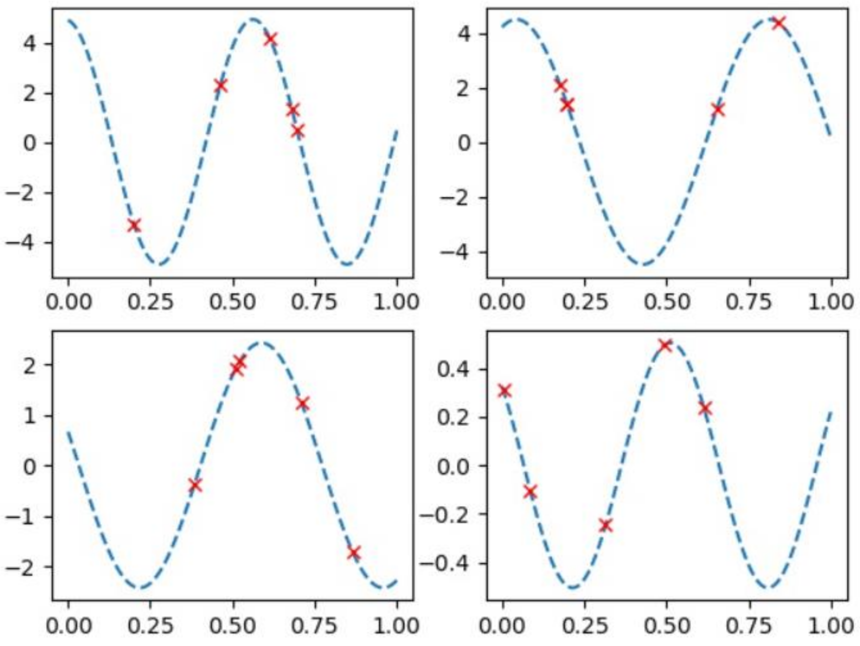
Recasting Gradient-Based Meta-Learning as Hierarchical Bayes

Content

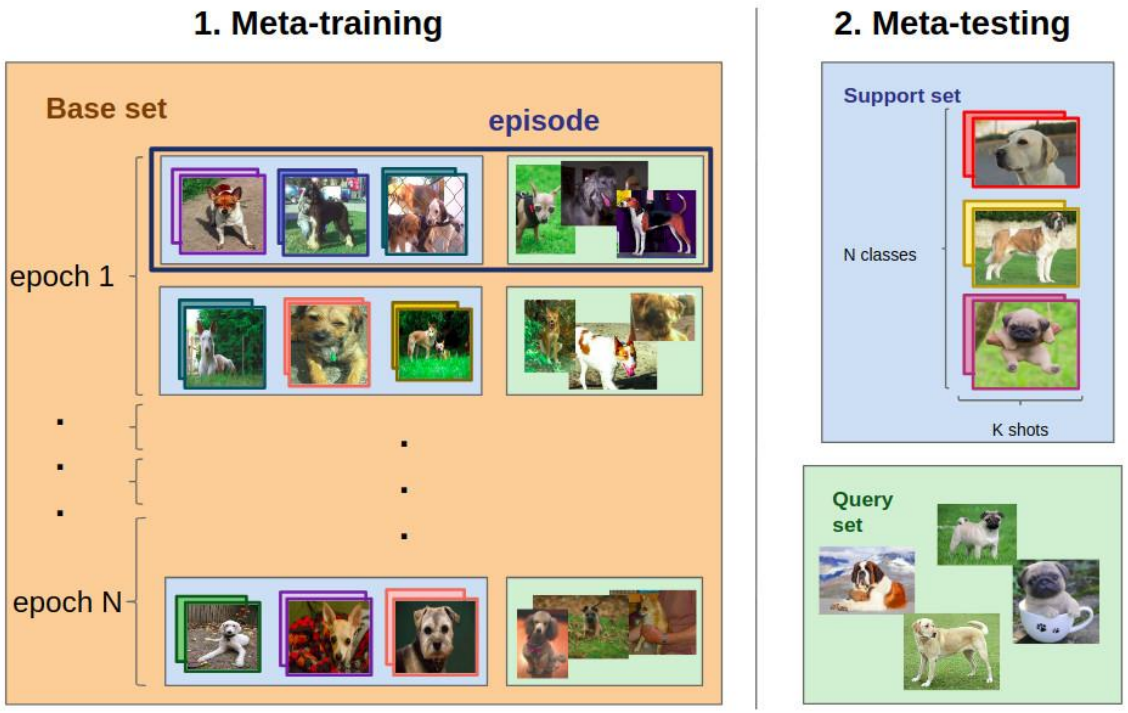
- Gradient-Based Meta-Learning
  - Finn C, et al. Model-agnostic meta-learning for fast adaptation of deep networks, ICML, 2017.
- Bayesian Meta-Learning
  - Ravi S, et al. Amortized bayesian meta-learning, ICLR. 2019.
- Connection of two approach
  - Grant E, et al. Recasting gradient-based meta-learning as hierarchical bayes, ICML, 2018.

Few-Shot Learning

Basic units of Meta-Learning are tasks, instead of samples



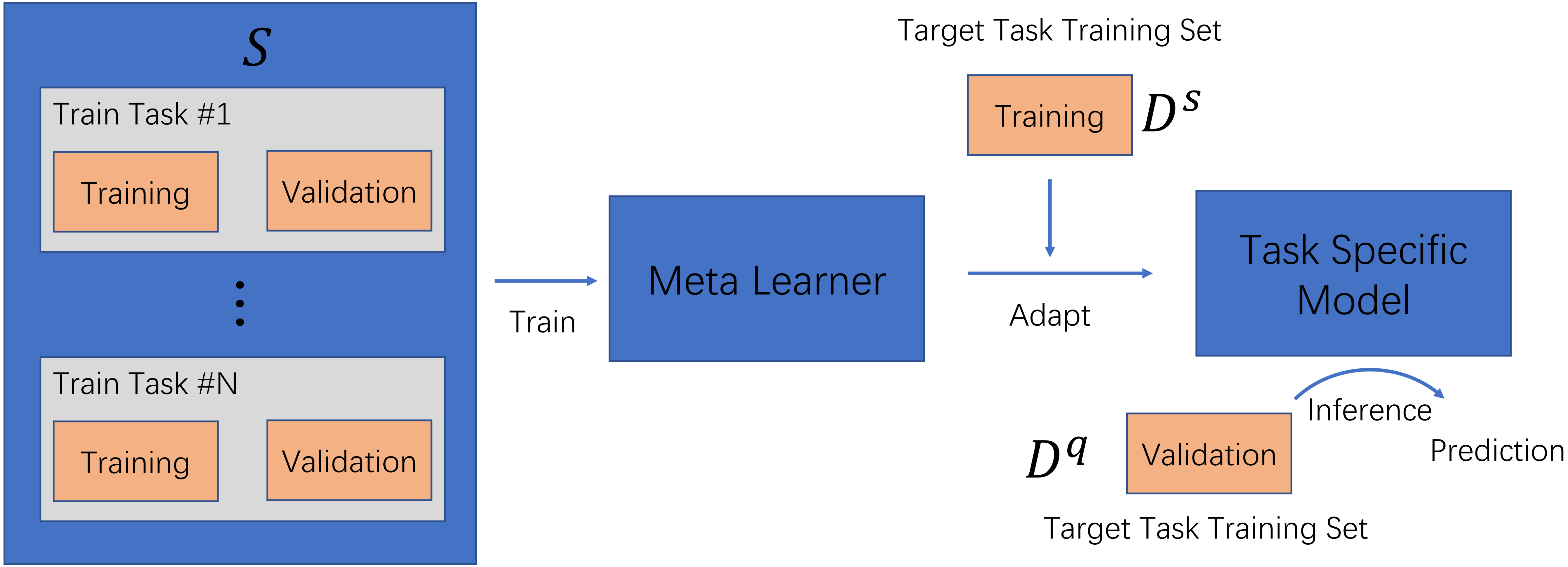
Few-Shot Regression



Few-Shot Classification

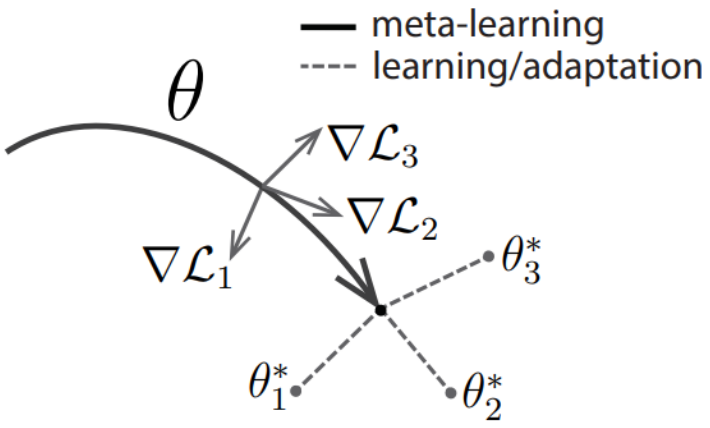
Meta Learning

Meta Learning focus on how to adapt to new task quickly
By learning from a large collection of similar tasks



Gradient-Based Meta Learning

MAML learns an initialization that quickly adapt to different tasks



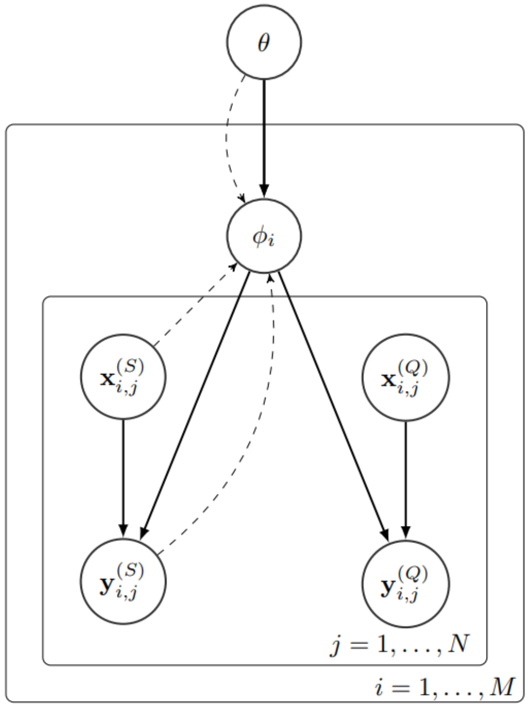
$$\min_{\theta_0} \sum_{(D^s,D^q)} \sum_{x,y\sim D^q} L[f_{\theta^*}(x), y]$$

$$\theta^* = \theta_0 - \alpha\nabla_\theta \sum_{x',y'\sim D^s} L[f_\theta(x'), y']$$

Finn C, et al. Model-agnostic meta-learning for fast adaptation of deep networks, ICML, 2017

Bayesian Meta Learning

Bayes ML learns a generative model that generates tasks.



$$\max_{\psi} \sum_{(D^s, D^q)} \log p(D^s, D^q)$$

Variational Lower Bound

$$\log \left[\prod_{i=1}^{M} p(\mathcal{D}_i) \right] = \log \left[\int p(\theta) \left[\prod_{i=1}^{M} \int p(\mathcal{D}_i|\phi_i) p(\phi_i|\theta) \, d\phi_i \right] d\theta \right]$$

$$\geq \mathbb{E}_{q(\theta;\psi)} \left[\log \left(\prod_{i=1}^{M} \int p(\mathcal{D}_i|\phi_i) p(\phi_i|\theta) \, d\phi_i \right) \right] - \mathbf{KL}(q(\theta;\psi) \| p(\theta))$$

$$= \mathbb{E}_{q(\theta;\psi)} \left[\sum_{i=1}^{M} \log \left(\int p(\mathcal{D}_i|\phi_i) p(\phi_i|\theta) \, d\phi_i \right) \right] - \mathbf{KL}(q(\theta;\psi) \| p(\theta))$$

$$\geq \mathbb{E}_{q(\theta;\psi)} \left[\sum_{i=1}^{M} \mathbb{E}_{q(\phi_i;\lambda_i)} \left[\log p(\mathcal{D}_i|\phi_i) \right] - \mathbf{KL}(q(\phi_i;\lambda_i) \| p(\phi_i|\theta)) \right] - \mathbf{KL}(q(\theta;\psi) \| p(\theta))$$

Ravi S, et al. Amortized bayesian meta-learning, ICLR. 2019

Recast MAML as Bayes-ML

MAML is a variant of Bayes ML, using truncated-GD as MAP inference.

Corollary (Santos, 1996) Consider a quadratic approximation to the objective

$$\ell(\boldsymbol{\phi}) \approx \tilde{\ell}(\boldsymbol{\phi}) := \tfrac{1}{2}\|\boldsymbol{\phi} - \boldsymbol{\phi}^*\|^2_{\mathbf{H}^{-1}} + \ell(\boldsymbol{\phi}^*)$$

Then the k-step truncated gradient decent

$$\boldsymbol{\phi}_{(k)} = \boldsymbol{\phi}_{(k-1)} - \mathcal{B}\,\nabla_{\boldsymbol{\phi}}\,\tilde{\ell}(\boldsymbol{\phi}_{(k-1)})$$

is the solution to the MAP problem

$$\min\left(\|\boldsymbol{\phi} - \boldsymbol{\phi}^*\|^2_{\mathbf{H}^{-1}} + \|\boldsymbol{\phi}_{(0)} - \boldsymbol{\phi}\|^2_{\mathbf{Q}}\right)$$

with $\boldsymbol{Q} = O\Lambda^{-1}\left((I - B\Lambda)^{-k} - I\right)O^T$, where B is a diagonal matrix that results from a simultaneous diagonalization of $\mathbf{H}$ and $\mathcal{B}$

Bayes ML

$$\log\left[\prod_{i=1}^{M} p(\mathcal{D}_i)\right] = \log\left[\int p(\theta)\left[\prod_{i=1}^{M}\int p(\mathcal{D}_i|\phi_i)p(\phi_i|\theta)\,d\phi_i\right]d\theta\right]$$

Point Estimate

MAML

$$-\log p\left(\mathbf{X} \mid \boldsymbol{\theta}\right) \approx \sum_j\left[-\log p\left(\mathbf{x}_{j_{N+1}}, \ldots \mathbf{x}_{j_{N+M}} \mid \hat{\phi}_j\right)\right]$$

$$\hat{\phi}_j = \boldsymbol{\theta} + \alpha\,\nabla_{\boldsymbol{\theta}} \log p(\mathbf{x}_{j_n} \mid \boldsymbol{\theta})$$

MAP Inference

Grant E, et al. Recasting gradient-based meta-learning as hierarchical bayes, ICML, 2018

Conclusion

- Direct algorithm design provides more flexibility towards specific problem, while Bayes approach offers interpretability and uncertainty estimation.
- Their connection helps algorithm design and probabilistic tool selection.