1

Algorithm 1 MAML

Randomly initialize $\boldsymbol{\mu}_\theta, \boldsymbol{\sigma}_\theta^2$
while not done **do**
  for task $\mathcal{T}_i \sim p(\mathcal{T})$ **do**
    Draw support set $\mathcal{D}_i^S = \{(\boldsymbol{x}_j, \boldsymbol{y}_j)\}_{j=1\ldots K}$ from $\mathcal{T}_i$
    sample $\theta \sim \mathcal{N}(\boldsymbol{\mu}_\theta, \boldsymbol{\sigma}_\theta^2)$
    Adapt parameters $\theta_i = \theta - \alpha \nabla_\theta \mathcal{L}_i(\theta, \mathcal{D}_i^S)$
    Draw test samples $\mathcal{D}_i^Q = \{(\boldsymbol{x}_j, \boldsymbol{y}_j)\}$ from $\mathcal{T}_i$
  end for
  Meta-Update: $\boldsymbol{\mu}_\theta \leftarrow \boldsymbol{\mu}_\theta - \beta \nabla_{\boldsymbol{\mu}_\theta} \sum_i \mathcal{L}_i(\theta_i, \mathcal{D}_i^Q)$
  Meta-Update: $\boldsymbol{\sigma}_\theta^2 \leftarrow \boldsymbol{\sigma}_\theta^2 - \beta \nabla_{\boldsymbol{\sigma}_\theta^2} \sum_i \mathcal{L}_i(\theta_i, \mathcal{D}_i^Q)$
end while

MAML

2 graphical models

$$
\begin{aligned}
&p(\boldsymbol{\theta}, \boldsymbol{\phi}_{1:T}, \boldsymbol{x}_{1:T,1:N+M}, \boldsymbol{y}_{1:T,1:N+M}) \\
&= \left(\prod_{i=1}^{T} \left(\prod_{j=1}^{N+M} p(\boldsymbol{x}_{1:T,1:N+M}, \boldsymbol{y}_{1:T,1:N+M} | \phi_i) \right) p(\phi_i|\theta) \right) p(\theta)
\end{aligned}
$$

What does it mean to have observed data? In Meta Training we basically have given all samples. In meta testing we dont have the test sample targets. Classical MAML:

$$
p(y_i^{test}|x_i^{test}, x_i^{tr}, y_i^{tr}) = \int p(y_i^{test}|x_i^{test}, \phi_i) p(\phi_i|x_i^{tr}, y_i^{tr}) d\phi_i
$$

3 first lower bound equation

General Variational Bayes:

$$\log p(x) = \mathop{\mathbb{E}}_{z\sim q}[\log \frac{p(x|z)p(z)}{q(z)}] + KL\big(q(z)\|p(z|x)\big) \tag{1}$$

Mapping from general VB to the Probabilistic MAML case:

$$\begin{align}
p(x) &= p(y_i^{test}|x_i^{tr,test}, y_i^{tr}) \tag{2}\\
p(z) &= p(\phi_i, \theta|x_i^{tr,test}, y_i^{tr}) \tag{3}\\
p(x|z) &= p(y_i^{test}|\phi_i, x_i^{tr,test}, y_i^{tr}) \tag{4}\\
q(z) &= q_\psi(\phi_i|\theta, x_i^{tr,test}, y_i^{tr,test})q_\psi(\theta|x_i^{tr,test}, y_i^{tr,test}) \tag{5}
\end{align}$$

Lower Bound for Probabilistic MAML:

$$\begin{align}
&\log p(y_i^{test}|x_i^{tr,test}, y_i^{tr}) \tag{6}\\
&\geq \mathop{\mathbb{E}}_{\theta,\phi_i\sim q_\psi}[\log p(y_i^{test}|\phi_i, x_i^{tr,test}, y_i^{tr}) + \log p(\phi_i, \theta|x_i^{tr,test}, y_i^{tr})] \tag{7}\\
&- \mathop{\mathbb{E}}_{\theta,\phi_i\sim q_\psi}[\log q_\psi(\phi_i|\theta, x_i^{tr,test}, y_i^{tr,test})q_\psi(\theta|x_i^{tr,test}, y_i^{tr,test})] \tag{8}\\
&= \mathop{\mathbb{E}}_{\theta,\phi_i\sim q_\psi}[\log p(y_i^{test}|\phi_i, x_i^{test}) + \color{red}{\log p(\phi_i, \theta, y_i^{tr}|x_i^{tr}) - \log p(y_i^{tr}|x_i^{tr})}] \tag{9}\\
&+ \mathcal{H}(q_\psi(\phi_i|\theta, x_i^{tr,test}, y_i^{tr,test})) + \mathcal{H}(q_\psi(\theta|x_i^{tr,test}, y_i^{tr,test})) \tag{10}
\end{align}$$

Only the red part:

$$\begin{align}
\color{red}{\log p(y_i^{tr}|x_i^{tr})} &= \text{const} \tag{11}\\
\color{red}{\log p(\phi_i, \theta, y_i^{tr}|x_i^{tr})} &= \log p(y_i^{tr}|x_i^{tr}, \phi_i, \theta)p(\phi_i, \theta|x_i^{tr}) \tag{12}\\
&= \log p(y_i^{tr}|x_i^{tr}, \phi_i) + \log p(\phi_i|\theta) + \log p(\theta) \tag{13}
\end{align}$$

4

4.1 Gradient-based Meta Learning with variational Inference

In LLAMA we used deterministic $p(\theta)$ and did gradient descent on $log(y^{tr}|x^{tr}, \theta)$ to get the next ϕ_{k+1}. Now we use structured variational inference to approximate $p(\phi_i, \theta)$ by $q_i(\phi_i|\theta)q_i(\theta)$.We can avoid storing two distributions for each task by parameterizing one distribution family.

$$p(\phi_i, \theta) \approx q_\psi(\phi_i|\theta, x_i^{\text{tr,test}}, y_i^{\text{tr,test}})q_\psi(\theta|x_i^{\text{tr,test}}, y_i^{\text{tr,test}}) \tag{14}$$

We set up variational inference lower bound we want to maximize. We choose prior $p(\theta) = \mathcal{N}(\mu_\theta, \sigma_\theta^2)$ and $p(\phi_i|\theta) = \mathcal{N}(\theta, \sigma_?^2)$. For the inference networks we choose

$$q_\psi(\phi_i|\theta, x_i^{\text{tr,test}}, y_i^{\text{tr,test}}) = \mathcal{N}\Big(\mu_\theta + \gamma_q \nabla_{\mu_\theta} \log p(y_i^{\text{tr,test}}|x_i^{tr,test}, \mu_\theta), v_q\Big) \tag{15}$$

for the other inference network this can be done as well, but only for meta-training. in meta-testing we don't have y_i^{test} so we need to do something different.

Overview According to the left model we have:

$$\phi_i \sim p(\phi_i | x_i^{\text{tr}}, y_i^{\text{tr}}) \propto \int p(y_i^{\text{tr}} | x_i^{\text{tr}}, \phi_i) p(\phi_i|\theta) p(\theta) d\theta \tag{16}$$

which is totally intractable, so we use point approximation of ϕ in the next section.

5 Probabilistic MAML with Hybrid Inference

we use maml to compute $p(\phi_i | x_i^{tr}, y_i^{tr}, \theta) \approx \delta(\phi - \phi^*)$ where ϕ^* is obtained by gradient descent over $\log p(y_i^{tr} | x_i^{tr}, \theta)$. Again we define Lower Bound, but now only over θ as ϕ^* is now deterministic.

$$\log\,(y_i^{\text{test}} | x_i^{\text{tr+test}}, y_i^{\text{tr}}) \geq \text{Lower Bound}(\psi) \tag{17}$$

$$= E_{\theta \sim q_\psi}[\log\, p(y_i^{\text{test}} | x_i^{\text{test}}, \phi^*) + \log\, p(\theta)] + H(q_\psi(\theta | x_i^{\text{test}}, y_i^{\text{test}})) \tag{18}$$

$$= E_{\theta \sim q_\psi}[\log\, p(y_i^{\text{test}} | x_i^{\text{test}}, \phi^*)] - KL(q_\psi(\theta | x_i^{\text{test}}, y_i^{\text{test}}) \| p(\theta)) \tag{19}$$

In the end $q_\psi(\theta | x_i^{\text{test}}, y_i^{\text{test}})$ approximates $p(\theta | x_i^{\text{tr,test}}, y_i^{\text{tr,test}})$.