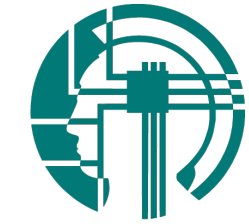


Meta-Learned Models of Cognition

Marcel Binz

September 2, 2022

Max Planck Institute for Biological Cybernetics
Computational Principles of Intelligence Lab

Computational Models

Computational models are immensely useful for improving our understanding of the human mind.

Traditionally, such computational models have been **hand-designed** by expert researchers.

Examples:

- Bayesian models
- Cognitive architectures
- Connectionist networks

Meta-Learning

Meta-learning:

- Framework for learning computational models of learning from data.
- Radically different approach for constructing computational models.

Meta-Learning

Psychologists have recently started to apply meta-learning to study human learning.

Meta-learned models capture a wide range of empirically observed phenomena that could not be explained otherwise:

- Reproduce human biases in probabilistic reasoning [Dasgupta et al., 2020].
- Discover heuristic decision-making strategies used by people [Binz et al., 2022].
- Generalize compositionally on language tasks in a human-like manner [Lake, 2019].

Meta-Learning

Recent result from the machine learning community:

Ortega et al. (2019). "Meta-learning of sequential strategies."

Meta-learning can be used to construct Bayes-optimal learning algorithms.

Links meta-learning to an already well-established framework: the **rational analysis of cognition** [Anderson, 2013].

Rational Analysis

Rational analysis:

1. Specify the goal of the agent and the environment it interacts with.
2. Derive the Bayes-optimal solution for the task at hand.
3. Test model predictions against empirical data.
4. Modify assumptions and repeat the whole process if needed.

Tenenbaum. (2021). Homepage.

Explains “why cognition works, by viewing it as an approximation to ideal statistical inference given the structure of natural tasks and environments”.

Goals of this tutorial:

1. Introduce the general ideas behind meta-learning.
2. Highlight its close connections to the Bayesian inference.
3. Walk through a simple implementation in PyTorch.
4. Discuss advantages and disadvantages to other frameworks.

Meta-Learning

Learning

The prefix *meta-* is generally used in a self-referential sense.

Meta-learning therefore refers to learning about learning.

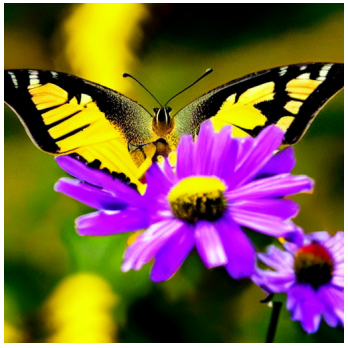
First establish a common definition of *learning*.

Learning

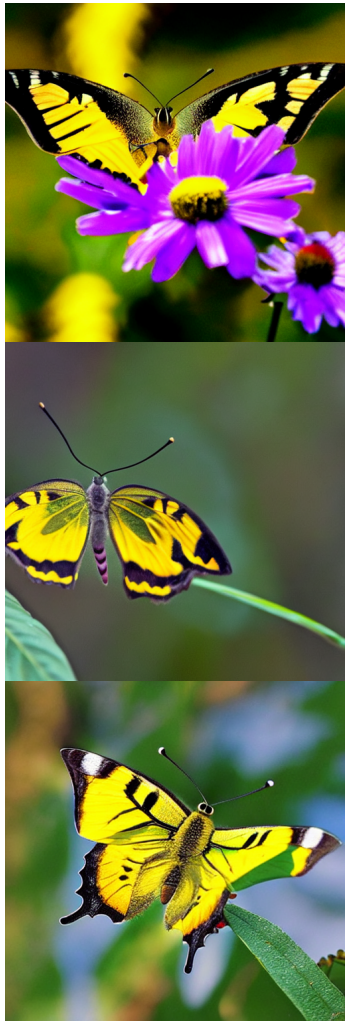
Mitchell (1997). "Machine learning."

For a given task, training experience, and performance measure, an algorithm is said to learn if its performance at the task improves with experience.

Learning

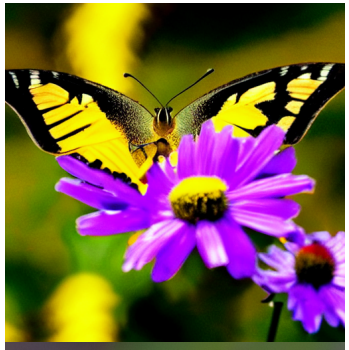


Learning

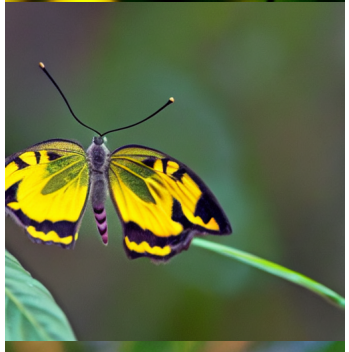


Learning

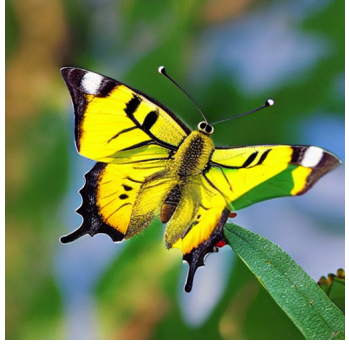
x_k



10cm



15cm



13cm

Learning

Task: predict the size of the next observed insect x_{t+1}.

Training experience: three exemplars with lengths of $x_1 = 10\text{cm}$, $x_2 = 15\text{cm}$, and $x_3 = 13\text{cm}$.

Performance measure: mean squared error or negative log-loss.

Bayesian Inference

Rational analysis of cognition: compare human behavior to that of an optimal learning algorithm.

No learning algorithm is better than another when averaged over all possible problems [Wolpert, 1996].

We first have to make additional assumptions about the to-be-solved problem.

Bayesian Inference

Example:

1. Each observed length x_k is sampled from a normal distribution with mean μ and standard deviation σ, i.e., $x_k \sim \mathcal{N}(\mu, \sigma)$.
2. The mean length μ cannot be observed directly, but the standard deviation σ is known.
3. The unobserved mean length is sampled from a standard normal distribution which remains constant over time, i.e., $\mu \sim \mathcal{N}(0, 1)$.

Bayesian inference provides the way of making optimal predictions under such assumptions.

Bayesian Inference

Prior $p(\mu)$: defines an agent's initial beliefs about possible parameter values before observing any data.

Likelihood $p(x_{1:t}|\mu)$: captures the agent's knowledge about how data is generated for a given set of parameters.

In our example:

$$p(\mu) = \mathcal{N}(\mu; 0, 1)$$

$$p(x_{1:t}|\mu) = \prod_{k=1}^{t} p(x_k|\mu) = \prod_{k=1}^{t} \mathcal{N}(x_k; \mu, \sigma)$$

Bayesian Inference

Predictive posterior distribution $p(x_{t+1}|x_{1:t})$: can be used to make probabilistic predictions about future observations.

First, compute the posterior distribution over parameters:

$$p(\mu|x_{1:t}) = \frac{p(x_{1:t}|\mu)p(\mu)}{\int p(x_{1:t}|\mu)p(\mu)d\mu}$$

Then, average over all possible parameter values:

$$p(x_{t+1}|x_{1:t}) = \int p(x_{t+1}|\mu)p(\mu|x_{1:t})d\mu$$

Bayesian Inference

Not always possible to find analytical expressions for the predictive posterior distribution.

With assumptions that we made in our example, it has an analytical solution.

More on this in the code example later.

Bayesian Inference

Multiple arguments justify Bayesian inference as a normative procedure:

- Dutch book arguments
- Free energy minimization
- Performance-based justifications

Focus on performance-based justifications.

These can be used to derive meta-learning algorithms that learn approximations to Bayesian inference.

Bayesian Inference

Performance-based justifications assert that no learning algorithm can be better than Bayesian inference on a certain performance measure.

Aitchison (1975). "Goodness of prediction fit."

The predictive posterior distribution is the distribution that maximizes the log-likelihood of future observations when averaged over the generating distribution.

$$p(x_{t+1}|x_{1:t}) = \arg\max_{q} \mathbb{E}_{p(\mu, x_{1:t+1})} \left[\log q(x_{t+1}|x_{1:t})\right]$$

Proof is available here.

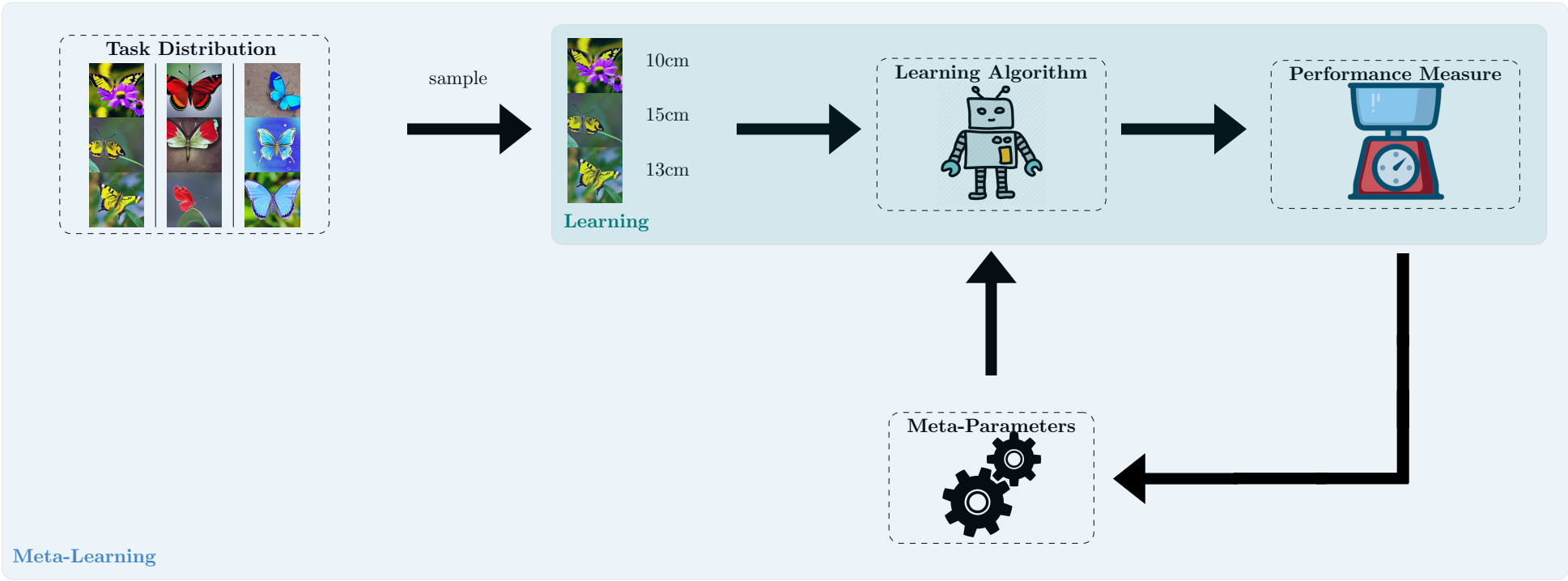
Meta-Learning

Schaul & Schmidhuber (2010). "Metalearning."

A meta-learning algorithm is any algorithm that uses its experience to change certain aspects of a learning algorithm, or the learning method itself, such that the modified learner is better than the original learner at learning from additional experience.

1. Decide on a base learning algorithm.
2. Determine which of its aspects can be modified.
3. Adapt the system on a set of similar tasks to maximize some measure of performance.

Meta-Learning



Meta-Learning

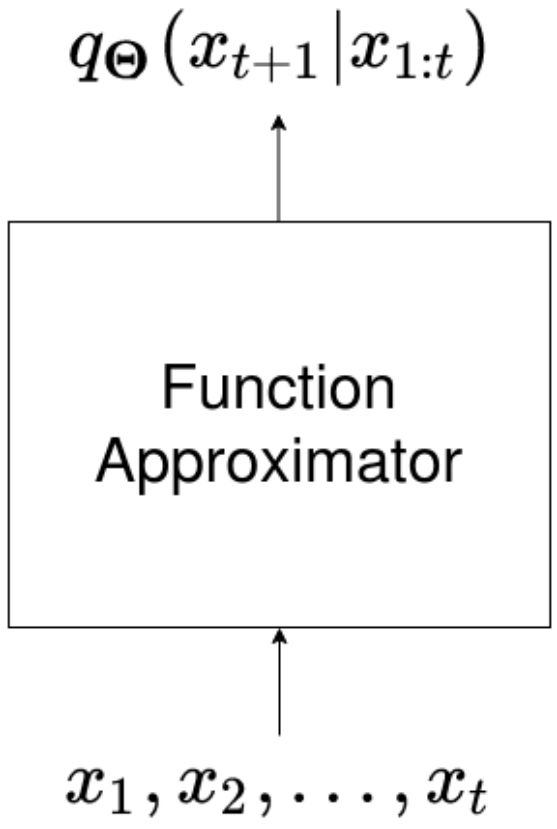
Possible to meta-learn:

- Hyperparameters for a base learning algorithm, such as learning rates, batch sizes, or the number of training epochs.
- Initial parameters of a neural network that is trained via stochastic gradient descent.
- Prior distributions in a probabilistic graphical model.
- Entire learning algorithms.

This tutorial focuses on the latter approach.

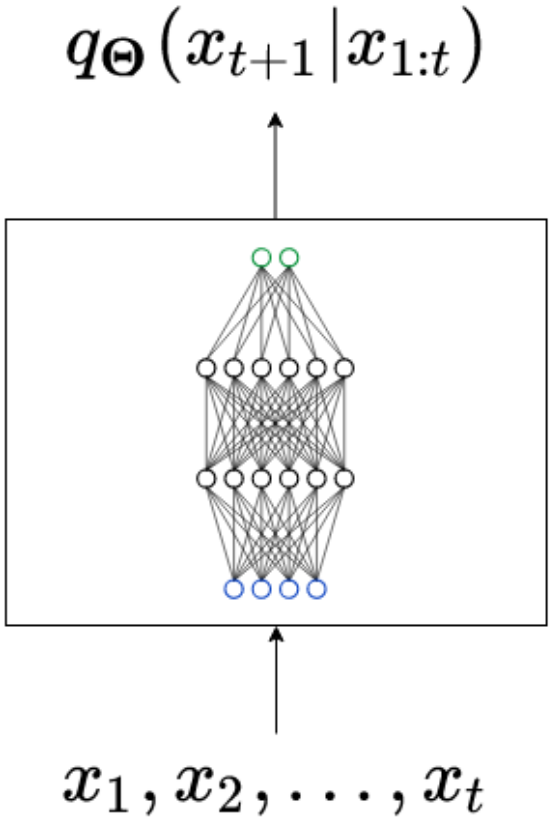
Meta-Learning

We teach a general-purpose function approximator to do this inference:



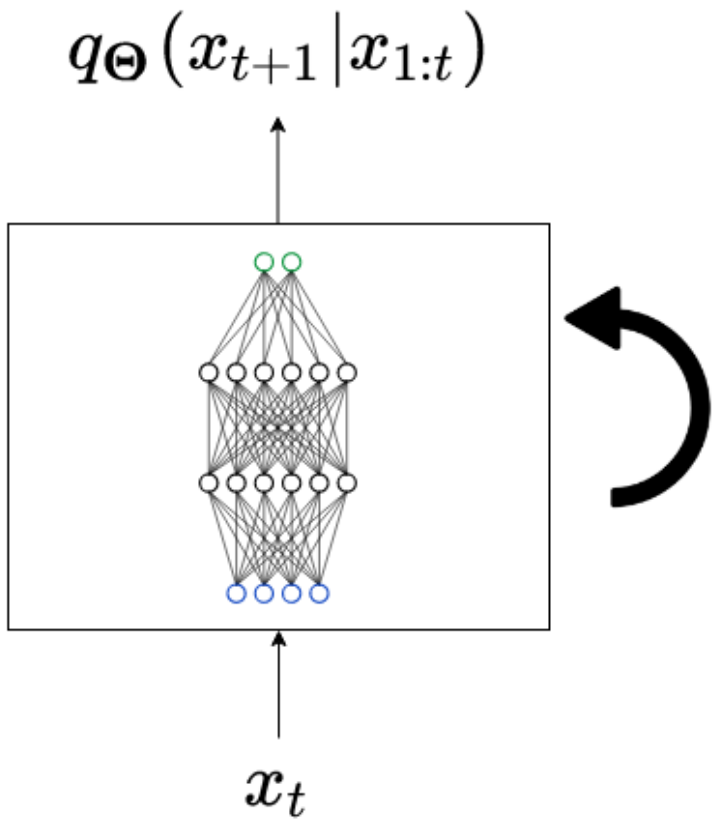
Meta-Learning

Commonly realized through some form of neural network:



Meta-Learning

Recurrent networks are the natural choice for sequences with varying length:



Meta-Learning

How to train this network?

The previous theorem suggests a training procedure:

$$p(x_{t+1}|x_{1:t}) = \arg\max_q \mathbb{E}_{p(\mu, x_{1:t+1})} \left[\log q(x_{t+1}|x_{1:t})\right]$$

Meta-Learning

How to train this network?

The previous theorem suggests a training procedure:

$$\boldsymbol{\Theta}^* = \arg\max_{\boldsymbol{\Theta}} \mathbb{E}_{p(\mu, x_{1:t+1})} \left[\log q_{\boldsymbol{\Theta}}(x_{t+1} | x_{1:t}) \right]$$

Replace the search over all possible distributions with a search over meta-parameters $\boldsymbol{\Theta}$.

Meta-Learning

How to train this network?

The previous theorem suggests a training procedure:

$$\Theta^* \approx \arg\max_{\Theta} \frac{1}{M} \sum \log q_{\Theta}(x_{t+1}|x_{1:t})$$

Train the network using a sample-based approximation of this performance measure.

Meta-Learning

Algorithm:

1. Sample a batch of tasks from the data-generating distribution.
2. Pass the sequence of observations through the network and generate the approximate predictive posterior distribution.
3. Perform a gradient step on the final objective from the last slide.

Meta-Learning

Initially, the network implements a random “learning” algorithm.

During meta-learning, the network adapts its weights through repeated interactions with sampled tasks.

Implements an approximation to exact Bayesian inference after meta-learning is completed.

To perform an inference, we only have to perform a forward-pass through the network.

No further weight updates are required.

Meta-Learning

In our example scenario, meta-learning would involve many tasks that require predicting the length of a species (not just one).

The model can acquire useful inductive biases during meta-learning, such as:

- Prior and likelihood are normally distributed.
- Noise variance is a constant.
- Exact values for the prior mean and variance.

or, it can learn other inductive biases when necessary.

Tool or Theory

Do we want to understand the process of meta-learning itself?

or

Do we want to use it purely as a methodological tool to construct Bayes-optimal learning algorithms?

Tool or Theory?

This tutorial focused on using meta-learning as a **methodological tool** to construct Bayes-optimal learning algorithms.

But, understanding the process of meta-learning may also be interesting.

Psychologists have been researching this question since the 1940s (e.g. [Harlow, 1949]).

How does meta-learning happen at evolutionary and developmental time-scales?

Meta-Reinforcement Learning

The same ideas can be extended to the reinforcement learning setting.

Learn a history-dependent policy $\pi_{\boldsymbol{\Theta}}(a_t|h_t)$ that maximizes reward instead of a predictive posterior distribution:

$$\boldsymbol{\Theta}^* = \arg\max_{\boldsymbol{\Theta}} \mathbb{E}_{p(\omega_r,\omega_s)\prod p(r_t|s_t,a_t,\omega_r)p(s_{t+1}|s_t,a_t,\omega_s)\pi_{\boldsymbol{\Theta}}(a_t|h_t)} \left[\sum_{t=1}^{H} r_t\right]$$

Trained on a distribution of Markov Decision Processes (MDPs) instead of supervised learning problems.

Example PyTorch Implementation

Why Not Bayesian Inference?

Why Not Bayesian Inference?

We have just established that it is possible to meta-learn Bayes-optimal learning algorithms.

Open questions:

- What additional explanatory power does the meta-learning framework offer?
- Why should one not just stick to tried-and-tested Bayesian inference?

Next:

- 4 arguments in favor of meta-learning.
- 3 disadvantages compared to Bayesian inference.

Intractable Inference

Bayesian inference is hard:

- It becomes intractable very quickly.
- It is only possible to find closed-form expressions for certain combinations of prior and likelihood.

Intractable Inference

Meta-learning approaches does not require an explicit calculation of the exact predictive posterior distribution.

Forward-pass through the network provides an approximation to the predictive posterior distribution.

Argument 1

Meta-learning can produce approximately optimal learning algorithms even if exact Bayesian inference is computationally intractable.

Intractable Inference

We need to determine the functional form of the predictive posterior distribution.

The chosen form may deviate from the true form of the predictive posterior distribution.

This may lead to approximation errors.

We can keep the approximation error small by choosing flexible functional forms, such as mixtures of Gaussians or discretized distributions with small bin sizes.

Intractable Inference

This feature is shared with many other methods for approximate inference:

- Variational inference
- Markov Chain Monte Carlo methods
- ...

It is ultimately an empirical question which of these approximations provides a better description of human learning.

Unspecified Problems

Posing the correct inference problem can be hard in the first place.

We need to specify the prior and the likelihood.

Easy for artificial problems (“small worlds”).

Hard for many real-world problems (“large worlds”).

Unspecified Problems

Unspecified priors or likelihoods are not a problem for meta-learning.

Meta-learning only requires sampled tasks from the data-generating distribution.

If you have a data-set of tasks $\mathcal{D} = \{x_1^i, x_2^i, \ldots, x_T^i\}_{i \leq K}$, you can apply meta-learning.

Argument 2

Meta-learning can produce optimal learning algorithms even if it is not possible to phrase the corresponding inference problem in the first place.

Unspecified Problems

This is a much weaker requirement.

Most approximate inference schemes cannot be applied in this setting.

Unique feature of the meta-learning framework.

Opens up totally new avenues for constructing Bayes-optimal learning algorithms.

Resource Rationality

Bayesian inference has been successfully applied to model human behavior across a number of domains:

- Perception
- Motor control
- Everyday judgements
- Logical reasoning

But, there are also well-documented deviations from Bayesian reasoning!

Resource Rationality

People only attempt to approximate Bayesian inference.

They are subject to limited computational resources.

Resource rationality: make optimal use of limited computational resources.

Resource Rationality

Limited computational resources can be readily incorporated in a meta-learned algorithm.

We can simply make the neural network less complex.

Argument 3

Meta-learning makes it easy to manipulate a learning algorithm's complexity and can therefore be used to construct resource-rational models of learning.

Resource Rationality

Many approximate inference schemes can also be cast as resource-rational algorithms.

Typically implement some form of computational complexity.

Computational complexity

Time or space required to *execute* an algorithm.

Meta-learning also allows to restrict an algorithm's algorithmic complexity.

Algorithmic complexity

Storage space needed to *implement* an algorithm.

Neuroscience

Bayesian inference is often argued to be biologically implausible.

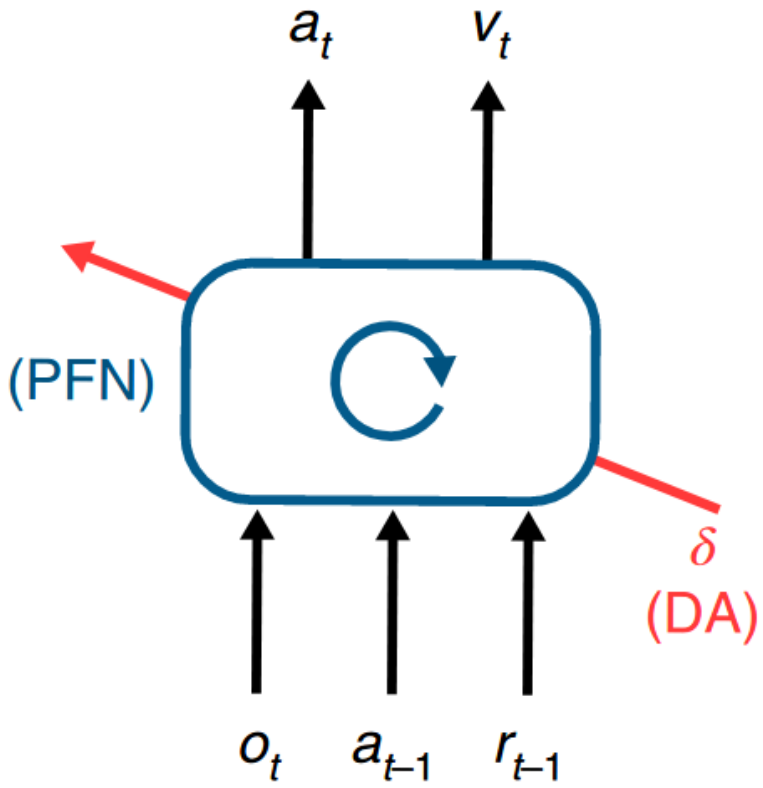
We may use biologically plausible architectures, e.g. spiking neural networks.

Meta-learning allows us to integrate neuroscientific insights into the rational analysis of cognition.

Links Marr's computational and implementational levels.

Neuroscience

Prefrontal circuits may constitute a meta-reinforcement learning system [Wang et al., 2018]:



Neuroscience

Incorporate neuroscientific insights into computational models.

Understand the brain through the meta-learning framework.

Argument 4

Meta-learning offers a powerful lens through which to view brain structure and function.

Why Isn’t Everything Meta-Learned?

Analytical solutions are useful to derive empirical predictions.

Disadvantage 1

Neural networks are black-box systems that don’t provide analytical solutions.

Harder to derive empirical predictions.

To do so, additional tools and analyses are required.

Why Isn't Everything Meta-Learned?

Training a neural network via meta-learning is complex and takes time.

Sometimes feel more like a dark art than a scientific endeavor.

Modifying assumptions in the data-generating distribution requires retraining the model from scratch.

Disadvantage 2

Reiterating different assumptions about the environment – a crucial step in the rational analysis of cognition – is time-consuming in the meta-learning setting.

Why Isn't Everything Meta-Learned?

Disadvantage 3

There is no guarantee that the fully converged model actually implements a Bayes-optimal learning algorithm.

We can compare to analytical solutions for simple cases.

But, in general it is impossible to verify that a meta-learned algorithm is optimal.

There are reported cases in which meta-learning failed to find the Bayes-optimal solution for more complicated problems.

Why Isn't Everything Meta-Learned?

Meta-learning and Bayesian inference each have their individual strengths and weaknesses.

In some situations traditional Bayesian models are more appropriate.

But, meta-learning broadens our available toolkit.

Meta-learned models do not replace Bayesian inference but complement them.

Summary

Meta-learning: framework for learning learning algorithms from data.

Produces learning algorithms that mimic Bayesian inference.

Four arguments in favor of the meta-learning framework:

1. Works even if exact Bayesian inference is intractable.
2. Works even if it is not possible to phrase the inference problem.
3. Makes it easy to manipulate a learning algorithm's complexity.
4. Makes it possible to integrate neuroscientific insights into the rational analysis of cognition.

Conclusion

If you like rational analysis or Bayesian models of cognition, you should like meta-learning!

Ortega (2020). Twitter.

> Meta-learning "brings back Bayesian statistics within deep learning without even trying—no latents, no special architecture, no special cost function, nada."

Thank you!

Dominik Endres

Samuel Gershman

Eric Schulz

Ishita Dasgupta

Akshay Jagadish

Jane Wang

Matthew Botvinick

References i


Anderson, J. R. (2013).
The adaptive character of thought.
Psychology Press.


Binz, M., Gershman, S. J., Schulz, E., and Endres, D. (2022).
Heuristics from bounded meta-learned inference.
Psychological review.


Dasgupta, I., Schulz, E., Tenenbaum, J. B., and Gershman, S. J. (2020).
A theory of learning to infer.
Psychological review, 127(3):412.


Harlow, H. F. (1949).
The formation of learning sets.
Psychological review, 56(1):51.

References ii


Lake, B. M. (2019).
Compositional generalization through meta sequence-to-sequence learning.
Advances in Neural Information Processing Systems, 32:9791–9801.


Wang, J. X., Kurth-Nelson, Z., Kumaran, D., Tirumala, D., Soyer, H., Leibo, J. Z., Hassabis, D., and Botvinick, M. (2018).
Prefrontal cortex as a meta-reinforcement learning system.
Nature neuroscience, 21(6):860–868.


Wolpert, D. H. (1996).
The lack of a priori distinctions between learning algorithms.
Neural computation, 8(7):1341–1390.