1 Step from Eq.2 to Eq.3

$$p(X^c|\theta) = \prod_j \int p(x_{j_1}, ..., x_{j_N}|\phi_j)\, p(\phi_j|\theta)\, \mathrm{d}\phi_j \tag{1}$$

$$p(X^t|\theta) = \prod_j \int p(x_{j_{N+1}}, ..., x_{j_{N+M}}|\phi_j)\, p(\phi_j|\theta)\, \mathrm{d}\phi_j \tag{2}$$

$$\begin{align}
-\log\, p(X^t|\theta) &= \sum_j -\log\, \int p(x_{j_{N+1}}, ..., x_{j_{N+M}}|\phi_j)\, p(\phi_j|\theta)\, \mathrm{d}\phi_j \tag{3}\\
&\approx \sum_j -\log\, \int p(x_{j_{N+1}}, ..., x_{j_{N+M}}|\phi_j)\, \delta(\phi_j - \hat{\phi}_j)\, \mathrm{d}\phi_j \tag{4}\\
&= \sum_j -\log\, p(x_{j_{N+1}}, ..., x_{j_{N+M}}|\hat{\phi}_j) \tag{5}
\end{align}$$

2 Introduction

Cross Entropy Loss for Standard Classification

- network outputs a probability vector $f_\theta(\underline{x}) \in [0,1]^d$ where we have d classes. This can be interpreted as $p_\theta(c|\underline{x}) = f_\theta(\underline{x})[c]$ for $c \in \{1, ..., d\}$
- target $\underline{c}_i \in \{0,1\}^d$ is a one-hot encoded vector
- cross-entropy loss for dataset $\mathcal{D} = \{(\underline{x}_i, \underline{c}_i)\}_{i=1...N}$

$$\mathcal{L}(\theta, \mathcal{D}) = -\sum_{i=1}^{N} \underline{c}_i^T \log(f_\theta(\underline{x}_i)) = -\sum_{i=1}^{N} \log\, p_\theta(c_i|\underline{x}_i)$$

3 Meta-Learning Formulation

3.1 MAML

3.2 Hierarchical Bayes Inference

this is theory. p is not our network but some unknown density we want to optimize $\theta^* = \mathrm{argmax}_\theta\, p(\mathcal{D}|\theta)$. where $\mathcal{D}$ includes the observed data of all tasks.

$$p(\mathcal{D}|\theta) = \prod_i \int p(\mathcal{D}_i|\phi_i)\, p(\phi_i|\theta)\, \mathrm{d}\phi_i \tag{6}$$

Algorithm 1 MAML

Randomly initialize θ
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 $\mathcal{T}_i$
    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: $\theta \leftarrow \theta - \beta \nabla_\theta \sum_i \mathcal{L}_i(\theta_i, \mathcal{D}_i^Q)$
end while

Meta-Update:

$$\frac{\partial \mathcal{L}}{\partial \theta} = \frac{\partial \mathcal{L}}{\partial \theta_i} \frac{\partial \theta_i}{\partial \theta} = \nabla_{\theta_i} \mathcal{L}(\theta_i, \mathcal{D}_i^Q)(I - \alpha \nabla_\theta \nabla_\theta \mathcal{L}(\theta, \mathcal{D}_i^S))$$

4 Linking MAML and Hierarchical Bayes

The marginalization over ϕ_i is not tractable so we simply use a point estimate $\hat{\phi}_i$ for each task.

$$\log p(\mathcal{D}|\theta) \approx \sum_i \log p(\mathcal{D}_i|\hat{\phi}_i) \tag{7}$$

For standard MAML we use $\hat{\phi}_i = \theta - \alpha \nabla_\theta \mathcal{L}(\theta, \mathcal{D}_i)$. (Skip rest of 3.1 and 3.2)

5 actual Algorithm to improve MAML

5.1 Laplace Method of Integration

For the advanced version we use Laplace approximation.

$$\int p(\mathcal{D}_i|\phi_i)\, p(\phi_i|\theta)\, \mathrm{d}\phi_i \approx p(\mathcal{D}_i|\phi_i^*)\, p(\phi_i^*|\theta)\, \frac{1}{\sqrt{\det(\frac{1}{2\pi} H_i)}} \tag{8}$$

where ϕ_i^* is is a mode (local maximum) of the integrand and $H_i = \nabla^2_{\phi_i^*} \mathcal{L}(\phi_i^*, \mathcal{D}_i)$ is the Hessian matrix of the loss at ϕ_i^*.

Lightweight Laplace Approximation for Meta-Adaption

Algorithm 2 LLAMA

Randomly initialize θ
while not done **do**
  for task $\mathcal{T}_i \sim p(\mathcal{T})$ **do**
    Draw support set $\mathcal{D}_i^S = \{(\boldsymbol{x}_j, \boldsymbol{y}_j)\}_{j=1...K}$ from $\mathcal{T}_i$
    Adapt parameters $\theta_i = \theta - \alpha \nabla_\theta \mathcal{L}_i(\theta, \mathcal{D}_i^S)$
    estimate quadratic curvature $H_i = \nabla^2_{\theta_i} \mathcal{L}_i(\theta_i, \mathcal{D}_i^S)$
    Draw test samples $\mathcal{D}_i^Q = \{(\boldsymbol{x}_j, \boldsymbol{y}_j)\}$ from $\mathcal{T}_i$
  end for
  Meta-Update: $\theta \leftarrow \theta - \beta \nabla_\theta \sum_i \mathcal{L}_i(\theta_i, \mathcal{D}_i^Q)$ $+\eta \log \det(H_i)$
end while

Production:
for new Tasks $\mathcal{T}_i \sim p(\mathcal{T})$ **do**
  calculate $\mathcal{D}_i^S$, θ_i, H_i as in meta-training
  draw $\theta_{sample} \sim \mathcal{N}(\theta_i, H_i^{-1})$
  use $f_{\theta_{sample}}$ for new production data
end for

6 Powerpoint

Conceptually we can view the problem of finding a good theta as finding the MLE estimate

$$\theta^* = \underset{\theta}{\operatorname{argmax}}\ p(\mathcal{D}|\theta) \tag{9}$$

$$= \underset{\theta}{\operatorname{argmax}} \prod_i \int p(\mathcal{D}_i|\phi_i)\, p(\phi_i|\theta)\, \mathrm{d}\phi_i \tag{10}$$

where $\mathcal{D}$ includes the observed data of all tasks. We cannot compute this integrals. To make them tractable we employ a point estimate $\hat{\phi}_i$ which we obtain from MAML. $p(\phi_i|\theta)$ acts like a dirac impulse.

$$\theta^* = \underset{\theta}{\operatorname{argmin}} -\log\, p(\mathcal{D}|\theta) \approx \underset{\theta}{\operatorname{argmin}} -\sum_i \log\, p(\mathcal{D}_i|\hat{\phi}_i) \tag{11}$$

So we are minimizing the cross entropy loss.

for classification this is

$$\mathcal{L}(\phi, \mathcal{D}) = -\sum_j \log\, f_\phi(c_j|x_j) = -\sum_j \log\, p(c_j|x_j, \phi) = -\log\, p(\mathcal{D}|\phi) \tag{12}$$

for a model f_ϕ dataset $\mathcal{D} = \{(x_j, c_j)\}_{j=1...N}$ where $c_j \in \{1, ..., C\}$ for C different classes.

For regression this is

$$\mathcal{L}(\phi, \mathcal{D}) = \sum_j (f_\phi(x_j) - y_j)^2 = ??? = -\log\, p(\mathcal{D}|\phi)$$

where $y_j \in \mathbb{R}^d$
Instead of only using a point estimate we also can use Laplace Approximation:

Point estimate:	$p(\mathcal{D}\vert\phi) \approx p(\mathcal{D}_i\vert\hat{\phi}_i)$
Laplace Approximation:	$p(\mathcal{D}\vert\phi) \approx p(\mathcal{D}_i\vert\hat{\phi}_i)\, p(\hat{\phi}_i\vert\theta) \frac{1}{\sqrt{\det(\frac{H_i}{2\pi})}}$

with Hessian $H_i = -\nabla_\phi^2 \log\, p(\mathcal{D}_i|\phi)p(\phi|\theta)|_{\phi=\hat{\phi}_i}$

and the point estimate θ^* from MAML.
We can sample $\theta \sim \mathcal{N}(\theta^*, H^{-1})$ where

$$H = \nabla_\theta^2\, \mathcal{L}(\theta)|_{\theta=\theta^*} \tag{13}$$