

---

# Direct Policy Gradients: Direct Optimization of Policies in Discrete Action Spaces

---

Guy Lorberbom  
Technion

Chris J. Maddison  
Oxford & DeepMind

Nicolas Heess  
DeepMind

Tamir Hazan  
Technion

Daniel Tarlow  
Google Research, Brain Team

## Abstract

Direct optimization [24] is an appealing approach to differentiating through discrete quantities [35, 19]. Rather than relying on REINFORCE or continuous relaxations of discrete structures, it uses optimization in discrete space to compute gradients through a discrete argmax operation. In this paper, we develop reinforcement learning algorithms that use direct optimization to compute gradients of the expected return in environments with discrete actions. We call the resulting algorithms *direct policy gradient* algorithms and investigate their properties, showing that there is a built-in variance reduction technique and that a parameter that was previously viewed as a numerical approximation can be interpreted as controlling risk sensitivity. We also tackle challenges in algorithm design, leveraging ideas from A\* Sampling [21] to develop a practical algorithm. Empirically, we show that the algorithm performs well in illustrative domains, and that it can make use of domain knowledge about upper bounds on return-to-go to speed up training.

## 1 Introduction

Direct optimization [24] is a promising but relatively underutilized approach to computing gradients through a discrete argmax operation. McAllester et al. [24] introduce the method and apply it to structured prediction in the supervised learning setting. Song et al. [35] use it to train deep networks with application-specific losses. Lorberbom et al. [19] apply it to generative learning and show that it improves over relaxations based on Gumbel-Softmax [23, 12] in Variational Autoencoders with discrete hidden states.

The questions in this paper are the following: How can we apply direct optimization to reinforcement learning (RL) of policies with discrete action spaces? How does the resulting algorithm compare to standard algorithms from the literature? As we will show, there is significant depth to these questions.

Our first contribution is to adapt the idea in Lorberbom et al. [19] to the RL setting. This yields an expression for a policy gradient that differs from the standard ones and suggests a different algorithmic approach, of replacing the sampling of trajectories with the optimization of trajectories according to a noisy objective function. It is interesting because it yields a stable update where parameters are only updated if the optimization finds an improvement in a certain sense, and it naturally allows incorporating domain knowledge (like in A\* search) to speed up its computation of a policy gradient.

Our second contribution is addressing algorithmic challenges. At first glance, the method appears to require generating exponentially many random variables for each update. However, we will show how to resolve the problem using ideas from A\* Sampling of Maddison et al. [21]. We develop a practical algorithm and explore issues that arise in its design.

Our third contribution is developing new understandings and connections. We show that there is an interpretation in terms of Stochastic Optimal Control [31], although there is a variance reduction idea implicit in the new expression that to our knowledge has not previously appeared in policy gradient formulations. This analysis also yields a new understanding of alternative direct loss updates from McAllester et al. [24], termed “towards good” and “away from bad”. While equivalent in a limit, we show that they have different risk-sensitive behavior in the numerical approximation used in practice.

Finally, we evaluate the new algorithm in environments that highlight its distinguishing properties. In total, this work gives a new perspective on the fundamental problem of computing a policy gradient and opens the door to many future directions.

## 2 Preliminaries

**The reinforcement learning problem.** We consider a standard problem of RL, in which an agent interacts with a Markov Decision Process (MDP) for a finite number of steps<sup>1</sup> and attempts to maximize its reward. At any given time  $t \geq 0$  the environment is in some state  $s_t \in \mathcal{S}$  in the given state space  $\mathcal{S}$ ; there is a fixed initial state  $s_0 \in \mathcal{S}$ . At each time  $t$  the agent interacts with the environment by taking an action  $a_t$  from a finite set of actions  $a_t \in \mathcal{A}$  according to a policy parameterized by  $\theta \in \mathbb{R}^d$ ,  $\pi_\theta(a_t | s_t)$ . The environment then follows a transition distribution  $p(r_t, s_{t+1} | s_t, a_t)$  over rewards  $r_t$  and next states  $s_{t+1}$  given previous state  $s_t$  and action  $a_t$ . The agent interacts with the environment in this way for  $T > 0$  steps generating a sequence of states  $\mathbf{s} = (s_1, \dots, s_T)$ , actions  $\mathbf{a} = (a_0, \dots, a_{T-1})$ , and rewards  $\mathbf{r} = (r_0, \dots, r_{T-1})$ . This corresponds to the following generative model,

$$\begin{aligned} a_t &\sim \pi_\theta(\cdot | s_t) \text{ for } t \in \{0, \dots, T-1\} \\ r_t, s_{t+1} &\sim p(\cdot, \cdot | a_t, s_t) \text{ for } t \in \{0, \dots, T-1\} \end{aligned} \quad (1)$$

given  $s_0 \in \mathcal{S}$ . Taken together this defines the following joint distribution,

$$p_\theta(\mathbf{a}, \mathbf{s}, \mathbf{r}) = \prod_{t=0}^{T-1} \pi_\theta(a_t | s_t) p(r_t, s_{t+1} | s_t, a_t). \quad (2)$$

The sum of rewards  $r_t$  over an interaction is called the return, and the goal of the agent is to maximize the expected return over its policy parameters,  $\max_{\theta \in \mathbb{R}^d} \mathbb{E}_{\mathbf{a}, \mathbf{s}, \mathbf{r} \sim p_\theta} \left[ \sum_{t=0}^{T-1} r_t \right]$ .

**Gumbel-max reparameterizations.** A random variable  $G \sim \text{Gumbel}(m)$  is Gumbel-distributed with location  $m$  if  $p(G \leq g) = \exp(-\exp(-g + m))$ . The Gumbel-max trick is a way of casting sampling from a softmax as an argmax computation by using the fact that if  $G(i)$  are drawn i.i.d. as  $\text{Gumbel}(m_i)$ , then  $i^* = \operatorname{argmax}_i G(i) \sim \exp(m_i) / \sum_{i'} \exp m_{i'}$ . Moreover,  $G^* = \max_i G(i) \sim \text{Gumbel}(\log \sum_{i'} \exp m_{i'})$  and  $i^*$  and  $G^*$  are independent random variables. See [9, 21, 20].

## 3 Direct Policy Gradient

Here we introduce direct optimization to approximate the gradient of the expected return of a policy. The result is our *direct policy gradient (DirPG)* expression, which we name as such because it uses the ideas of direct optimization to compute a policy gradient. Full algorithms are developed in Sec. 5.

**Sampling Trajectories as Optimization on State-Reward Trees.** Our approach depends on a reparameterization of the standard model (1) that separates the environmental stochasticity from the stochasticity in an agent’s choices. The high-level idea is that for each realization of the MDP, we could have pre-sampled the state transitions for all sequences of actions and structured the results in a tree, assuming the ability to reset the environment to a previously visited state. Then afterwards, by instantiating stochasticity in the agent’s policy, choose a path from the root to a leaf to get a trajectory distributed the same as the standard MDP model. It is not tractable to instantiate the tree up front, and our eventual algorithms will lazily instantiate the needed subtrees, but this view will help exposition.

<sup>1</sup>Technically, everything in the paper works with an unbounded numbers of steps as long as trajectories terminate with probability 1, but we assume a maximum number of steps to simplify some parts of the exposition.

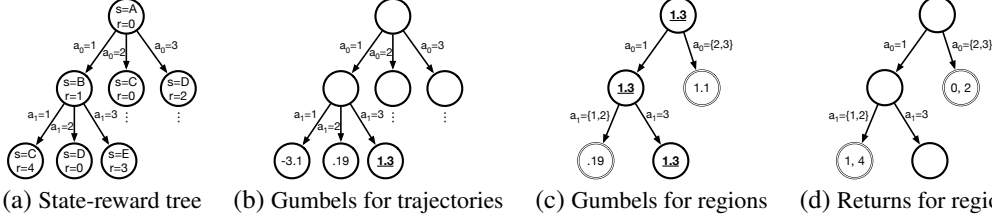


Figure 1: Example state-reward tree and associated values. **(a)** A state-reward tree for  $T = 2$  timesteps. Subtrees under  $a_0 \in \{2, 3\}$  are omitted for space. **(b)** Gumbel values  $G_\theta(\mathbf{a}; \Gamma, \mathbf{S})$  associated with each trajectory  $\mathbf{a}$ . The trajectory with maximum value (underlined) is  $\mathbf{a}_{opt}$ . **(c)** State of the search tree after sampling  $\mathbf{a}_{opt}$ . Gumbels for regions are also included. Nodes on the queue are drawn with double outline. **(d)** Return-so-far and upper bound on return-to-go (respectively) for nodes on the queue.

In more detail, consider the  $|\mathcal{A}|^T$  possible sequences of actions  $\mathbf{a} \in \mathcal{A}^T$ . These can be organized into a perfect  $|\mathcal{A}|$ -ary tree, whose root is associated with the empty sequence, and each node of depth  $t > 0$  is associated with a prefix  $(a_0, \dots, a_{t-1})$  of length  $t$ . Each valid action sequence  $\mathbf{a} \in \mathcal{A}^T$  corresponds to a leaf node. With each node  $\mathbf{a} \in \mathcal{A}^t$  for  $t \geq 0$  we can associate the random variables  $r_{\mathbf{a}}, s_{\mathbf{a}}$ . If  $\mathbf{a} = \emptyset$ , then  $r_{\mathbf{a}} = 0, s_{\mathbf{a}} = s_0$ . Otherwise,  $r_{\mathbf{a}}, s_{\mathbf{a}} \sim p(\cdot, \cdot \mid a_t, s_{\tilde{\mathbf{a}}})$  where  $\tilde{\mathbf{a}} = (a_0, \dots, a_{t-1})$  is the prefix of  $\mathbf{a}$ . See Fig. 1 (a). Taken together, this defines a distribution over two random variables per node of the tree. We call this the *state-reward tree* and denote it  $\mathbf{S} = (r_{\mathbf{a}}, s_{\mathbf{a}} \mid \mathbf{a} \in \mathcal{A}^t, 0 \leq t \leq T)$  with distribution  $P$ . The advantage of this view is that we can formulate the simulation of  $(\mathbf{a}, \mathbf{s}, \mathbf{r})$  according to (2) in two steps; first the simulation of a random state-reward tree  $\mathbf{S}$ . Then, treating  $\mathbf{S}$  as a deterministic environment, the simulation of a trajectory on  $\mathbf{S}$  using the policy  $\pi_\theta$  to choose actions. Specifically, define the conditional distribution over action sequences given a state-reward tree as

$$\Pi_\theta(\mathbf{a} \mid \mathbf{S}) = \prod_{t=0}^{T-1} \pi_\theta(a_t \mid s_{(a_0 \dots a_{t-1})}). \quad (3)$$

Given  $\mathbf{S}$  and  $\mathbf{a} \in \mathcal{A}^T$ , there is one sequence of states and rewards, corresponding to a traversal of the state-reward tree taking action  $a_t$  at depth  $t - 1$ . Moreover  $\mathbf{a} \sim \Pi_\theta(\cdot \mid \mathbf{S})$  has exactly the marginal distribution as  $\mathbf{a}$  in  $p_\theta$  of (2). Now we can reparameterize the sampling of  $\mathbf{a}$  using Gumbel-max:

$$\Gamma(\mathbf{a}) \sim \text{Gumbel}(0) \quad G_\theta(\mathbf{a}; \Gamma, \mathbf{S}) = \log \Pi_\theta(\mathbf{a} \mid \mathbf{S}) + \Gamma(\mathbf{a}) \quad \mathbf{a}^* = \arg\max_{\mathbf{a}} G_\theta(\mathbf{a}; \Gamma, \mathbf{S}). \quad (4)$$

$G_\theta$  are then distributed as Gumbels with shifted locations and  $\mathbf{a}^*$  is a sample from (3). We define the return of a trajectory  $\mathbf{a}$  on  $\mathbf{S}$ ,  $R(\mathbf{a}, \mathbf{S}) = \sum_{t=0}^{T-1} r_{(a_0, \dots, a_{t-1})}$ , where the dependence on  $\mathbf{S}$  comes implicitly through the  $r$ 's. Putting everything together, we get

$$\mathbb{E}_{\mathbf{a}, \mathbf{s}, \mathbf{r} \sim p_\theta} \left[ \sum_{t=0}^{T-1} r_t \right] = \mathbb{E}_{\mathbf{S} \sim P} [\mathbb{E}_{\mathbf{a} \sim \Pi_\theta(\cdot \mid \mathbf{S})} [R(\mathbf{a}, \mathbf{S})]] = \mathbb{E}_{\mathbf{S} \sim P, \Gamma} [R(\mathbf{a}^*, \mathbf{S})]. \quad (5)$$

**Direct Policy Gradient.** The above reparameterization allows us to extend [19] to the RL setting. The derivation starts by defining a *direct objective*  $D_\theta$  and *prediction generating function*  $f$ :

$$D_\theta(\mathbf{a}; \Gamma, \mathbf{S}, \epsilon) = G_\theta(\mathbf{a}; \Gamma, \mathbf{S}) + \epsilon R(\mathbf{a}, \mathbf{S}), \quad (6)$$

$$f(\theta, \epsilon) = \mathbb{E}_{\mathbf{S} \sim P, \Gamma} \left[ \max_{\mathbf{a}} \{D_\theta(\mathbf{a}; \Gamma, \mathbf{S}, \epsilon)\} \right], \quad (7)$$

$$\mathbf{a}^*(\epsilon) = \arg\max_{\mathbf{a}} D_\theta(\mathbf{a}; \Gamma, \mathbf{S}, \epsilon). \quad (8)$$

When clear from context, we drop the explicit dependence on noise terms  $\mathbf{S}$  and  $\Gamma$  for brevity. Differentiating  $f$  with respect to  $\epsilon$  and  $\theta$  in either order and evaluating at  $\epsilon = 0$  yields the same value, because  $f$  is smooth [19] (or see [35] for an alternative proof). Thus,

$$\frac{\partial}{\partial \theta_i} \mathbb{E} [R(\mathbf{a}^*(0), \mathbf{S})] = \frac{\partial^2 f(\theta, \epsilon)}{\partial \theta_i \partial \epsilon} \bigg|_{\epsilon=0} = \frac{\partial^2 f(\theta_i, \epsilon)}{\partial \epsilon \partial \theta_i} \bigg|_{\epsilon=0} = \frac{\partial}{\partial \epsilon} \mathbb{E} \left[ \frac{\partial}{\partial \theta_i} \log \Pi_\theta(\mathbf{a}^*(\epsilon) \mid \mathbf{S}) \right] \bigg|_{\epsilon=0}. \quad (9)$$

Note that if  $\epsilon = 0$  then (8) reduces to (4) and is a trajectory sampled from the current policy. When  $\epsilon$  deviates from 0, (8) is a trajectory that is close to a sample from the current policy but that has higher or lower return, where the strength and direction of this pull comes from the magnitude and sign of  $\epsilon$ .

A finite-difference approximation in  $\epsilon$  of the RHS of (9) yields the direct policy gradient (DirPG),

$$\frac{1}{\epsilon} \mathbb{E}_{\mathbf{S} \sim P, \Gamma} [\nabla_{\theta} \log \Pi_{\theta} (\mathbf{a}^*(\epsilon) \mid \mathbf{S}) - \nabla_{\theta} \log \Pi_{\theta} (\mathbf{a}^*(0) \mid \mathbf{S})]. \quad (10)$$

Following terminology of [24] we name  $\mathbf{a}_{opt} = \mathbf{a}^*(0)$  as the optimum in Eq. 4, and  $\mathbf{a}_{dir} = \mathbf{a}^*(\epsilon)$  as the trajectory that defines the update direction. Because the LHS of (9) is the gradient of the expected return, DirPG approaches the standard policy gradient as  $\epsilon \rightarrow 0$ .

**Algorithms.** The general form of algorithms we consider is given in Algorithm 1. The basis is a TrajectoryGenerator that generates pairs of trajectories  $\mathbf{a}$  and associated direct objectives  $D_{\theta}(\mathbf{a}; \epsilon)$  via a search over trajectories. The algorithm is not executable until we describe how to implement the trajectory generator and lazily instantiate  $\mathbf{S}$  and  $\Gamma$  in Sec. 5. The first step of Algorithm 1 is to find  $\mathbf{a}_{opt}$  and  $d_{opt} = D_{\theta}(\mathbf{a}_{opt}; 0)$  and initialize  $\mathbf{a}_{dir} = \mathbf{a}_{opt}, d_{dir} = d_{opt}$ . Our generators in Sec. 5 naturally produce  $\mathbf{a}_{opt}$  and  $d_{opt}$  as the first result, so we assume that behavior. The algorithm then applies heuristic search to find a trajectory  $\mathbf{a}_{dir}$  with direct objective  $d_{dir}$  better than  $d_{opt}$  (lines 5-13). If no improvement is found before a budget is exceeded, then  $\mathbf{a}_{opt}$  is equal to  $\mathbf{a}_{dir}$  and the result of line 14 is a zero gradient. One option is to terminate the search upon finding any improvement (line 9). This is desirable because it automatically adapts the search budget as training progresses. At first it is easy to improve over  $\mathbf{a}_{opt}$  (a sample from a random policy), but more search is needed after training for longer. Given enough budget and no early termination, the algorithm exactly implements (10).

---

**Algorithm 1** Direct Policy Gradient (General Form)

---

```

1:  $\mathbf{S} \sim P(\mathbf{S})$ 
2:  $\Gamma(\mathbf{a}) \sim \text{Gumbel}(0)$  for all  $\mathbf{a}$ 
3: trajectories = TrajectoryGenerator( $\mathbf{S}, \Gamma, \epsilon$ )
4:  $\mathbf{a}_{opt}, d_{opt} \leftarrow \mathbf{a}_{dir}, d_{dir} \leftarrow \text{trajectories.next}()$ 
5: while budget not exceeded do
6:    $\mathbf{a}_{cur}, d_{cur} \leftarrow \text{trajectories.next}()$ 
7:   if  $d_{cur} > d_{dir}$  then
8:      $\mathbf{a}_{dir}, d_{dir} \leftarrow \mathbf{a}_{cur}, d_{cur}$ 
9:     if terminate on first improvement then
10:       break
11:   end if
12: end while
13: end while
14: return  $\frac{1}{\epsilon} \nabla_{\theta} [\log \pi_{\theta} (\mathbf{a}_{dir} \mid \mathbf{S}) - \log \pi_{\theta} (\mathbf{a}_{opt} \mid \mathbf{S})]$ 

```

---

**Limitation.** The expectation in (10) is over variables that determine the realization of noise in the environment and policy, which are shared between  $\mathbf{a}_{dir}$  and  $\mathbf{a}_{opt}$ . To compute an update, we need to hold fixed the realization of noise in the environment and consider alternative actions, which assumes the ability to reset the environment to previously visited states  $s \in \mathcal{S}$ . However, in principle one can train a policy using these algorithms in simulation and then execute the learned policy in the real world where it is not possible to reset the environment.

## 4 Properties

In this section we develop key properties of the DirPG update. These are derived by developing an alternate interpretation of DirPG as the gradient of some other function, namely

$$l(\theta, \epsilon) = \mathbb{E}_{\mathbf{S} \sim P} \left[ \frac{1}{\epsilon} \log \left( \mathbb{E}_{\mathbf{a} \sim \Pi_{\theta}(\cdot \mid \mathbf{S})} [\exp(\epsilon R(\mathbf{a}, \mathbf{S}))] \right) \right], \quad (11)$$

$$\nabla_{\theta} l(\theta, \epsilon) = \frac{1}{\epsilon} \mathbb{E}_{\mathbf{S} \sim P} \left[ \mathbb{E}_{\mathbf{a} \sim P_R(\cdot \mid \mathbf{S})} [\nabla_{\theta} \log \Pi_{\theta} (\mathbf{a} \mid \mathbf{S})] - \mathbb{E}_{\mathbf{a} \sim \Pi_{\theta}(\cdot \mid \mathbf{S})} [\nabla_{\theta} \log \Pi_{\theta} (\mathbf{a} \mid \mathbf{S})] \right], \quad (12)$$

where  $P_R(\mathbf{a} \mid \mathbf{S}) \propto \Pi_{\theta} (\mathbf{a} \mid \mathbf{S}) \exp(\epsilon R(\mathbf{a}, \mathbf{S}))$ . The derivation is in the Appendix. This reveals an interpretation of DirPG as having a built-in control variate and risk-sensitive behavior when  $\epsilon \neq 0$ .

**Control Variate Interpretation.** The key step in the derivation of (10) from (12) is reparameterizing the expectations in (12) using Gumbel-max and expressing the samples in terms of (8):

$$= \frac{1}{\epsilon} \mathbb{E}_{\mathbf{S} \sim P} [\mathbb{E}_{\Gamma} [\nabla_{\theta} \log \Pi_{\theta} (\mathbf{a}^*(\epsilon) \mid \mathbf{S})] - \mathbb{E}_{\Gamma} [\nabla_{\theta} \log \Pi_{\theta} (\mathbf{a}^*(0) \mid \mathbf{S})]]. \quad (13)$$

Having expressed both expectations in terms of Gumbel noise  $\Gamma$  with the same distribution, we can use common random numbers to recover the direct policy gradient (10).

The last term of (12) has expected value of 0. The benefit of including it only becomes apparent in (10), where we can interpret it as a control variate. The optimization problems that define  $\mathbf{a}_{dir}$

and  $\mathbf{a}_{opt}$  differ only in value of  $\epsilon$ , so for small  $\epsilon$  we expect the solutions to have similar features and correlated score functions. When this is the case, control variates reduce the variance of the overall gradient estimate. To our knowledge, this formulation of control variate is novel, though at a high level it resembles other uses of control variates in machine learning [32].

**Risk-sensitivity.** The objective (11) is closely related to a classical objective in risk-sensitive control [29, 11, 8],  $\log \mathbb{E} [\exp(\epsilon R(\mathbf{a}, \mathbf{S}))] / \epsilon$ . For  $\epsilon > 0$ , optimal policies under the classical objective prefer high risk strategies as long as high rewards have some positive probability. For  $\epsilon < 0$ , optimal policies prefer low risk strategies that avoid placing probability on low rewards. (11) has an important difference. Following [8, 22], we take a Taylor expansion of  $\exp(t)$  and  $\log(1+t)$  at  $t=0$  to get

$$l(\theta, \epsilon) = \mathbb{E}_{\mathbf{S} \sim P, \mathbf{a} \sim \Pi_\theta(\cdot | \mathbf{S})} [R(\mathbf{a}, \mathbf{S})] + \frac{\epsilon}{2} \mathbb{E}_{\mathbf{S} \sim P} [\text{var}_{\mathbf{a} \sim \Pi_\theta(\cdot | \mathbf{S})} (R(\mathbf{a}, \mathbf{S}))] + \mathcal{O}(\epsilon^2), \quad (14)$$

where we use the notation  $\text{var}_{\mathbf{a} \sim \Pi_\theta(\cdot | \mathbf{S})} (R(\mathbf{a}, \mathbf{S}))$  to mean the conditional variance of  $R(\mathbf{a}, \mathbf{S})$  given  $\mathbf{S}$ . Note that expected conditional variance is not equal to the joint variance, which makes this objective different from the typical risk-sensitive analysis. If the second term were simply the variance under the joint, then the agent is sensitive to variance in return regardless of whether it was due to stochasticity in the environment or in the policy. In (14), we see that the agent only seeks out or suppresses “controllable risk,” which is variance in return created due to stochasticity in its policy.

**Analysis of Approximate Update.** We prove in the Appendix that one can search just until first improvement in Algorithm 1 and converge to the same result as fully searching the space in multi-armed bandits with deterministic rewards. This provides a sanity check on the approximation and formalizes a notion of “Gumbel-approximate-max” sampling that is good enough for learning.

## 5 Concrete Algorithms to Search for $\mathbf{a}_{opt}$ and $\mathbf{a}_{dir}$

**Search Space.** The search over  $\mathbf{S}$  for  $\mathbf{a}_{opt}$  and  $\mathbf{a}_{dir}$  is structured into a search tree over sets of action sequences that share a common prefix that we refer to as *regions*. Region  $\mathcal{R}(\tilde{\mathbf{a}}, \mathcal{B}; \mathbf{S})$  is the set of trajectories that start with prefix  $\tilde{\mathbf{a}} = (a_0, \dots, a_{t-1})$  and then take a next action from  $\mathcal{B} \subseteq \mathcal{A}$ . The root of the search tree  $\mathcal{R}(\emptyset, \mathcal{A})$  is the set of all trajectories. An example search tree is shown in Fig. 1 (c). The root region (top) represents the set of all trajectories and its right child represents the set of trajectories  $\{\mathbf{a} : a_0 \in \{2, 3\}\}$ .

The search tree is expanded by choosing a region  $\mathcal{R} = \mathcal{R}(\tilde{\mathbf{a}}, \mathcal{B}; \mathbf{S})$  from a search queue and a next action  $a_t \in \mathcal{B}$ .  $\mathcal{R}$  is split into two child regions. The first appends  $a_t$  to the prefix and then takes any next action; i.e.,  $\mathcal{R}_1 = \mathcal{R}(\tilde{\mathbf{a}} \oplus a_t, \mathcal{A})$ . The second leaves the prefix unchanged and eliminates  $a_t$  as a possible next action; i.e.,  $\mathcal{R}_2 = \mathcal{R}(\tilde{\mathbf{a}}, \mathcal{B} \setminus \{a_t\})$ . If  $s_{\tilde{\mathbf{a}} \oplus a_t}$  is a terminal state then  $\mathcal{R}_1$  contains a single trajectory and is not expanded further. If  $\mathcal{B} \setminus \{a_t\}$  is empty, then  $\mathcal{R}_2$  can be discarded. In Fig. 1 (c), the first split chose  $a_0 = 1$  and created regions  $\mathcal{R}_1 = \mathcal{R}((1), \mathcal{A})$  and  $\mathcal{R}_2 = \mathcal{R}(\emptyset, \mathcal{A} \setminus \{1\})$ .

**Lazily Sampling  $\mathbf{S}$ .** Because  $\mathbf{S}$  is exponentially large in  $T$ , we sample it lazily as we expand the search tree. When creating  $\mathcal{R}_1$  based on action  $a_t$ , we can simply call the environment’s step function  $r_{\tilde{\mathbf{a}} \oplus a_t}, s_{\tilde{\mathbf{a}} \oplus a_t} \sim p(\cdot, \cdot | a_t, s_{\tilde{\mathbf{a}}})$  to realize a node of  $\mathbf{S}$  with the right distribution.

**Searching for large  $G_\theta$ .** We have shown how to construct the search tree given choices of  $a_t$ , but ultimately we want to choose  $a_t$  that lead to trajectories with large  $D_\theta$ . Here we show how to optimally extend any prefix  $\tilde{\mathbf{a}}$  into a trajectory  $\mathbf{a}$  that has maximum value of  $G_\theta(\mathbf{a}; \mathbf{S}, \Gamma)$ . The key ideas come from the Top-Down and A\* sampling algorithms of Maddison et al. [21]. Our extension to the RL setting appears in Algorithm 2, which is a trajectory generator that can be used in Algorithm 1. Code can be found in the Supplementary Materials. The basis of the algorithm is a *Gumbel Process*, which—in our terms—is a (consistent) assignment of Gumbel-distributed random variables to all subsets of trajectories. The marginal distribution of the Gumbel in a region is  $G_\theta(\mathcal{R}; \mathbf{S}, \Gamma) \sim \text{Gumbel}(\log \Pi_\theta(\mathcal{R} | \mathbf{S}))$  where  $\Pi_\theta(\mathcal{R} | \mathbf{S}) = \sum_{\mathbf{a} \in \mathcal{R}} \Pi_\theta(\mathbf{a} | \mathbf{S})$ , and consistency constraints enforce that  $G_\theta(\mathcal{R}_1 \cup \mathcal{R}_2) = \max(G_\theta(\mathcal{R}_1), G_\theta(\mathcal{R}_2))$ .

To avoid the need for up-front instantiation of noise, we assume that Algorithm 2 is given an environment *env* and action space  $\mathcal{A}$  instead of a pre-instantiated  $\mathbf{S}$  and  $\Gamma$ . It then lazily samples the needed parts of  $\mathbf{S}$  and  $\Gamma$ . Line 8 adds a node to  $\mathbf{S}$ , and we assume the additions persist

in Algorithm 1 so that gradient calculations can use them. Lazy instantiation of  $\Gamma$  comes from executing the Top-Down algorithm, which constructs a Gumbel Process from root to leaves. The algorithm begins by sampling  $G_\theta(\mathcal{R})$  for the root region  $\mathcal{R}$  that contains all trajectories (line 4). Then trajectories are divided as in our search space above (line 17 corresponds to  $\mathcal{R}_1$ ; lines 9-13 correspond to  $\mathcal{R}_2$ ), and upon creation of new regions, their  $G_\theta$  values are sampled conditional upon the parent's  $G_\theta$  value. These conditionals either copy the parent's value (line 17) or are Truncated Gumbel distributions (line 11). For the algorithm to be tractable, we need to compute  $\Pi_\theta(\mathcal{R} | \mathcal{S})$  (line 10), which can be done by pushing the sum inwards through the shared prefix:  $\Pi_\theta(\mathcal{R}(\tilde{\mathbf{a}}, \mathcal{B}; \mathcal{S}) | \mathcal{S}) = \left( \prod_{t'=0}^{t-1} \pi_\theta(a_{t'} | s_{(a_0, \dots, a_{t'-1})}) \right) \sum_{a \in \mathcal{B}} \pi_\theta(a | s_{\tilde{\mathbf{a}}})$ .

There are subtle conceptual differences to [21] and to an alternative version in Kim et al. [13]. In short, the version from [13] is closer to what is needed, but both [21] and [13] would roll-out an entire trajectory for each region expanded and thus make less efficient use of interactions with the environment. We discuss these details and how Algorithm 2 fixes the problem in the Appendix. A minor difference is that we yield pairs of  $(\mathbf{a}, D_\theta(\mathbf{a}; \epsilon))$  (line 15) as opposed to  $(\mathbf{a}, G_\theta(\mathbf{a}))$  pairs.

Algorithm 2 generates trajectories and associated Gumbels from a consistent realization of a Gumbel process, but the order in which they are generated depends on the order in which regions are removed from the queue. If we use a priority queue with priority  $G_\theta(\mathcal{R})$ , then the algorithm will yield pairs in descending order of  $G_\theta(\mathbf{a})$ , which also means that the first pair yielded will be  $(\mathbf{a}_{opt}, D_\theta(\mathbf{a}_{opt}; \epsilon))$ . We assume regions are always prioritized this way until the first yield so that line 4 in Algorithm 1 is correct. We are then free to change the priority function as in the next subsection and reorder the queue. However if we do not, then this gives an alternative implementation of the "stochastic beams" of [15] (generating trajectories with largest  $G_\theta(\mathbf{a})$ ), though Algorithm 2 is not a beam search.

---

**Algorithm 2** Top-Down Sampling  $\mathbf{a}$ 


---

```

1: In: environment  $env$ , actions  $\mathcal{A}$ ,  $\epsilon$ .
2: Out: Stream of  $(\mathbf{a}, G_\theta(\mathbf{a}))$  pairs.
3:  $Q, \mathcal{S} \leftarrow$  Queue, StateRewardTree
4:  $Q.push(\emptyset, \mathcal{A}, \text{Gumbel}(0))$ 
5: while  $Q$  is not empty do
6:    $\tilde{\mathbf{a}}, \mathcal{B}, G \leftarrow Q.pop()$ 
7:    $\mathbf{a} \leftarrow \text{Sample } \pi_\theta(\mathbf{a} | s_{\tilde{\mathbf{a}}}) \mathbb{1}\{\mathbf{a} \in \mathcal{B}\}$ 
8:    $s_{\tilde{\mathbf{a}} \oplus \mathbf{a}}, r_{\tilde{\mathbf{a}} \oplus \mathbf{a}} \leftarrow env.step(\mathbf{a}, s_{\tilde{\mathbf{a}}})$ 
9:   if  $\mathcal{B} \setminus \{\mathbf{a}\}$  is not empty then
10:     $\mu \leftarrow \log \Pi_\theta(\mathcal{R}(\tilde{\mathbf{a}}, \mathcal{B} \setminus \{\mathbf{a}\}) | \mathcal{S})$ 
11:     $G' \leftarrow \text{TruncGumbel}(\mu, G)$ 
12:     $Q.push(\tilde{\mathbf{a}}, \mathcal{B} \setminus \{\mathbf{a}\}, G')$ 
13:   end if
14:   if  $s_{\tilde{\mathbf{a}} \oplus \mathbf{a}}$  is terminal then
15:     yield  $(\tilde{\mathbf{a}} \oplus \mathbf{a}, G + \epsilon R(\tilde{\mathbf{a}} \oplus \mathbf{a}, \mathcal{S}))$ 
16:   else
17:      $Q.push(\tilde{\mathbf{a}} \oplus \mathbf{a}, \mathcal{A}, G)$ 
18:   end if
19: end while

```

---

**Searching for large  $D_\theta$  using  $A^*$  sampling.** Our final algorithm prioritizes regions on the queue using the return achieved so far and (if available) an upper bound on the return-to-go. It is the same as Algorithm 2, except before pushing a region on the queue (lines 4, 12, 17), we compute a priority for a region based on all the terms in (6). Let  $L(\mathcal{R}) = \sum_{t'=0}^{t-1} r_{(a_0, \dots, a_{t'-1})}$  be the reward accumulated so far by the prefix and  $U(\mathcal{R}) \geq \sum_{t'=t}^T r_{(a_0, \dots, a_{t'-1})}$  be an upper bound on the return-to-go for any trajectory in region  $\mathcal{R}$ . An example appears in Fig. 1 (d). Recall the  $G_\theta(\mathcal{R})$  computed during the search is the maximum  $G_\theta$  for any trajectory in the region. We can then upper bound  $D_\theta(\mathcal{R}; \epsilon) = \max_{\mathbf{a} \in \mathcal{R}} D_\theta(\mathbf{a}; \epsilon) \leq G_\theta(\mathcal{R}) + \epsilon \cdot (L(\mathcal{R}) + U(\mathcal{R}))$ . We can also prune regions from the search if their upper bound is worse than  $D_\theta(\mathbf{a}; \epsilon)$  for the best  $\mathbf{a}$  found so far. Using the upper bound as a priority yields a stochastic version of  $A^*$  search (i.e., it is  $A^*$  Sampling [21]). In practice, we have found it better to use a priority like  $G_\theta(\mathcal{R}) + \epsilon \cdot (L(\mathcal{R}) + \alpha U(\mathcal{R}))$  for  $0 \leq \alpha < 1$ . See Sec. 6.3.

## 6 Experiments

### 6.1 Combinatorial Bandits

We first experiment with combinatorial bandits and compare DirPG to Upper Confidence Bound (UCB) algorithms [2, 5]. The environment is defined by a graph  $G = (V, E)$  where  $V = \{1, \dots, n\}$  is the set of nodes and  $E \subseteq V \times V$  is the set of undirected edges. For each edge  $e \in E$  there is a real-valued parameter  $\mu_e$  that determines a per-edge reward distribution as  $r_e \sim \text{Uniform}(0, 2\mu_e)$ . An agent queries the environment for the reward of a tree  $r_{\mathcal{T}} = \sum_{e \in \mathcal{T}} r_e$ . Fresh realizations of  $r_e$  are drawn for each episode. UCB algorithms end an episode after a single interaction, while DirPG uses multiple interactions per episode (at the cost of seeing fewer realizations). As baselines, we use a privileged "semi-bandit" version of UCB that observes per-edge rewards and a "full bandit" version

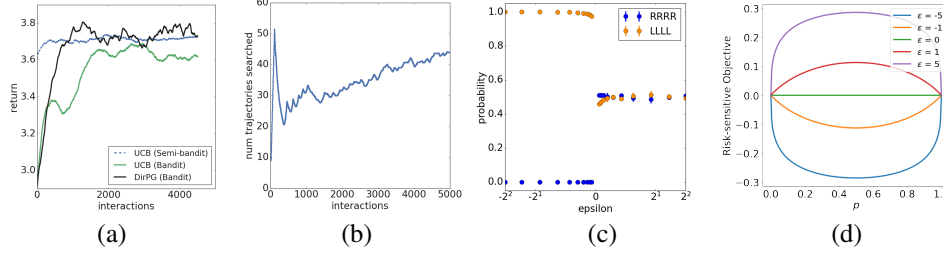


Figure 2: Combinatorial bandits and risk sensitivity results. (a) Moving average return vs number of interactions. (b) Number of steps needed to find  $\mathbf{a}_{dir}$ . (c) DeepSea results showing learned  $\Pi(\text{LLLL})$  (safe) and  $\Pi(\text{RRRR})$  (risky) vs epsilon. (d) Quadrature evaluation of (11) for the Gaussian choice problem for varying  $\epsilon$ .

that assumes the per-tree rewards are attributed evenly to the edges, i.e.,  $r_e = \frac{r_{\mathcal{T}}}{n-1}$ . Both baselines choose a tree at time  $t$  by computing a maximum spanning tree given upper confidence bound edge costs  $u_e = \hat{\mu}_e + \frac{1.5 \log t}{c_e}$  where  $\hat{\mu}_e$  is the average per-edge reward for edge  $e$  and  $c_e$  is the number of times edge  $e$  has been chosen.

To apply DirPG, we let  $\mathbf{a}$  be a sequence of  $|E|$  binary decisions of whether to include each edge in the spanning tree. Learnable parameters  $\theta_e$  determine the probability of inclusion via  $\sigma(\theta_e)$  where  $\sigma$  is the sigmoid function. The environment helps to construct a spanning tree by presenting a legal set of actions at each step (see Appendix). It is still possible to generate an invalid spanning tree, in which case we continue searching over trajectories in descending order of  $G_{\theta}(\mathbf{a})$  until finding a valid tree. After producing a full tree  $\mathcal{T}$ , we observe  $r_{\mathcal{T}}$ . To compute  $\mathbf{a}_{dir}$ , we give a budget of 100 interactions and use priority  $G_{\theta}(\mathbf{a})$  in the search, enabling the early termination option in Algorithm 1.

Results appear in Fig. 2 (a). The plots show the moving average return as a function of interactions averaged over 10 runs. The black curve evaluates samples from the policy  $\mathbf{a}_{opt}$ , which is noisier due to there being fewer realizations. DirPG with bandit feedback is competitive with a UCB variant using semi-bandit feedback and that it outperforms the bandit feedback variant. Fig. 2 (b) shows the number of steps taken to find an improvement. Aside from initial noise due to the moving average, the number of interactions used in the search automatically grows as learning progresses.

## 6.2 DeepSea

Here we empirically study the risk-sensitive behavior analyzed in Sec. 4. We use an adaptation of the DeepSea environment from [27] and vary  $\epsilon$ , which controls risk sensitivity. The environment is a 5x5 grid where the agent starts from the top-left cell and the goal is in the bottom-right. The agent has a choice of left (L) or right (R) at each step. If the agent chooses L, it gets 0 reward and moves down and left. If it chooses R, it gets a reward sampled from  $\mathcal{N}(1, 1)$  if transitioning to the bottom-right corner and otherwise  $-\frac{1}{3}$ . This is interesting because any policy that is a mixture of LLLL and RRRR has optimal return (mixture of 0,  $\mathcal{N}(0, 1)$  respectively), but the policies have different variance and thus we expect the choice of  $\epsilon$  to affect what the agent learns.

In Fig. 2 (c) we train policies with a range of  $\epsilon$  values for 400,000 episodes to ensure convergence and plot the probability assigned to trajectories LLLL and RRRR in the learned policy. For  $\epsilon < 0$ , most mass is put on LLLL, which has no variance and is thus favorable to a risk-avoiding agent. For  $\epsilon > 0$ , mass is split evenly, which has highest controllable risk. To further illustrate this, we used numerical integration to compute (11) for a simplified ‘‘Gaussian choice’’ setting where an agent chooses to take a reward sampled from  $\mathcal{N}(0, 1)$  with probability  $p$  and 0 reward with probability  $1 - p$ . Fig. 2 (d) shows that the risk-seeking objective favors ‘‘controllable risk’’ created due to stochasticity in the agent’s policy but not variance created due to stochasticity in the environment, as discussed in Sec. 4.

## 6.3 MiniGrid

In our final experiments we use the **MiniGrid-MultiRoom-N6-v0** environment [7] to study how to prioritize nodes within the search for  $\mathbf{a}_{dir}$ . MiniGrid is a partially observable grid-world where the agent observes an egocentric  $7 \times 7$  grid around its current location and has the choice of 7 actions including moving right, left, forward, or toggling doors. We use environments of  $25 \times 25$  grids with

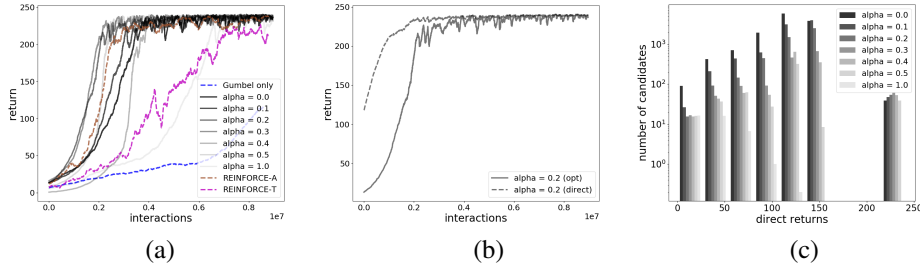


Figure 3: Minigrid results. (a) Return vs number of interactions. (b) Direct objective of  $\mathbf{a}_{dir}$  and  $\mathbf{a}_{opt}$  vs iteration. (c) Histograms showing quality-vs-quantity tradeoff for various search priorities.

a series of 6 connected rooms separated by doors that need to be opened. Intermediate rewards are given for opening doors and reaching a final goal state. As baselines we compare to REINFORCE and the cross entropy method. Details on their implementation are in the Appendix.

We explore variations on how to set the priority of nodes in the search for  $\mathbf{a}_{dir}$ . First, in the “Gumbel only” priority, we use just  $G_\theta(\mathcal{R})$  as a region’s priority. In the others, we use  $G_\theta(\mathcal{R}; \mathbf{S}, g) + \epsilon(L(\mathcal{R}) + \alpha U(\mathcal{R}))$ , where  $U$  is based on the Manhattan distance to the goal and the number of unopened doors. Setting  $\alpha = 0$  trades off enumerating by descending order of  $G_\theta(\mathcal{R}; \mathbf{S}, g)$  with favoring prefixes that have already achieved high return. Setting  $\alpha = 1$  yields A\* search. Fig. 3 (a) shows average return versus training episode.  $\alpha = 0$  provides good results, and increasing  $\alpha$  up to  $\alpha = 0.3$  gives improved performance. Beyond that, performance degrades, with  $\alpha = 1$  performing worst.

To better understand this, we partially trained a model for 1.2M interactions and then froze the parameters and ran several searches for the same number of interactions but with different priority functions. Fig. 3 (c) shows the results. For smaller  $\alpha$ , more trajectories are finished to completion but the returns achieved are worse. As  $\alpha$  increases, fewer full trajectories are found but they have better returns, but past  $\alpha = 0.4$  not enough full trajectories are found, and both the quality and the quantity shrink. Thus, setting  $\alpha$  too high leads to “breadth-first behavior” where too much time is spent exploring prefixes and not completing trajectories. A good heuristic should have some tendency towards “depth-first behavior,” rolling out some promising trajectories to the end even if they are unlikely to be optimal. In Fig. 3 (b), we show the relationship between  $D_\theta(\mathbf{a}_{dir})$  and  $D_\theta(\mathbf{a}_{opt})$  over the course of learning. This shows  $\mathbf{a}_{dir}$  “pulling up”  $\mathbf{a}_{opt}$  and that  $\mathbf{a}_{dir}$  does not need to find a trajectory with the optimal return in order to provide signal for the policy to improve.

## 7 Related Work

As discussed in Sec. 4 the objective (12) bears some similarity to the objectives from the body of work casting RL as probabilistic inference, in particular in Expectation-Maximization (EM) Policy Search methods [28, 36, 30, 17, 18, 25, 6, 1, 4]. Broadly, these methods alternate a step akin to posterior inference that improves the trajectory distribution with an update to the policy parameters using in an EM formulation. In this context our work could be interpreted as an incremental variant [26] of Monte Carlo EM [16] to yield an update similar to the first term in (12). Relative to this approach, our novelty would be the introduction of the control variate, the approximate optimization of the argmax function when drawing the sample, and the adaptation of A\* sampling to guide the sampling.

The basis of our formulation is a reparameterization that is similar to [10], except they focus on continuous actions and otherwise develop a very different approach. The most prominent example of search in RL is Monte Carlo Tree Search (MCTS) [14, 3]. On its own, MCTS is quite different from our approach, but it becomes similar when search results are distilled into a policy as in [34]. However, we are not aware of results showing that MCTS can be used to directly compute a policy gradient. Another related use of search trees is the *vine* method from [33], which leverages a simulator’s ability to reset to previous states to construct a tree over trajectories. Multiple roll-outs are created from tree nodes, and common random numbers are used across the roll-outs to reduce variance.



## 8 Discussion

We have presented a new method for computing a policy gradient and studied its properties from theoretical and empirical perspectives. This also provides new understandings of direct loss optimization in terms of variance reduction and risk-sensitivity. One limitation is that in its current form, the algorithm only learns in an episodic framework and from complete trajectories. We are currently exploring how this limitation could be removed. Our experiments so far have been geared towards understanding the algorithm and its important degrees of freedom. We are eager to take these learnings and apply them to real-world applications where search and heuristics (upper bounds) have traditionally been successful; perhaps in domains with a navigation component or program synthesis.

## References

- [1] Abbas Abdolmaleki, Jost Tobias Springenberg, Yuval Tassa, Rémi Munos, Nicolas Heess, and Martin A. Riedmiller. Maximum a posteriori policy optimisation. *CoRR*, abs/1806.06920, 2018.
- [2] Peter Auer. Using confidence bounds for exploitation-exploration trade-offs. *Journal of Machine Learning Research*, 3(Nov):397–422, 2002.
- [3] Cameron B Browne, Edward Powley, Daniel Whitehouse, Simon M Lucas, Peter I Cowling, Philipp Rohlfshagen, Stephen Tavener, Diego Perez, Spyridon Samothrakis, and Simon Colton. A survey of monte carlo tree search methods. *IEEE Transactions on Computational Intelligence and AI in games*, 4(1):1–43, 2012.
- [4] Lars Buesing, Theophane Weber, Yori Zwols, Sebastien Racaniere, Arthur Guez, Jean-Baptiste Lespiau, and Nicolas Heess. Woulda, coulda, shoulda: Counterfactually-guided policy search. *arXiv preprint arXiv:1811.06272*, 2018.
- [5] Nicolo Cesa-Bianchi and Gábor Lugosi. Combinatorial bandits. *Journal of Computer and System Sciences*, 78(5):1404–1422, 2012.
- [6] Yevgen Chebotar, Mrinal Kalakrishnan, Ali Yahya, Adrian Li, Stefan Schaal, and Sergey Levine. Path integral guided policy search. *CoRR*, abs/1610.00529, 2016.
- [7] Maxime Chevalier-Boisvert, Dzmitry Bahdanau, Salem Lahlou, Lucas Willems, Chitwan Saharia, Thien Huu Nguyen, and Yoshua Bengio. BabyAI: First steps towards grounded language learning with a human in the loop. In *International Conference on Learning Representations*, 2019. URL <https://openreview.net/forum?id=rJeXC0cYX>.
- [8] Stefano Coraluppi. *Optimal control of Markov decision processes for performance and robustness*. PhD thesis, University of Maryland, 1997.
- [9] Emil Julius Gumbel. *Statistical theory of extreme values and some practical applications: a series of lectures*, volume 33. US Government Printing Office, 1954.
- [10] Nicolas Heess, Gregory Wayne, David Silver, Timothy Lillicrap, Tom Erez, and Yuval Tassa. Learning continuous control policies by stochastic value gradients. In *Advances in Neural Information Processing Systems*, pages 2944–2952, 2015.
- [11] Ronald A Howard and James E Matheson. Risk-sensitive markov decision processes. *Management science*, 18(7):356–369, 1972.
- [12] Eric Jang, Shixiang Gu, and Ben Poole. Categorical Reparameterization with Gumbel-Softmax. In *International Conference on Learning Representations*, 2017.
- [13] Carolyn Kim, Ashish Sabharwal, and Stefano Ermon. Exact sampling with integer linear programs and random perturbations. In *Thirtieth AAAI Conference on Artificial Intelligence*, 2016.
- [14] Levente Kocsis and Csaba Szepesvári. Bandit based monte-carlo planning. In *European conference on machine learning*, pages 282–293. Springer, 2006.

- [15] Wouter Kool, Herke van Hoof, and Max Welling. Stochastic beams and where to find them: The gumbel-top-k trick for sampling sequences without replacement. *arXiv preprint arXiv:1903.06059*, 2019.
- [16] Richard A Levine and George Casella. Implementations of the monte carlo em algorithm. *Journal of Computational and Graphical Statistics*, 10(3):422–439, 2001.
- [17] Sergey Levine and Vladlen Koltun. Variational policy search via trajectory optimization. In *Advances in Neural Information Processing Systems*, pages 207–215, 2013.
- [18] Sergey Levine, Chelsea Finn, Trevor Darrell, and Pieter Abbeel. End-to-end training of deep visuomotor policies. *Journal of Machine Learning Research*, 17(39):1–40, 2016.
- [19] Guy Lorberbom, Andreea Gane, Tommi Jaakkola, and Tamir Hazan. Direct Optimization through arg max for Discrete Variational Auto-Encoder. *arXiv e-prints*, art. arXiv:1806.02867, Jun 2018.
- [20] Chris J. Maddison and Daniel Tarlow. Gumbel machinery. <https://cmaddis.github.io/gumbel-machinery>. Accessed: 2019-05-21.
- [21] Chris J. Maddison, Daniel Tarlow, and Tom Minka. A\* Sampling. In *Advances in Neural Information Processing Systems* 27, 2014.
- [22] Chris J. Maddison, Dieterich Lawson, George Tucker, Nicolas Heess, Arnaud Doucet, Andriy Mnih, and Yee Whye Teh. Particle value functions. *arXiv preprint arXiv:1703.05820*, 2017.
- [23] Chris J. Maddison, Andriy Mnih, and Yee Whye Teh. The Concrete Distribution: A Continuous Relaxation of Discrete Random Variables. In *International Conference on Learning Representations*, 2017.
- [24] David A McAllester, Tamir Hazan, and Joseph Keshet. Direct loss minimization for structured prediction. In *Advances in Neural Information Processing Systems*, pages 1594–1602, 2010.
- [25] William Montgomery and Sergey Levine. Guided policy search as approximate mirror descent. *CoRR*, abs/1607.04614, 2016. URL <http://arxiv.org/abs/1607.04614>.
- [26] Radford M Neal and Geoffrey E Hinton. A view of the em algorithm that justifies incremental, sparse, and other variants. In *Learning in graphical models*, pages 355–368. Springer, 1998.
- [27] Brendan O’Donoghue. Variational bayesian reinforcement learning with regret bounds. *arXiv preprint arXiv:1807.09647*, 2018.
- [28] Jan Peters, Katharina Mülling, and Yasemin Altün. Relative entropy policy search. In *Proceedings of the Twenty-Fourth AAAI Conference on Artificial Intelligence (AAAI)*, 2010.
- [29] John W Pratt. Risk aversion in the small and in the large. *Econometrica*, 32(1/2):122–136, 1964.
- [30] Konrad Rawlik, Marc Toussaint, and Sethu Vijayakumar. On stochastic optimal control and reinforcement learning by approximate inference. In *(R:SS 2012)*, 2012. *Runner Up Best Paper Award*.
- [31] Konrad Rawlik, Marc Toussaint, and Sethu Vijayakumar. On stochastic optimal control and reinforcement learning by approximate inference. In *Twenty-Third International Joint Conference on Artificial Intelligence*, 2013.
- [32] Geoffrey Roeder, Yuhuai Wu, and David K Duvenaud. Sticking the landing: Simple, lower-variance gradient estimators for variational inference. In *Advances in Neural Information Processing Systems*, pages 6925–6934, 2017.
- [33] John Schulman, Sergey Levine, Pieter Abbeel, Michael Jordan, and Philipp Moritz. Trust region policy optimization. In *International Conference on Machine Learning*, pages 1889–1897, 2015.

- [34] David Silver, Thomas Hubert, Julian Schrittwieser, Ioannis Antonoglou, Matthew Lai, Arthur Guez, Marc Lanctot, Laurent Sifre, Dharmashan Kumaran, Thore Graepel, et al. A general reinforcement learning algorithm that masters chess, shogi, and go through self-play. *Science*, 362(6419):1140–1144, 2018.
- [35] Yang Song, Alexander Schwing, Raquel Urtasun, et al. Training deep neural networks via direct loss minimization. In *International Conference on Machine Learning*, pages 2169–2177, 2016.
- [36] M. Toussaint and A.J. Storkey. Probabilistic inference for solving discrete and continuous state Markov Decision Processes. In *Proceedings of 23rd International Conference on Machine Learning (ICML 2006)*, 2006.

## A Proofs & Additional Details on Properties

### A.1 Direct Policy Gradient as the derivative of another function.

The direct policy gradient update (10) is the derivative of (11). To derive (12), first divide by 1,

$$l(\theta, \epsilon) = \frac{1}{\epsilon} \mathbb{E}_{\mathbf{S} \sim P} \left[ \log \frac{\sum_{\mathbf{a}} \exp \{ \log \Pi_{\theta}(\mathbf{a} | \mathbf{S}) + \epsilon R(\mathbf{a}, \mathbf{S}) \}}{\sum_{\mathbf{a}} \exp \{ \log \Pi_{\theta}(\mathbf{a} | \mathbf{S}) \}} \right], \quad (15)$$

and differentiate to get

$$\nabla_{\theta} l(\theta, \epsilon) = \frac{1}{\epsilon} \mathbb{E}_{\mathbf{S} \sim P} \left[ \mathbb{E}_{\mathbf{a} \sim P_R(\cdot | \mathbf{S})} [\nabla_{\theta} \log \Pi_{\theta}(\mathbf{a} | \mathbf{S})] - \mathbb{E}_{\mathbf{a} \sim \Pi_{\theta}(\cdot | \mathbf{S})} [\nabla_{\theta} \log \Pi_{\theta}(\mathbf{a} | \mathbf{S})] \right], \quad (16)$$

where  $p_R(\mathbf{a} | \mathbf{S}) = \frac{1}{Z_R} \exp \{ \log \Pi_{\theta}(\mathbf{a} | \mathbf{S}; \theta) + \epsilon R(\mathbf{a}, \mathbf{S}) \}$ .

### A.2 Proof of Correctness of Gumbel-approx-max in Deterministic Multi-armed Bandits

Suppose we have  $N$  arms, each with a fixed but unknown reward  $R(i)$  and that arms are ordered according to their reward so  $R(i) > R(j)$  iff  $i > j$ , and  $\epsilon > 0$ . Let the following:

- $\pi_{\theta}(i) \propto \exp \theta_i$  be the probability of arm  $i$  under a softmax policy parameterized by  $\theta$ ,
- $G_{\theta}(i) \sim \text{Gumbel}(\theta_i)$
- $D_{\theta}(i, \epsilon) = G_{\theta}(i) + \epsilon R(i)$  be the direct objective
- $i_{opt} = \arg \max_i D_{\theta}(i, 0)$
- $i_{dir} = \arg \max_i D_{\theta}(i, \epsilon)$

Finally, let  $i_{approx}$  be the value of  $i_{direct}$  arising from running Algorithm 1 using  $G_{\theta}(i)$  as priority. That is, we iterate over  $i$  in descending order of  $G_{\theta}(i)$  until we find an  $i$  such that  $D_{\theta}(i, \epsilon) > D_{\theta}(i_{opt}, \epsilon)$  or we have enumerated all  $i$ , in which case we set  $i_{approx} = i_{opt}$ .

We prove that learning using  $i_{approx}$  in place of  $i_{dir}$  still leads to learning the optimal policy.

**Lemma 1.**  $i_{direct} \geq i_{approx} \geq i_{opt}$ .

*Proof.* To prove  $i_{approx} \geq i_{opt}$ , observe that by definition we have  $D_{\theta}(i_{approx}, \epsilon) \geq D_{\theta}(i_{opt}, \epsilon)$  and  $G_{\theta}(i_{opt}) \geq G_{\theta}(i_{approx})$ . This implies

$$G_{\theta}(i_{approx}) + \epsilon R(i_{approx}) \geq G_{\theta}(i_{opt}) + \epsilon R(i_{opt}) \quad (17)$$

$$\epsilon R(i_{approx}) - \epsilon R(i_{opt}) \geq G_{\theta}(i_{opt}) - G_{\theta}(i_{approx}) \geq 0. \quad (18)$$

Thus  $R(i_{approx}) \geq R(i_{opt})$  and  $i_{approx} \geq i_{opt}$ .

To prove  $i_{dir} \geq i_{approx}$  observe that we must have  $G_{\theta}(i_{approx}) \geq G_{\theta}(i_{dir})$ , because otherwise we would have encountered  $i_{dir}$  before  $i_{approx}$  when iterating  $i$ 's, and because  $D_{\theta}(i_{dir}, \epsilon) \geq D_{\theta}(i_{approx}, \epsilon)$  by definition, we would have chosen  $i_{dir}$  as  $i_{approx}$  when we encountered it.

So we have  $G_{\theta}(i_{approx}) - G_{\theta}(i_{dir}) \geq 0$ , which implies

$$G_{\theta}(i_{dir}) + \epsilon R(i_{dir}) \geq G_{\theta}(i_{approx}) + \epsilon R(i_{approx}) \quad (19)$$

$$\epsilon R(i_{dir}) - \epsilon R(i_{approx}) \geq G_{\theta}(i_{approx}) - G_{\theta}(i_{dir}) \geq 0 \quad (20)$$

$$(21)$$

Thus  $R(i_{dir}) \geq R(i_{approx})$  and  $i_{dir} \geq i_{approx}$ .  $\square$

**Lemma 2.** *We're at a stationary point iff  $i_{\text{direct}} = i_{\text{opt}}$  (or  $i_{\text{approx}} = i_{\text{opt}}$ ) almost surely.*

*Proof.* In one direction, if  $i_{\text{direct}} = i_{\text{opt}}$  almost surely, then DirPG updates on 0 almost surely. In the other direction, suppose for the sake of contradiction that there is some realization of  $G_\theta$  where  $i_{\text{direct}}$  is not equal to  $i_{\text{opt}}$ . By Lemma 1,  $i_{\text{direct}} > i_{\text{opt}}$ . Then the gradient vector will have a positive entry for  $\theta_{i_{\text{direct}}}$  and a negative entry for  $\theta_{i_{\text{opt}}}$ . In order to be at a stationary point, other realizations of  $G_\theta$  need to cancel these contributions. Because of Lemma 1, however, it is only possible to simultaneously decrement the gradient vector at  $i$  and increment it at  $j$  if  $j > i$ . The only way to decrement the previously incremented entry for  $i_{\text{direct}}$  would be to increment an even larger entry, and the only way to increment the previously decremented entry for  $i_{\text{opt}}$  would be to decrement an even smaller entry. Thus, there is no way to cancel gradients if any entry is nonzero, and thus the only way to get a zero gradient is if  $i_{\text{direct}} = i_{\text{opt}}$  for all realizations of  $G_\theta$ . In Lemma 1 we have  $i_{\text{direct}} \geq i_{\text{approx}} \geq i_{\text{opt}}$ , so the same argument holds for  $i_{\text{approx}}$ .  $\square$

**Proposition 1.** *The stationary points assuming exact optimization of  $i_{\text{direct}}$  are the same as the stationary points assuming approximate optimization to get  $i_{\text{approx}}$ .*

*Proof.* By Lemma 2, all stationary points assuming exact optimization have  $i_{\text{direct}} = i_{\text{opt}}$  for all realizations of  $G_\theta$ . By Lemma 1, in each of these realizations we have  $i_{\text{direct}} \geq i_{\text{approx}} \geq i_{\text{opt}}$ . Thus, for all realizations we have  $i_{\text{approx}} = i_{\text{opt}}$  and thus we are at a stationary point assuming approximate search. In the other direction, Lemma 2 implies that all stationary points assuming approximate optimization have  $i_{\text{approx}} = i_{\text{opt}}$  almost surely. The only way for this to happen is that in trying to find  $i_{\text{approx}}$  we exhaustively iterated over all arms and found no improvement. Thus,  $i_{\text{direct}}$  could not have been an improvement and  $i_{\text{direct}} = i_{\text{opt}}$  almost surely.  $\square$

### A.3 Intuition as Steepest Ascent.

(10) computes a gradient of the expected return by increasing the log probability of  $\mathbf{a}_{\text{dir}}$  and decreasing the log probability of  $\mathbf{a}_{\text{opt}}$ . If we interpret  $\delta(\mathbf{a}_{\text{opt}}, \mathbf{a}) = G_\theta(\mathbf{a}_{\text{opt}}; \Gamma, \mathbf{S}) - G_\theta(\mathbf{a}; \Gamma, \mathbf{S})$  as a distance between  $\mathbf{a}_{\text{opt}}$  and candidate  $\mathbf{a}$  (it is non-negative for all  $\mathbf{a}$ ), then we can write  $D_\theta(\mathbf{a}; \Gamma, \mathbf{S}, \epsilon) = \epsilon R(\mathbf{a}, \mathbf{S}) - \delta(\mathbf{a}_{\text{opt}}, \mathbf{a}) + G_\theta(\mathbf{a}_{\text{opt}}; \Gamma, \mathbf{S})$ . The last term can be dropped and we can exponentiate both sides without changing the argmax, which gives an equivalent expression  $\mathbf{a}_{\text{dir}} = \text{argmax}_{\mathbf{a}} \frac{\exp \epsilon R(\mathbf{a}, \mathbf{S})}{\exp \delta(\mathbf{a}_{\text{opt}}, \mathbf{a})}$ . In this sense,  $\mathbf{a}_{\text{dir}}$  defines the (discrete) direction of steepest ascent in  $\exp R$  away from  $\mathbf{a}_{\text{opt}}$ , where trajectories  $\mathbf{a}$  that have lower probability under the model and/or that get unlucky draws of  $\Gamma(\mathbf{a})$  are considered further away. This gives intuition of computing a gradient by finding the direction of steepest ascent directly in trajectory space.

## B A\* Sampling with Lazily-constructed Argmaxes

**Gumbel Processes.** To evaluate  $D_\theta(\mathbf{a}, \epsilon)$ , which defines  $\mathbf{a}_{\text{opt}}$  and  $\mathbf{a}_{\text{dir}}$ , we need to sample a  $G_\theta(\mathbf{a})$  value for each complete trajectory encountered during the search. It is not possible to generate  $G_\theta(\mathbf{a})$  for each  $\mathbf{a}$  before starting the search, because there may be exponentially (or even infinitely) many possible trajectories. Another option would be to expand the search tree independently of  $G_\theta$  values and then sample  $G_\theta(\mathbf{a})$  via (4) for each singleton region encountered during the search. This would produce  $G_\theta$  values with the right distribution, but it is also a non-starter because we are precisely interested in biasing the search towards trajectories with large  $G_\theta$  values.

The solution to this problem comes from Maddison et al. Instead of only assigning  $G_\theta$  values to trajectories, we also assign them to regions. To assign random variables to overlapping regions in a consistent way, Maddison et al. introduce the *Gumbel Process*. A Gumbel process is defined in terms of a sample space  $\Omega$  and measure  $\mu$ . In our case,  $\Omega = \mathcal{A}^T$  is the set of all length  $T$  trajectories and  $\mu$  assigns probabilities to any subset  $\mathcal{R} \subseteq \mathcal{A}^T$  as  $\mu(\mathcal{R} | \mathbf{S}) = \sum_{\mathbf{a} \in \mathcal{R}} \Pi_\theta(\mathbf{a} | \mathbf{S})$ . A Gumbel Process is then defined as the set  $\{G(\mathcal{R}) | \mathcal{R} \subseteq \Omega\}$  where the following properties hold:

1.  $G(\mathcal{R}) \sim \text{Gumbel}(\log \mu(\mathcal{R}))$ ,
2.  $\mathcal{R}_1 \cap \mathcal{R}_2 = \emptyset \implies G(\mathcal{R}_1) \perp G(\mathcal{R}_2)$ ,
3.  $G(\mathcal{R}_1 \cup \mathcal{R}_2) = \max(G(\mathcal{R}_1), G(\mathcal{R}_2))$ .

That is, (1) the  $G$  values are marginally distributed as Gumbels with location given by the log measure of the region, (2) the random variables for disjoint regions are independent, and (3) the random variable in the union of two regions is equal to the max of the random variables in the two regions. A fourth property is implied by the first three, which we state in our context:

$$4. X(\mathcal{R}) = \operatorname{argmax}_{\mathbf{a} \in \mathcal{R}} G(\mathbf{a}) \sim 1\{\mathbf{a} \in \mathcal{R}\} \Pi_\theta(\mathbf{a} \mid \mathcal{S}).$$

That is, the argmax trajectory  $X(\mathcal{R})$  in a region is distributed according to  $\Pi_\theta(\cdot \mid \mathcal{S})$  that is masked out to only give support to  $\mathcal{R}$ . Finally, an important property that comes from Gumbel distributions is that  $G(\mathcal{R})$  and  $X(\mathcal{R})$  are independent random variables [21]. This means that we are free to interleave the sampling of  $X$  and  $G$  as we please, and it will be leveraged in the algorithms in the following sections.

**Top-Down Sampling.** Conceptually, if we had sampled  $G_\theta(\mathbf{a})$  for all  $\mathbf{a}$ , then the rest of the Gumbel process would be determined by  $G_\theta(\mathcal{R}) = \max_{\mathbf{a} \in \mathcal{R}} G_\theta(\mathbf{a})$ . However, Maddison et al. show that assuming  $\mu$  is computable for all regions, a Gumbel Process can be constructed lazily in a “top-down” fashion, first sampling  $G(\Omega)$ , and then recursively subdividing regions  $\mathcal{R}_0$  and sampling  $G$ ’s for the child regions conditional upon the value of  $G(\mathcal{R}_0)$ . Specifically, they divide  $\mathcal{R}_0$  into three disjoint regions:  $\mathcal{R}_1, \mathcal{R}_2$ , and  $\{X(\mathcal{R}_0)\}$  such that  $\mathcal{R}_0 = \mathcal{R}_1 \cup \mathcal{R}_2 \cup \{X(\mathcal{R}_0)\}$ . They show that for  $i \in \{1, 2\}$  the conditional distribution of  $G(\mathcal{R}_i)$  given previous splits in the tree is  $\text{TruncGumbel}(\log \mu(\mathcal{R}_i), G(\mathcal{R}_0))$  and  $G(\{X(\mathcal{R}_0)\}) = G(\mathcal{R}_0)$ .

Under our choice of regions,  $\mu(\mathcal{R} \mid \mathcal{S}) = \sum_{\mathbf{a} \in \mathcal{R}} \Pi_\theta(\mathbf{a} \mid \mathcal{S})$  can indeed be computed efficiently as

$$\Pi_\theta(\mathcal{R}(\tilde{\mathbf{a}}, \mathcal{B}; \mathcal{S}) \mid \mathcal{S}) = \left( \prod_{t'=0}^{t-1} \pi_\theta(a_{t'} \mid s_{(a_0, \dots, a_{t'-1})}) \right) \sum_{\mathbf{a} \in \mathcal{B}} \pi_\theta(\mathbf{a} \mid s_{\tilde{\mathbf{a}}}). \quad (22)$$

$\mathcal{B}$  is the set of actions that can be taken after the prefix  $\tilde{\mathbf{a}}$ .

If all prefixes eventually terminate with probability 1, then it is possible to apply one step of Top-Down Sampling to sample trajectories. To split a region  $\mathcal{R}_0 = \mathcal{R}(\tilde{\mathbf{a}}, \mathcal{B})$ , we would sample  $X(\mathcal{R}_0) \sim 1\{\mathbf{a} \in \mathcal{R}_0\} \pi_\theta(\mathbf{a} \mid s_{\tilde{\mathbf{a}}})$ . This is straightforward because it is essentially conditioning on a prefix in an autoregressive model. Specifically, start with  $\tilde{\mathbf{a}}$ , sample  $a_t \sim 1\{a_t \in \mathcal{B}\} \pi_\theta(a_t \mid s_{\tilde{\mathbf{a}}})$ , and then sample a completion according to

$$\prod_{t'=t+1}^T \pi_\theta(a_{t'} \mid s_{(a_0, \dots, a_{t'-1})}) \quad (23)$$

However, recursing would be problematic because we do not have a way of splitting  $\mathcal{R}_0 \setminus \{X(\mathcal{R}_0)\}$  into two regions that can compactly be represented as a prefix plus legal set of next actions. To address a similar issue, Kim et al. propose a modified split criteria that divides a region  $\mathcal{R}_0$  into two regions. Roughly the idea is to group together  $\mathcal{R}_1 \cup \{X(\mathcal{R}_0)\}$  from above into one region, and  $\mathcal{R}_2$  as the other region.

Applying the idea to our setting (which is slightly different because we support  $|\mathcal{A}| > 2$ ), to split a region  $\mathcal{R}_0 = \mathcal{R}(\tilde{\mathbf{a}}, \mathcal{B})$ , we assume inductively that we have already sampled  $G(\mathcal{R}_0)$  and  $X(\mathcal{R}_0)$ . Let prefix  $\tilde{\mathbf{a}}$  have  $t$  states and  $X(\mathcal{R}_0) = (a_0, \dots, a_{t-1})$ . Note that  $X(\mathcal{R}_0) \in \mathcal{R}_0$  by definition, so  $\tilde{\mathbf{a}}$  is a prefix of  $X(\mathcal{R}_0)$  and  $a_t \in \mathcal{B}$ . We can then define  $\mathcal{R}_1 = \mathcal{R}(\tilde{\mathbf{a}} \oplus a_t, \mathcal{A})$  and  $\mathcal{R}_2 = \mathcal{R}(\tilde{\mathbf{a}}, \mathcal{B} \setminus \{a_t\})$ . We then need  $G$  and  $X$  for the new regions. First,  $X(\mathcal{R}_0) \in \mathcal{R}_1$ , so it must be the case that it continues to be the argmax when considering a smaller region. Thus  $\mathcal{R}_1$  “inherits” the parent’s max and argmax:  $G(\mathcal{R}_1) = G(\mathcal{R}_0)$  and  $X(\mathcal{R}_1) = X(\mathcal{R}_0)$ . Creating a child region that does not contain the parent argmax follows the same logic as in standard Top-Down sampling:  $G(\mathcal{R}_2) \sim \text{TruncGumbel}(\log \mu(\mathcal{R}_2), G(\mathcal{R}_0))$ , and we can sample  $X(\mathcal{R}_2) \sim 1\{\mathbf{a} \in \mathcal{R}_2\} \pi_\theta(\mathbf{a} \mid s_{\tilde{\mathbf{a}}})$  as described in the previous subsection.

**Top-Down Sampling Trajectories.** Adapting the search space structure from Kim et al. makes it practical to implement Top-Down sampling for trajectories. However, the algorithm is wasteful in its interactions with the environment, particularly if trajectories can be long, because  $X(\mathcal{R})$  is instantiated fully for each region that is put on the queue. This would also prevent applying the

algorithm at all if trajectories are of infinite length. We develop a further modification that addresses these issues.

Our idea is to use a similar search space as Kim et al. but to lazily sample  $X(\mathcal{R})$ . The key observation is that the full value of  $X(\mathcal{R})$  is never used when splitting regions. Paired with the fact that maxes and argmaxes are independent, this means that we are free to only maintain prefixes of  $X(\mathcal{R})$  and sample extensions when they are needed. Using the same notation as above, we just need samples of the next action  $a_t$  to define the split. In fact, we can do away with explicitly maintaining  $X$ 's in the algorithm altogether. They can be recovered when we encounter a singleton region as the only trajectory in the region. The resulting algorithm is our Modified Top-Down algorithm and appears in Algorithm 2.

## C Additional Experimental Details

### C.1 Combinatorial Bandits

If adding an edge would create a cycle, the only legal action is to not add the edge. If there are  $k$  steps left and only  $n - k - 1$  edges so far, the only legal action is to add the edge. If there is only one legal action, we take it with probability 1.

### C.2 DeepSea

The policy model is a linear layer which gets as input one-hot vector of size 5x5 and outputs log probability for each action [FC(number of states, number of actions)]. We used Adam optimizer with a learning rate of 0.001

### C.3 Minigrid

The observations are provided as a tensor of shape 7x7x3. Each of the  $7 \times 7$  tiles is encoded using 3 integer values: one describing the type of object contained in the cell, one describing its color, and a flag indicating whether doors are open or closed. In addition, the agent's orientation is also provided as one-hot vector of size 4.

The policy model consists of 3 convolutional layers and one linear layer on top of them.  $Conv1(3, 32) \rightarrow ReLU \rightarrow Conv2(32, 48) \rightarrow ReLU \rightarrow Conv3(48, 64)$ . The linear layer gets as input a concatenation of orientation vector and the output of the convolutional layers, namely  $FC(64 + 4, 7)$ . The output of the linear layer is the log-probabilities of possible action. We used Adam optimizer with a learning rate of 0.001. We used the same architecture for our algorithm and the baselines.

We trained the model for 9M iterations, with a maximum of 3000 iterations per episode. In our algorithm we used the interactions budget for searching for direct candidates. In REINFORCE and cross-entropy method algorithms we used the interactions budget to sample 30 independent trajectories (100 steps trajectories). For REINFORCE we averaged the gradients of the 30 trajectories before updating the policy model. For the cross-entropy method we averaged  $\nabla_{\theta} \log \Pi_{\theta}(a | S)$  over the best 2 out of 30 trajectories.

We consider two versions of REINFORCE algorithm. The first is the standard trajectory-level  $\nabla \mathbb{E}_{a, s, r \sim p_{\theta}} \left[ \sum_{t=0}^{T-1} r_t \right] = \sum_{t=0}^{T-1} \nabla_{\theta} \log \pi_{\theta}(a_t | s_t) \sum_{i=0}^{T-1} r_i$ . However, the variance of the trajectory-level is high. The other version is an action-level which consider only the future rewards and serves as a variance reduction technique  $\nabla \mathbb{E}_{a, s, r \sim p_{\theta}} \left[ \sum_{t=0}^{T-1} r_t \right] = \sum_{t=0}^{T-1} \nabla_{\theta} \log \pi_{\theta}(a_t | s_t) \sum_{i=t}^{T-1} r_i$