework (Hessel et al., 2017) which is the current state of the art in performance.

REFERENCES

Dimitri P Bertsekas. *Dynamic Programming and Optimal Control: Volume 2*. Athena Scientific, 1995.

Vivek S Borkar. *Stochastic approximation: a dynamical systems viewpoint*, volume 48. Springer, 2009.

Greg Brockman, Vicki Cheung, Ludwig Pettersson, Jonas Schneider, John Schulman, Jie Tang, and Wojciech Zaremba. Openai gym. *arXiv preprint arXiv:1606.01540*, 2016.

Thomas M Cover and Joy A Thomas. *Elements of information theory*. John Wiley & Sons, second edition, 2006.

Roy Fox, Ari Pakman, and Naftali Tishby. Taming the noise in reinforcement learning via soft updates. In *Proceedings of the Thirty-Second Conference on Uncertainty in Artificial Intelligence*, pp. 202–211. AUAI Press, 2016.

Tim Genewein, Felix Leibfried, Jordi Grau-Moya, and Daniel Alexander Braun. Bounded rationality, abstraction, and hierarchical decision-making: An information-theoretic optimality principle. *Frontiers in Robotics and AI*, 2:27, 2015.

Jordi Grau-Moya, Felix Leibfried, Tim Genewein, and Daniel A Braun. Planning with information-processing constraints and model uncertainty in markov decision processes. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pp. 475–491. Springer, 2016.

Tuomas Haarnoja, Haoran Tang, Pieter Abbeel, and Sergey Levine. Reinforcement learning with deep energy-based policies. In *International Conference on Machine Learning*, pp. 1352–1361, 2017.

Tuomas Haarnoja, Aurick Zhou, Pieter Abbeel, and Sergey Levine. Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. *arXiv preprint arXiv:1801.01290*, 2018.

Matteo Hessel, Joseph Modayil, Hado Van Hasselt, Tom Schaul, Georg Ostrovski, Will Dabney, Dan Horgan, Bilal Piot, Mohammad Azar, and David Silver. Rainbow: Combining improvements in deep reinforcement learning. *arXiv preprint arXiv:1710.02298*, 2017.

Felix Leibfried, Jordi Grau-Moya, and Haitham B Ammar. An information-theoretic optimality principle for deep reinforcement learning. *arXiv preprint arXiv:1708.01867*, 2017.

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

Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A Rusu, Joel Veness, Marc G Bellemare, Alex Graves, Martin Riedmiller, Andreas K Fidjeland, Georg Ostrovski, et al. Human-level control through deep reinforcement learning. *Nature*, 518(7540):529, 2015.

Martin Puterman. Markov decision processes: Discrete stochastic dynamic programming. 1994.

Konrad Rawlik, Marc Toussaint, and Sethu Vijayakumar. On stochastic optimal control and reinforcement learning by approximate inference. In *Robotics: science and systems*, volume 13, pp. 3052–3056, 2012.

Jonathan Rubin, Ohad Shamir, and Naftali Tishby. Trading value and information in mdps. In *Decision Making with Imperfect Decision Makers*, pp. 57–74. Springer, 2012.

John Schulman, Xi Chen, and Pieter Abbeel. Equivalence between policy gradients and soft q-learning. *arXiv preprint arXiv:1704.06440*, 2017.

Susanne Still and William Bialek. How many clusters? an information-theoretic perspective. *Neural computation*, 16(12):2483–2506, 2004.

Susanne Still and Doina Precup. An information-theoretic approach to curiosity-driven reinforcement learning. *Theory in Biosciences*, 131(3):139–148, 2012.

R Sutton and A Barto. *Reinforcement learning*. MIT Press, Cambridge, 1998.

Yee Teh, Victor Bapst, Wojciech M Czarnecki, John Quan, James Kirkpatrick, Raia Hadsell, Nicolas Heess, and Razvan Pascanu. Distral: Robust multitask reinforcement learning. In *Advances in Neural Information Processing Systems*, pp. 4496–4506, 2017.

Naftali Tishby, Fernando C Pereira, and William Bialek. The information bottleneck method. *The 37th Annual Allerton Conference on Communication, Control, and Computing.*, 1999.

Hado Van Hasselt, Arthur Guez, and David Silver. Deep reinforcement learning with double q-learning. 2016.

A APPENDIX

Proposition 2. Let $\mu_\pi(s)$ and $\rho_\pi(a)$ be the stationary distribution over states and actions under policy π according to Definitions (1) and (2). Then the stationary mutual information defined as $I_f(\mu_\pi, \pi, \rho_\pi) := \min_\rho I_f(\mu_\pi \pi, \rho)$ can also be written as

$$I_f(\mu_\pi, \pi, \rho_\pi) = \min_\rho \lim_{\gamma \to 1} (1-\gamma) \sum_{t=0}^{\infty} \gamma^t I_f(\nu_t, \pi, \rho) \tag{13}$$

where $\nu_t(s)$ is defined as in (7).

Proof. Following similar steps as in (Bertsekas, 1995, p.186) we have

$$\begin{aligned} I_f(\mu_\pi, \pi, \rho_\pi) &= \min_\rho \lim_{N \to \infty} \frac{1}{N} \sum_{t=0}^{N-1} I_f(\nu_t, \pi, \rho) \\ &= \min_\rho \lim_{N \to \infty} \lim_{\gamma \to 1} \frac{\sum_{t=0}^{N-1} \gamma^t I_f(\nu_t, \pi, \rho)}{\sum_{t=0}^{N} \gamma^t} \\ &= \min_\rho \lim_{\gamma \to 1} \lim_{N \to \infty} \frac{\sum_{t=0}^{N} \gamma^t I_f(\nu_t, \pi, \rho)}{\sum_{t=0}^{N} \gamma^t} \\ &= \min_\rho \lim_{\gamma \to 1} (1-\gamma) \sum_{t=0}^{\infty} \gamma^t I_f(\nu_t, \pi, \rho) \end{aligned}$$

where we used Proposition 3 (shown next) in the first equality and where the limits in the third equality can be interchanged due to the Monotone Convergence Theorem. □

Proposition 3. *Let $\mu_\pi(s)$ and $\rho_\pi(a)$ be the stationary distribution over states and actions under policy π according to Definitions* (1) *and* (2)*. Then the stationary mutual information defined as $I_f(\mu_\pi, \pi, \rho_\pi)$ can also be written as*

$$I_f(\mu_\pi, \pi, \rho_\pi) = \min_\rho \lim_{N \to \infty} \frac{1}{N} \sum_{t=0}^{N-1} \sum_{s_t} \nu_t(s_t) KL(\pi(\cdot|s_t) || \rho(\cdot)) \tag{14}$$

where $\nu_t(s)$ is defined as in (7)*.*

Proof. Let $\mathcal{I}_\rho(s) := \mathrm{KL}(\pi(\cdot|s)||\rho(\cdot))$ and $\mathcal{I}_{\rho_\pi}(s) := \mathrm{KL}(\pi(\cdot|s)||\rho_\pi(\cdot))$. Note that both previous quantities are bounded for all t. Then

$$
\begin{aligned}
&\min_\rho \lim_{N\to\infty} \frac{1}{N} \sum_{t=0}^{N-1} \sum_s \nu_t(s)\mathcal{I}_\rho(s) = \\
&= \min_\rho \lim_{N\to\infty} \frac{1}{N} \left(\sum_{t=0}^{K} \sum_s \nu_t(s)\mathcal{I}_\rho(s) + \sum_{t=K+1}^{N-1} \sum_s \nu_t(s)\mathcal{I}_\rho(s) \right) \\
&= \min_\rho \lim_{N\to\infty} \frac{1}{N} \sum_{t=K+1}^{N-1} \sum_s \nu_t(s)\mathcal{I}_\rho(s) \\
&= \min_\rho \lim_{N\to\infty} \frac{1}{N} \sum_{t=K+1}^{N-1} \sum_s \mu_\pi(s)\mathcal{I}_\rho(s) \\
&= \min_\rho \sum_s \mu_\pi(s)\mathcal{I}_\rho(s) \lim_{N\to\infty} \frac{1}{N} \sum_{t=K+1}^{N-1} 1 \\
&= \sum_s \mu_\pi(s)\mathcal{I}_\mu(s) \lim_{N\to\infty} \frac{N-K}{N} \\
&= \sum_s \mu_\pi(s)\mathcal{I}_\mu(s) \\
&= I_f(\mu_\pi, \pi, \rho_\pi)
\end{aligned}
$$

where we assumed that $\nu_t(s) = \mu_\pi(s)$ for all $t > K$ and finite large enough K. □

Parameter	Value
Frame size	$[84, 84]$
Frame skip	4
History frames (in s)	4
Reward clipping	$[-1., 1.]$
Max environment steps	27000
Target update frequency (train. steps)	10000
Training update frequency (env. steps)	4
Batch size	32
Memory capacity	10^6
γ	0.99
α_ρ	$2 \cdot 10^{-6}$
α_Q	$2 \cdot 10^{-5}$
α_β	$3.3 \cdot 10^{-6}$
β_0	0.01

Table 2: Hyperparameters for MIRL. DQN and SQL have the same parameter values as in this table for the used parameters.

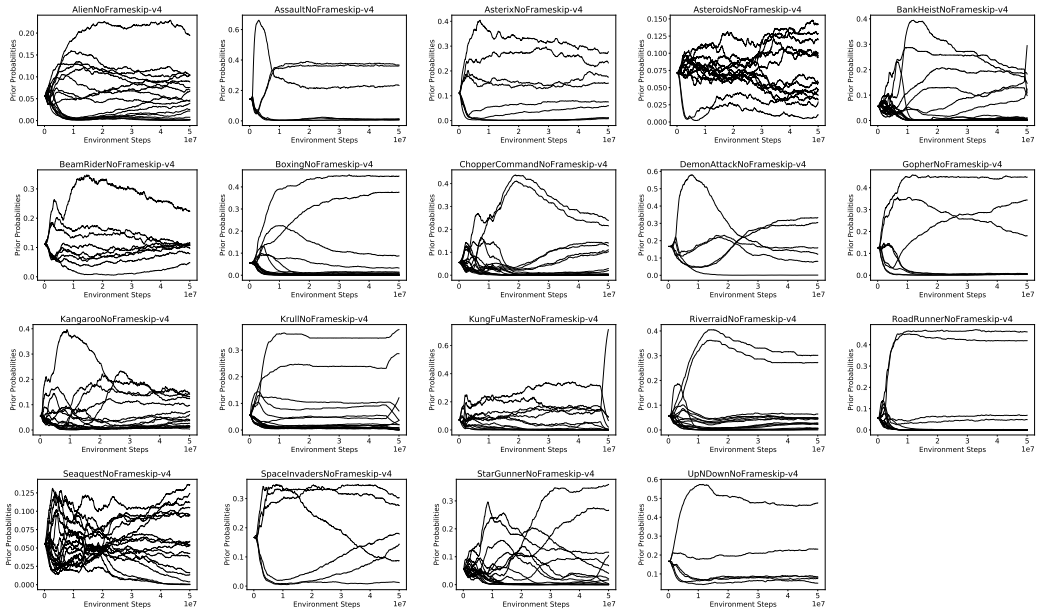

Figure 2: Prior Evolution for all games.

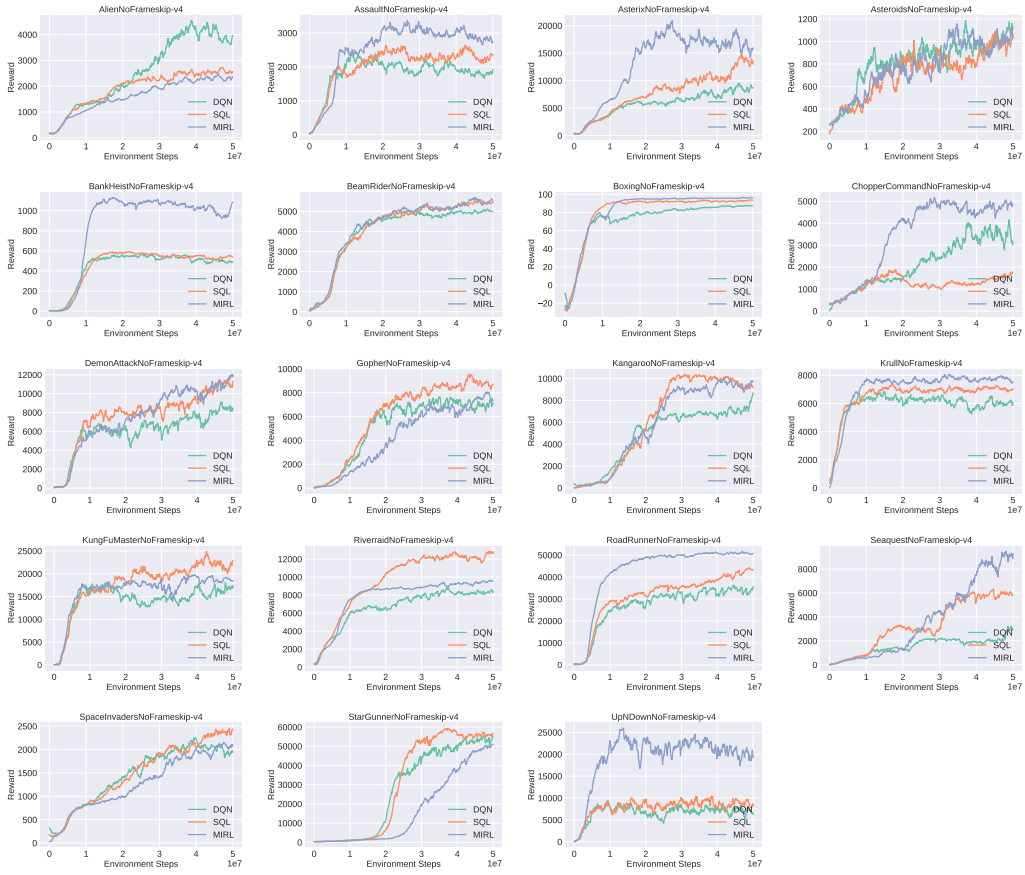

Figure 3: Scores for all games on the evaluation snapshots.