Learning Exploration Policies for Model-Agnostic Meta-Reinforcement Learning

Anonymous Authors[1]

Abstract

Meta-Reinforcement learning approaches aim to develop learning procedures that can adapt quickly to a distribution of tasks with the help of a few examples. Developing efficient exploration strategies capable of finding the most useful samples becomes critical in such settings. Existing approaches to finding efficient exploration strategies add auxiliary objectives to promote exploration by the pre-update policy, however, this makes the adaptation using a few gradient steps difficult as the pre-update (exploration) and post-update (exploitation) policies are quite different. Instead, we propose to explicitly model a separate exploration policy for the task distribution. Having two different policies gives more flexibility in training the exploration policy and also makes adaptation to any specific task easier. We show that using self-supervised or supervised learning objectives for adaptation stabilizes the training process and also demonstrate the superior performance of our model compared to prior works in this domain.

1. Introduction

Reinforcement learning (RL) approaches have seen many successes in recent years, from mastering the complex game of Go (Silver et al., 2017) to even discovering molecules (Olivecrona et al., 2017). However, a common limitation of these methods is their propensity to overfitting on a single task and inability to adapt to even slightly perturbed configuration (Zhang et al., 2018). On the other hand, humans have this astonishing ability to learn new tasks in a matter of minutes by using their prior knowledge and understanding of the underlying task mechanics. Drawing inspiration from human behaviors, researchers have proposed to incorporate multiple inductive biases and heuristics to help the models learn quickly and generalize to unseen scenarios. However, despite a lot of effort it has been difficult to approach human levels of data efficiency and generalization.

Meta-RL tries to address these shortcomings by learning these inductive biases and heuristics from the data itself. These inductive biases or heuristics can be induced in the model in various ways like optimization, policy initialization, loss function, exploration strategies, etc. Recently, a class of policy initialization based meta-learning approaches have gained attention like Model Agnostic Meta-Learning (MAML) (Finn et al., 2017). MAML finds a good initialization for a policy that can be adapted to a new task by fine-tuning with policy gradient updates from a few samples of that task.

Given the objective of meta RL algorithms to adapt to a new task from a few examples, efficient exploration strategies are crucial for quickly finding the optimal policy in a new environment. Some recent works (Gupta et al., 2018a;b) have tried to address this problem by using latent variables to model the distribution of exploration behaviors. Another set of approaches (Stadie et al., 2018; Rothfuss et al., 2018) focus on improving the credit assignment of the meta learning objective to the pre-update trajectory distribution. However, all these prior works use one or few policy gradient updates to transition from pre- to post-update policy. This limits the applicability of these methods to cases where the post-update (exploitation) policy is similar to the pre-update (exploration) policy and can be obtained with only a few updates. Also, for cases where pre- and post-update policies are expected to exhibit different behaviors, large gradient updates may result in training instabilities and lack of convergence. To address this problem, we propose to explicitly model a separate exploration policy for the distribution of tasks. The exploration policy is trained to find trajectories that can lead to fast adaptation of the exploitation policy on the given task. This formulation provides much more flexibility in training the exploration policy. In the process, we also establish that, in order to adapt as quickly as possible to the new task, it is often more useful to use self-supervised or supervised learning approaches, where possible, to get more effective updates.

[1]Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country.

Preliminary work. Under review by the International Conference on Machine Learning (ICML). Do not distribute.

2. Background

2.1. Meta-Reinforcement Learning

Unlike RL which tries to find an optimal policy for a single task, meta-RL aims to learn a policy that can generalize to a distribution of tasks. Each task $\mathcal{T}$ sampled from the distribution $\rho(\mathcal{T})$ corresponds to a different Markov Decision Process (MDP). These MDPs have similar state and action space but might differ in the reward function or the environment dynamics. The goal of meta RL is to quickly adapt the policy to any task $\mathcal{T} \sim \rho(\mathcal{T})$ with the help of few examples from that task.

2.2. Credit Assignment in Meta-RL

Finn et al. (2017) introduced MAML - a gradient-based meta-RL algorithm that tries to find a good initialization for a policy which can be adapted to any task $\mathcal{T} \sim \rho(\mathcal{T})$ by fine-tuning with one or more gradient updates using the sampled trajectories of that task. MAML maximizes the following objective function:

$$J(\theta) = \mathbb{E}_{\mathcal{T}\sim\rho(\mathcal{T})}\left[\mathbb{E}_{\boldsymbol{\tau}'\sim P_{\mathcal{T}}(\boldsymbol{\tau}'|\theta')}\left[R(\boldsymbol{\tau}')\right]\right]$$
$$\text{with} \quad \theta' := U(\theta, \mathcal{T}) = \theta + \alpha \nabla_\theta \mathbb{E}_{\boldsymbol{\tau}\sim P_{\mathcal{T}}(\boldsymbol{\tau}|\theta)}\left[R(\boldsymbol{\tau})\right] \quad (1)$$

where U is the update function that performs one policy gradient ascent step to maximize the expected reward $R(\boldsymbol{\tau})$ obtained on the trajectories $\boldsymbol{\tau}$ sampled from task $\mathcal{T}$. Rothfuss et al. (2018) showed that the gradient of the objective function $J(\theta)$ can be written as:

$$\nabla_\theta J(\theta) = \mathbb{E}_{\mathcal{T}\sim\rho(\mathcal{T})}\left[\mathbb{E}_{\substack{\boldsymbol{\tau}\sim P_{\mathcal{T}}(\boldsymbol{\tau}|\theta) \\ \boldsymbol{\tau}'\sim P_{\mathcal{T}}(\boldsymbol{\tau}'|\theta')}}\left[\nabla_\theta J_{\text{post}}(\boldsymbol{\tau}, \boldsymbol{\tau}') + \nabla_\theta J_{\text{pre}}(\boldsymbol{\tau}, \boldsymbol{\tau}')\right]\right]$$

$$\nabla_\theta J_{\text{post}}(\boldsymbol{\tau}, \boldsymbol{\tau}') = \underbrace{\nabla_{\theta'} \log \pi_\theta(\boldsymbol{\tau}') R(\boldsymbol{\tau}')}_{\nabla_{\theta'} J^{\text{outer}}} \underbrace{\left(I + \alpha R(\boldsymbol{\tau}) \nabla^2_\theta \log \pi_{\theta'}(\boldsymbol{\tau})\right)}_{\text{transformation from } \theta' \text{ to } \theta}$$

$$\nabla_\theta J_{\text{pre}}(\boldsymbol{\tau}, \boldsymbol{\tau}') = \alpha \nabla_\theta \log \pi_\theta(\boldsymbol{\tau}) \left(\underbrace{\left(\nabla_\theta \log \pi_\theta(\boldsymbol{\tau}) R(\boldsymbol{\tau})\right)^\top}_{\nabla_\theta J^{\text{inner}}} \underbrace{\left(\nabla_{\theta'} \log \pi_{\theta'}(\boldsymbol{\tau}') R(\boldsymbol{\tau}')\right)}_{\nabla_{\theta'} J^{\text{outer}}} \right)$$

where, $\nabla_\theta J_{\text{post}}(\boldsymbol{\tau}, \boldsymbol{\tau}')$ optimizes θ to increase the likelihood of the trajectories $\boldsymbol{\tau}'$ that lead to higher returns given some trajectories $\boldsymbol{\tau}$. In other words, this term does not optimize θ to yield trajectories $\boldsymbol{\tau}$ that lead to good adaptation steps. That is infact, done by the second term $\nabla_\theta J_{\text{pre}}(\boldsymbol{\tau}, \boldsymbol{\tau}')$. It optimizes for the pre-update trajectory distribution, $P_{\mathcal{T}}(\boldsymbol{\tau}|\theta)$, i.e, increases the likelihood of trajectories $\boldsymbol{\tau}$ that lead to good adaptation steps.

During optimization, MAML only considers $J_{\text{post}}(\boldsymbol{\tau}, \boldsymbol{\tau}')$ and ignores $J_{\text{pre}}(\boldsymbol{\tau}, \boldsymbol{\tau}')$. Thus MAML finds a policy that adapts quickly to a task given relevant experiences, however, the policy is not optimized to gather useful experiences from the environment that can lead to fast adaptation.

ProMP (Rothfuss et al., 2018) analyzes this issue with MAML and incorporates $\nabla_\theta J_{\text{pre}}(\boldsymbol{\tau}, \boldsymbol{\tau}')$ term in the update as well. They propose to use DICE (Foerster et al., 2018) to allow causal credit assignment on the pre-update trajectory distribution, however, the gradients computed by DICE suffer from high variance estimates. To remedy this, they proposed a low variance (and slightly biased) approximation of the DICE based loss that leads to stable updates.

However, the pre-update and post-update policies are often expected to exhibit very different behaviors, i.e, exploration and exploitation behaviors respectively. In such cases, transitioning a single policy from pure exploration phase to pure exploitation phase via policy gradient updates will require multiple steps. Unfortunately, this significantly increases the computational and memory complexities of the algorithm. Furthermore, it may not even be possible to achieve this transition via few gradient updates. This raises an important question: **Do we really need to use the pre-update policy for exploration as well? Can we use a separate policy for exploration?**

Another problem with these approaches stems from the high variance nature of the inner loop policy gradient updates resulting in both $\nabla_\theta J_{\text{pre}}$ and $\nabla_\theta J_{\text{post}}$ terms being high variance. Note that $\nabla_\theta J_{\text{pre}}$ computes the inner product between pre-update and post-update policy gradients. Since, the two policy gradient terms have been computed using two different policies (i.e pre-update and post-update policies) on two different trajectory distributions, the resulting updates are highly likely to be unstable. ProMP tries to stabilize the updates by adding trust region constraints, however, that doesn't completely remedy the issue. This motivates another question: **Do we really need the pre-update gradients to be policy gradients? Can we use a different objective in the inner loop to get more stable updates?**

3. Method

We now describe our proposed model that addresses the queries we raised in the previous section.

Using separate policies for pre-update and post-update sampling: The straightforward solution to the first problem mentioned above is to use a separate exploration policy μ_ϕ responsible for collecting trajectories for the inner loop updates. Following that, post-update policy π_θ can be used to collect trajectories for performing the outer loop updates. Unfortunately, this is not as simple as it sounds. To understand this, let's look at the inner loop updates:

$$U(\theta, \mathcal{T}) = \theta + \alpha \nabla_\theta \mathbb{E}_{\boldsymbol{\tau} \sim P_\mathcal{T}(\boldsymbol{\tau}|\theta)} \left[R(\boldsymbol{\tau}) \right]$$

When the exploration policy is used for sampling trajectories, we need to perform importance sampling. The update would thus become:

$$U(\theta, \mathcal{T}) = \theta + \alpha \nabla_\theta \mathbb{E}_{\boldsymbol{\tau} \sim Q_\mathcal{T}(\boldsymbol{\tau}|\phi)} \left[\frac{P_\mathcal{T}(\boldsymbol{\tau}|\theta)}{Q_\mathcal{T}(\boldsymbol{\tau}|\phi)} R(\boldsymbol{\tau}) \right]$$

where $P_\mathcal{T}(\boldsymbol{\tau}|\theta)$ and $Q_\mathcal{T}(\boldsymbol{\tau}|\phi)$ represent the trajectory distribution sampled by π_θ and μ_ϕ respectively. The above update, being off-policy, results in high variance estimates when the two trajectory distributions are quite different from each other, thus making it infeasible to use it in the current form. To address this instability issue, let us consider the second question we raised before.

Using a self-supervised/supervised objective for the inner loop update step: The instability in the inner loop updates arises due to the high variance nature of the policy gradient update. One potential solution is to use some form of self-supervised or supervised learning objective in place of policy gradient in the inner loop to ensure that the updates are more stable. We propose to use a network for predicting some task (or MDP) specific property like reward function, expected return or value. During the inner loop update, the network updates its parameters by minimizing its prediction error on the given task. Unlike prior meta-RL works where the task adaptation in the inner loop is done by policy gradient updates, here, we perform gradient descent on a supervised loss objective function resulting in stability during the adaptation phase.

3.1. Model

Our proposed model comprises of three modules, the exploration policy $\mu_\phi(s)$, the exploitation policy $\pi_{\theta,z}(s)$, and the self-supervision network $M_{\beta,z}(s,a)$. Note that $M_{\beta,z}$ and $\pi_{\theta,z}$ share a set of parameters z while containing their own set of parameters β and θ respectively.

The agent first collects a set of trajectories $\boldsymbol{\tau}$ using its exploration policy μ_ϕ for each task $\mathcal{T} \sim \rho(\mathcal{T})$. It then updates the shared parameter z by minimizing the regression loss $\left(M_{\beta,z}(s,a) - M^t(s,a)\right)^2$:

$$z' = U(z, \mathcal{T}) = z - \alpha \nabla_z \mathbb{E}_{\boldsymbol{\tau} \sim Q_\mathcal{T}(\boldsymbol{\tau}|\phi)} \left[\sum_{t=0}^{H-1} \left(M_{\beta,z}(s_t, a_t) - \overline{M}(s_t, a_t) \right)^2 \right]$$

where, $\overline{M}(s,a)$ is the target which can be any of the task specific quantities as mentioned above. We further modify the above equation by multiplying the DICE operator to simplify the gradient computation w.r.t ϕ. This eliminates the need to apply the policy gradient trick to expand the above expression for gradient computation. The update then becomes:

$$z' = U(z, \mathcal{T}) = z - \alpha \nabla_z \mathbb{E}_{\boldsymbol{\tau} \sim Q_\mathcal{T}(\boldsymbol{\tau}|\phi)} \left[\sum_{t=0}^{H-1} \left(\prod_{t'=0}^{t} \frac{\mu_\phi(a_{t'}|s_{t'})}{\perp(\mu_\phi(a_{t'}|s_{t'}))} \right) \left(M_{\beta,z}(s_t, a_t) - \overline{M}(s_t, a_t) \right)^2 \right]$$

where $\perp$ is the stop gradient operator. After obtaining the updated parameters z', the agent samples trajectories $\boldsymbol{\tau}'$ using its updated exploitation policy $\pi_{\theta,z'}$. Note that our model enables the agent to learn a generic exploitation policy $\pi_{\theta,z}$ for the task distribution which can then be adapted to any specific task by updating z to z' as shown above. Effectively, z' encodes the necessary information regarding the task that helps an agent in adapting its behavior to maximize its expected return.

The collected trajectories are then used to perform a policy gradient update for all the parameters z, θ, ϕ and β using the following objective:

$$J(z', \theta) = \mathbb{E}_{\mathcal{T} \sim \rho(\mathcal{T})} \left[\mathbb{E}_{\boldsymbol{\tau}' \sim P_\mathcal{T}(\boldsymbol{\tau}'|\theta, z')} \left[R(\boldsymbol{\tau}') \right] \right]$$

The gradients of $J(z', \theta)$ w.r.t. ϕ are shown in Eq. 2 (see appendix). Although the gradients are unbiased, they still have very high variance. To solve this problem, we draw inspiration from Mnih et al. (2016) and replace the return R_t^μ (see Eq. 3 in appendix) with an advantage estimate A_t^μ (see 4 in appendix). Due to space constraints we describe these formulations in more detail in appendix.

4. Experiments

We have evaluated our proposed model on the environments used by Finn et al. (2017) and Rothfuss et al. (2018). Specifically, we have used HalfCheetahFwdBack, HalfCheetahVel and AntFwdBack environments for the dense reward tasks and a 2D point environment proposed in Rothfuss et al. (2018) for sparse rewards. We noticed that the dense reward environments used in Rothfuss et al. (2018) and Finn et al. (2017) had different reward functions. So we chose to stick to the environments used in Finn et al. (2017) for our experiments.

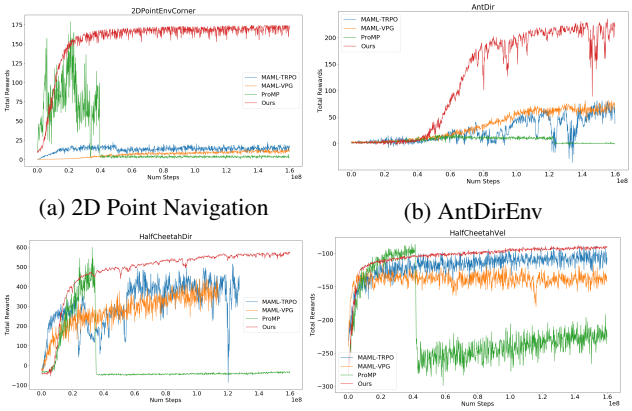
We treat the shared parameter z as a learnable latent embedding. The exploitation policy $\pi_{\theta,z}(s)$ and the self-supervision network $M_{\beta,z}(s,a)$ concatenates z with their respective inputs. All the three networks have the same architecture as that of the policy network in Finn et al. (2017). For the self-supervision network, we use the returns calculated at each time step of the exploration policy as the self-supervision signal for the adaptation phase.

We compare our model against MAML-TRPO, MAML-VPG (vanilla policy gradient) and ProMP. Note that we used the same hyperparameters as specified in their official

repositories. For our model we use the hyperparameters used in Finn et al. (2017). Also, we restrict ourselves to a single adaptation step in all environments, but it can be easily extended to multiple gradient steps as well. We used Adam (Kingma & Ba, 2014) optimizer with a learning rate of $7e-4$ for all parameters except z which uses $4e-5$.

4.1. Results

The performance plots for all four algorithms are shown in Fig. 1. In all the environments, our proposed method outperforms others in terms of asymptotic performance and sample complexity. As evident from the The training is much more stable for our method especially when compared with ProMP . This can also be noted from the much lower variance of our updates compared to the baselines. Although ProMP manages to reach similar peak performance to our method in HalfCheetahFwdBack, HalfCheetahVel and 2DPointEnvCorner, the training itself is pretty unstable indicating the inherent fragility of their updates. Our algorithm particularly shines in 2DPointEnvCorner where the reward is sparse. In this environment, the agent needs to perform efficient exploration and use the sparse reward trajectories to perform stable updates both of which are salient aspects of our algorithm. Similarly, in AntDir, the rewards are largely uninformative in the initial stages of training given the large state space. Thus, here as well, exploration and stable updates become critical thus allowing our algorithm to perform much better than the baseline.



(a) 2D Point Navigation (b) AntDirEnv

(c) HalfCheetahDirEnv (d) HalfCheetahVelEnv

Figure 1. Comparison of our method with 3 baseline methods in sparse and dense reward environments.

Ablations: We perform several ablation experiments to analyze the impact of different components of our algorithm on 2D point navigation task. Fig. 2 shows the performance plots for 5 different variants:

VPG-Inner loop : The semi-supervised/supervised loss in the inner loop is replaced with the vanilla policy gradient loss as in MAML while using the exploration policy to sample the pre-update trajectories. As expected, this model performs poorly due to the high variance off-policy updates in the inner loop.

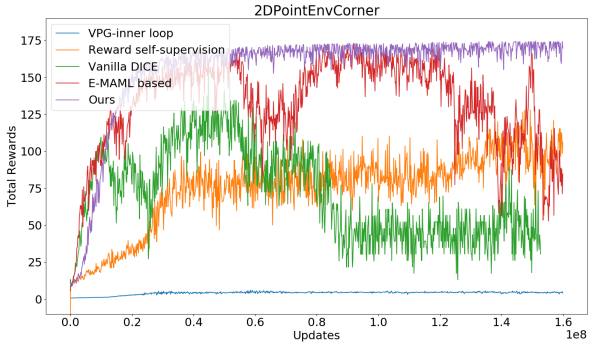


Figure 2. Ablation results

Reward Self-Supervision : A reward based self-supervised objective is used instead of return based self-supervision. This variant performs reasonably well but struggles to reach peak performance since the task is sparse reward.

Vanilla DICE : We directly use the dice gradients to perform updates on ϕ instead of using the low variance gradient estimator. The high variance dice gradients lead to unstable training as can be seen from the plots.

E-MAML Based : Used an E-MAML (Stadie et al., 2018) type objective to compute the gradients w.r.t ϕ instead of using DICE. Although this variant manages to reach peak performance, it is unstable due to the lack of causal credit assignment.

Ours : Used the low variance estimate of the dice gradients to compute updates for ϕ along with return based self-supervision for inner loop updates. Our model reaches peak performance and exhibits stable training due to low variance updates.

5. Conclusion and Future Work

Unlike conventional meta-RL approaches, we proposed to explicitly model a separate exploration policy for the task distribution. Having two different policies gives more flexibility in training the exploration policy and also makes adaptation to any specific task easier. We showed that, through various experiments on both sparse and dense reward tasks, our model outperforms previous works while also being very stable during training. This validates that using self-supervised techniques increases the stability of these updates thus allowing us to use a separate exploration policy to collect the initial trajectories. As future work, we plan to experiment on more environments and add auxiliary objectives for the exploration agent. Furthermore, we also intend to incorporate the PPO inspired modifications made by ProMP to its objective. This would allow us to perform multiple outer-loop updates to the agents after each rollout leading to faster convergence of our model.

References

Duan, Y., Chen, X., Houthooft, R., Schulman, J., and Abbeel, P. Benchmarking deep reinforcement learning for continuous control. In *International Conference on Machine Learning*, pp. 1329–1338, 2016.

Finn, C., Abbeel, P., and Levine, S. Model-agnostic meta-learning for fast adaptation of deep networks. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pp. 1126–1135. JMLR. org, 2017.

Foerster, J., Farquhar, G., Al-Shedivat, M., Rocktäschel, T., Xing, E. P., and Whiteson, S. Dice: The infinitely differentiable monte-carlo estimator. *arXiv preprint arXiv:1802.05098*, 2018.

Gupta, A., Eysenbach, B., Finn, C., and Levine, S. Unsupervised meta-learning for reinforcement learning. *arXiv preprint arXiv:1806.04640*, 2018a.

Gupta, A., Mendonca, R., Liu, Y., Abbeel, P., and Levine, S. Meta-reinforcement learning of structured exploration strategies. In *Advances in Neural Information Processing Systems*, pp. 5307–5316, 2018b.

Kingma, D. P. and Ba, J. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Mnih, V., Badia, A. P., Mirza, M., Graves, A., Lillicrap, T., Harley, T., Silver, D., and Kavukcuoglu, K. Asynchronous methods for deep reinforcement learning. In *International conference on machine learning*, pp. 1928–1937, 2016.

Olivecrona, M., Blaschke, T., Engkvist, O., and Chen, H. Molecular de-novo design through deep reinforcement learning. *Journal of cheminformatics*, 9(1):48, 2017.

Rothfuss, J., Lee, D., Clavera, I., Asfour, T., and Abbeel, P. Promp: Proximal meta-policy search. *arXiv preprint arXiv:1810.06784*, 2018.

Silver, D., Schrittwieser, J., Simonyan, K., Antonoglou, I., Huang, A., Guez, A., Hubert, T., Baker, L., Lai, M., Bolton, A., et al. Mastering the game of go without human knowledge. *Nature*, 550(7676):354, 2017.

Stadie, B. C., Yang, G., Houthooft, R., Chen, X., Duan, Y., Wu, Y., Abbeel, P., and Sutskever, I. Some considerations on learning to explore via meta-reinforcement learning. *arXiv preprint arXiv:1803.01118*, 2018.

Zhang, C., Vinyals, O., Munos, R., and Bengio, S. A study on overfitting in deep reinforcement learning. *arXiv preprint arXiv:1804.06893*, 2018.

6. Appendix

We described our proposed model in Section 3. The pseudo-code of our algorithm is shown in algorithm 6.

while not converge **do**
  Sample a batch of tasks $\mathcal{T}_i \sim \rho(\mathcal{T})$
  for all sampled tasks $\mathcal{T}_i$ **do**
    Collect pre-update trajectories $\boldsymbol{\tau}_{\mu}^{\mathcal{T}_i}$ using $\mu_\phi(s)$
    Update z by minimizing $\left(M_{\beta,z}(s,a) - \bar{M}(s,a)\right)^2$:

$$z' = z - \alpha \nabla_z \mathbb{E}_{\boldsymbol{\tau}_{\mu}^{\mathcal{T}_i} \sim Q_{\mathcal{T}_i}(\boldsymbol{\tau}|\phi)} \left[\sum_{t=0}^{H-1} \left(\prod_{t'=0}^{t} \frac{\mu_\phi(a_{t'}|s_{t'})}{\perp(\mu_\phi(a_{t'}|s_{t'}))} \right) \left(M_{\beta,z}(s,a) - \bar{M}(s,a)\right)^2 \right]$$

    where $\left(\prod_{t'=0}^{t} \frac{\mu_\phi(a_{t'}|s_{t'})}{\perp(\mu_\phi(a_{t'}|s_{t'}))}\right)$ is the DICE operator
    Collect post-update trajectory $\boldsymbol{\tau}_{\pi}^{\mathcal{T}_i}$ using $\pi_{\theta,z'}(s)$
  Policy gradient update to optimize all the parameters z, θ, ϕ, β

Although, in principle, the above algorithm is what we want to implement, implementing it as is leads to high variance dice gradients. This consequently leads to high variance gradients for the ϕ's. To alleviate this, we apply some variance reduction techniques described as follows.

The vanilla dice gradients can be written as follows:

$$\nabla_\phi J(z',\theta) = \mathbb{E}_{\mathcal{T}\sim\rho(\mathcal{T})} \Bigg[\mathbb{E}_{\boldsymbol{\tau}\sim Q_{\mathcal{T}}(\boldsymbol{\tau}|\phi)} \sum_{t=0}^{H-1} \alpha \nabla_\phi \log \mu_\phi(s_t) \left[\sum_{t'=t}^{H-1} \left(\mathbb{E}_{\boldsymbol{\tau}'\sim P_{\mathcal{T}}(\boldsymbol{\tau}'|\theta,z')} \left(\nabla_{z'} \log \pi_{\theta,z'}(\boldsymbol{\tau}') R(\boldsymbol{\tau}')\right)^\top \right) \left(\nabla_z \left(M_{\beta,z}(s_t,a_t) - \overline{M}(s_t,a_t)\right)^2 \right) \right] \Bigg]$$

The above expression can be seen as a policy gradient update:

$$\nabla_\phi J(z',\theta) = \mathbb{E}_{\mathcal{T}\sim\rho(\mathcal{T})} \left[\mathbb{E}_{\boldsymbol{\tau}\sim Q_{\mathcal{T}}(\boldsymbol{\tau}|\phi)} \sum_{t=0}^{H-1} \alpha \nabla_\phi \log \mu_\phi(s_t) R_t^\mu \right] \tag{2}$$

with returns

$$R_t^\mu = \left[\sum_{t'=t}^{H-1} \left(\mathbb{E}_{\boldsymbol{\tau}'\sim P_{\mathcal{T}}(\boldsymbol{\tau}'|\theta,z')} \left(\nabla_{z'} \log \pi_{\theta,z'}(\boldsymbol{\tau}') R(\boldsymbol{\tau}')\right)^\top \right) \left(\nabla_z \left(M_{\beta,z}(s_t,a_t) - \overline{M}(s_t,a_t)\right)^2 \right) \right] \tag{3}$$

Thus, instead of directly applying the gradients computed above, we compute the advantages A_t^μ using the returns R_t^μ and apply the following gradients.

$$\nabla_\phi \hat{J}(z',\theta) = \mathbb{E}_{\mathcal{T}\sim\rho(\mathcal{T})} \left[\mathbb{E}_{\boldsymbol{\tau}\sim Q_{\mathcal{T}}(\boldsymbol{\tau}|\phi)} \sum_{t=0}^{H-1} \alpha \nabla_\phi \log \mu_\phi(s_t) A_t^\mu \right]$$

where,

$$A_t^{\mu} = r_t^{\mu} + V_{t+1}^{\mu} - V_t^{\mu} \tag{4}$$

where V_t^{μ} is computed using using a linear feature baseline (Duan et al., 2016) fitted on the returns R_t^{μ}. And r_t^{μ} is given by,

$$r_t^{\mu} = \Big(\mathbb{E}_{\boldsymbol{\tau}' \sim P_{\mathcal{T}}(\boldsymbol{\tau}'|\theta, z')} \left(\nabla_{z'} \log \pi_{\theta, z'}(\boldsymbol{\tau}') R(\boldsymbol{\tau}') \right)^{\top} \Big) \Big(\nabla_z \left(M_{\beta, z}(s_t, a_t) - \overline{M}(s_t, a_t) \right)^2 \Big) \Big) \tag{5}$$