2022 American Control Conference (ACC)
Atlanta, USA, June 8-10, 2022

A Probabilistic Perspective on Risk-sensitive Reinforcement Learning

Erfaun Noorani and John S. Baras

Abstract— **Robustness is a key enabler of real-world applications of Reinforcement Learning (RL). The robustness properties of risk-sensitive controllers have long been established. We investigate risk-sensitive Reinforcement Learning (as a generalization of risk-sensitive stochastic control), by theoretically analyzing the risk-sensitive exponential (exponential of the total reward) criteria, and the benefits and improvements the introduction of risk-sensitivity brings to conventional RL. We provide a probabilistic interpretation of (I) the risk-sensitive exponential, (II) the risk-neutral expected cumulative reward, and (III) the maximum entropy Reinforcement Learning objectives, and explore their connections from a probabilistic perspective. Using Probabilistic Graphical Models (PGM), we establish that in the RL setting, maximization of the risk-sensitive exponential criteria is equivalent to maximizing the probability of taking an optimal action at all time-steps during an episode. We show that the maximization of the standard risk-neutral expected cumulative return is equivalent to maximizing a lower bound, particularly the Evidence lower Bound, on the probability of taking an optimal action at all time-steps during an episode. Furthermore, we show that the maximization of the maximum-entropy Reinforcement Learning objective is equivalent to maximizing a lower bound on the probability of taking an optimal action at all time-steps during an episode, where the lower bound corresponding to the maximum entropy objective is tighter and smoother than the lower bound corresponding to the expected cumulative return objective. These equivalences establish the benefits of risk-sensitive exponential objective and shed lights on previously postulated regularized objectives, such as maximum entropy. The utilization of a PGM model, coupled with exponential criteria, offers a number of advantages (e.g. facilitate theoretical analysis and derivation of bounds).**

I. Introduction

Reinforcement Learning (RL) studies a sequential decision problem under uncertainty in which an agent interacts with an unknown stochastic environment by choosing actions sequentially, based on its observations up the decision time, so as to optimize some desired system performance measure [1]. In classical Reinforcement Learning, the objective is to optimize expectation of some long run objective, such as discounted or undiscounted cumulative reward, that is to say, classical RL algorithms optimize a risk-neutral objective. Classical RL has been investigated for many years and is now well understood from the perspective of optimization and control. It has several weaknesses, with the most well-known ones being sensitivity to initial conditions, often unstable behavior, and non-robustness.

E. Noorani and J. Baras are with the Department of Electrical and Computer Engineering and the Institute for System Research (ISR) at the University of Maryland College Park, College Park, MD, USA. {enoorani,baras}@umd.edu

*E. Noorani is a Clark Doctoral Fellow at the Clark School of Engineering. Research partially supported by ONR grant N00014-17-1-2622, by a grant from the Army Research Lab and by the Clark Foundation.

To mitigate some of these shortcomings, regularized performance measures, such as maximum entropy RL objective, have been postulated which augment the classical risk-neutral objective with a regularization term; an entropy term in the case of maximum entropy objective [2]–[5]. For example, Soft Q-learning [4] and Soft Actor-critic (SAC) [5] algorithms use the Maximum Entropy regularized objective as the desired system performance measure. The policies learned by the maximum entropy objective have been seen to have typically better exploration and generalization characteristics, prevent premature convergence to sub-optimal policies, and lead to more robust policies. The motivation for Maximum Entropy RL has been an open question and been investigated in the literature from different perspectives, see [6] and the references therein.

The risk-neutral classical RL objectives are stemmed from the axioms of utility theory [7]. However, Prospect theory [8] has long established that humans do not decide only based on maximizing the expected value of some utility function, but rather they also consider the risk of their decisions. Thus giving rise to risk-sensitive (risk-aware) objectives. Small, albeit growing, number of results have been investigating risk-sensitive Reinforcement Learning, which incorporates some notion of risk, e.g. higher moments of return beyond expectation, into the desired system performance measure to be optimized [9]–[13]. A particular choice of such risk-sensitive objective is exponential criteria which have a long history in risk-sensitive control [14]–[16]. These criteria are of importance due to the known solution structure of risk-sensitive stochastic control and its equivalence to robust output feedback control and a dynamic game formulation of robust output feedback control which established the robustness properties of the risk-sensitive exponential criteria [17]–[21].

Conventionally, the RL problem is modeled as a Markov Decision Process (MDP) [22], where the reward signal, defining the task at hand, is an extrinsic signal. Probabilistic Graphical Model (PGM) [23], [24] offers an alternative modeling framework with a rich set of tools for inference for RL problems. RL has been previously formulated as a probabilistic inference problem on graphs using PGM, see a recent tutorial on this subject in [25]. In modeling Reinforcement Learning problems using PGM, the reward signal induces a distribution over random variables and is intrinsic to the model. While the MDP framework provides a powerful framework for modeling uncertainty, PGM provides a probabilistic perspective that may lead to a better understanding of the problem and ultimately more effective algorithms.

In this work, we investigate risk-sensitive reinforcement

learning (as a generalization of risk-sensitive stochastic control), and theoretically analyze the risk-sensitive RL. We present an interpretation of risk-sensitive Reinforcement Learning with exponential criteria by embedding it into a graph using a Probabilistic Graphical Model framework. We show that, for an RL problem with negative (resp. positive) reward structure, i.e. bounded reward, the maximization of the exponential criteria with positive (resp. negative) risk parameter is equivalent to maximizing the probability of taking an optimal action at all time-steps during an episode. We then explore the connection between risk-sensitive exponential criteria and the risk-neutral expected cumulative reward objective and show that optimizing the expected cumulative reward is equivalent to maximization of the Evidence Lower Bound on the probability of taking an optimal action at all time-steps during an episode. Furthermore, we show that the maximization of the maximum entropy objective is equivalent to the maximization of a lower bound on the probability of taking an optimal action at all time-steps during an episode, which is tighter and smoother than the Evidence lower Bound corresponding to expected cumulative return. That is to say, the maximization of expected cumulative reward and maximum entropy objectives can be interpreted as attempts towards approximately maximizing the probability of taking an optimal action at all time-steps during an episode. There is a general agreement that RL, and in particular risk-sensitive and robust RL, needs theoretical and mathematical proofs as foundations. The novelty and emphasis of the paper is a theoretical formulation and analysis of risk-sensitive RL, the benefits and improvements the introduction of risk-sensitivity brings to conventional RL, and its relation to other RL objectives. Probabilistic Graphical Model has been employed to illustrate the advantage of the risk-sensitive criteria and establish the relation between the risk-sensitive, regularized, and risk-neutral RL from a probabilistic point of view, and not to develop algorithms based on such models. The utilization of a PGM model together with exponential criteria offers a number of advantages (e.g. facilitate theoretical analysis and derivation of bounds).

II. Preliminaries

A. RL and Markov Decision Processes

Reinforcement Learning problem is conventionally modeled as a Markov Decision Process (MDP) [22]. An MDP is a tuple $\mathscr{M}=(\mathscr{S},\mathscr{A},p_1,P,r)$, where $\mathscr{S}$ is the state space and $\mathscr{A}$ is the action space, which in general may each be discrete or continuous; p_1 is the initial state distribution; $P:\mathscr{S}\times\mathscr{A}\to\Delta(\mathscr{S})$ is the transition kernel , which is in general unknown, where $\Delta(\mathscr{S})$ denotes the space of probability distributions on $\mathscr{S}$. The function $r:\mathscr{S}\times\mathscr{A}\to\mathbb{R}$ is the reward function, defining a task. In such environment, the behavior of an RL-agent is characterized by its policy. A (randomized) policy $\pi(\cdot|s)$ is a probability distribution over action space given state, which prescribes the probability of taking an action $a\in\mathscr{A}$ when in state $s\in\mathscr{S}$.

At each time-step t, when in state s_t, the RL-agent executes an action a_t according to a differentiable parametrized policy (possibly a Neural Network) $\pi(\cdot|s_t;\theta)$ where $\theta\in\mathbb{R}^d$ is a vector of d parameters. Upon the execution of the action, the system transitions to a successor state s_{t+1} according to transition probability $p(s_{t+1}|s_t,a_t)$, and the agent receives a reward $r_t:=r(s_t,a_t)$.

A trajectory τ is a sequence of states and actions

$$\tau=(s_1,a_1,s_2,a_2,\cdots,s_{|\tau|-1},a_{|\tau|-1},s_{|\tau|}).$$

The agents' policy and the system transition probabilities induce a trajectory distribution, a probability distribution over the sequence of states and actions, i.e. space of possible trajectories. The probability distribution induced over the space of trajectories by following the policy parameterized by θ is denoted by $\rho_\theta(\tau)$ and is given by

$$\rho_\theta(\tau)=p_1\prod_{t=1}^{T}\pi(a_t|s_t;\theta)p(s_{t+1}|s_t,a_t) \tag{1}$$

In the setting of finite-horizon RL, the agent aims to find a policy, so as to maximize the system performance measure over a given horizon denoted by T. Particularly, we consider the risk-neutral objective of expected cumulative return, the Maximum Entropy objective, and the risk-sensitive exponential (exponential of total reward) criteria.

In classical RL, the desired performance measure of the system is usually a risk-neutral objective, i.e., expectation of some long-run objective. A common example of a risk-neutral objective in RL literature is expected (undiscounted) cumulative reward, i.e.

$$J(\theta):=\mathbb{E}_{\tau\sim\rho_\theta}\Big[R(\tau)\Big] \tag{2}$$

where the expectation is taken under policy's trajectory distribution (parameterized by θ), i.e., $s_1\sim p_1$, $a_t\sim\pi(\cdot|s_t;\theta)$ and $s_{t+1}\sim p(\cdot|s_t,a_t)$; $R(\tau):=\sum_{t=1}^{T}r_t$ is the (undiscounted) sum of all rewards during an episode.

If needed, a discount factor can be incorporated into the model by modifying the transition dynamics, so that at each time-step, the system transitions to a terminal absorbing state with probability $1-\gamma$ and follows the original dynamics with probability γ.

Maximum Entropy is another popular RL objective, which augments the risk-neutral RL objective with an entropy regularization, i.e.,

$$J_{ent}(\theta):=\mathbb{E}_{\tau\sim\rho_\theta}\Big[R(\tau)\Big]+\lambda\mathbb{E}_{\substack{s_1\sim p_1\\ s_t\sim p(\cdot|s_{t-1},a_{t-1})}}\Big[\sum_{t=1}^{T}\mathscr{H}^\pi(a_t|s_t)\Big] \tag{3}$$

where $\mathscr{H}^\pi(a_t|s_t)=-\mathbb{E}_{a_t\sim\pi(\cdot|s_t;\theta)}\Big[\log\pi(a_t|s_t;\theta)\Big]$ is the entropy of policy π in state s_t, and the regularization weight λ is a real value non-negative constant. The weight λ is a design parameter that controls the level of regularization. The second term is the expected sum of entropy along the system trajectory. Maximum-entropy RL objective helps with exploration, prevents pre-mature convergence to sub-optimal policies, and provides better generalization, which leads to more robust policies.

In risk-sensitive RL, some notion of risk, e.g., higher moments of return, has been incorporated into the desired system

performance measure. Exponential criteria is a particular example of such a risk-sensitive objective, i.e.,

$$J_\beta(\theta) := \mathbb{E}_{\tau\sim\rho_\theta}\Big[\beta \exp\{\beta R(\tau)\}\Big] \tag{4}$$

where $\beta \in \mathbb{R}$ is a constant design parameter. Note that the exponential criteria approaches the risk-neutral objective as the risk parameter β approaches zero, i.e., $\lim_{\beta\to 0} J_\beta(\theta)=J(\theta)$.

The exponential criteria is well-studied in the context of risk-sensitive control [14]–[16]. Mathematical tractability of the exponential criteria, coupled with its intuitive appeal, makes it an attractive choice of performance measure.

B. RL and Probabilistic Graphical Models

It has been shown that Reinforcement Learning problem can be modeled as a Probabilistic Graphical Model (PGM) with factors of the form $p(s_{t+1}|s_t,a_t)$ by modeling relationship between state s_t, action a_t and successor-state s_{t+1}, and introduction of a fictitious binary optimality variable denoted by o_t at each time-step to incorporate the notion of reward into the graphical model, see figure 1 [25]. By conditioning on the optimality variables being true, one can infer the most probable policy.

The optimality variable is equal to one, $o_t=1$, if time-step t is optimal and equal to zero, $o_t=0$, if time-step t is not optimal (hence the name optimality variable). For brevity, we use o_t and o'_t to denote $o_t=1$ and $o_t=0$, respectively. We also use $O_{1:T}$ to denote the event that the optimal action was taken at each time-step during an episode, i.e., $O_{1:T}=(o_1,\cdots,o_T)$. The choice of the probability distribution of the optimality variable conditioned on the state-action pair $p(o_t|a_t,s_t) = p(o_t=1|a_t,s_t)$ defines the meaning of optimality and hence the task in hand.

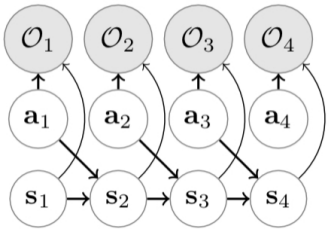

Fig. 1. Reinforcement Learning as a Graphical Model

The joint probability of observing trajectory τ and being optimal at all time-steps is given by

$$p(O_{1:T},\tau) = p_1 \prod_{t=1}^{T} p(a_t|s_t)p(s_{t+1}|s_t,a_t)p(o_t|a_t,s_t) \tag{5}$$

where the action prior is denoted by $p(a_t|s_t)$. We assume that the action prior $p(a_t|s_t)$ is a constant corresponding to a uniform distribution over the actions space. This assumption does not introduce any loss of generality, because any non-uniform prior $p(a_t|s_t)$ can be incorporated instead into $p(o_t|s_t,a_t)$ via the reward function as we shall see.

III. A Probabilistic Interpretation of Reinforcement Learning Objectives

In this section, we analyze the risk-sensitive exponential criteria and the benefits it brings to RL algorithms by offering a probabilistic interpretation of maximization of the exponential criteria (cf. Eq. (4)), the cumulative expected reward (cf. Eq. (2)), and the maximum entropy (3)) objectives and exploring the connection between them from a probabilistic perspective. In subsection III-A, we establish that the maximization of the risk-sensitive exponential criteria is equivalent to maximizing the probability of taking an optimal action at all time-steps during an episode. In subsection III-B, we show that the maximization of expected cumulative reward objective is equivalent to maximizing a lower bound on the probability of being optimal at all time-steps during an episode, that is to say, RL algorithms that optimize the expected cumulative reward objective attempt to approximately maximize the probability of being optimal at all time-steps during an episode. Subsection III-C provides an interpretation of maximum entropy RL and discusses the connections to exponential and risk-neutral objectives.

To that end, we first state our assumptions. We assume that the rewards are always negative. Note that the assumption of a negative reward structure is not a restrictive assumption. If the reward is bounded, i.e., $r{:}\mathscr{S}\times\mathscr{A}\to[r_{min},r_{max}]$, we can always satisfy this assumption by construction of an equivalent reward via subtracting the maximum reward. In other words, all results in this paper are invariant to subtraction of a constant from the reward function.

A. Risk-sensitive RL with Exponential Criteria

In this subsection, we aim to provide a probabilistic interpretation of the Risk-sensitive RL with exponential criteria (cf. Eq. 4) by casting the risk-sensitive RL problem into a graphical model.

We formally state our results in the following theorem, state a remark about the theorem, and end this section with the proof of our theorem.

Theorem 1: Under the assumption of negative reward structure, the maximization of the risk-sensitive exponential criteria of Eq. (4) with a positive risk parameter β,i.e., $J_\beta(\theta)=\mathbb{E}_{\tau\sim\rho_\theta}\Big[\beta\exp\{\beta R(\tau)\}\Big]$, is equivalent to maximization of the probability of being optimal at all time-steps during an episode $p(O_{1:T})=p(o_1=1,\cdots,o_T=1)$ for the choice of $p(o_t|s_t,a_t)=\pi(a_t|s_t;\theta)e^{\beta r_t}$ with the positive temperature parameter $1/\beta$, that is to say,

$$\operatorname*{argmax}_\theta J_\beta(\theta) = \operatorname*{argmax}_\theta p(O_{1:T})$$

Note that analogues results hold for positive reward structure and negative β parameter.

Remark 1: Theorem 1 provides a probabilistic view of Risk-sensitive RL with exponential criteria and justifies the choice of these criteria as a reasonable objective for an RL agent.

Proof: [Proof of Theorem 1] From the premise of the theorem, we have $p(o_t|s_t,a_t)=\pi(a_t|s_t;\theta)e^{\beta r_t}$

where β is a positive constant. Also, recall that $p(O_{1:T},\tau)=p_1\prod_{t=1}^{T}p(a_t|s_t)p(s_{t+1}|s_t,a_t)p(o_t|s_t,a_t)$ (cf. Eq. (5)), and $\rho_\theta(\tau)=p_1\prod_{t=1}^{T}\pi(a_t|s_t;\theta)p(s_{t+1}|s_t,a_t)$ (cf. Eq. (1)). Thus, using the property of exponentials, we have

$$p(O_{1:T},\tau) = p_1\prod_{t=1}^{T}p(a_t|s_t)p(s_{t+1}|s_t,a_t)p(o_t|a_t,s_t) \quad (6)$$

$$\propto \rho_\theta(\tau)e^{\beta R(\tau)}. \quad (7)$$

where $R(\tau)=\sum_{t=1}^{T}r_t$. The proportionality follows from the fact that the action prior $p(a_t|s_t)$ is a uniform distribution over the action space and is a constant for any state-action pair. By taking the integral (sum) of the joint probability $p(\tau,O_{1:T})$ with respect to all possible trajectories, we have

$$p(O_{1:T}) = \int_\tau p(\tau,O_{1:T})d\tau$$
$$\propto \int_\tau \rho_\theta(\tau)e^{\beta R(\tau)}d\tau \quad \propto \mathbb{E}_{\tau\sim\rho_\theta}\Big[\beta e^{\beta R(\tau)}\Big]$$

The first equality follows from the definition of marginal distribution, that is, for any two random variables X and Y, $p(X)=\int_Y p(X,Y)dy$. The proportionality in the second line is a straightforward use of Eq. (6) and noting that the proportionality constant is the same for any trajectory τ. The last proportionality follows immediately from the definition of expectation and the positivity of parameter β. Thus, the equivalence is established. ■

B. Risk-neutral Expected Cumulative Reward Objective

In this subsection, we show that the maximization of expected cumulative reward is equivalent to maximizing a lower bound on the probability of being optimal at all time-steps during an episode. To that end, we first state three lemmas that we will use to establish our results. Lemmas 1 and 2 give the Evidence Lower Bound for $p(O_{1:T})$, and multiple approaches to prove them have been suggested in the literature. For the convenience of reader and completeness, we include one approach for showing the bound here [26]. The proof for lemma 3 is of our own. Then, we formally state our results in Theorem 2 with the proof immediately following the theorem. We end this section with a brief discussion.

Lemma 1: The following equality holds:

$$\log p(O_{1:T}) = \mathbb{E}_{\tau\sim\rho_\theta}\Big[\log p(O_{1:T}|\tau)\Big] + D\Big(\rho_\theta(\tau)\|p(\tau|O_{1:T})\Big)$$

where $D(Q,P)$ is the KL-divergence between the probability distributions Q and P.

Proof: [Proof of Lemma 1]

$$D(\rho_\theta(\tau)\|p(\tau|O_{1:T})) = \mathbb{E}_{\tau\sim\rho_\theta}\Big[\log\frac{\rho_\theta(\tau)}{p(\tau|O_{1:T})}\Big]$$
$$= -\mathbb{E}_{\tau\sim\rho_\theta}\Big[\log\frac{p(\tau|O_{1:T})}{\rho_\theta(\tau)}\Big]$$
$$= -\mathbb{E}_{\tau\sim\rho_\theta}\Big[\log\frac{p(\tau,O_{1:T})}{\rho_\theta(\tau)} - \log p(O_{1:T})\Big]$$
$$= -\mathbb{E}_{\tau\sim\rho_\theta}\Big[\log\frac{p(\tau,O_{1:T})}{\rho_\theta(\tau)}\Big] + \log p(O_{1:T})$$
$$= -\mathbb{E}_{\tau\sim\rho_\theta}\Big[\log p(O_{1:T}|\tau)\Big] + \log p(O_{1:T})$$

The first line follows from the definition of Kullback-Leibler divergence. The second and third lines follow form properties of logarithm, and the forth line follows straightforward from the fact that $p(O_{1:T})$ is a constant with respect to τ. The last line follows from definition of conditional probability. Thus, by rearranging the terms, one can obtain the equality. ■

Lemma 2: The following lower bound holds on the probability of taking an optimal action at all time-steps during an episode $p(O_{1:T})$:

$$\log p(O_{1:T}) \geq \mathbb{E}_{\tau\sim\rho_\theta}\Big[\log p(O_{1:T}|\tau)\Big]$$

Proof: [Proof of Lemma 2]It follows straightforward from lemma 1 and non-negativity of KL-divergence. ■

The lower bound presented in lemma 2 is the Evidence Lower Bound on the probability of taking an optimal action at all time-steps during an episode, i.e.,

$$L = \mathbb{E}_{\tau\sim\rho_\theta}\Big[\log p(O_{1:T}|\tau)\Big] \quad (8)$$

Lemma 3: The Evidence Lower bound can be expressed as:

$$L = \beta\mathbb{E}_{\tau\sim\rho_\theta}\Big[R(\tau)\Big] + \sum_{t=1}^{T}\log p(a_t|s_t) \quad (9)$$

Proof: [Proof of lemma 3] It follows straightforward from lemma 2 that the Evidence lower bound is

$$L = \mathbb{E}_{\tau\sim\rho_\theta}\Big[\log p(O_{1:T},\tau)\Big] - \mathbb{E}_{\tau\sim\rho_\theta}\Big[\log\rho_\theta(\tau)\Big]$$
$$= \mathbb{E}_{\tau\sim\rho_\theta}\Big[\beta R(\tau)\Big] + \mathbb{E}_{\tau\sim\rho_\theta}\Big[\sum_{t=1}^{T}\log p(a_t|s_t)\Big]$$
$$= \beta\mathbb{E}_{\tau\sim\rho_\theta}\Big[R(\tau)\Big] + \sum_{t=1}^{T}\log p(a_t|s_t)$$

The first line follows from definition of conditional probability. The second line is obtained by substituting Eq's (1) and (5), substituting of $p(o_t|s_t,a_t)=\pi(a_t|s_t;\theta)e^{\beta r_t}$ in Eq. (5), and then using the sum property of logarithm. ■

Now, we are ready to state and proof our theorem.

Theorem 2: The maximization of expected cumulative reward objective, i.e.,

$$J(\theta) := \mathbb{E}_{\tau\sim\rho_\theta}\Big[R(\tau)\Big]$$

is equivalent to maximization of the Evidence Lower Bound on the probability of being optimal at all time-steps during an episode, i.e., $L=\mathbb{E}_{\tau\sim\rho_\theta}\Big[\log p(O_{1:T}|\tau)\Big]$.

Proof: It follows straightforward from lemmas 2 and 3 and positivity of β that

$$L \propto \mathbb{E}_{\tau\sim\rho_\theta}\Big[R(\tau)\Big] + (1/\beta)\sum_{t=1}^{T}\log p(a_t|s_t) \quad (10)$$

By noting that the action prior $p(a_t|s_t)$ is a uniform distribution over action space, and consequently, $\sum_{t=1}^{T}p(a_t|s_t)$ is a constant and does not depend on θ, we can see that the optimization of the Evidence Lower Bound is equivalent to maximizing the expected cumulative reward. ■

Theorem 2 can be interpreted to state that the expected cumulative reward objective attempts to approximately optimize the probability of being optimal at all time steps during

an episode by maximizing a lower bound, particularly the Evidence Lower Bound, on the probability of being optimal at all time steps during an episode.

Theorems 1 and 2 show that the optimization of the risk-sensitive exponential criteria is equivalent to maximizing the joint probability of taking and optimal action at all time-steps during an episode, resulting in a risk-seeking behavior, while the expected cumulative reward is an attempt to approximately solve this optimization by optimizing a lower bound on the probability of taking an optimal action at all time-steps.

C. Maximum Entropy Objective

We explore the connections of the maximum entropy objective with exponential criteria and expected cumulative return objective, and offer a mathematical and intuitive explanation for the maximum entropy objective, which justifies why maximum entropy objective results in more robust and improved policies. In particular, we show that the maximization of maximum entropy objective is an attempt to approximately solve a multi-objective optimization using scalarization method, which in turn is equivalent to maximization of a tighter and smoother lower bound on the probability of taking an optimal action at all time-steps during an episode than the Evidence Lower Bound corresponding to the maximization of the risk-neutral objective.

To that end, we first prove a lemma that in conjunction with our previous lemmas and theorems will help to establish the maximum entropy's connection to exponential and expected cumulative return objectives. We end this section by summarizing the connection between these objectives.

Lemma 4: The KL-divergence term in lemma 1 can be expressed in terms of the policy's entropy as follows

$$\begin{aligned} KL &= D\Big(\rho_\theta(\tau)\|p(\tau|O_{1:T})\Big) \\ &= -\mathbb{E}_{s_{t+1}\sim p(s_{t+1}|s_t,a_t)}\Big[\sum_{t=1}^{T}\mathscr{H}^\pi(\cdot|s_t)\Big] - \sum_{t=1}^{T}\log p(a_t|s_t) \end{aligned}$$

Proof: [proof of Lemma 4]

$$\begin{aligned} KL &= D\Big(\rho_\theta(\tau)\|p(\tau|O_{1:T})\Big) \\ &= \mathbb{E}_{\tau\sim\rho_\theta}\Big[\log\frac{\rho_\theta(\tau)}{p(\tau|O_{1:T})}\Big] \\ &= \mathbb{E}_{\tau\sim\rho_\theta}\Big[\log\frac{\rho_\theta(\tau)p(O_{1:T})}{p(O_{1:T},\tau)}\Big] \\ &= \mathbb{E}_{\tau\sim\rho_\theta}\Big[\log\frac{\prod_t\pi(a_t|s_t;\theta)}{\prod_t p(a_t|s_t)}\Big] \\ &= \sum_{t=1}^{T}\mathbb{E}_{\tau\sim\rho_\theta}\Big[\log\pi(a_t|s_t;\theta)\Big] - \sum_{t=1}^{T}\log p(a_t|s_t) \end{aligned}$$

The first line follows from the definition of Kullback–Leibler (KL) divergence. The second line follows for the fact that for any two random variables X and Y, $p(X|Y)=p(X,Y)p(X)/p(Y)$. The third equality follows from substituting $\rho_\theta(\tau)$ and $p(O_{1:T},\tau)$ using Eq's (1) and (5), and noting that the $p(O_{1:T})=\prod_{t=1}^{T}p(o_t|a_t,s_t)$.

By noting that $\mathbb{E}_{\tau\sim\rho_\theta}[\cdot]$ is an equivalent notation for $\mathbb{E}_{s_t\sim p(s_{t+1}|s_t;a_t)}\mathbb{E}_{a_t\sim\pi(a_t|s_t;\theta)}[\cdot]$, we have

$$\begin{aligned} KL &= \sum_{t=1}^{T}\mathbb{E}_{s_t\sim p(s_{t+1}|s_t;a_t)}\Big[\mathbb{E}_{a_t\sim\pi(a_t|s_t;\theta)}\Big[\log\pi(a_t|s_t;\theta)\Big]\Big] \\ &\quad - \sum_{t=1}^{T}\log p(a_t|s_t) \\ &= -\mathbb{E}_{s_t\sim p(s_{t+1}|s_t;a_t)}\Big[\sum_{t=1}^{T}\mathscr{H}^\pi(\cdot|s_t)\Big] - \sum_{t=1}^{T}\log p(a_t|s_t) \end{aligned}$$

where $\mathscr{H}^\pi(\cdot|s_t)=-\mathbb{E}_{a_t\sim\pi(a_t|s_t;\theta)}\Big[\log\pi(a_t|s_t;\theta)\Big]$ is the definition of shannon entropy. ■

Now, we are ready to explore the maximum entropy connections to the exponential and expected cumulative return objectives. Using lemma 1, we can see that the gap between the logarithm of probability of taking an optimal action at all time steps during an episode $\log p(O_{1:T})$, which is the objective that the exponential criteria optimize, and the Evidence lower bound L (cf. Eq. (8)), which is the objective that expected cumulative return optimizes, is

$$G = D\Big(\rho_\theta(\tau)\|p(\tau|O_{1:T})\Big)$$

Theorem 3: Under the assumption of negative reward structure, the maximization of the maximum entropy objective of Eq. (3),i.e.,

$$J_{ent}(\theta) := \mathbb{E}_{\tau\sim\rho_\theta}\Big[R(\tau)\Big] + \lambda\mathbb{E}_{\substack{s_1\sim p_1\\ s_t\sim p(\cdot|s_{t-1},a_{t-1})}}\Big[\sum_{t=1}^{T}\mathscr{H}^\pi(a_t|s_t)\Big]$$

is equivalent to maximizing a linear scalarization, i.e. a weighted sum of the objective functions, of the following multi-objective optimization

$$\max_\theta \quad (L,-G)$$

where $L=\mathbb{E}_{\tau\sim\rho_\theta}\Big[\log p(O_{1:T}|\tau)\Big]$ (cf. Eq. (8)) is the Evidence Lower Bound on the probability of taking an optimal action at all time-steps during an episode, and $G=D\Big(\rho_\theta(\tau)\|p(\tau|O_{1:T})\Big)$ (cf. Eq. (11)) is the gap between the Evidence Lower Bound and the log probability.

Proof: [Proof of Theorem 3] Using lemma 4, we have

$$G = -\mathbb{E}_{s_t\sim p(s_{t+1}|s_t,a_t)}\Big[\sum_{t=1}^{T}\mathscr{H}^\pi(\cdot|s_t)\Big] - \sum_{t=1}^{T}\log p(a_t|s_t) \quad (11)$$

Also, recall that using lemma 3, we have the Evidence Lower Bound $L=\beta\mathbb{E}_{\tau\sim\rho_\theta}\Big[R(\tau)\Big]+\sum_{t=1}^{T}\log p(a_t|s_t)$.

The proof is complete by scalarizing the multi-objective optimization as a single-objective optimization with a scalarization weight corresponding to the regularization weight $\lambda\beta$, and noting that the action prior $p(a_t|s_t)$ is a uniform distribution and therefore is constant with respect to θ. ■

Theorem 3 shows that the maximum entropy objective is equivalent to a linear scalarization of the multi-objective optimization involving simultaneously maximizing the Evidence Lower Bound on the probability of taking an optimal action

at all time-steps during an episode and minimizing the gap between the log probability and the Evidence Lower Bound.

Thus, the maximization of maximum entropy objective can be thought of as an attempt to trade-off the tightness of the lower bound on the log probability of taking an optimal action at all time-steps during an episode and the optimization of the lower bound, effectively trying to find and optimize a smooth lower bound, tighter than the Evidence Lower Bound, on the log probability $\log p(O_{1:T})$.

The solution to the maximum entropy objective is a Pareto optimal solution of the multi-objective problem corresponding to the given regularization weight λ. Given this perspective, one can use alternative methods to solve the multi-objective optimization, which might lead to more effective algorithms.

To summarize, we showed the optimization of the risk-sensitive exponential criteria is equivalent to maximizing the joint probability of taking an optimal action at all time-steps during an episode, while the expected cumulative reward and maximum entropy objectives are attempts to approximately solve this optimization by optimizing a lower bound on the probability of taking an optimal action at all time-steps, where the bound optimized by maximum entropy objective is a tighter and more smooth lower bound on the probability of being optimal at all time-steps than the bound optimized by the standard expected cumulative reward.

IV. Conclusion

We presented a probabilistic perspective on risk-sensitive RL with exponential criteria by casting the Reinforcement Learning problem into a graph using Probabilistic Graphical Models. We showed that Reinforcement Learning with exponential criteria has the intuitive interpretation of optimizing the probability of taking an optimal action at all time-steps during an episode. We also presented a probabilistic interpretation of risk-neutral expected cumulative reward. We established that risk-neutral expected cumulative reward Reinforcement Learning is equivalent to optimization of the Evidence lower bound on the probability of taking an optimal action at all time-steps. We also showed that the maximum entropy objective attempts to maximize a lower bound, tighter and more smooth than the Evidence lower bound, on the probability of taking an optimal action at all time-steps during an episode. That is precisely why the optimization of maximum entropy results in more robust policies.

References

[1] R. S. Sutton and A. G. Barto, *Reinforcement Learning: An Introduction*. Cambridge, MA: MIT Press, 2018.

[2] R. J. Williams and J. Peng, "Function Optimization using Connectionist Reinforcement Learning Algorithms," *Connection Science*, vol. 3, no. 3, pp. 241–268, 1991.

[3] B. D. Ziebart, A. Maas, J. A. Bagnell, and A. K. Dey, "Maximum Entropy Inverse Reinforcement Learning," in *Proceedings of the 23rd National Conference on Artificial Intelligence - Volume 3*, ser. AAAI'08. AAAI Press, 2008, p. 1433–1438.

[4] T. Haarnoja, H. Tang, P. Abbeel, and S. Levine, "Reinforcement Learning with Deep Energy-Based Policies," in *Proceedings of the 34th International Conference on Machine Learning - Volume 70*, ser. ICML'17. JMLR.org, 2017, p. 1352–1361.

[5] T. Haarnoja, A. Zhou, P. Abbeel, and S. 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*, ser. Proceedings of Machine Learning Research, J. Dy and A. Krause, Eds., vol. 80. Stockholmsmässan, Stockholm Sweden: PMLR, 10–15 Jul 2018, pp. 1861–1870. [Online]. Available: http://proceedings.mlr.press/v80/haarnoja18b.html

[6] B. Eysenbach and S. Levine, "If MaxEnt RL is the Answer, What is the Question?" *arXiv preprint arXiv:1910.01913*, 2019.

[7] O. Morgenstern and J. Von Neumann, *Theory of games and economic behavior*. Princeton university press, 1953.

[8] A. Tversky and D. Kahneman, "Advances in prospect theory: Cumulative representation of uncertainty," *Journal of Risk and uncertainty*, vol. 5, no. 4, pp. 297–323, 1992.

[9] Y. Chow and M. Ghavamzadeh, "Algorithms for CVaR Optimization in MDPs," *Advances in Neural Information Processing Systems*, vol. 27, pp. 3509–3517, 2014.

[10] L. A. Prashanth, "Policy Gradients for CVaR-Constrained MDPs," in *Algorithmic Learning Theory*. Cham: Springer International Publishing, 2014, pp. 155–169.

[11] A. Tamar, "Risk-Sensitive and Efficient Reinforcement Learning Algorithms," 2015.

[12] B. Liu, J. Liu, and K. Xiao, "R2PG: Risk-Sensitive and Reliable Policy Gradient," in *The Workshops of the The Thirty-Second AAAI Conference on Artificial Intelligence, New Orleans, Louisiana, USA, February 2-7, 2018*, ser. AAAI Workshops, vol. WS-18. AAAI Press, 2018, pp. 682–687.

[13] D. Nass, B. Belousov, and J. Peters, "Entropic Risk Measure in Policy Search," in *2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2019, pp. 1101–1106.

[14] D. Jacobson, "Optimal Stochastic Linear Systems With Exponential Performance Criteria and Their Relation to Deterministic Differential Games," *IEEE Transactions on Automatic Control*, vol. 18, no. 2, pp. 124–131, 1973.

[15] J. Speyer, J. Deyst, and D. Jacobson, "Optimization of Stochastic Linear Systems With Additive Measurement and Process Noise Using Exponential Performance Criteria," *IEEE Transactions on Automatic Control*, vol. 19, no. 4, pp. 358–366, 1974.

[16] P. Kumar and J. Van Schuppen, "On The Optimal Control of Stochastic Systems With an Exponential-of-integral Performance Index," *Journal of mathematical analysis and applications*, vol. 80, no. 2, pp. 312–332, 1981.

[17] M. R. James, J. S. Baras, and R. J. Elliott, "Risk-sensitive Control and Dynamic Games for Partially Observed Discrete-time Nonlinear Systems," *IEEE Transactions on Automatic Control*, vol. 39, no. 4, pp. 780–792, 1994.

[18] J. S. Baras and M. R. J. , "Robust and Risk-sensitive Output Feedback Control for Finite State Machines and Hidden Markov Models," *Journal of Mathematical Systems, Estimation, and Control*, vol. 7, no. 3, pp. 371–374, 1997.

[19] M. R. James and J. S. Baras, "Robust H_∞ output feedback control for nonlinear systems," *IEEE Transactions on Automatic Control*, vol. 40, no. 6, pp. 1007–1017, 1995.

[20] M. R. James and J. Baras, "Partially Observed Differential Games, Infinite-Dimensional Hamilton–Jacobi–Isaacs Equations, and Nonlinear H_∞ Control," *SIAM Journal on Control and Optimization*, vol. 34, no. 4, pp. 1342–1364, 1996.

[21] J. S. Baras and N. S. Patel, "Robust Control of Set-valued Discrete-time Dynamical Systems," *IEEE Transactions on Automatic Control*, vol. 43, no. 1, pp. 61–75, 1998.

[22] M. L. Puterman, *Markov Decision Processes: Discrete Stochastic Dynamic Programming*. John Wiley & Sons, 2014.

[23] M. J. Wainwright and M. I. Jordan, *Graphical Models, Exponential Families, and Variational Inference*. Now Publishers Inc, 2008.

[24] D. Koller and N. Friedman, *Probabilistic graphical models: principles and techniques*. MIT press, 2009.

[25] S. Levine, "Reinforcement learning and control as probabilistic inference: Tutorial and review," *arXiv preprint arXiv:1805.00909*, 2018.

[26] X. Yang, "Understanding The Variational Lower Bound," 2017.

[27] A. T. Dotan Di Castro and S. Mannor, "Policy Gradients With Variance Related Risk Criteria," in *Proceedings of the 29th International Conference on Machine Learning, Edinburgh, Scotland, UK*, 2012.