Multi Agent Meta Reinforcement Learning on Neural Networks

Alexander D. Cai
Harvard School of Engineering
Harvard College
Cambridge, MA 02138
alexcai@college.harvard.edu

Oliver Cheng
Department of Mathematics
Harvard College
Cambridge, MA 02138
ocheng@college.harvard.edu

Abstract

The abstract paragraph should be indented ½ inch (3 picas) on both the left- and right-hand margins. Use 10 point type, with a vertical spacing (leading) of 11 points. The word **Abstract** must be centered, bold, and in point size 12. Two line spaces precede the abstract. The abstract must be limited to one paragraph.

1 Introduction to Meta-Reinforcement Learning

In recent years there has been a growing amount of excitement about *meta-learning* in order to solve a wider set of problems. This is often used in reinforcement learning tasks. In a standard reinforcement learning task, we have a policy that governs how an agent transitions between states in an environment. There are rewards that the agent can receive, and the optimal policy is one that maximizes rewards. We can formalize this by defining a reinforcement learning regime as a Markov Decision Process (MDP) where S is the set of states, A is the set of actions, T is the transition distribution (how states transition given an action), and R is the reward, which is a function of the state. Let π_θ be the policy, which is parameterized by some θ. Thus for some "task" $\mathcal{T}$, the goal is thus to maximize the objective

$$\mathcal{J}_{\mathcal{T}}(\theta) = \mathbb{E}_{a_t \sim \pi_\theta(s_t), s_{t+1} \sim T(s_t, a_t)} \left(\sum_t \gamma^t R(s_t, a_{t-1}) \right).$$

The task $\mathcal{T}$ determines the rewards, and hence the policy that is optimized depend heavily on what the task $\mathcal{T}$ is.

Reinforcement learning is used to solve many problems such as Go, Chess, Atari games, etc. where the environment is unknown and is structured in a way which an agent must discover an optimal strategy. This framework gives us a way to view human beings and the nervous system as a reinforcement learning program where we humans are the agent. (Un)Fortunately, this is a bit too basic of a way of modeling human beings and the brain, as reinforcement learning networks have not taken over humanity yet (citation needed). One of the reasons is that rewards and the task that an agent needs to optimize a policy for is constantly changing in the real world. Thus we introduce the idea of meta-learning, where rather than optimizing a single policy for a single task $\mathcal{T}$, we optimize for a learning strategy on how to optimize a policy for a distribution of tasks $p(\mathcal{T})$. This idea is motivated by the idea in psychology of "learning how to learn" and is conceptually closer to how the brain works (that one prefrontal cortex Deepind paper.). With meta-learning, due to the way we are improving the way we are finding the policy, this is a way we can introduce the idea of "past experiences" to help train policy for tasks that are similar.

Preprint. Under review.

2 Implementation of Meta-Learning MARL

A very basic way to model a brain is as a simple fixed architecture feedforward neural network where each neuron is its own agent and we are optimizing a multi-agent reinforcement learning problem. Then, to imitate a model-free dopamine system which determines which synapses are reinforced, we consider a policy neural network which will update the weights for the neural net. The state space of this system is essentially the possible weights that this neural net can have. Each agent is therefore each node of the neural network, and the agent is essentially in control of the extent to which they transmit their activation. In other words, the agent corresponding to the i-th neuron of the ℓ-th layer, $x_i^{(\ell)}$, has control over the weights $\begin{bmatrix} W_{1,i}^{(\ell)} \\ \vdots \\ W_{R_{\ell+1},i}^{(\ell)} \end{bmatrix}$. Actions correspond to changing these activations. This is the inner loop of learning. On the other hand, in order to improve the policy, we use PPO on the loss from each episode of learning. Reward is equivalent to negative loss, and hence the system is rewarded for smaller loss.

3 Meta-Reinforcement Learning in the Brain

Basically try and repeat the thing in Box 4 of https://www.sciencedirect.com/science/article/pii/S1364661319300610

maybe this is garbage lol

In the Theory of Neural Computation, we can see reinforcement-learning-esque regimes. In particular researchers have found that the dopamine system seems to follow the TD algorithm, a reinforcement learning algorithm. In particular, the error function of the TD algorithm operates in the same way to how dopamine neurons learn their firing rates in order to associate stimuli with future rewards. In this case, the dopamine neuron is the agent and it can change its firing rate as an action.

4 Comparison to Known Learning Rules

We compare this meta-learning (learning of the best learning rule) to accepted learning rules in the literature. Stochastic gradient descent (SGD) converges the quickest, but due to the weight transport problem, is not biologically plausible (citation needed). Biologically plausible alternatives involve learning from a global error signal, and involve perturbation-type learning rules. Global error signals have been observed in the brain, but it is still under much research exactly how the brain uses these error signals in order to learn and update neurons. We will look at node (NP) and weight perturbation (WP). We compared how our new model did for teacher-student learning and learning to classify the MNIST dataset.

The idea behind node or weight perturbation is to take some perturbation ξ of either the weights or the nodes, and then see how that changes the objective or loss. The learning rule update then corresponds to scaling this perturbation by whether or not it increases or decreases the loss. Consider a general neural network with L layers and a data set $\{\mathbf{x}_i^{(0)}, \mathbf{y}_i\}_{i=1}^N$. Let $\mathbf{W}_\ell \in \mathbb{R}^{R_{\ell+1} \times R_\ell}$ where R_ℓ is the number of neurons in layer ℓ. We define the forward pass as

$$\mathbf{X}^{(\ell)} = \sigma\left(\mathbf{X}^{(\ell-1)}\mathbf{W}_{\ell-1}^\top\right). \tag{1}$$

Where σ is some non-linearity such as ReLU. We use MSE loss for student teacher

$$E = \frac{1}{NR_L}\frac{1}{2}\|\mathbf{X}^{(L)} - \mathbf{Y}\|^2.$$

and for classification into $C = R_L$ classes let

$$p(\mathbf{x}_i^{(L)} = r) = \frac{\exp x_{j,r}^{(L)}}{\sum_{j=1}^C \exp x_{i,j}^{(L)}}$$

we use cross entropy loss

$$E = -\frac{1}{N}\sum_{i=1}^N \sum_{r=1}^C y_r \log p(\mathbf{x}_i^{(L)} = r)$$

In SGD, we use backpropagation to update the weight parameters by

$$\Delta_{SGD}\mathbf{W}_\ell = \eta\frac{\partial E}{\partial \mathbf{W}_\ell},$$

where the chain rule means we need to "transport" these values backwards through the neural network. On the other hand, in WP/NP, we have two forward passes. For the first, we define the same as 1, however for NP, we will let the perturbation be a vector $\xi \in \mathbb{R}^{N\times R_\ell}$ where all entries are distributed i.i.d $\mathcal{N}(0,\sigma)$. We can then define the forward pass as

$$\widetilde{\mathbf{X}}^{(\ell)} = \sigma\left(\widetilde{\mathbf{X}}^{(\ell-1)}\mathbf{W}_{\ell-1}^\top + \xi^{(\ell)}\right).$$

While for WP we take $\mathbf{\Xi}^{(\ell)} \in \mathbb{R}^{R_{\ell+1}\times R_\ell}$ where all entries are distributed i.i.d $\mathcal{N}(0,\sigma)$.

$$\hat{\mathbf{X}}^{(\ell)} = \sigma\left(\hat{\mathbf{X}}^{(\ell-1)}\left(\mathbf{W}_{\ell-1} + \mathbf{\Xi}^{(\ell-1)}\right)^\top\right).$$

We can then define two errors for each NP and WP respectively. For MSE

$$E_N = \frac{1}{NR_L}\frac{1}{2}\|\widetilde{\mathbf{X}}^{(L)} - \mathbf{Y}\|^2 \qquad E_W = \frac{1}{NR_L}\frac{1}{2}\|\hat{\mathbf{X}}^{(L)} - \mathbf{Y}\|^2.$$

For classification

$$E_N = -\frac{1}{N}\sum_{i=1}^{N}\sum_{r=1}^{C} y_r \log p(\widetilde{\mathbf{x}}_i^{(L)} = r) \qquad E_W = -\frac{1}{N}\sum_{i=1}^{N}\sum_{r=1}^{C} y_r \log p(\hat{\mathbf{x}}_i^{(L)} = r)$$

From here we can simply transmit the global error signal $E - E_N$ or $E - E_W$ and update our weights according to this error signal. In particular, we update parameters as follows for NP

$$\Delta_{NP}\mathbf{W}_\ell = \frac{\eta}{\sigma}(E - E_N)\sum_{i=1}^{R_{\ell+1}}\mathbf{x}_i^{(\ell)}\xi^{(\ell+1)\top}$$

While for WP we update as

$$\Delta_{WP}\mathbf{W}_\ell = \frac{\eta}{\sigma}(E - E_W)\mathbf{\Xi}^{(\ell)}$$

The idea is that perturbations which cause decrease the objective will be added to the parameters of the model. In expectation we have that up to a constant

$$\langle\Delta_{NP}\mathbf{W}_\ell\rangle_{\xi^{(\ell)}} = \langle\Delta_{WP}\mathbf{W}_\ell\rangle_{\mathbf{\Xi}^{(\ell)}} = \Delta_{SGD}\mathbf{W}_\ell.$$

Node and weight perturbation are often used in small teacher-student learning cases where there is only a single perceptron. Recent work by Hiratani et al. explores node perturbation in neural networks with a hidden layer, and finds computational and stability issues that suggest biological implausibility (cite hiratani et al). Nonetheless, these perturbation techniques remain an interesting way to investigate possible learning rules for neurons.

For the student-teacher model

Going to MNIST, the problem becomes much more noisy. In order to decrease noise, we first preprocessed by projecting the data onto the first 10 principal components ($> 99\%$ of the variance), and decreased the learning rate by a factor of 100. It was also necessary to increase the hidden layer width $L_h = 1000$, as otherwise the model did not have sufficient complexity. As observed by Hiratani et al. the amount of training time required for these global error signal models to converge for MNIST is much higher, and took too long for the scope of this project. The results are detailed in Fig XXXXX.

5 Citations, figures, tables, references

These instructions apply to everyone.

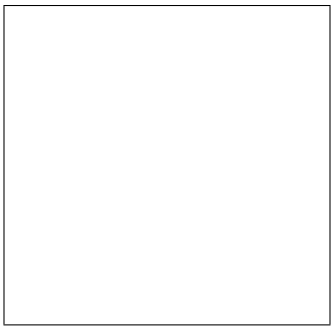

Figure 1: Sample figure caption.

5.1 Citations within the text

The `natbib` package will be loaded for you by default. Citations may be author/year or numeric, as long as you maintain internal consistency. As to the format of the references themselves, any style is acceptable as long as it is used consistently.

The documentation for `natbib` may be found at

`http://mirrors.ctan.org/macros/latex/contrib/natbib/natnotes.pdf`

Of note is the command `\citet`, which produces citations appropriate for use in inline text. For example,

```
\citet{hasselmo} investigated\dots
```

produces

Hasselmo, et al. (1995) investigated…

If you wish to load the `natbib` package with options, you may add the following before loading the `neurips_2020` package:

```
\PassOptionsToPackage{options}{natbib}
```

If `natbib` clashes with another package you load, you can add the optional argument `nonatbib` when loading the style file:

```
\usepackage[nonatbib]{neurips_2020}
```

As submission is double blind, refer to your own published work in the third person. That is, use "In the previous work of Jones et al. [4]," not "In our previous work [4]." If you cite your other papers that are not widely available (e.g., a journal paper under review), use anonymous author names in the citation, e.g., an author of the form "A. Anonymous."

5.2 Footnotes

Footnotes should be used sparingly. If you do require a footnote, indicate footnotes with a number[1] in the text. Place the footnotes at the bottom of the page on which they appear. Precede the footnote with a horizontal rule of 2 inches (12 picas).

Note that footnotes are properly typeset *after* punctuation marks.[2]

5.3 Figures

All artwork must be neat, clean, and legible. Lines should be dark enough for purposes of reproduction. The figure number and caption always appear after the figure. Place one line space before the figure

[1] Sample of the first footnote.

[2] As in this example.

Table 1: Sample table title

	Part	
Name	Description	Size (μm)
Dendrite	Input terminal	$\sim$100
Axon	Output terminal	$\sim$10
Soma	Cell body	up to 10^6

caption and one line space after the figure. The figure caption should be lower case (except for first word and proper nouns); figures are numbered consecutively.

You may use color figures. However, it is best for the figure captions and the paper body to be legible if the paper is printed in either black/white or in color.

5.4 Tables

All tables must be centered, neat, clean and legible. The table number and title always appear before the table. See Table 1.

Place one line space before the table title, one line space after the table title, and one line space after the table. The table title must be lower case (except for first word and proper nouns); tables are numbered consecutively.

Note that publication-quality tables *do not contain vertical rules.* We strongly suggest the use of the `booktabs` package, which allows for typesetting high-quality, professional tables:

`https://www.ctan.org/pkg/booktabs`

This package was used to typeset Table 1.

Broader Impact

Authors are required to include a statement of the broader impact of their work, including its ethical aspects and future societal consequences. Authors should discuss both positive and negative outcomes, if any. For instance, authors should discuss a) who may benefit from this research, b) who may be put at disadvantage from this research, c) what are the consequences of failure of the system, and d) whether the task/method leverages biases in the data. If authors believe this is not applicable to them, authors can simply state this.

Use unnumbered first level headings for this section, which should go at the end of the paper. **Note that this section does not count towards the eight pages of content that are allowed.**

Acknowledgments and Disclosure of Funding

Use unnumbered first level headings for the acknowledgments. All acknowledgments go at the end of the paper before the list of references. Moreover, you are required to declare funding (financial activities supporting the submitted work) and competing interests (related financial activities outside the submitted work). More information about this disclosure can be found at: `https://neurips.cc/Conferences/2020/PaperInformation/FundingDisclosure`.

Do **not** include this section in the anonymized submission, only in the final paper. You can use the `ack` environment provided in the style file to autmoatically hide this section in the anonymized submission.

References

References follow the acknowledgments. Use unnumbered first-level heading for the references. Any choice of citation style is acceptable as long as you are consistent. It is permissible to reduce the

font size to `small` (9 point) when listing the references. **Note that the Reference section does not count towards the eight pages of content that are allowed.**

[1] Alexander, J.A. & Mozer, M.C. (1995) Template-based algorithms for connectionist rule extraction. In G. Tesauro, D.S. Touretzky and T.K. Leen (eds.), *Advances in Neural Information Processing Systems 7*, pp. 609–616. Cambridge, MA: MIT Press.

[2] Bower, J.M. & Beeman, D. (1995) *The Book of GENESIS: Exploring Realistic Neural Models with the GEneral NEural SImulation System.* New York: TELOS/Springer–Verlag.

[3] Hasselmo, M.E., Schnell, E. & Barkai, E. (1995) Dynamics of learning and recall at excitatory recurrent synapses and cholinergic modulation in rat hippocampal region CA3. *Journal of Neuroscience* **15**(7):5249-5262.