An Analysis of Representation Learning

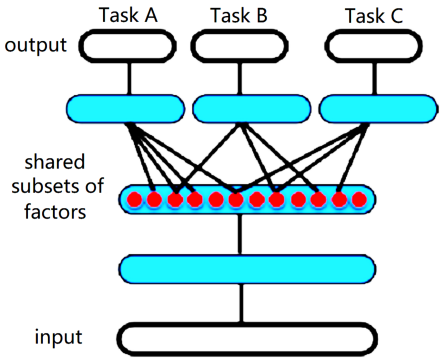
Rui Ai, Hongrui Chen, Chengxin Gong, Kaicheng Shao

School of Mathematical Sciences, Peking University

Dec,22nd

Background and Motivation

- Representation learning is an important scheme in machine learning and achieved success in many applications.
- Example: pre-trained neural network for image recognition
- Intuition: learning underlying structure to reduce sample complexity
- Lack of theoretical understanding



Set up

Data Assumptions:

- T source tasks $1, \cdots, T$ and a target task 0
- Feature Map: $\Phi \in \mathcal{F}^k : \mathbb{R}^d \to \mathbb{R}^k$ selects k features from data
- n_t i.i.d. samples in task t: $y_{t,i} = w_t^\top \Phi(x_{t,i}) + \text{noise}, \ x_{t,i} \sim \rho$

Algorithm:

- Source task training: $\hat{\Phi} \leftarrow \min_{\|\hat{\Phi}\|_{\mathcal{F}^k} \le 1, \|\hat{w}_t\|_2 \le 1} \frac{1}{T} \sum_{t=1}^{T} \frac{1}{n_t} \sum_{i=1}^{n_t} (y_{t,i} - (\hat{w}_t)^\top \hat{\Phi}(x_{t,i}))^2.$
- Target task training: $\hat{w}_0 \leftarrow \min_{\|\hat{w}_0\| \le 1} \frac{1}{n_0} \sum_{i=1}^{n_0} (y_{0,i} - (\hat{w}_0)^\top \hat{\Phi}(x_{0,i}))^2.$

[Du et al., 2021][1] and [Tripuraneni et al., 2021] [2] also consider this setting.

[1] Du, S. S., Hu, W., Kakade, S. M., Lee, J. D., and Lei, Q. (2021). Few-shot learning via learning the representation, provably.

[2] Tripuraneni, N., Jin, C., and Jordan, M. I. (2021). Provable meta- learning of linear representations.

Theoretical Results

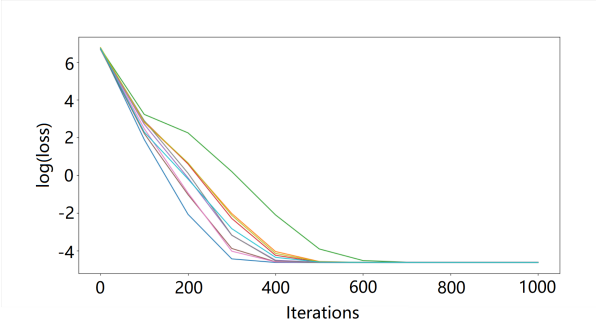
- Excess error: $R_t(\hat{\Phi}, \hat{w}_t) = \mathbb{E}_{x,y}\left((\hat{w}_t)^\top \hat{\Phi}(x) - y_t\right)^2 - \mathbb{E}_{x,y}\left(w_t^{*\top} \Phi(x) - y_t\right)^2$
- $R_0(\hat{\Phi}, \hat{w}_0) = R_0(\hat{\Phi}, w_0^*) + (R_0(\hat{\Phi}, \hat{w}_0) - R_0(\hat{\Phi}, w_0^*))$, where $w_t^* = \operatorname{argmin}_w R_t(\hat{\Phi}, w)$.
- If $w_0^\top \Phi(x) = \sum_{t=1}^{T} a_t w_t^\top \Phi(x)$, then $R_0(\hat{\Phi}, w_0^*) \leq \sum_{t=1}^{T} a_t^2 \sum_{t=1}^{T} R_t(\hat{\Phi}, \hat{w}_t)$
- If $d := \inf_{a_1, \cdots, a_T \in \mathbb{R}} \|w_0^\top \Phi(x) - \sum_{t=1}^{T} a_t w_t^\top \Phi(x)\|_{L_2(\rho)}$, then $R_0(\hat{\Phi}, w_0^*) \geq d$.

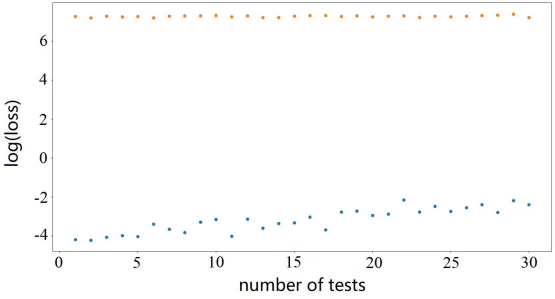
Theorem (Informally stated)

Suppose that $w_0^\top \Phi(x) = \sum_{t=1}^{T} a_t w_t^\top \Phi(x)$, with probability at least $1 - \delta$, the generalization excess error of the target task learned by the two-stage procedure is bounded as

$$R_0(\hat{\Phi}, \hat{w}_0) \lesssim (1 + \sigma)\left(\frac{1}{\sqrt{n_0}} + \sum_{t=1}^{T} a_t^2 \sum_{t=1}^{T} \hat{\mathcal{G}}_{n_t}(\mathcal{F}_1)\right) + C_\delta$$

Experiment Results





Conclusion

In the setting of multi-tasks representation learning, we provide a comprehensive analysis on

- the conditions we require for obtaining a "good" representation.
- how sample complexity of the target task decreases using a "good" representation

Future work:

- More reasonable setting for understanding representation learning
- Representation learning beyond linear prediction functions

Thank You For Listening.