Learning Multimodal Transition Dynamics for Model-Based Reinforcement Learning: Abstract.

Thomas M. Moerland, Joost Broekens and Catholijn M. Jonker
Dep. of Computer Science, Delft University of Technology, The Netherlands

1 Introduction

In this work[1] we study how to learn stochastic, multimodal transition dynamics in reinforcement learning tasks. Model-based RL is an important class of RL algorithms that learns and utilizes transition dynamics to enhance data efficiency and target exploration. However, many tasks environments inherently have *stochastic* transition dynamics. We therefore require methods to approximate such complex distributions, while they should also scale to higher-dimensions. In this paper we study conditional variational inference in (deep) neural networks as a principled method to solve this challenge.

2 Conditional Variational Inference

Our goal is to learn a generative model of a (possibly multimodal) distribution $p(y|x)$. We assume the generative process is actually conditioned on some unobserved latent variables z: $p(y|x) = \int p(y|z,x)p(z|x)dz$. The stochastic latent variables z provide the flexibility to predict complex marginal outcome distributions $p(y|x)$. However, the z variables are unobserved and the posterior over z, $p(z|y,x)$, is analytically intractable in most models of interest (for example deep non-linear neural networks).

However, the parameters of this distribution can be efficiently approximated with Stochastic Gradient Variation Bayes (SGVB) [1]. We may first derive a variational lower bound $\mathcal{L}(y|x)$ on our data likelihood $p(y|x)$:

$$\log p(y|x) \geq \mathbb{E}_{z\sim q(\cdot|x,y)}[\log p_\theta(y|z,x)] - D_{\mathrm{KL}}[q_\phi(z|x,y)\|p_\phi(z|x)] = \mathcal{L}(y|x;\theta,\phi) \tag{1}$$

where θ denotes the parameters in a generative network $p_\theta(y|x,z)$, ϕ denotes the parameters in an *inference network* $q_\phi(z|y,x)$ and prior $p_\phi(z|x)$, and D_{KL} denotes the Kullback-Leibler (KL) divergence. The parametric inference network approximates the intractable true posterior over z, while the KL divergence term ensures that the inference network $q_\theta(z|y,x)$ does not diverge too much from the prior $p_\phi(z|x)$. This acts as a regularizer, and ensures that we can at test time (when we do not observe y) sample from $p(z|x)$ instead of $q(z|x,y)$.

[1] Originally published as: Moerland TM, Broekens J, Jonker CM. Learning Multimodal Transition Dynamics for Model-Based Reinforcement Learning. In *Scaling-Up Reinforcement Learning (SURL) Workshop* @ ECML. 2017.

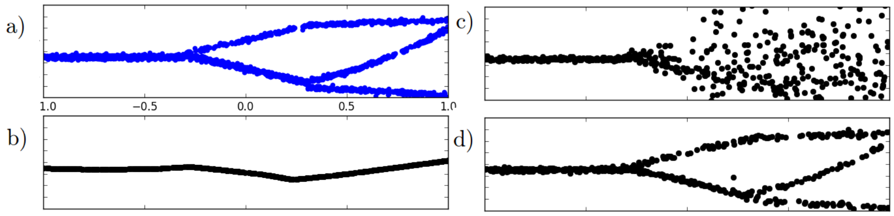

Fig. 1. Comparison of samples from the models produced by multi-layer perceptron (MLP) and conditional variational inference (CVI) networks after training for 30,000 mini-batches. a) Ground truth (artificial) data. b) MLP with deterministic prediction. c) MLP with stochastic inputs. d) CVI with discrete latent z variables.

While the original variational auto-encoder was introduced as a generative model for $p(y)$, the RL setting requires us to condition the entire generative process on the previous state x (i.e. in prior, inference and recognition network). Then, the objective in Eq. 1 can be trained on a single computational graph through the *reparametrization* trick, for details see [1]. For this work we experiment with different types of continuous and discrete latent variables z, for more details see the original paper.

3 Experiments

We illustrate these ideas on an artificial transition function (Fig. 1a) with unimodal (Fig. 1a, left part), bimodal (middle), and trimodal (right) dynamics. Fig. 1b shows the predictions of a default MLP trained on mean-squared error, which erroneously fits the conditional expectation. Since CVI uses additional noise input to the network (by sampling z), we also compare to a MLP with additional stochastic inputs. Without the inference network, the model is unable to map the input noise to the correct output distribution (Fig. 1c). In contrast, the model with CVI (Fig. 1d) accurately learns the different types of multi-modality, while also correctly predicting the deterministic part of the function. Please refer to the original paper for experiments on reinforcement learning tasks.

4 Conclusion

Our results show that conditional variational inference in deep neural networks successfully predicts multimodal distributions, but also robustly ignores these for deterministic parts of the transition dynamics. Due to the flexibility of neural networks as black-box function approximators, these results are applicable to a variety of RL tasks, and are a key preliminary for model-based RL in stochastic domains.

References

1. Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.