THEODON PROJECT NOTES

FLORIN GOGIANU, TUDOR BERARIU

Abstract

This document should work as a lab journal of sorts containing expositions of the ideas being tested, experimental results and their interpretation.

CONTENTS

1 ACTIVE LEARNING, THE BAYESIAN WAY

Saturday, Jun 22, 2019

In this note I will try to informally recall the origins of this research project and then focus on principled prioritization keys in Active Learning with a focus on Bayesian Active Learning.

1.1 How I came up with *uncertainty*-based prioritization?

I first came to think about novel prioritization schemes while reading the work done by Mattar and Daw 2018 on models explaining the different regimes of hippocampal replay. The authors argue that prioritizing experiences in a `DYNA` framework, based on a key composed of a `Gain` term measuring how much the expected return improves with the change in policy induced by learning from a given transition and a `Need` term which is the discounted number of times the agent is expected to visit a target state given the current state (think *successor representations*), resulting in an `ExpectedValueofBackup` $=$ `Gain` $\times$ `Need` measure. The interplay between these two terms, the authors argue, can model the disconnected observations of *forward, backward* and *offline replay* in a simulated spatial navigation task.

At the time I was still thinking about replacing ER with a parametrized model and the work of Mattar and Daw 2018 seemed relevant. About the same time I was having the first contacts with the Posterior-Sampling RL literature for exploration (Osband, Aslanides, et al. 2018; Russo et al. 2017) and reproducing the various PER implementation so the idea of using the *epistemic uncertainty* of the agent for selecting valuable transitions came naturally. **Basically it just seemed like a fun comparison**.

But why the variance of the predictive distribution as a measure for the uncertainty of the agent? Well, in PSRL you usually don't compute explicitly the uncertainty of the estimator but rely on Thompson sampling to select at the beginning of each episode a policy from the posterior distribution according to the probability they are optimal – in this setup exploration is guided by the variance of the sample policies (Osband, Russo, et al. 2013). So I did what every respectable researcher does: I turned to a blog post (Gal 2015) I had in the back of my head for some time where I recall seeing a definition for the epistemic uncertainty of the a deep neural network – the predictive variance of the model with parameters sampled from the posterior.

1.2 Getting formal?

While seemingly fun and with some degree of empirical evidence, prioritizing experiences by their epistemic uncertainty isn't really theoretically sound. It can easily be seen that learning from the transition with the highest uncertainty we are not guaranteed that the new policy will lead to a higher expected return. However that is also the case with `TD-error` prioritization.

Following Houlsby et al. 2011 we note that from a Bayesian perspective, identifying the best transitions to learn from means *reducing the number of hypotheses as fast as possible*, which is another way of saying to reduce the entropy of the posterior distribution:

$$\underset{\mathcal{D}}{\operatorname{argmin}}\, \mathbb{H}[\boldsymbol{\theta} \mid \mathcal{D}] = \int p(\boldsymbol{\theta} \mid \mathcal{D}) \log p(\boldsymbol{\theta} \mid \mathcal{D}) d\boldsymbol{\theta}.$$

This can be greedily approximated by finding the transition that maximises the decrease in expected posterior entropy:

$$\underset{\boldsymbol{x}}{\operatorname{argmax}}\, \mathbb{H}[\boldsymbol{\theta} \mid \mathcal{D}] - \mathbb{E}_{y \sim p(y \mid \boldsymbol{x}, \mathcal{D})} \left[\mathbb{H}[\boldsymbol{\theta} \mid y, \boldsymbol{x}, \mathcal{D}] \right] \tag{1}$$

Houlsby et al. 2011 points out that while some works use this objective directly, this is not feasible for non-trivial models. The authors further claim that Eqn. (1), maximizing the decrease in entropy of the model given some $\boldsymbol{x}$, is equivalent to the conditional mutual information between predictions and the model parameters. That is how much information about the model parameters we gain from y.

Recall one of the alternate forms we can derive from the definition of the *Mutual Information*:

$$\begin{aligned}
\mathbb{I}[X, Y] &= \mathbb{E}_{x,y \sim p(x,y)} \left[\log \frac{p(x,y)}{p(x)p(y)} \right] \\
&= \mathbb{E}_{x,y \sim p(x,y)} \left[\log \frac{p(x,y)}{p(x)} \right] - \mathbb{E}_{x,y \sim p(x,y)} \left[\log p(y) \right] \\
&= \mathbb{E}_{x \sim p(x)} \left[\mathbb{E}_{y \sim p(y)} \left[\log p(y) \right] \right] - \sum_{y} \left(\sum_{x} p(x,y) \right) \log p(y) \\
&= \mathbb{E}_{x \sim p(x)} \left[\mathbb{H}[Y \mid X = x] \right] - \mathbb{E}_{y \sim p(y)} \left[\log p(y) \right] \\
&= -\mathbb{H}[Y \mid X] + \mathbb{H}[Y] \\
&= \mathbb{H}[Y] - \mathbb{H}[Y \mid X].
\end{aligned}$$

Similarly we can arrive at the following objective based on the conditional mutual information between $\mathbb{I}\left[\boldsymbol{\theta}, y \mid \boldsymbol{x}, \mathcal{D}\right]$:

$$\underset{\boldsymbol{x}}{\operatorname{argmax}} \mathbb{H}[y \mid \boldsymbol{x}, \mathcal{D}] - \mathbb{E}_{\boldsymbol{\theta} \sim p(\boldsymbol{\theta} \mid \mathcal{D})}\left[\mathbb{H}[y \mid \boldsymbol{x}, \boldsymbol{\theta}]\right] \tag{2}$$

Using this objective in our value-based RL setting, we need to replace:

$$\begin{aligned} p(y \mid \boldsymbol{x}, \mathcal{D}) &= p\big(Q(s, a \mid \mathcal{D})\big) \\ &= \int p\big(Q(s, a \mid \boldsymbol{\theta})\big)\, p(\boldsymbol{\theta} \mid \mathcal{D})\, d\boldsymbol{\theta} \end{aligned}$$

Following Gal et al. 2017 derivation for the classification setting we arrive at the following objective:

$$\begin{aligned} \mathbb{I}\left[Q(s, a), \boldsymbol{\theta}\right] = &- \int_{\mathrm{Dom}(Q_{s,a})} p\big(Q(s, a \mid \mathcal{D})\big) \log p\big(Q(s, a \mid \mathcal{D})\big)\, dQ_{s,a} \\ &+ \mathbb{E}_{\boldsymbol{\theta} \sim p(\boldsymbol{\theta} \mid \mathcal{D})}\left[\int_{\mathrm{Dom}(Q_{s,a})} p\big(Q(s, a \mid \boldsymbol{\theta})\big) \log p\big(Q(s, a \mid \boldsymbol{\theta})\big)\, dQ_{s,a}\right] \end{aligned} \tag{3}$$

I'm not sure how intelligible is this last equation so this is how I understand it. The first term is the entropy of the $Q(s, a)$ estimate when sampling from the posterior. We will call this the Monte-Carlo estimate:

$$Q^{MC}(s, a) = \frac{1}{T} \sum_{t} Q(s, a \mid \boldsymbol{\theta}_t) \tag{4}$$

The second problem in computing this entropy is $p(Q(s, a))$. What do we mean by the probability of the state-action value function? This implies keeping a distribution over the returns. Since a Gaussian assumption is not really good for this, I believe we can use a distributional algorithm. This way we can compute the entropy of the $Q(s, a)$ distribution, when $Q(s, a)$ is actually a Monte-Carlo estimate $Q^{MC}(s, a)$.

The second term is simply an expected value of the entropy of $Q(s, a)$ given samples from the posterior.

I don't know exactly how to simplify this. To sum-up, we implement a Bayesian Categorical-DQN and compute the values above, yay.

1.3 Other prioritization keys

Gal et al. 2017 reviews some other prioritization measures I will mention here. The notation is based on a classification task because I am lazy.

- *Max Entropy* – the example with the largest predictive entropy is picked. In MLE this is the entropy of the softmax distribution.
- Maximise the *Variation Ratios* $1 - \max_y p(y \mid x, \mathcal{D})$
- Maximize the mean standard deviation. This resembles the prioritization measure we used so far.

REFERENCES

Gal, Yarin

2015 *What my Deep Model Doesn't Know*, https://www.cs.ox.ac.uk/people/yarin.gal/website/blog_3d801aa532c1ce.html (visited on 06/22/2019).

Gal, Yarin, Riashat Islam, and Zoubin Ghahramani

2017 "Deep Bayesian Active Learning with Image Data", in *ICML*.

Houlsby, Neil, Ferenc Huszar, Zoubin Ghahramani, and Máté Lengyel

2011 "Bayesian Active Learning for Classification and Preference Learning", *CoRR*, abs/1112.5745.

Mattar, Marcelo G and Nathaniel D Daw

2018 "Prioritized memory access explains planning and hippocampal replay", *Nature Neuroscience*, 21, 11, p. 1609.

Osband, Ian, John Aslanides, and Albin Cassirer

2018 "Randomized prior functions for deep reinforcement learning", in *Advances in Neural Information Processing Systems*, pp. 8617-8629.

Osband, Ian, Daniel Russo, and Benjamin Van Roy

2013 "(More) Efficient Reinforcement Learning via Posterior Sampling", in *Advances in Neural Information Processing Systems 26*, ed. by C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger, Curran Associates, Inc., pp. 3003-3011, http://papers.nips.cc/paper/5185-more-efficient-reinforcement-learning-via-posterior-sampling.pdf.

Russo, Daniel, Benjamin Van Roy, Abbas Kazerouni, and Ian Osband

2017 "A Tutorial on Thompson Sampling", *CoRR*, abs/1707.02038, arXiv: 1707.02038, http://arxiv.org/abs/1707.02038.