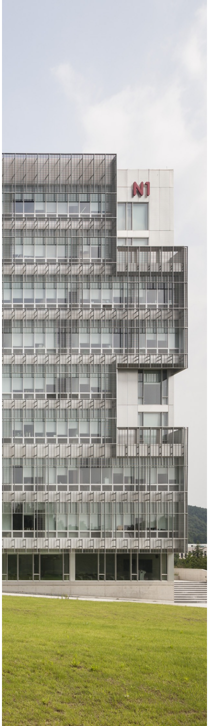


Theoretical Analysis of Weak-to-Strong Generalization

Neurips 2024

Hunter Lang, David Sontag, Aravindan Vijayaraghavan

KAIST

Problem Setup

Notation

- $\boldsymbol{x} \sim \mathcal{D}$, assume input space $\mathcal{X}$ is discrete.
- Ground-truth function $y : \mathcal{X} \to \mathcal{Y} = \{1, \ldots, k\}$
- Pseudo-labeler (Teacher) $\tilde{y} : \mathcal{X} \to \mathcal{Y} \cup \{\varnothing\} \to$ Label is either in $\mathcal{Y}$ or does not exist.
- Covered set $S = \{\boldsymbol{x} | \tilde{y}(\boldsymbol{x}) \neq \varnothing\}$: subset of $\mathcal{X}$ that has a pseudo-label.
- Uncovered set $T = \{\boldsymbol{x} | \tilde{y}(\boldsymbol{x}) = \varnothing\} = \mathcal{X} \setminus S$
- Partitioning based on the ground-truth: $\mathcal{X}_i = \{\boldsymbol{x} | y(\boldsymbol{x}) = i\}$, $S_i = S \cap \mathcal{X}_i$, $T_i = T \cap \mathcal{X}_i$
  $\to \{\mathcal{X}_i\}, \{S_i\}$, and $\{T_i\}$ are partitions of $\mathcal{X}$, S, and T respectively.
- Partitioning S_i: $S_i^{good} : \{\boldsymbol{x} \in S_i | \tilde{y}(\boldsymbol{x}) = y(\boldsymbol{x})\}$, $S_i^{bad} : \{\boldsymbol{x} \in S_i | \tilde{y}(\boldsymbol{x}) \neq y(\boldsymbol{x})\}$
- Error rate of pseudo-labeler: $\alpha_i = \mathbb{P}(S_i^{bad} | S_i)$, assume $\alpha_i \in (0, 1/2)$ for all i.
- Strong model hypothesis class: $\mathcal{F}$

Problem Setup.

- Training on the covered set targeted by the weak label: $f^* = \arg\min_{\mathcal{F}} \mathrm{err}(f, \tilde{y} | S)$.
- Goal: Obtain upper bounds on the error: $\mathrm{err}(f^*, y | \mathcal{X})$.

Illustrative Example: Sentiment Classifier

Settings

- $\mathcal{X}$: Text documents, $\mathcal{Y} = \{-1, +1\}$.
- Pseudo-label

$$\tilde{y} = \begin{cases} +1, & \text{if } \texttt{`incredible'} \in \boldsymbol{x}, \\ -1, & \text{if } \texttt{`horrible'} \in \boldsymbol{x}, \\ \varnothing, & \text{otherwise.} \end{cases}$$

  Assume 'incredible' and 'horrible' never co-occur.
- $\alpha_{-1} = \alpha_{+1} = \alpha$.

Existing Error bounds

- Proposition 3.1 (Bound from [1]): $\tilde{f} = \mathbb{P}(S) \cdot 4\alpha(1-\alpha) + \mathbb{P}(T)$

  $\rightarrow$ No pseudo-label correction in $\alpha < 3/4$.
- Proposition 3.2 (Bound from [2]): $\mathbb{P}(S)$ should be larger than $2/3$.

  $\rightarrow$ Cannot explain weak-to-strong generalization in the low-coverage regime.

$\Rightarrow$ Existing studies fail to explain weak-to-strong generalization.

Neighborhood and η-Robustness

Neighborhood $\mathcal{N}$

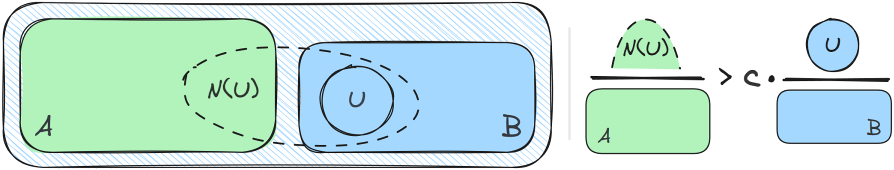
- General definition: Function that maps each point $\boldsymbol{x}$ to a set of points $\mathcal{N}(\boldsymbol{x}) \in \mathcal{X}$.
- Symmetric assumption: $\boldsymbol{x} \in \mathcal{N}(\boldsymbol{x}') \Leftrightarrow \boldsymbol{x}' \in \mathcal{N}(\boldsymbol{x})$.
- Examples: $\mathcal{N}(x) = \{\boldsymbol{x}' : \|\varphi(\boldsymbol{x}) - \varphi(\boldsymbol{x}')\| \leq r\}$ for some rep. function $\varphi : \mathcal{X} \to \mathbb{R}^d$.

η-Robustness

- $r(f, \boldsymbol{x}) = \mathbb{P}(f(\boldsymbol{x}') \neq f(\boldsymbol{x}) | \boldsymbol{x}' \in \mathcal{N}(\boldsymbol{x}))$
- η-robust at $\boldsymbol{x}$: $r(f, \boldsymbol{x}) \leq \eta$.
- $R_\eta(f) = \{\boldsymbol{x} : r(f, \boldsymbol{x}) \leq \eta\}$: set of η-robust points for f.
- Average-case robustness: Classifier gives the same labels for most of their neighbors.

$$\mathbb{E}_{\boldsymbol{x} \sim \mathcal{D}, \boldsymbol{x}' \sim \mathcal{D}|\mathcal{N}(\boldsymbol{x})}[f(\boldsymbol{x}) \neq f(\boldsymbol{x}')] \leq \gamma$$

is η-robust in probability at least $1 - \gamma/\eta$.



Expansion: For fixed sets $A, B \in \mathcal{X}$, $\mathbb{P}_x$ satisfies (c, q)-expansion on (A, B) if

$$\forall U \in B \text{ with } \mathbb{P}(U|B) > q, \quad \mathbb{P}(\mathcal{N}(U)|A) > c\mathbb{P}(U|B).$$

Expansion of a set collection (Relaxed version): Suppose $\mathcal{M}$ is a collection of subsets of B. $\mathcal{M}$ is (c, q)-expansion on (A, B) if

$$\forall U \in \mathcal{M} \text{ with } \mathbb{P}(U|B) > q, \quad \mathbb{P}(\mathcal{N}(U)|A) > c\mathbb{P}(U|B).$$

Adversarially Robust Models ($\eta = 0$)

Expansion

- "Bad" points (incorrect pseudolabels) have many "good" neighbors.
- "Uncovered" points (no pseudolabels) have many "good" neighbors.

Idea: Student model with "robust" on the neighborhoods

Theorem 4.1 (Pseudo-label correction, informal.)

The true error of f on covered set S_i satisfies:

$$err(f, y|S_i) \leq err(f, \tilde{y}|S_i) + \alpha_i \left(1 - \frac{3}{2}c\right).$$

Trivial bounds: $err(f, y|S_i) \leq err(f, \tilde{y}|S_i) + \alpha_i \rightarrow$ tighter than trivial bounds.

Theorem 4.2 (Coverage Expansion, informal.)

The true error of f on uncovered set T_i satisfies:

$$err(f, y|T_i) \leq err(\bar{R}(f)|T_i) + \max\left(q, \frac{err(f, \tilde{y}|S_i) - c\alpha_i}{c(1 - 2\alpha_i)}\right).$$

KAIST

References

[1] D. Fu, M. Chen, F. Sala, S. Hooper, K. Fatahalian, and C. Ré, "Fast and three-rious: Speeding up weak supervision with triplet methods," in *International conference on machine learning*, pp. 3280–3291, PMLR, 2020.

[2] C. Wei, K. Shen, Y. Chen, and T. Ma, "Theoretical analysis of self-training with deep networks on unlabeled data," *arXiv preprint arXiv:2010.03622*, 2020.