Entropy Regularization with Discounted Future State Distribution in Policy Gradient Methods

Riashat Islam
McGill University, Mila
School of Computer Science
riashat.islam@mail.mcgill.ca

Raihan Seraj
McGill University, CIM
Electrical and Computer Engineering
raihan.seraj@mail.mcgill.ca

Pierre-Luc Bacon
Stanford University
Computer Science
plbacon@cs.stanford.edu

Doina Precup
McGill University, Mila
School of Computer Science
Google Deepmind
dprecup@cs.mcgill.ca

Abstract

The policy gradient theorem is defined based on an objective with respect to the initial distribution over states. In the discounted case, this results in policies that are optimal for one distribution over initial states, but may not be uniformly optimal for others, no matter where the agent starts from. Furthermore, to obtain unbiased gradient estimates, the starting point of the policy gradient estimator requires sampling states from a normalized discounted weighting of states. However, the difficulty of estimating the normalized discounted weighting of states, or the stationary state distribution, is quite well-known. Additionally, the large sample complexity of policy gradient methods is often attributed to insufficient exploration, and to remedy this, it is often assumed that the restart distribution provides sufficient exploration in these algorithms. In this work, we propose exploration in policy gradient methods based on maximizing entropy of the discounted future state distribution. The key contribution of our work includes providing a practically feasible algorithm to estimate the normalized discounted weighting of states, i.e, the *discounted future state distribution*. We propose that exploration can be achieved by entropy regularization with the discounted state distribution in policy gradients, where a metric for maximal coverage of the state space can be based on the entropy of the induced state distribution. The proposed approach can be considered as a three time-scale algorithm and under some mild technical conditions, we prove its convergence to a locally optimal policy. Experimentally, we demonstrate usefulness of regularization with the discounted future state distribution in terms of increased state space coverage and faster learning on a range of complex tasks.

1 Introduction

Exploration in policy optimization methods is often tied to exploring in the policy parameter space. This is primarily achieved by adding noise to the gradient when following stochastic gradient ascent. More explicit forms of exploration within the state and action space include policy entropy regularization. This promotes stochasticity in policies, thereby preventing premature convergence to deterministic policies [Mnih et al., 2016a, Schulman et al., 2017]. Such regularization schemes play the role of smoothing out the optimization landscape in non-convex policy optimization problems [Ahmed et al., 2018]. Deep reinforcement learning algorithms have had enormous success with

Optimization Foundations for Reinforcement Learning Workshop at NeurIPS 2019, Vancouver, Canada.

entropy regularized policies, commonly known as maximum entropy RL framework [Ziebart, 2010]. These approaches ensure exploration in the action space, which indirectly contributes to exploration in the state space, but do not explicitly address the issue of state space exploration. This leads us to the question : *how do we regularize policies to obtain maximal coverage in the state space?*

One of the metrics to measure coverage in state space is the entropy of the *discounted* future state distribution, as proposed in [Hazan et al., 2018]. In their work, they prove that using the entropy of discounted future state distribution as a reward function, we can achieve improved coverage of the state space. Drawing inspiration from this idea, and to provide a practically feasible construct, we first propose an approach to estimate the discounted future state distribution. We then provide an approach for efficient exploration in policy gradient methods, to reduce sample complexity, by regularizing policy optimization based on the entropy of the discounted future state distribution. The implication of this is that the policy gradient algorithm yields policies that improve state space coverage by maximizing the entropy of the discounted future state distribution induced by those policies as an auxiliary regularized objective. This distribution takes into account when various states are visited in addition to which states are visited. The main contribution of our work is to provide a practically feasible way to estimate the discounted future state distribution with a density estimator. Furthermore, we show that regularizing policy gradients with the entropy of this distribution can improve exploration. To the best of our knowledge, there are no previous works that provide a practical realization for estimating and regularizing with the entropy of the discounted state distribution. It is worthwhile to note that the estimation of the discounted/stationary state distribution is not readily achievable in practice. This is because the stationary distribution requires an estimate based on rollouts, as in value function estimates, under a given policy π. In contrast, the discounted state distribution requires estimation of discounted occupancy measures for the various states. Since the discounted occupancy measure is purely a theoretical construct, it is not possible to sample from this distribution using rollouts. In order to use this as an entropy regularizer, we also need the discounted or stationary distributions to be explicitly dependent on the policy parameters, which is not straightforward in practice.

To address this, we estimate the state distribution by separately training a density estimator based on sampled states in the rollout. The crucial step here is that, we use a density estimator that is explicitly a function of the policy parameters θ. In other words, our density estimator takes as input, the parameters θ of the policy itself (for instance, weights of a policy neural network) through which we now obtain an estimate of $p_\theta : \theta \mapsto \Delta(S)$, where $p_\theta(s)$ is the occupancy probability (discounted or otherwise) of state s. We use a variational inference based density estimator, which can be trained to maximize a variational lower bound to the the log-likelihood of $p_\theta(s)$. As a result, we can obtain an estimation of d_{π_θ} since in case of stationary distributions, we have $\log p_\theta(s) = \log d_{\pi_\theta}(s)$. Estimation of d_{π_θ} under any policy π_θ requires collecting a large number of samples from the rollout. Instead of this, we can use ideas from multi-scale stochastic algorithms to learn this in an online manner. Hence, we require a separate time-scale for training the density estimator, in addition to learning the policy and value functions in policy gradient based approaches. We formally state and prove the corresponding three time-scale algorithm.

2 State Distribution in Policy Gradient Methods

Policy gradient theorem [Sutton et al., 1999] for the starting state formulation are given for an initial state distribution α, where the exact solution for the discounted objective is given by $J\theta = \alpha^T v_\theta = \alpha^T(\mathbf{I} - \gamma P_\theta)^{-1} r_\theta$. In [Sutton et al., 1999], this is often known as the *discounted weighting of states* defined by $d^T_{\alpha,\gamma,\pi} = \alpha^T(\mathbf{I} - \gamma P_\theta)^{-1}$, where in the average reward case this reaches a stationary distribution implying that the process is independent of the initial states. However, the discounted weighting of states is not a distribution, or a stationary distribution in itself, since the rows of the matrix $(\mathbf{I} - \gamma P_\theta)^{-1}$ do not sum to 1. The normalized version of this is therefore often considered, commonly known as the *discounted future state distribution* [Kakade, 2003] or the discounted state distribution [Thomas, 2014]. Detailed analysis of the significance of the state distribution in policy gradient methods is further given in [Bacon, 2018].

$$\bar{d}_{\alpha,\gamma,\pi} = (1-\gamma) d_{\alpha,\gamma,\pi} = (1-\gamma)\alpha^T(\mathbf{I} - \gamma \mathbf{P}_\pi)^{-1} = (1-\gamma)\alpha^T \sum_{t=0}^{\infty} \gamma^t P_\pi(s_t = s) \tag{1}$$

Given an infinite horizon MDP, and a stationary policy $\pi(a, s)$, equation (1) is the γ discounted future state distribution, i.e., the normalized version for the discounted weighting of states. We can draw samples from this distribution, by simulating π and accepting each state as the sample with a probability $(1-\gamma)$. With the discounted future state distribution, the equivalent policy gradient objective can therefore be given by $J(\theta) = \bar{\mathbf{d}}_{\alpha,\gamma,\theta}^{T} \mathbf{r}_{\theta}$. In practice, we want to express the policy gradient theorem with an expectation that we can estimate by sampling, but since the discounted weighting of states $d_{\alpha,\gamma,\pi}$ is not a distribution over states, we often use the normalized counterpart of the discounted weighting of states $\bar{d}_{\alpha,\gamma,\pi}$ and correct the policy gradient with a factor of $\frac{1}{(1-\gamma)}$.

$$\nabla_{\theta} J(\theta) = \frac{1}{(1-\gamma)} \mathbb{E}_{\bar{d}_{\alpha,\gamma,\theta}, a \sim \pi_{\theta}} [\nabla_{\theta} \log \pi_{\theta}(a, s) Q_{\pi_{\theta}}(s, a)] \tag{2}$$

However, since the policy gradient objective is defined with respect to an initial distribution over states, the resulting policy is not *optimal* over the entire state space, ie, not *uniformly optimal*, but are rather optimal for one distribution over the initial states but may not be optimal for a different starting state distribution. This often leads to the large sample complexity of policy gradient methods [Kakade, 2003] where a large number of samples may be required for obtaining good policies. The lack of exploration in policy gradient methods may often lead to large sample complexity to obtain accurate estimates of the gradient direction. It is often assumed that the restart, or starting state distribution in policy gradient method provides sufficient exploration. In this work, we tackle the exploration problem in policy gradient methods by explicitly using the entropy of the discounted future state distribution. We show that even for the starting state formulation of policy gradients, we can construct the normalized discounted future state distribution, where instead of sampling from this distribution (which is hard in practice, since sampling requires discounting with $(1-\gamma)$, we instead regularize policy optimization with the entropy $\mathbb{H}((1-\gamma)\mathbf{d}_{\alpha,\gamma,\theta})$

3 Entropy Regularization with Discounted Future State Distribution

The key idea behind our approach is to use regularization with the entropy of the state distribution in policy gradient methods. In policy optimization based methods, the state coverage, or the various times different states are visited can be estimated from the state distribution induced by the policy. This is often called the discounted (future) state distribution, or the normalized discounted weighting of states. In this work, our objective is to promote exploration in policy gradient methods by using the entropy of the discounted future state distribution $d_{\alpha,\gamma,\pi}$ (which we will denote as d_{π_θ}) where α is the distribution over the initial states and to explicitly highlight that this distribution is dependent on the changes in the policy parameters θ, and we propose a practically feasible algorithm for estimating and regularizing policy gradient methods with the discounted state distribution for exploration and reducing sample complexity. We propose the following state distribution entropy regularized policy gradient objective: $\tilde{J}(\theta) = \mathbb{E}_{\pi_\theta}\left[\sum_{t=0}^{\infty} \gamma^t r(S_t, A_t) \;\middle|\; S_0\right] + \lambda \mathbb{H}(d_{\pi_\theta})$, where d_{π_θ} is the discounted state distribution induced by the policy π. We can estimate $\nabla_\theta J(\theta)$ while using stochastic policies from [Sutton et al., 1999] or deterministic policies from [Silver et al., 2014]. The regularized policy gradient objective is: $\nabla_\theta \tilde{J}(\theta) = \nabla_\theta J(\theta) + \lambda \nabla_\theta \mathbb{H}(d_{\pi_\theta})$.

Entropy of the *discounted* state distribution $\mathbb{H}(d_{\alpha,\gamma,\pi})$: The discounted state distribution $d_{\alpha,\gamma,\pi_\theta}$ can be computed as: $d_{\alpha,\gamma,\pi_\theta}(s) = (1-\gamma)\alpha^T \sum_{t=0}^{\infty} \gamma^t P(S_t = s), \quad \forall s \in \mathcal{S}$

We note that this is a theoretical construct and we cannot sample from this distribution, since it would require sampling each state with a probability $(1-\gamma)$ such that the accepted state is then distributed according to $d_{\alpha,\gamma,\pi_\theta}$. However, we can modify the state distribution $p(s)$ to a weighted distribution $\tilde{p}(s)$ as follows. We estimate $p(s)$ from samples as: $p(s) = \frac{1}{T} \sum_{t=0}^{T} \mathbb{1}(S_t = s)$, where the weight of each sample is $1/T$. To estimate $\tilde{p}(s)$, we use an importance sampling weighting of $(1-\gamma)\gamma^t$ to yield:

$$\tilde{p}(s) \overset{(a)}{=} \frac{(1-\gamma)}{T} \sum_{t=0}^{T} \gamma^t \mathbb{1}(S_t = s) \overset{(b)}{=} (1-\gamma) \sum_{t=0}^{T} (\gamma^t P(S_t = s \mid S_0)) \overset{(c)}{\approx} d_{\gamma,\pi_\theta}(s), \tag{3}$$

where (a) follows from the importance sampling approach, (b) follows from the fact that $\frac{\mathbb{1}(S_t=s)}{T} = P(S_t = s \mid S_o)$ and (c) follows from above where the approximation is due to the finite truncation of the infinite horizon trajectory. Note that due to this finite truncation, our estimate of d_{γ,π_θ} will be sub-stochastic. Therefore, we can estimate the entropy of this distribution as:

$$\mathbb{H}(d_{\alpha,\gamma,\pi_\theta}) \approx -\frac{1}{T}\sum_{t=0}^{T} \log \tilde{p}(S_t). \tag{4}$$

Entropy of the *stationary* state distribution $\mathbb{H}(d_{1,\pi_\theta})$: For the average reward case with infinite horizon MDPs, we can similarly compute the entropy of the *stationary* state distribution. The stationary distribution d_{1,π_θ} is a solution of the following fixed point equation satisfying $d_{1,\pi_\theta} = P_{\pi_\theta}^{\intercal} d_{1,\pi_\theta}$, where P_{π_θ} is the transition probability matrix corresponding to policy π_θ. In practice, this is the long term state distribution under policy π_θ, which is denoted as $p(s)$. In infinite horizon problems, the stationary state distribution is indicative of the majority of the states visited under the policy. We expect the stationary state distribution to change slowly, as we adapt the policy parameters (especially for a stochastic policy). Hence, we assume that the states have mixed, as we learn the policy over several iterations. In practice, instead of adding a mixing time specifically, we can use different time-scales for learning the policy and estimating the stationary state distribution. The entropy of the stationary state distribution can therefore be computed as :

$$\mathbb{H}(d_{1,\pi_\theta}) \overset{(a)}{=} -\sum_{s\in\mathcal{S}} d_{1,\pi_\theta}(s)\log(d_{1,\pi_\theta}(s)) \overset{(b)}{\approx} -\frac{1}{T}\sum_{t=0}^{T}\log d_{1,\pi_\theta}(S_t) \overset{(c)}{=} -\frac{1}{T}\sum_{t=0}^{T}\log p(S_t), \tag{5}$$

where T is a finite number of time-steps after which an infinite horizon episode can be truncated due to discounting. In deriving (5), (a) follows from the definition of entropy, (b) follows by assuming ergodicity, which allows us to replace an expectation over the state space with an expectation over time under all policies. The approximation here is due to the finite truncation of the infinite horizon to T. Step (c) follows from the density estimation procedure.

Estimating the entropy of discounted future state distribution: In practice, we use a neural density estimator for estimating the discounted state distribution, based on the states induced by the policy π_θ. The training samples for the density estimator is obtained by rolling out trajectories under the policy π_θ. We train a variational inference based density estimator (similar to a variational auto-encoder) to maximize variational lower bound $\log p(s)$, where for the discounted case, we denote this as $\log \tilde{p}(s)$, as given in (3) and (4). We therefore obtain an approximation to the entropy of discounted future state distribution which can be used in the modified policy gradient objective, where for the *discounted* case, with stochastic policies [Sutton et al., 1999], we have

$$\tilde{J}(\theta) = \mathbb{E}_{\pi_\theta}\Big[\sum_{t=0}^{\infty}\gamma^t r(S_t, A_t) - \lambda \log d_{\alpha,\gamma,\pi_\theta}(s_t) \;\Big|\; S_0\Big] \tag{6}$$

The objective in the *stationary* case can be obtained by substituting the $d_{\alpha,\gamma,\pi_\theta}(s_t)$ with $d_{1,\pi_\theta}(s_t)$ in (6). The neural density estimator is independently parametrized by ϕ, and is a function that maps the policy parameters θ to a state distribution. The loss function for this density estimator is the KL divergence between $\mathtt{KL}(q_\phi(Z \mid \theta)||p(Z|\theta))$. The training objective for our density estimator in the *stationary* case is given by :

$$\mathcal{L}_\gamma(\phi,\theta) = (1-\gamma)\gamma^k \mathbb{E}_{q_\phi(Z|\theta)}\big[\log p_\phi(S|\theta)\big] - KL\big(q_\phi(Z|\theta)||p(\theta)\big) \tag{7}$$

Equation (7) gives the expression for the loss function for training the state density estimator (which is the variational inference lower bound loss for estimating $log(p(s))$,i.e.,ELBO Kingma and Welling [2013]. Here θ are the parameters of the policy network π_θ, ϕ are the parameters of the density estimator. The novelty of our approach is that the density estimator takes as input the parameters of the policy network directly (similar to hypernetworks Krueger et al. [2017], Ha et al. [2016]). The encoder then maps the policy parameters θ into the latent space Z given by $q_\phi(Z \mid \theta)$ with a Gaussian prior over the policy parameters θ. During implementation we feed the parameters of the last two layers of the policy network, assuming the previous layers extract the relevant state features and the last two layers map the obtained features to a distribution over actions. Hence θ only comprises of the weights of these last two layers ensuring computation tractability. We take this approach since the discounted future state distribution is a function of the policy parameters θ. Our overall gradient objective with the regularized update is therefore given by $\tilde{J}(\theta) = \mathbb{E}_{\pi_\theta}\Big[\sum_{t=0}^{\infty}\gamma^t r(S_t, A_t) - \lambda\mathcal{L}_\gamma(\phi,\theta)\Big]$, where

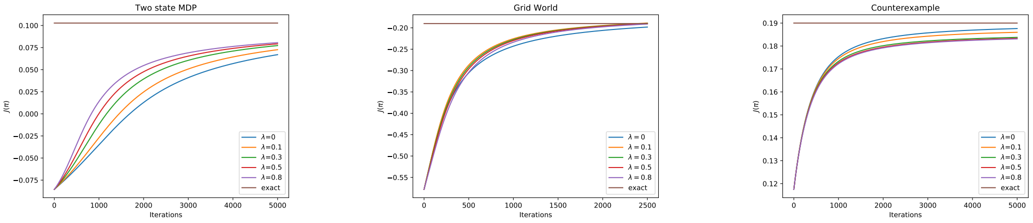
$\mathcal{L}_\gamma(\phi, \theta)$ directly depends on the policy parameters θ. This gives the regularized policy gradient update with the entropy of the discounted future state distribution, for stochastic policies as :

$$\nabla_\theta \tilde{J}(\theta) = \mathbb{E}_{\pi_\theta}\Big[\nabla_\theta \log \pi(A_t \mid S_t) Q^\pi(S_t, A_t) - \lambda \nabla_\theta \mathcal{L}_\gamma(\phi, \theta).\Big], \text{where } \mathcal{L}_\gamma(\phi, \theta) = (1-\gamma)\gamma^t \mathcal{L}(\phi, \theta) \tag{8}$$

4 Experiments

In this section, we demonstrate our approach based on entropy regularization with the normalized discounted weighting of states, also known as the discounted future state distribution. Our method can be applied on top of *any* existing RL algorithm. In all our experiments, we use γ-*StateEnt* (or *Discounted StateEnt*) for denoting entropy regularization with the discounted future state distribution, and *StateEnt* for denoting the unnormalized counterpart of the discounted weighting of states.

Entropy regularization in Exact Policy Gradients with $\mathbb{H}(d_\pi)$: We first verify that entropy regularization with *exact* discounted future state distribution $\mathbb{H}(d_{\pi_\theta})$ can lead to benefits in policy optimization when used as a regularizer. We demonstrate this on three toy domains, varying the amount λ of state distribution regularization, in the case where we can compute *exact policy gradient* given by $J(\pi) = (I - \gamma P_\pi)^{-1} R$. In all these examples, the optimal solution can be found with value iteration algorithm.



(a) (b) (c)

Figure 1: State distribution entropy regularized exact policy gradient can lead to a better converged solution on a simple two state MDP (taken from [Dadashi et al., 2019] (Figure (a)) and a discrete GridWorld domain (Figure (b)). The regularized objective has a faster convergence rate compared to the unregularized objective (with $\lambda = 0.0$). Interestingly, in toy MDPs where there exists state aliasing, as in Figure (c) for MDP taken from the counterexample domain [Imani et al., 2018], we find that state distribution entropy does not lead to significant improvements. This is an interesting result justifying that state space exploration may not necessarily be needed in all MDPs, especially when states are aliased

Toy Domains: Having verified our hypothesis in figure 1, we now present our approach based on separately learning a density estimator for the state distribution, on tabular domains with actor-critic algorithms. We compare our results for both the *discounted* state distributions and the unnormalized counterpart, with a baseline actor-critic (with $\lambda = 0.0$ for the regularizer). Figure 2 summarizes our results.

Complex Sparse Reward GridWorld Tasks: We demonstrate the usefulness of our approach, with entropy of stationary (denoted StateEnt) and discounted (denoted γ StateEnt) state distributions, on sparse reward complex gridworld domains. These are hard exploration tasks, where the agent needs to pass through slits and walls to reach the goal state (placed at the top right corner of the grid). We use REINFORCE [Williams, 1992] as the baseline algorithm, and for all comparisons use standard policy entropy regularization (denoted PolicyEnt for baseline).

Continuous Control Tasks: We extend our proposed regularized policy gradient objective on standard continuous control Mujoco domains [Todorov et al., 2012]. First, we examine the significance of the state distribution entropy regularizer in DDPG algorithm [Lillicrap et al., 2016]. In DDPG,

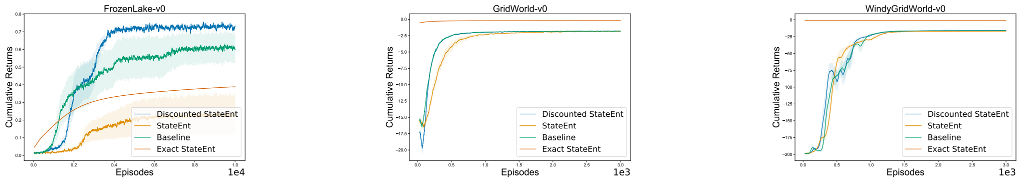
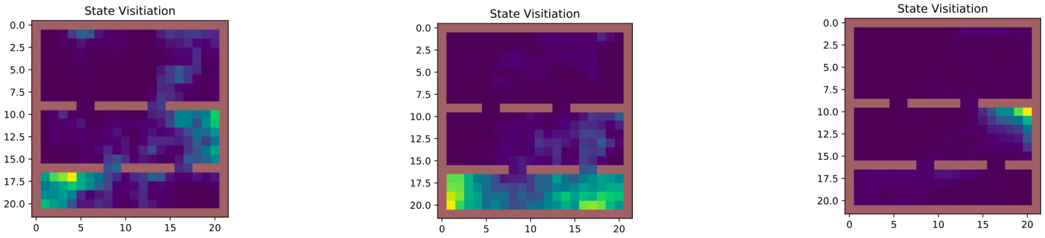


Figure 2: We show benefits of state distribution entropy regularization on toy domains, especially hard exploration tabular tasks such as FrozenLake. In all the tasks, we find that regularization with entropy of *discounted* state distribution performs significantly better compared to baseline and regularization with the unnormalized state distribution. In all tasks, we use $\lambda = 0.1$ for our methods.



(a) γ StateEnt

(b) StateEnt

(c) PolicyEnt

Figure 3: State space coverage on complex sparse reward double-slit-double-wall gridworld domains. Figure shows that regularization with the *discounted* state distribution indeed has more useful effects in terms of exploration and state space coverage compared to regularization with policies. We also find that state space coverage is more with entropy of *discounted* state distribution compared to *stationary* state distribution. All state visitation heat maps are shown after only 1000 time steps of initial training phase.

policy entropy regularization cannot be used due to existence of deterministic policies [Silver et al., 2014]. In Figure 4, we show that by inducing policies to maximize state space coverage, we can enhance exploration that leads to significant improvements on standard benchmark tasks, especially in environments where exploration in the state space plays a key role (e.g HalfCheetah environment)

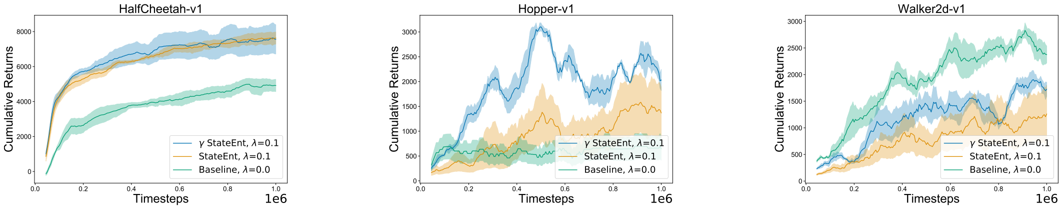


Figure 4: Significant performance improvements with state entropy regularization with DDPG, especially in tasks like HalfCheetah-v1 and Hopper-v1 where exploration plays a key role. In all the experiments, we use a state entropy regularization coefficient of $\lambda = 0.1$ for our approach, and $\lambda = 0.0$ for the baseline DDPG. Experiment results are averaged over 10 random seeds [Henderson et al., 2018]

5 Summary and Discussion

In this work, we provided a practically feasible algorithm for entropy regularization with the state distributions in policy optimization. We present a practically feasible algorithm, based on estimating the discounted future state distribution, for both episodic and infinite horizon environments. The key to our approach relies on using a density estimator for the state distribution d_{π_θ}, which is a direct function of the policy parameters θ itself, such that we can regularize policy optimization to induce policies that can maximize state space coverage. We demonstrate the usefulness of this approach on a wide range of tasks, starting from simple toy tasks to sparse reward gridworld domains, and eventually extending our results to a range of continuous control suites. We re-emphasize that our approach gives a practically convenient handle to deal with the *discounted* state distribution, that are difficult to work with in practice. In addition, we provided a proof of convergence of our method as a three time-scale algorithm, where learning a policy depends on both a value function and a state distribution estimation.

References

Volodymyr Mnih, Adrià Puigdomènech Badia, Mehdi Mirza, Alex Graves, Timothy P. Lillicrap, Tim Harley, David Silver, and Koray Kavukcuoglu. Asynchronous methods for deep reinforcement learning. In *Proceedings of the 33nd International Conference on Machine Learning, ICML 2016, New York City, NY, USA, June 19-24, 2016*, pages 1928–1937, 2016a. URL `http://jmlr.org/proceedings/papers/v48/mniha16.html`.

John Schulman, Pieter Abbeel, and Xi Chen. Equivalence between policy gradients and soft q-learning. *CoRR*, abs/1704.06440, 2017. URL `http://arxiv.org/abs/1704.06440`.

Zafarali Ahmed, Nicolas Le Roux, Mohammad Norouzi, and Dale Schuurmans. Understanding the impact of entropy on policy optimization. *CoRR*, abs/1811.11214, 2018. URL `http://arxiv.org/abs/1811.11214`.

Brian D. Ziebart. *Modeling Purposeful Adaptive Behavior with the Principle of Maximum Causal Entropy*. PhD thesis, Pittsburgh, PA, USA, 2010. AAI3438449.

Elad Hazan, Sham M. Kakade, Karan Singh, and Abby Van Soest. Provably efficient maximum entropy exploration. *CoRR*, abs/1812.02690, 2018. URL `http://arxiv.org/abs/1812.02690`.

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 12, [NIPS Conference, Denver, Colorado, USA, November 29 - December 4, 1999]*, pages 1057–1063, 1999.

Sham M. Kakade. On the sample complexity of reinforcement learning. 2003.

Philip Thomas. Bias in natural actor-critic algorithms. In *Proceedings of the 31th International Conference on Machine Learning, ICML 2014, Beijing, China, 21-26 June 2014*, pages 441–448, 2014. URL `http://proceedings.mlr.press/v32/thomas14.html`.

David Silver, Guy Lever, Nicolas Heess, Thomas Degris, Daan Wierstra, and Martin A. Riedmiller. Deterministic policy gradient algorithms. In *Proceedings of the 31th International Conference on Machine Learning, ICML 2014, Beijing, China, 21-26 June 2014*, pages 387–395, 2014. URL `http://jmlr.org/proceedings/papers/v32/silver14.html`.

Diederik P. Kingma and Max Welling. Auto-encoding variational bayes. *CoRR*, abs/1312.6114, 2013. URL `http://arxiv.org/abs/1312.6114`.

David Krueger, Chin-Wei Huang, Riashat Islam, Ryan Turner, Alexandre Lacoste, and Aaron Courville. Bayesian hypernetworks. *arXiv preprint arXiv:1710.04759*, 2017.

David Ha, Andrew Dai, and Quoc V Le. Hypernetworks. *arXiv preprint arXiv:1609.09106*, 2016.

Robert Dadashi, Adrien Ali Taïga, Nicolas Le Roux, Dale Schuurmans, and Marc G. Bellemare. The value function polytope in reinforcement learning. *CoRR*, abs/1901.11524, 2019. URL `http://arxiv.org/abs/1901.11524`.

Ehsan Imani, Eric Graves, and Martha White. An off-policy policy gradient theorem using emphatic weightings. In *Advances in Neural Information Processing Systems*, pages 96–106, 2018.

Ronald J. Williams. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine Learning*, 8:229–256, 1992. doi: 10.1007/BF00992696. URL `https://doi.org/10.1007/BF00992696`.

Emanuel Todorov, Tom Erez, and Yuval Tassa. Mujoco: A physics engine for model-based control. In *IROS*, pages 5026–5033. IEEE, 2012. ISBN 978-1-4673-1737-5. URL `http://dblp.uni-trier.de/db/conf/iros/iros2012.html#TodorovET12`.

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 *4th International Conference on Learning Representations, ICLR 2016, San Juan, Puerto Rico, May 2-4, 2016, Conference Track Proceedings*, 2016. URL `http://arxiv.org/abs/1509.02971`.

Peter Henderson, Riashat Islam, Philip Bachman, Joelle Pineau, Doina Precup, and David Meger. Deep reinforcement learning that matters. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), the 30th innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-18), New Orleans, Louisiana, USA, February 2-7, 2018*, pages 3207–3214, 2018. URL `https://www.aaai.org/ocs/index.php/AAAI/AAAI18/paper/view/16669`.

Ronald J Williams and Jing Peng. Function optimization using connectionist reinforcement learning algorithms. *Connection Science*, 3(3):241–268, 1991.

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*, pages 1928–1937, 2016b.

Tuomas Haarnoja, Aurick Zhou, Pieter Abbeel, and Sergey Levine. Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. In *Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018*, pages 1856–1865, 2018. URL `http://proceedings.mlr.press/v80/haarnoja18b.html`.

Anirudh Goyal, Riashat Islam, Daniel Strouse, Zafarali Ahmed, Matthew Botvinick, Hugo Larochelle, Sergey Levine, and Yoshua Bengio. Infobot: Transfer and exploration via the information bottleneck. *CoRR*, abs/1901.10902, 2019. URL `http://arxiv.org/abs/1901.10902`.

Alexandre Galashov, Siddhant M Jayakumar, Leonard Hasenclever, Dhruva Tirumala, Jonathan Schwarz, Guillaume Desjardins, Wojciech M Czarnecki, Yee Whye Teh, Razvan Pascanu, and Nicolas Heess. Information asymmetry in kl-regularized rl. *International Conference on Learning Representations*, 2019. doi: arXiv:1905.01240. URL `https://arxiv.org/abs/1905.01240`.

Philip Bachman, Riashat Islam, Alessandro Sordoni, and Zafarali Ahmed. Vfunc: a deep generative model for functions. *arXiv preprint arXiv:1807.04106*, 2018.

Benjamin Eysenbach, Abhishek Gupta, Julian Ibarz, and Sergey Levine. Diversity is all you need: Learning skills without a reward function. *International Conference on Learning Representations*, 2019.

Deepak Pathak, Pulkit Agrawal, Alexei A Efros, and Trevor Darrell. Curiosity-driven exploration by self-supervised prediction. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 16–17, 2017.

Marc G. Bellemare, Sriram Srinivasan, Georg Ostrovski, Tom Schaul, David Saxton, and Rémi Munos. Unifying count-based exploration and intrinsic motivation. In *Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016, December 5-10, 2016, Barcelona, Spain*, pages 1471–1479, 2016. URL `http://papers.nips.cc/paper/6383-unifying-count-based-exploration-and-intrinsic-motivation`.

Georg Ostrovski, Marc G. Bellemare, Aäron van den Oord, and Rémi Munos. Count-based exploration with neural density models. In *Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017*, pages 2721–2730, 2017. URL `http://proceedings.mlr.press/v70/ostrovski17a.html`.

Vijay R Konda and John N Tsitsiklis. Actor-critic algorithms. In *Advances in neural information processing systems*, pages 1008–1014, 2000.

V.S. Borkar. *Stochastic Approximation: A Dynamical Systems Viewpoint*. Texts and Readings in Mathematics. Hindustan Book Agency, 2009. ISBN 9788185931852. URL `https://books.google.ca/books?id=avOFtgAACAAJ`.

Harold Kushner and G George Yin. *Stochastic approximation and recursive algorithms and applications*, volume 35. Springer Science & Business Media, 2003.

Ziyu Wang, Victor Bapst, Nicolas Heess, Volodymyr Mnih, Rémi Munos, Koray Kavukcuoglu, and Nando de Freitas. Sample efficient actor-critic with experience replay. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*, 2017. URL `https://openreview.net/forum?id=HyM25Mqel`.

Scott Fujimoto, Herke van Hoof, and David Meger. Addressing function approximation error in actor-critic methods. *arXiv preprint arXiv:1802.09477*, 2018.