Probabilistic Model-Agnostic Meta Learning
Chelsea Finn*, Kelvin Xu*, Sergey Levine

Motivation
1. Learning a prior is one avenue to enable few-shot learning, which is a key aspect of human intelligence.
2. For some tasks however, task ambiguity in the few shot domain is unavoidable. For safety critical tasks (e.g., in medical domains), modeling ambiguity is key.

This Work
Extends model-agnostic meta learning [1] by modeling uncertainty in a principled manner using variational inference in a graphical model formulation of meta-learning (also see [2])

Background: the Meta-Learning Problem
- Assume access to a set of tasks drawn from the task distribution \(T \sim p(T)\) which can be split into a training set \(D_T^i := \{(x_{i,j}^s, y_{i,j}^s) | y_s\}\) and test set \(D_{T^i}^t := \{(x_{i,j}^t, y_{i,j}^t) | y_t\}\)
- The goal of the meta-learner is to learn the prior \(p(\theta)\) which can be split into a training \(\theta \approx \frac{\sum y \log p(\theta)}{\sum \log p(\theta)}\) and test \(\theta \approx \frac{\sum y \log p(\theta)}{\sum \log p(\theta)}\) distribution.

Model-Agnostic Meta-Learning (MAML) [1]

Graphical Model of Gradient Based Meta-Learning
- MAML can be interpreted as approximate inference in the graphical model to the left

\[
p(Y_{i}^{\text{test}} | x_{i}^{\text{test}}, y_{i}^{\text{tr}}, x_{i}^{\text{test}}) = f(y_{i}^{\text{test}} | x_{i}^{\text{test}}, y_{i}^{\text{tr}}, \theta) \approx p(y_{i}^{\text{test}} | x_{i}^{\text{test}}, \phi_{i}^{*})
\]
where \(\phi_{i}^{*}\) is the MAP approximation [2]

Variational Approximation
- The previous approximation \(p(\phi_{i} | x_{i}^{\text{test}}, y_{i}^{\text{tr}}, \theta) \approx \delta(\phi_{i} = \phi_{i}^{*})\) leads to the graphical model on the left from which we can derive a variational lower bound on the approximate log-likelihood

\[
\log p(y_{i}^{\text{test}} | x_{i}^{\text{test}}, x_{i}^{\text{tr}}, y_{i}^{\text{tr}}) \geq \mathcal{L}(\phi_{i} | y_{i}^{\text{test}}, y_{i}^{\text{tr}}) + \mathcal{L}(\theta | y_{i}^{\text{test}}, y_{i}^{\text{tr}})
\]
where \(\mathcal{L}(\phi_{i} | y_{i}^{\text{test}}, y_{i}^{\text{tr}}) = \mathcal{N}(\mu_{\phi_{i}}, \gamma_{\phi_{i}} \log p(y_{i}^{\text{test}} | x_{i}^{\text{test}}, \mu_{\phi_{i}}; v_{\phi_{i}})\)

Additional dependencies
- In the transformed graphical model above, \(x_{i}^{\text{tr}}, y_{i}^{\text{tr}}, \theta\) are conditionally independent
- Since we have only a very crude approximation from the MAP approximation, these independences may not hold
- Allow to model to compensate by learning a task specific prior \(p(\theta_{i} | x_{i}^{\text{tr}}, y_{i}^{\text{tr}}) = \mathcal{N}(\mu_{\theta_{i}}, \gamma_{\theta_{i}} \log p(y_{i}^{\text{test}} | x_{i}^{\text{test}}, \mu_{\theta_{i}}; v_{\theta_{i}})\)

Probabilistic Latent Model for Incorporating Priors and Uncertainty in Few-Shot Learning (PLATIPUS): Pseudocode

Algorithm 1 Meta-training, differences from MAML in red

Require: \(T\) tasks
1: Initialize \(\theta \sim (\mu_{\theta}, \sigma_{\theta}, \gamma_{\theta}, \gamma_{\theta})\)
2: while not done do
3: Sample batch of tasks \(T_{i} \sim p(T)\)
4: for all \(T_{i}\) do
5: \(D_{T_{i}}^{i}, D_{T_{i}}^{t} \sim p(T_{i})\)
6: Evaluate \(V_{\theta}(\mu_{\theta}, D_{T_{i}}^{i})\)
7: Sample \(\theta \sim q_{\phi_{i}} = \mathcal{N}(\mu_{\theta} - \gamma_{\theta} \nabla_{\mu_{\theta}} \mathcal{L}(\mu_{\theta}, D_{T_{i}}^{i}), v_{\theta})\)
8: Evaluate \(V_{\theta}(\theta, D_{T_{i}}^{t})\)
9: Compute adapted parameters with gradient descent:
\(\phi_{i} = \theta - \alpha \nabla_{\phi_{i}} \mathcal{L}(\theta, D_{T_{i}}^{t})\)
10: Let \(p(\theta_{i} | D_{T_{i}}^{i}) = \mathcal{N}(\mu_{\theta_{i}}, \gamma_{\theta_{i}} \log p(y_{i}^{\text{test}} | x_{i}^{\text{test}}, \mu_{\theta_{i}}; v_{\theta_{i}})\)
11: Compute \(V_{\theta}(\theta_{i} | D_{T_{i}}^{i})\)
12: Update \(\theta\) using Adam

Algorithm 2 Meta-testing
Require: training data \(D_{T}\) new task \(T\)
Require: learned \(\Theta\)
1: Sample \(\theta\) from the prior \(p(\theta | D_{T})\)
2: Evaluate \(V_{\theta}(\theta, D_{T})\)
3: Compute adapted parameters with gradient descent:
\(\phi_{i} = \theta - \alpha \nabla_{\phi_{i}} \mathcal{L}(\theta, D_{T})\)

Future Directions
- Studying how ambiguity and uncertainty can guide data-acquisition
- Data-dependent posterior variances for tasks with differing levels of uncertainty

Results
1-shot Classification
5-shot Regression

Ambiguous Attributes

Mini-Imagenet

- Slight boost over MAML, while comparable to other approaches, (*similar architecture)