Paper: On the Global Optimality of Model-Agnostic Meta-Learning [4]

Jiaxin Liu

Group Reading

May 24, 2021

Overview

Meta-Learning

- **Meta-Learning**: ‘learning-to-learn’.
  - **Mechanistic view**: model that can read in an entire dataset and make predictions for new datapoints.
  - **Probabilistic view**: extract prior information from a set of (meta-training) tasks that allows efficient learning of new tasks.

Meta-Learning

- **Meta-Learning**: 'learning-to-learn'.
  - **Mechanistic view**: model that can read in an entire dataset and make predictions for new datapoints.
  - **Probabilistic view**: extract prior information from a set of (meta-training) tasks that allows efficient learning of new tasks.
- Incorporate additional data?
  - $D = \{(x_1, y_1), \ldots, (x_k, y_k)\}$
  - $D_{meta-train} = \{D_1, \ldots, D_n\}, D_{meta-test} = \{D_1, \ldots, D_m\}$

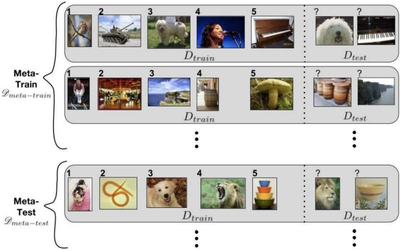

Figure: Example for meta learning. [3]

Meta-Learning

- Meta-learning problem: given data from $T_1, \ldots, T_n$, quickly solve new task T_{test}.
- Key assumption: meta-training tasks and meta-test task drawn i.i.d from same task distribution
- Multi-task learning, transfer learning and the meta-learning problem.

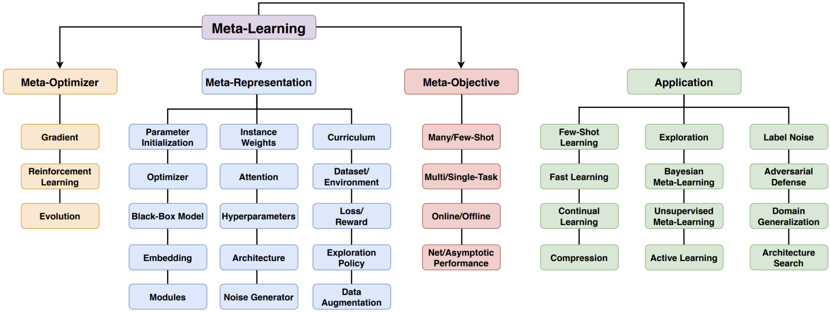

Figure: Overview of the meta-learning landscape. [2]

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

1. Fine-tuning: $\phi \longleftarrow \theta - \alpha \nabla_\theta L(\theta, D^{tr})$
   - θ: pre-trained parameters;
   - D^{tr}: training data for new task.

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

1. Fine-tuning: $\phi \longleftarrow \theta - \alpha \nabla_\theta L(\theta, D^{tr})$
   - θ: pre-trained parameters;
   - D^{tr}: training data for new task.
2. MAML: fine-tune with small amount of data during the test time.
   - $\min_\theta \sum_i L(\theta - \alpha \nabla_\theta L(\theta, D_i^{tr}), D_i^{ts})$
   - θ: parameter vector being meta-learned
   - θ_i^*: optimal parameter vector for task i.

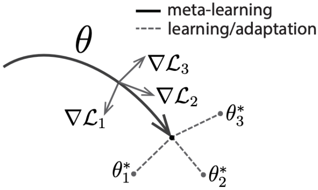

Figure: Diagram of MAML [1].

MAML

Key idea: acquire θ_i^* through optimization.

Algorithm 1 Model-Agnostic Meta-Learning

Require: $p(\mathcal{T})$: distribution over tasks
Require: α, β: step size hyperparameters
1: randomly initialize θ
2: **while** not done **do**
3:     Sample batch of tasks $\mathcal{T}_i \sim p(\mathcal{T})$
4:     **for all** $\mathcal{T}_i$ **do**
5:         Evaluate $\nabla_\theta \mathcal{L}_{\mathcal{T}_i}(f_\theta)$ with respect to K examples
6:         Compute adapted parameters with gradient descent: $\theta_i' = \theta - \alpha \nabla_\theta \mathcal{L}_{\mathcal{T}_i}(f_\theta)$
7:     **end for**
8:     Update $\theta \leftarrow \theta - \beta \nabla_\theta \sum_{\mathcal{T}_i \sim p(\mathcal{T})} \mathcal{L}_{\mathcal{T}_i}(f_{\theta_i'})$
9: **end while**

Formal Definition

- **For a subtask T_i:**
  - Learning tasks $\{T_i\}_{i\in[n]} \overset{i.i.d}{\sim} \imath$
  - Hypothesis class H, a distribution D over Z.
  - Loss function $l : H \times Z \mapsto \mathbb{R}$.
  - Risk for a subtask: $R(h) = \mathbb{E}_{z\sim D}[l(h,z)]$

Formal Definition

- **For a subtask T_i:**
  - Learning tasks $\{T_i\}_{i \in [n]} \overset{i.i.d}{\sim} \imath$
  - Hypothesis class H, a distribution D over Z.
  - Loss function $l : H \times Z \mapsto \mathbb{R}$.
  - Risk for a subtask: $R(h) = \mathbb{E}_{z \sim D}[l(h,z)]$
- **For meta-learner:**
  - $\bar{L}(\theta) = \mathbb{E}_{T \sim \imath}[R_T(h)]$
  - $L(\theta) = \frac{1}{n}\sum_{i=1}^{n} R_{T_i}(h)$
  - MAML: $L(\theta) = \frac{1}{n}\sum_{i=1}^{n} R_i(h_{\theta - \eta \nabla_\theta R_i(h_\theta)})$
- Notation:
  - $L_p(v) - \text{norm} : \|f(\cdot)\|_{p,v} := \{\int_X f^p(x) dv(x)\}^{1/p}$
  - $L_2(\rho) - \text{inner product} :< f, g >_H := \int_X f(x) \cdot g(x) d\rho$

Meta-SL

The goal of the supervised learning subtask (D_i, l, H).

$$h_i^* = \arg\min_{h \in H} R_i(h) = \arg\min_{h \in H} \mathbb{E}_{z \sim D_i}[l(h, z)]$$

where parameterize H by H_θ with a feature mapping $\phi : X \mapsto \mathbb{R}^d$

$$H_\theta = \{h_\theta(\cdot) = \phi(\cdot)^\top \theta : \theta \in \mathbb{R}^d\}$$

Meta-SL

The goal of the supervised learning subtask (D_i, l, H).

$$h_i^* = \arg\min_{h \in H} R_i(h) = \arg\min_{h \in H} \mathbb{E}_{z \sim D_i}[l(h, z)]$$

where parameterize H by H_θ with a feature mapping $\phi : X \mapsto \mathbb{R}^d$

$$H_\theta = \{h_\theta(\cdot) = \phi(\cdot)^\top \theta : \theta \in \mathbb{R}^d\}$$

Meta-objective

$$L(\theta) = \frac{1}{n} \sum_{i=1}^{n} R_i(h_{\theta_i}), \text{ where } h_{\theta_i} = h_{\theta - \eta \nabla_\theta R_i(h_\theta)}.$$

Minimizing $L(\theta)$ uses gradient descent

$$\theta_{l+1} \leftarrow \theta_l - \alpha_l \cdot \nabla_\theta L(\theta_l), \text{ for } l = 0, \ldots, T-1.$$

Frechet Differentiability

Definition 1: Frechet Differentiability

Let H be a Banach space with the norm $\|\cdot\|_H$. A functional $R : H \mapsto \mathbb{R}$ is Frechet differentiable at $h \in H$ if it holds for a bounded linear operator $A : H \mapsto \mathbb{R}$ that

$$\lim_{h_1 \in H, \|h_1\|_H \to 0} \frac{|R(h+h_1) - R(h) - A(h_1)|}{\|h_1\|_H} \to 0.$$

We define A as the F-derivative of R at $h \in H$ and

$$D_h R(\cdot) = A(\cdot) = < \cdot, a_h >_H, \text{ where } a_h(x) = \frac{\delta R}{\delta h}(x), \forall x \in X, h \in H$$

Example 1

$f : \mathbb{R} \to \mathbb{R}. f(x) = x^2$

Convex and Differentiable Risk

Assumption 1: (Convex and Differentiable Risk)

We assume for all $i \in [n]$ that the risk R_i is convex and Frechet differentiable on H.

Proposition 1: (Convex and Differentiable Risk)

Under Assumption 1, it holds for all $i \in [n]$ that

$$R_i(h_1) \geq R_i(h_2) + < \frac{\delta R_i}{\delta h_2}, h_1 - h_2 >_H, \forall h_1, h_2 \in H.$$

- Linear approximation for a convex function.

ε-stationary point

Definition 2 (descent direction)

We say that the direction s is a descent direction for the continuously differentiable function f at the point x if

$$g(x)^\top s < 0$$

$$f'(x;s) \stackrel{def}{=} \lim_{t\to 0} \frac{f(x+ts)-f(x)}{t} = g(x)^\top s$$

ε-stationary point

Definition 2 (descent direction)

We say that the direction s is a descent direction for the continuously differentiable function f at the point x if

$$g(x)^\top s < 0$$

$$f'(x;s) \overset{def}{=} \lim_{t \to 0} \frac{f(x+ts) - f(x)}{t} = g(x)^\top s$$

Definition 3 (ε-stationary point ω)

w be the ε-stationary point attained by meta-SL such that

$$\nabla_\omega L(\omega)^\top v \le \varepsilon, \quad \forall v \in \mathscr{B} = \{\theta \in \mathbb{R}^d : \|\theta\|_2 \le 1\}.$$

Goal

Theorem 1 (Optimality Gap of ε-Stationary Point).

Let θ^* be a global minimizer of $L(\theta)$. Also, let w be the ε-stationary point defined in Definition 3. Let $l(h_\theta(x),(x,y))$ be twice differentiable with respect to all $\theta \in \mathbb{R}^d$ and $(x,y) \in (X \times Y)$. Under Assumption 1, it holds for all $R > 0$ that

$$L(\omega) - L(\theta^*) \leq R \cdot \varepsilon + \|w\|_{M\cdot\rho} \cdot \inf_{v \in \mathscr{B}_R} \|u(\cdot) - \phi_{l,\omega}(\cdot)^\top v\|_{M\cdot\rho}$$

where we define $\mathscr{B}_R = \{\theta \in \mathbb{R}^d : \|\theta\|_2 \leq R\}$ and $\|w\|_{M\cdot\rho}$ is the $L_2(M\cdot\rho)$-norm of w.

Cont.

$$w(x,y,x') = \frac{1}{n} \cdot \sum_{i=1}^{n} (\delta R_i / \delta h_{\omega_i})(x') \cdot (dD_i/dM)(x,y)$$

$$u(x,y,x') = (\frac{1}{n} \cdot \sum_{i=1}^{n} (\delta R_i / \delta h_{\omega_i})(x') \cdot (h_{\omega_i}(x') - h_{\theta_i^*}(x')))/w(x,y,x')$$

$$\phi_{l,\omega}(x,y,x') = (I_d - \eta_\omega^2 l(\phi(x)^\top \omega, (x,y)))\phi(x')$$

where we define the mix distribution M over all the distributions $\{D_i\}_{i \in [n]}$

$$M(x,y) = \frac{1}{n} \sum_{i=1}^{n} D_i(x,y), \qquad \forall (x,y) \in X \times Y$$

Proof.

Theorem 1 (see notes) □

* Meta-SL with Squared Loss

Squared Loss

$$l(h,(x,y)) = (h(x)-y)^2, \qquad \forall h \in H, (x,y) \in X \times Y.$$

Proposition 2

We denote by $\bar{D}_i$ the marginal distribution of D_i over X. Let $D_i = \rho$ for all $i \in [n]$. For the squared loss l and $R_i = \mathbb{E}_{(x,y)\sim D_i}[l(h,(x,y))]$, it holds that

$$(\delta R_i / \delta h) = 2\mathbb{E}_{(x,y)\sim D_i}[h(x) - y | x = x'], \qquad \forall h \in H, x' \in X.$$

Corollary 1

For the squared loss l and $R > 0$, we have

$$L(\omega) - L(\theta^*) \leq R \cdot \varepsilon + 2\bar{R} \cdot \inf_{v \in \mathscr{B}} \|u - (K_\eta \cdot \phi)^\top (R \cdot v)\|_\rho.$$

Cont.

$$K_\eta = \mathbb{E}_{x\sim\rho}[I_d - 2\eta \cdot \phi(x)\phi(x)^\top],$$

$$u(x') = (\sum_{i=1}^{n}(\delta r_i/\delta_{\omega_i})(x') \cdot (h_{\omega_i}(x') - h_{\theta_i^*}(x')))/(\sum_{i=1}^{n} \delta R_i/\delta h_{\omega_i}(x')),$$

$$\bar{R} = \frac{1}{n} \cdot \sum_{i=1}^{n} R_i^{1/2}(h_{\omega_i}) = \frac{1}{n}\sum_{i=1}^{n}\{\mathbb{E}_{(x,y)\sim D_i}[(y - h_{\omega_i}(x))^2]\}^{1/2}$$

Proof.

Corollary 1 (see notes) □

Chelsea Finn, Pieter Abbeel, and Sergey Levine. "Model-agnostic meta-learning for fast adaptation of deep networks". In: *International Conference on Machine Learning*. PMLR. 2017, pp. 1126–1135.

Timothy Hospedales et al. "Meta-learning in neural networks: A survey". In: *arXiv preprint arXiv:2004.05439* (2020).

Sachin Ravi and Hugo Larochelle. "Optimization as a model for few-shot learning". In: (2016).

Lingxiao Wang et al. "On the global optimality of model-agnostic meta-learning". In: *International Conference on Machine Learning*. PMLR. 2020, pp. 9837–9846.