Context-Based Meta-Reinforcement Learning with Structured Latent Space

Hongyu Ren
Stanford University
hyren@cs.stanford.edu

Animesh Garg
University of Toronto
garg@cs.storonto.edu

Anima Anandkumar
Caltech
anima@caltech.edu

Abstract

Meta-reinforcement learning (meta-RL) allows agents to adapt quickly to unseen new tasks when trained on similar tasks. Given context information from a set of tasks, recent methods perform online probabilistic inference on the task at hand and leverage the estimated posterior to achieve transferable skills. Most of the context-based meta-RL algorithms use Gaussian latent variables to model the task distribution. However, complex tasks are not controlled by single variable in real world settings, which inherently cannot be modeled by isotropic Gaussian random variables. In this paper, we propose a structured latent space that combines Gaussian random variables with Dirichlet random variables. Inspired by the topic models, we assume the tasks are represented as a mixture of "base tasks" (modeled by the Dirichlet distribution) and style factors (modeled by Gaussian distribution), thus disentangling and factorizing the latent space. We show in our experiments that the proposed method provides more accurate task inference and a boost in performance in meta-learning benchmarks.

1 Introduction

Although reinforcement learning algorithms have achieved high performance in many tasks, even surpassed humans [1], they require vast amounts of samples/interactions with environment to conquer real world tasks. Meta-reinforcement learning (meta-RL) aims to train versatile agents that are capable of many tasks and can adapt quickly to new environments with a small amount of training steps. Context-based meta-RL algorithms [2–4] propose to explicitly model the task distribution, which shows superior performance than gradient-based meta-learning models [5]. Given a new task, the algorithm is able to identify tasks by online probabilistic inference, and then incorporates the posterior with the policy network so as to tackle different tasks of interest. Disentangling task inference from decision making in policy networks make the model desirable for off-policy meta-learning, where the policy network can leverage off-policy history data from the replay buffer, thus achieving 20-100X improvement in sample efficiency compared to gradient-based meta-learning models.

However, previous context-based meta-learning methods [2] use a simple latent space, e.g. isotropic Gaussian, to model the task distribution. It is not rich enough to model a family of complex tasks that go beyond easy tasks controlled by single variable, e.g. goal location, velocity, etc. As a concrete example, in the application of robot grasping, reaching and pulling are two independent tasks, while opening a drawer requires the ability to both reach and pull. This type of structure of tasks appear everywhere in the real world, which inherently cannot be modeled by simple distributions such as the Gaussian or the categorical distribution.

Here we propose to design a structured latent space to model task distribution based on Dirichlet topic models [6]. Previously Dirichlet topic models have been employed for document categorization where each document is modeled as a mixture of underlying "base topics". In the setting of meta-RL, we assume the tasks are represented as a mixture of "base tasks", we aim to model each task as a draw

33rd Conference on Neural Information Processing Systems (NeurIPS 2019), Vancouver, Canada.

from the Dirichlet distribution, soft-clustering the tasks, e.g. opening a drawer is assigned roughly half and half to reaching and pulling.

We also find that only using Dirichlet distribution is not sufficient to model the complex underlying task distributions. Besides the "cluster-like" property, there are also some style factors that are independent of task topics. Hence, we propose to augment Dirichlet random variables with Gaussian variables and show that our model with multi-modal latent space is capable of learning disentangled and interpretable policies in an unsupervised manner. The Dirichlet variables model the proportions to each category of topic while the Gaussian variables encode continuous latent factors such as style or preference.

In order to infer the parameters of the multi-modal posterior, we treat Dirichlet and Gaussian separately. While the Gaussians can be taken care of in a similar way as in [2], the Dirichlet posterior of a task is difficult to calculate directly. We approximate the Dirichlet distribution with logit-normal distribution, which can be easily calculated by taking the softmax of Gaussian variables [7, 8]. We validate our model in a 2D point-robot navigation task, where we show that our model is able to achieve both interpretable policy and higher empirical performance in test-task adaptation. We plan to further implement our model in several meta-learning benchmarks with complex task dependency, such as Meta-World [9] and Football-Academy [10]. We start in Section 2 to introduce recent advances in context-based meta-RL, then we introduce our method in Section 3, and the experimental results in Section 4.

2 Context-Based Meta-RL

In meta-RL, we assume a (multi-modal) distribution of tasks $p(\mathcal{T})$, where each task $\mathcal{T} \sim p(\mathcal{T})$ is a Markov decision process (MDP) and we further assume all the tasks in $p(\mathcal{T})$ share the same state and action space.

Context-based meta-RL algorithms [2–4] aim to capture the uncertainty of tasks by mapping the tasks to latent space with an encoder $q_\phi(z|\mathcal{T})$ (parameterized by ϕ). The policy $\pi_\theta(a|s,z)$ (parameterized by θ) are thus additionally conditioned on the latent variable z, disentangling task inference from decision making. The latent variable z is crucial in achieving fast adaptation, since it encodes the necessary information about the task $\mathcal{T}$. Inspired by amortized variational inference algorithms [11], the objective is to maximize the variational lower bound, which consists of a reconstruction loss and a KL divergence term, as shown in the Equation 1 below. The reconstruction term can be viewed as the difference between task and its reconstruction, or simply the reward the policy achieves under current encoder. The KL term is considered as a information bottleneck that limits the mutual information between $\mathcal{T}$ and z.

$$\mathbb{E}_{\mathcal{T}}[\mathbb{E}_{z\sim q_\phi(z|\mathcal{T})}[f(\mathcal{T},z) + D_{\mathrm{KL}}(q_\phi(z|\mathcal{T})||p(z)]] \tag{1}$$

where f is the reconstruction term and $p(z)$ is the prior distribution.

However, directly inferring the posterior of a task is difficult because of the abstract representation of MDPs. Pearl [2] proposes to collect n transition tuples from the task $\mathcal{T}$ and the encoder is trained to encode each transition tuple to the latent space, then the posterior of the task is calculated as the product of several independent factors. Note that a transition tuple, a.k.a. context is referred to as $c_i^{\mathcal{T}} = (s_i, a_i, r_i, s_i')$, representing the *state*, *action*, *reward*, *next state* at step i for task $\mathcal{T}$. The process of posterior estimation can be formulated as follows:

$$q_\phi(z|\mathcal{T}) \propto \Pi_{i=1}^n q_\phi(z|c_i)$$

3 Structured Latent Space for Task Inference

Real-world tasks demonstrate complex dependency and relationship between each other. Modeling the distribution of tasks with simple distributions such as isotropic Gaussian leads to sub-optimal task inference and thus limits model's ability to make fast adaptation, especially when the set of tasks are no longer controlled by a single continuous factor, such as velocity.

Inspired by the Dirichlet topic models [6], we propose to model the latent space using Dirichlet distribution. Dirichlet distribution has shown fitness in modeling multi-modal or proportional data, which aligns well with our prior over the distribution of tasks. For example, opening a drawer is

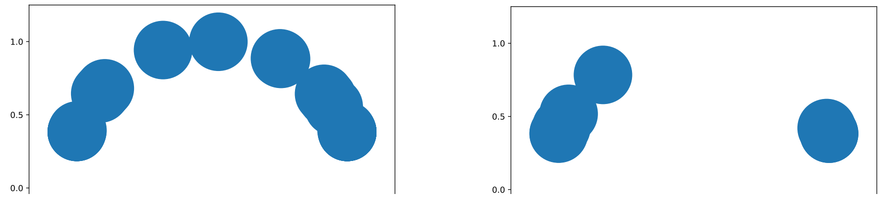

Figure 1: Left: the distribution of 30 training tasks; Right: the distribution of 10 test tasks. All 40 tasks follow the Dirichlet distribution.

related roughly half and half to reaching and pulling skills. Thus, we adopt a Dirichlet prior of dimension K in order to capture this multi-modality, especially model multiple peaks for each basic task, where K is pre-defined to represent the number of basic tasks.

3.1 Approximate Dirichlet with Logit-Normal

However, estimating the posterior parameter of a Dirichlet is nontrivial in context-based meta-RL because the product of the PDF of Dirichlet distribution is no longer proportional to a Dirichlet PDF in most cases. To be concrete, if the posterior is a Dirichlet distribution, the PDF of context c_j is: $q_\phi(z|c_j) = \frac{1}{B(\boldsymbol{\alpha}_j)} \prod_{i=1}^{K} z_i^{\alpha_{ji}-1}$, where $\boldsymbol{\alpha}_j = (\alpha_{j1}, \ldots, \alpha_{jK})$ and $\alpha_{ji} > 0 \, \forall i, j$. Assume we have n contexts, the product is calculated as follows, $\Pi_{j=1}^{n} q_\phi(z|c_j) = \frac{1}{\prod_{j=1}^{n} B(\boldsymbol{\alpha}_j)} \Pi_{i=1}^{K} z_i^{\sum_{j=1}^{n}(\alpha_{ji}-1)}$. The resulting posterior needs to follow the constraint $1 + \sum_{j=1}^{n}(\alpha_{ji} - 1) > 0$ in order to remain a Dirichlet distribution. However, this constraint is hard to impose on the encoder. A hard constraint such as normalizing every α_{ji} to be within $(\frac{n-1}{n}, \infty)$ is undesirable for the following two reasons: 1) it loses some degrees of freedom of the model because this constraint essentially is a sufficient and unnecessary condition; 2) the number of contexts n varies during the training procedure.

Considering the difficulty of directly "accumulating" Dirichlet distribution, we propose to circumvent the issue by approximating Dirichlet posterior with logit-normal distribution, where the parameters of the Dirichlet distribution is calculated as follows:

$$\alpha_k = \frac{1}{\Sigma_{kk}}(1 - \frac{2}{K} + \frac{e^{-\mu_k}}{K^2} \sum_{i}^{K} e^{-\mu_i}) \quad \forall\, k = 1, \ldots, K$$

In this case, the encoder q_ϕ outputs a Gaussian distribution and can be accumulated to calculate $q_\phi(z|\mathcal{T})$, then we directly take the softmax of the resulting Gaussian to transform it to logit-normal distribution, which can then be used to calculate the parameters of the Dirichlet posterior $q_\phi(z|\mathcal{T})$. We could directly calculate the KL divergence between $q_\phi(z|\mathcal{T})$ and $p(z)$. In order to make the gradient flow through the sampling procedure, we use the standard reparameterization tricks [11] to sample z from the logit normal distribution. The objective in Equation 1 can thus be optimized for our model with Dirichlet random variables.

3.2 Augment Dirichlet with Gaussian

Although Dirichlet distribution provides the model with the ability to better capture the task relationship by introducing several "base tasks", there are some invariant factors that are independent of tasks but associated with agent's preference, such as style. These style factors are usually continuous and represent another multi-modal side of how the task is performed by the agent, which cannot be modeled by Dirichlet distribution. Hence, we propose to augment the Dirichlet random variables with Gaussian random variables so that they can each capture their respective multi-modality, and the joint latent space can be expressive enough to learn interpretable and disentangled policy.

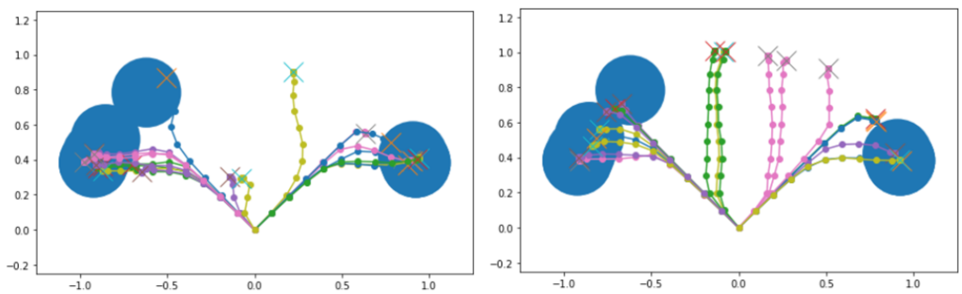

Figure 2: The trajectory of the learned agent with latent variable sampled directly from the prior distribution. The prior distribution of the left subfigure is 2 Dirichlet random variables (each with 4 dimension); the prior distribution of the right subfigure is 2 Dirichlet random variables (each with 4 dimension) combined with 2 continuous random variables from normal distribution.

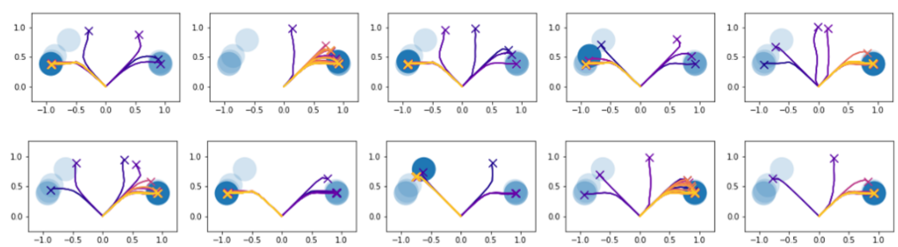

Figure 3: The trajectories of the learned policy on 10 test tasks, (dark blue denotes the goal location). By using posterior sampling, the agent can quickly adapt to unseen test tasks for all 10 test tasks. The trajectories with brighter color (yellow) means the agent uses more context information for posterior estimation.

4 Experimental Result

We validate the performance of our method on point-robot navigation task, where the agent aims to navigate to different goal locations on the edge of a half-circle. In order to create a multi-modal task distribution, we first pick two anchor positions on two sides of the half circle, and sample 40 goal locations as interpolations between the two anchors that follow Dirichlet distribution $Dir(0.2, 0.2)$. We randomly split 40 goals into 30 training tasks and 10 test tasks as shown in the Figure 1.

We compare several baselines including Pearl with categorical random variables and Gaussian random variables. We denote m categorical random variables with n classes as "Cat(m,n)", m Gaussian random variables as "Gau(m)", m Dirichlet random variables with n classes as "Dir(m,n)". The prior distribution for categorical distribution is uniform and the Gaussian is $\mathcal{N}(0, 1)$. To better model multi-modality, the Dirichlet prior is $Dir(0.2, \ldots, 0.2)$.

The quantitative results are shown in Table 1, our method that only uses Dirichlet random variables is able to achieve better performance than natural baselines that use Gaussian random variables or categorical random variables or combined. After we augment our model with additional two dimensional Gaussian random variables, our model is able to achieve higher performance. Some trajectories of sampling directly from prior distribution and from posterior distribution are shown in Figure 2 and Figure 3 respectively. When sampling from prior distribution, it means the model has no knowledge about the task, our model is able to cover all the possible goal locations. When the model accumulates context information, it can quickly adapt itself to the task at hand, and reach all test goals successfully.

Return	Gau(5)	Cat(2,4)	Cat(2,4)+Gau(2)	Dir(2,4)	Dir(2,4)+Gau(2)
Point-Robot	10.2	4.6	9.5	11.3	12.2

Table 1: The result of Pearl with different choices of prior distribution in point-robot navigation task.

5 Conclusion

We propose in this paper a structured latent space for context-based meta-reinforcement learning algorithms. Instead of modeling the task distribution with a simple isotropic Gaussian, we design a multi-modal latent space that consists of Dirichlet random variables and Gaussian random variables to model the complex relationship between tasks and some invariant factors respectively. Preliminary experiments on point-robot navigation tasks show that our method works better than baselines with Guassian random variables or categorical random variables. We plan to further explore its application in several meta-learning benchmarks, such as Meta-World [9] and Football-Academy [10].

References

[1] V. Mnih, K. Kavukcuoglu, D. Silver, A. Graves, I. Antonoglou, D. Wierstra, and M. Riedmiller, "Playing atari with deep reinforcement learning," *arXiv preprint arXiv:1312.5602*, 2013.

[2] K. Rakelly, A. Zhou, D. Quillen, C. Finn, and S. Levine, "Efficient off-policy meta-reinforcement learning via probabilistic context variables," *arXiv preprint arXiv:1903.08254*, 2019.

[3] J. Humplik, A. Galashov, L. Hasenclever, P. A. Ortega, Y. W. Teh, and N. Heess, "Meta reinforcement learning as task inference," *arXiv preprint arXiv:1905.06424*, 2019.

[4] L. Zintgraf, M. Igl, K. Shiarlis, A. Mahajan, K. Hofmann, and S. Whiteson, "Variational task embeddings for fast adapta-tion in deep reinforcement learning,"

[5] C. Finn, P. Abbeel, and S. Levine, "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.

[6] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," *Journal of machine Learning research*, vol. 3, no. Jan, pp. 993–1022, 2003.

[7] P. Hennig, D. Stern, R. Herbrich, and T. Graepel, "Kernel topic models," *arXiv preprint arXiv:1110.4713*, 2011.

[8] A. Srivastava and C. Sutton, "Autoencoding variational inference for topic models," *arXiv preprint arXiv:1703.01488*, 2017.

[9] T. Yu, D. Quillen, Z. He, R. Julian, K. Hausman, S. Levine, and C. Finn, "Meta-world: A benchmark and evaluation for multi-task and meta-reinforcement learning," 2019.

[10] K. Kurach, A. Raichuk, P. Stańczyk, M. Zajac, O. Bachem, L. Espeholt, C. Riquelme, D. Vincent, M. Michalski, O. Bousquet, *et al.*, "Google research football: A novel reinforcement learning environment," *arXiv preprint arXiv:1907.11180*, 2019.

[11] D. P. Kingma and M. Welling, "Auto-encoding variational bayes," *arXiv preprint arXiv:1312.6114*, 2013.