Task Relatedness-Based Generalization Bounds for Meta Learning

Jiechao Guan[1], Zhiwu Lu[2,*]

1. School of Information, Renmin University of China, Beijing, China

2. Gaoling School of Artificial Intelligence, Renmin University of China, Beijing, China

* Corresponding Author Email: {2014200990, luzhiwu}@ruc.edu.cn

Contents

- Background of Meta Learning Theory
- Task Relatedness-Based Bounds for Meta Learning
- Spectrally-Normalized Bounds for Meta Learning
- Conclusions and Future Works

Contents

- Background of Meta Learning Theory
- Task Relatedness-Based Bounds for Meta Learning
- Spectrally-Normalized Bounds for Meta Learning
- Conclusions and Future Works

Definition of Meta Learning Algorithm

- Generally speaking, meta learning is a mechanism that uses the information extracted from several observed tasks to improve the prediction performance of the learner on the unseen task.

- Any method that adjusts the parameters of an algorithm based on the experience made with other learning tasks can be regarded as a meta learning algorithm. --Maurer (JMLR 2005)

Assumptions of Meta Learning Theory

- The goal of single task learning is to select from the hypothesis space $\mathcal{H}$ a hypothesis h to minimize the expected loss of task P on sample space $Z = X \times Y$:

$$er_P(h) = \int_{X \times Y} l(h(x), y) \mathrm{d}P(x, y)$$

- The goal of meta learning is to select from the hypothesis space family H a hypothesis space $\mathcal{H}$ that contains a good solution to the task P sampled i.i.d. from the task environment Q. That is, meta learning focuses on minimizing the transfer risk on a novel task of a good hypothesis space $\mathcal{H}$:

$$er_Q(\mathcal{H}) = \int_{\mathcal{P}} \inf_{h \in \mathcal{H}} er_P(h) \mathrm{d}Q(P)$$

- The training (n, m)-sample $\boldsymbol{z} \in Z^{n \times m}$ is constructed by sampling n tasks from environment Q independently and then sampling m data from each task independently, the empirical risk on the (n, m)-sample of a space $\mathcal{H}$ is:

$$\hat{er}_{\mathbf{z}}(\mathcal{H}) = \frac{1}{n} \sum_{j=1}^{n} \inf_{h \in \mathcal{H}} \hat{er}_{\mathbf{z}_{:j}}(h)$$

- Meta learning theory aims to bound the gap: $\left| \hat{er}_{\mathbf{z}}(\mathcal{H}) - er_Q(\mathcal{H}) \right|$

Existing Model Complexity Bounds

- Covering number bound on transfer risk (Baxter JAIR 2000):

Theorem 3 *Let $(\mathcal{P}, Q)$ be an environment on the separable metric space $Z = X \times Y$. Let $\mathbf{z}$ be an (m, n)-sample generated by the process described in Section 3.1. Let $\mathbb{H} = \{\mathcal{H}\}$ be any permissible hypothesis space family. Then for any $\mathcal{H} \in \mathbb{H}$, $\epsilon \in (0, 1)$, with probability at least $1 - \delta$ over $\mathbf{z}$,*

$$|\hat{er}_{\mathbf{z}}(\mathcal{H}) - er_Q(\mathcal{H})| \leq \sqrt{\frac{64}{mn} \ln \frac{8\mathcal{C}(\epsilon/8, \mathbb{H}_l^n)}{\delta}} + \sqrt{\frac{64}{n} \ln \frac{8\mathcal{C}(\epsilon/8, \mathbb{H}^*)}{\delta}},$$

- When applying the above covering number bound to analyze deep neural networks, Baxter derives a parameter-count-based generalization bounds.
  - Bounding covering number with Pseudo-dimension V.
  - Since Pseudo-dimension V is of order O(Wlog(W)), where W is the number of total parameters of deep neural network, hence Baxters bound is linear of W and is vacuous when analyzing the overparameterized deep neural networks.

Contents

- Background of Meta Learning Theory
- Task Relatedness-Based Bounds for Meta Learning
- Spectrally-Normalized Bounds for Meta Learning
- Conclusions and Future Works

Our Task Relatedness Measurement

- A novel task relatedness measurement called “almost Π-relatedness“:

Definition 3 *(Almost Π-Related Tasks) Let Π be a set of transformations $\pi : Z \to Z$ and let P, P_1 be probability measures on $Z = X \times Y$. We say that P, P_1 are almost Π-related probability measures\tasks, if the following conditions are satisfied:*
(1) $\exists N, N_1 \subseteq Z$ such that $P(N) = P_1(N_1) = 0$, and
(2) $\exists \pi \in \Pi$, π is a one-to-one mapping from $(Z\backslash N, P)$ onto $(Z\backslash N_1, P_1)$. $\forall A \subseteq Z\backslash N$, A is P-measurable if and only if $\pi(A) = \{\pi(x,y)\big|(x,y) \in A\} \subseteq (Z\backslash N_1)$ is P_1-measurable, and
(3) $\int_{Z\backslash N} 1_A \mathrm{d}P = \int_{Z\backslash N_1} 1_{\pi(A)} \mathrm{d}P_1$, where 1_A is the indicator function on the set A, and
(4) the image $\pi(N)$ is a P_1-measurable set, and the inverse image $\pi^{-1}(N_1)$ is a P-measurable set.

- The almost Π −related task environment:

Definition 5 *(Almost Π-related Environment) In meta learning set up, an environment $(\mathcal{P}, Q)$ on $X \times Y$ is called an **almost Π-related environment**, if there exists a common probability measure P on $X \times Y$, such that for any measures $P_i \in \mathcal{P}(i \in \mathcal{I}, \mathcal{I}$ is the index set), P and P_i are almost Π-related in the sense of Definition 3.*

Our Task Relatedness-Based Bounds

- Our task relatedness-based covering number bound on transfer risk:

Theorem 3 *Let $(\mathcal{P}, Q)$ be an almost Π-related environment on the complete separable metric space $Z = X \times Y$. Let $\mathbf{z}$ be an (m, n)-sample generated by the process described in Section 3.1. Let $\mathbb{H} = \{\mathcal{H}\}$ be any permissible hypothesis space family. Then for any $\mathcal{H} \in \mathbb{H}$, $\epsilon \in (0, 1)$, with probability at least $1 - \delta$ over $\mathbf{z}$, we have*

$$|\hat{er}_{\mathbf{z}}(\mathcal{H}) - er_Q(\mathcal{H})| \leq \sqrt{\frac{64}{mn} \ln \frac{4\mathcal{N}(\epsilon/4, \mathbb{H}_l^n, d_{\bar{\mathbf{z}}})}{\delta}}.$$

- We extend the above bound to the representation learning setting:

Theorem 4 *Let $(\mathcal{P}, Q)$ be an almost Π-related environment on the complete separable metric space $Z = X \times Y$. Let $\mathbb{H} = \{\mathcal{H}\}$ be the set of hypothesis spaces of the form $\mathcal{H} = \mathcal{G} \circ f$, $f \in \mathcal{F}$. Then for any $\mathcal{H} \in \mathbb{H}$, any $0 < \epsilon < 1$, with probability at least $1 - \delta$ over $\mathbf{z}$, we have*

$$|\hat{er}_{\mathbf{z}}(\mathcal{H}) - er_Q(\mathcal{H})| \leq \sqrt{\frac{64}{mn} \ln \frac{4}{\delta}} + \sqrt{\frac{64}{mn} \left(\max_{1 \leq j \leq n} \ln \mathcal{N}\left(\frac{\epsilon}{8}, \mathcal{F}, d_{[P_{\mathbf{z},j}, \mathcal{G}_l]}\right) + \ln \mathcal{N}\left(\frac{\epsilon}{8}, \mathcal{G}_l, 2m\right)\right)}$$

Proof sketch: show that $\mathcal{N}(\epsilon_1 + \epsilon_2, \mathbb{H}_l^n, d_{\bar{\mathbf{z}}}) \leq \mathcal{N}(\epsilon_1, \bar{\mathcal{F}}, d_{[P_{\bar{\mathbf{z}}}, \mathcal{G}_l^n]}) \mathcal{N}(\epsilon_2, \mathcal{G}_l^n, 2m)$ and $\mathcal{N}(\epsilon, \mathcal{G}_l^n, 2m) \leq \mathcal{N}(\epsilon, \mathcal{G}_l, 2m)^n$

Contents

- Background of Meta Learning Theory
- Task Relatedness-Based Bounds for Meta Learning
- Spectrally-Normalized Bounds for Meta Learning
- Conclusions and Future Works

Our Spectrally-Normalized Bound

- A parameter-count-free bound for meta learning with deep neural network:

Theorem 5 *Let $(\mathcal{P}, Q)$ be an almost Π-related environment on the complete separable metric space $Z = X \times Y$. Let $\mathbb{H} = \{\mathcal{H}_{\mathcal{A}}\} = \{\mathcal{G} \circ f_{\mathcal{A}} : f_{\mathcal{A}} \in \mathcal{F}, \mathcal{A} = (A_1, ..., A_L), \|A_i\|_\sigma \leq s_i, \|A_i^\top - M_i^\top\|_{2,1} \leq b_i, i \in [L]\}$ be a hypothesis space family where each $\mathcal{H}_{\mathcal{A}}$ is of the form $\mathcal{H}_{\mathcal{A}} = \mathcal{G} \circ f_{\mathcal{A}} = \{g \circ f_{\mathcal{A}}(\cdot) : g = \sigma \circ W, W \in \mathbb{R}^{k \times d}, \|W^\top\|_{2,1} \leq \theta\}$, where σ is an element-wise function with Lipschitz constant θ_σ. Suppose that $\exists b > 0$, for any $x \in X \subseteq \mathbb{R}^{d_0}$, $\|x\|_2 \leq b$. Suppose that the loss function l satisfies two conditions: (1) when composed with g, $g_l(\cdot, y)$ is an α-Lipschitz function w.r.t. the norm $\|\cdot\|_2$, $\forall$ fixed $y \in Y$; (2) $\forall$ fixed $v \in \mathbb{R}^d$, fixed $y \in Y$, $\forall g = \sigma \circ W, g' = \sigma \circ W' \in \mathcal{G}$, $\exists \beta > 0$, such that $|g_l(v, y) - g'_l(v, y)| \leq \beta \|Wv - W'v\|_2$. Then for any $\mathcal{H}_{\mathcal{A}} \in \mathbb{H}$, for any $0 < \epsilon < 1/8$, with probability at least $1 - \delta$ over $\mathbf{z}$, we have*

$$|\hat{er}_{\mathbf{z}}(\mathcal{H}_{\mathcal{A}}) - er_Q(\mathcal{H}_{\mathcal{A}})| \leq \sqrt{\frac{64}{mn} \ln \frac{4}{\delta}} + \frac{8b \prod_{l=1}^{L} s_l \rho_l}{\epsilon \sqrt{mn}} \left[\alpha \sqrt{\ln(2D^2)} \left(\sum_{i=1}^{L} \left(\frac{b_i}{s_i}\right)^{\frac{2}{3}}\right)^{\frac{3}{2}} + \beta\theta \sqrt{n \ln(2dk)}\right].$$

- We further apply the above spectrally-normalized bounds for regression/binary classification/multi-class classification problems within the meta learning framework.

When Two Tasks are almost Π-related?

- We show that our proposed task-relatedness notation actually corresponds to the "almost isomorphism" between two measure spaces induced by the probability measures.

Definition 7 *(Almost Isomorphism) Let* $(Z_1, \mathcal{A}, \mu)$ *and* $(Z_2, \mathcal{B}, \nu)$ *be two measure spaces.*
(1)A point isomorphism π *of these spaces is a one-to-one mapping of* Z_1 *on to* Z_2 *such that* $\mu \circ \pi^{-1} = \nu$ *and* $\pi(\mathcal{A}) = \mathcal{B}$. *That is,* $\forall A \in \mathcal{A}, \pi(A) \in \mathcal{B}$, *and vice versa.*
(2) $(Z_1, \mathcal{A}, \mu)$ *and* $(Z_2, \mathcal{B}, \nu)$ *are called* almost isomorphic *if there exist sets* $N_1 \in \mathcal{A}_\mu, N_2 \in \mathcal{B}_\nu$ *with* $\mu(N_1) = \nu(N_2) = 0$ *and a point isomorphism* π *of the spaces* $Z_1 \backslash N_1$ *and* $Z_2 \backslash N_2$ *that are equipped with the restriction of the measures* μ *and* ν *and the complete* σ*-algebra* $\mathcal{A}_\mu$ *and* $\mathcal{B}_\nu$.

- The task-relatedness condition is satisfied when the sample space is a complete separable metric space.

Theorem 6 *Let* $(\mathcal{P}, Q)$ *be an environment on the complete separable metric space* Z. *Then for any atomeless* $P_i, P_j \in \mathcal{P}(i \neq j, i, j \in \mathcal{I})$, *the probability measure spaces* $(Z, \mathcal{B}_i, P_i)$ *and* $(Z, \mathcal{B}_j, P_j)$ *are almost isomorphic. In other words, the two measures* P_i *and* P_j *are almost* Π*-related.*

Contents

Takeaways

- By proposing a novel task relatedness concept, we provide a meta learning bound of order $O(\frac{1}{\sqrt{nm}})$, and further derive a parameter-count-free spectrally-normalized bounds for meta learning with deep neural networks.
- We rigorously demonstrate that the proposed task relatedness corresponds to the almost isomorphism between measure spaces of two tasks, and such relatedness assumption is satisfied when the sample space is a complete separable metric space.
- Our ongoing works include providing sharper meta learning bounds with non-i.i.d. task environment assumption or even without such environment assumption.

Thank you !