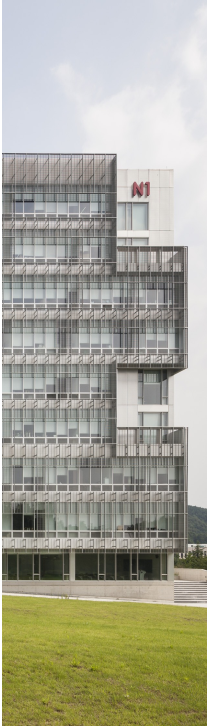



# Quantifying the Gain in Weak-to-Strong Generalization

Neurips 2024

Moses Charikar, Chirag Pabbaraju, Kirankumar Shiragur

KAIST

# Overview

**Main Question**: *weak-to-strong model > weak model?*

**Intuition**: Gain in weak-to-strong generalization $\approx$ Misfit between the weak and strong model

**Problem Setup**

- Data domain: $\mathbb{R}^d$
- Ground-truth representation function $h^* : \mathbb{R}^d \to \mathbb{R}^{d^*}$
- Target finetuning task: $f^* \circ h^*$
- Strong model $f_s \circ h_s$
  - Function class of the representation function $\mathcal{H}_s : \mathbb{R}^d \to \mathbb{R}^{d_s}$
  - Function class: $\mathcal{F}_s : \mathbb{R}^{d_s} \to \mathbb{R}$; assume that $\mathcal{F}_s$ is a *convex* set.
- Weak model $f_w \circ h_w$
  - Function class of the representation function $\mathcal{H}_w : \mathbb{R}^d \to \mathbb{R}^{d_w}$
- Distance between functions in $\mathcal{P}$: $d_{\mathcal{P}}(f, g) = \mathbb{E}_{x \sim \mathcal{P}}(f(x) - g(x))^2$



KAIST

# W2S Generalization under Realizability

**Theorem 1 (Realizability)**

Let $f_w \circ h_w$ be the function learnt by the weak model for some arbitrary function $f_w : \mathbb{R}^{d_w} \in \mathbb{R}$. Define an weak-to-strong model $f_{sw}$ as

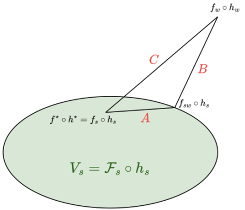
$$f_{sw} = \arg\min_{f \in \mathcal{F}_s} d_{\mathcal{P}}(f \circ h_s, f_w \circ h_w).$$

Assume that there exists $f_s \in \mathcal{F}_s$ such that $f_s \circ h_s = f^* \circ h^*$ (realizable). Then, we have

$$d_{\mathcal{P}}(f_{sw} \circ h_s, f^* \circ h^*) \leq d_{\mathcal{P}}(f_w \circ h_w, f^* \circ h^*) - d_{\mathcal{P}}(f_{sw} \circ h_s, f_w \circ h_w).$$

**Meaning of Theorem 1**: W2S model improves weak model by an amount equal to the *misfit*.

**Proof sketch.** $V_s := \{f \circ h_s : f \in \mathcal{F}_s\}$ is also a convex set.




- A: $d_{\mathcal{P}}(f_{sw} \circ h_s, f^* \circ h^*)$: first term
- B: $d_{\mathcal{P}}(f_{sw} \circ h_s, f_w \circ h_w)$: third term
- C: $d_{\mathcal{P}}(f_w \circ h_w, f^* \circ h^*)$: second term

Pythagorean theorem onto a convex set: $C \geq A + B$

# W2S Generalization under Non-Realizability

**Theorem 2 (Non-Realizability and Finite Samples)**

For a convex set of functions $\mathcal{F}_s$, define $f_s$ as

$$f_s = \arg\min_{f \in \mathcal{F}_s} d_{\mathcal{P}}(f \circ h_s, f^* \circ h^*),$$

and $\epsilon := d_{\mathcal{P}}(f_s, h_s, f^*, h^*)$ (Non-realizability). Suppose we obtain $n$ i.i.d. samples from the weak model; $\{(x_i, y_i)\}_{i=1}^n$, where $x_i \sim \mathcal{P}$ and $y_i = f_w \circ h_w(x_i)$. Define $\hat{f}_{sw}$ as

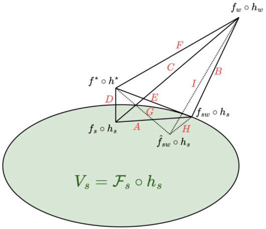
$$\hat{f}_{sw} = \arg\min_{f \in \mathcal{F}_s} \frac{1}{n} \sum_{i=1}^{n} (f \circ h_s(x_i) - y_i)^2. \quad \text{(Finite Sample)}$$

Assume that the range of $f^*$, $f_w$ and all functions in $\mathcal{F}$ is absolutely bounded. Then,we have that with probability at least $1-\delta$,

$$\begin{aligned} d_{\mathcal{P}}(\hat{f}_{sw} \circ h_s, f^* \circ h^*) \le\ & d_{\mathcal{P}}(f_w \circ h_w, f^* \circ h^*) - d_{\mathcal{P}}(\hat{f}_{sw} \circ h_s, f_w \circ h_w) \\ & + O(\sqrt{\epsilon}) + O\left(\frac{\mathcal{C}_{\mathcal{F}_s}}{n}\right)^{1/4} + O\left(\frac{\log(1/\delta)}{n}\right)^{1/4}, \end{aligned}$$

where $\mathcal{C}_{\mathcal{F}_s}$ is the complexity of the function class $\mathcal{F}_s$.

# Proof of Theorem 2




**Goal**: $F + O(\cdot) \geq G + I$

- $F$: $d_{\mathcal{P}}(f_w \circ h_w, f^* \circ h^*)$: second term
- $G$: $d_{\mathcal{P}}(\hat{f}_{sw} \circ h_s, f^* \circ h^*)$: first term
- $I$: $d_{\mathcal{P}}(\hat{f}_{sw} \circ h_s, f_w \circ h_w)$: third term

**Proof sketch.**

- $F$: $\sqrt{C} \leq \sqrt{D} + \sqrt{F} \rightarrow$ Triangle inequality
- $G$: $\sqrt{G} \leq \sqrt{E} + \sqrt{H} \rightarrow$ Triangle inequality
- $I$: $I \leq B + O\left(\sqrt{\frac{\mathcal{C}_{\mathcal{F}_s}}{n}}\right) + O\left(\sqrt{\frac{\log(1/\delta)}{n}}\right) \rightarrow$ Lemma 4
- $C, D, E, H, B$
  - $D = \epsilon$
  - $C \geq A + B \rightarrow$ Theorem 1
  - $I \geq H + B \rightarrow$ Theorem 1
  - $\sqrt{E} \leq \sqrt{A} + \sqrt{D} \rightarrow$ Triangle inequality

# Synthetic Experiment

## Experimental Setup

- Target data representation $h^* : \mathbb{R}^8 \to \mathbb{R}^{16}$; randomly initialized **5-layer** MLP with ReLU activations, with input dimension 8 and hidden layer dimension 16.
- Function space of strong(weak) model $\mathcal{F}_s : \mathbb{R}^{16} \to \mathbb{R}$: the class of linear functions
- Data distribution $\mathcal{P} = \mathcal{N}(0, \sigma^2 \boldsymbol{I})$, $\sigma = 500$.

## Representation Learning

- Pretraining: Obtain representation via training

1. Randomly sample $T$ finetuning tasks $f^{(1)} \ldots, f^{(t)} \in \mathcal{F}_s$.
2. Generate data $\{x_j^{(t)}, y_j^{(t)}\}_{j=1}^{N_r}$, where $x_j^{(t)} \sim \mathcal{P}$ and $y_j^{(t)} = f^{(t)} \circ h^*(x_j^{(t)})$.
3. Obtain $h_w$, $h_s$ as

$$h_k = \arg\min_{h \in \mathcal{H}_k} \frac{1}{TN_r} \sum_{t=1}^{T} \sum_{j=1}^{N_r} (f^{(t)} \circ h(x_j^{(t)}) - y_j^{(t)})^2, \quad (T = 10, N_r = 2000)$$

   where $\mathcal{H}_w$ and $\mathcal{H}_s$ be the classes of **2-layer** and **8-layer** neural networks, respectively.
4. Realizable setting: $h_s = h^*$.

# Synthetic Experiment

- Perturbations: Obtain representation via direct perturbations
  - $h_w, h_s$: perturb every parameter in $h^*$ by independent noise $\mathcal{N}(0, \sigma_w^2), \mathcal{N}(0, \sigma_s^2)$.
  - $\sigma_s < \sigma_w$: $h_s$ is closer approximation of $h^*$ than $h_w$.

**Weak Model Finetuning**: Fixed $h_w$ and $h_s$, find $f_w$

1. Randomly sample $M$ *new* finetuning tasks $f^{(1)}, \ldots, f^{(M)} \in \mathcal{F}_s$.
2. Generate data $\{x_j^{(i)}, y_j^{(i)}\}_{j=1}^{N_f}$, where $x_j^{(i)} \sim \mathcal{P}$ and $y_j^{(i)} = f^{(i)} \circ h^*(x_j^{(i)})$.
3. Obtain weak model $f_w^{(i)}$ as

$$f_w^{(i)} = \underset{f \in \mathcal{F}_s}{\arg\min} \frac{1}{N_f} \sum_{j=1}^{N_f} (f \circ h_w(x_j^{(i)}) - y_j^{(i)})^2, \quad (M = 100, N_f = 2000)$$

**Weak-to-Strong Supervision**: Train strong model from weakly labeled data

1. For each $i \in [M]$, generate data $\{\tilde{x}_j^{(i)}, \tilde{y}_j^{(i)}\}_{j=1}^{N_f}$, where $\tilde{x}_j^{(i)} \sim \mathcal{P}$ and $\tilde{y}_j^{(i)} = f_w^{(i)} \circ h_w(\tilde{x}_j^{(i)})$.
2. Obtain weak-to-strong model as

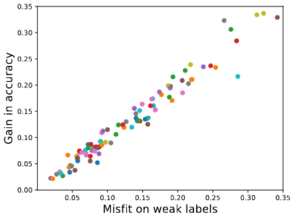
$$f_{sw}^{(i)} = \underset{f \in \mathcal{F}_s}{\arg\min} \frac{1}{N_f} \sum_{j=1}^{N_f} (f \circ h_s(\tilde{x}_j^{(i)}) - \tilde{y}_j^{(i)})^2$$

# Synthetic Experiment

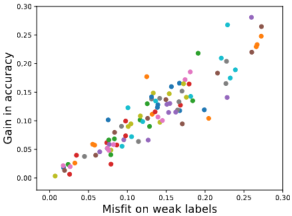
**Evaluation**

- $d_{\mathcal{P}}(f_{sw}^{(i)} \circ h_s, f^{(i)} \circ h^*)$: error of the weak-to-strong model on the true finetuning task
- $d_{\mathcal{P}}(f_{w}^{(i)} \circ h_w, f^{(i)} \circ h^*)$: error of the weak model on the true finetuning task
- $d_{\mathcal{P}}(f_{sw}^{(i)} \circ h_s, f_{w}^{(i)} \circ h_w)$: misfit error of the w2s model on the weakly label data

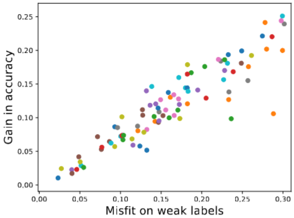
**Results**




(a) Realizable (pretraining).




(b) Non-realizable (pretraining).




(c) Non-realizable (perturbation).

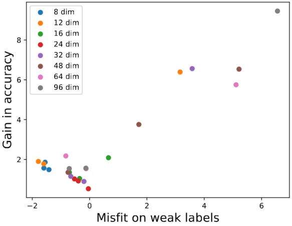
x-axis: $d_{\mathcal{P}}(f_{sw}^{(i)} \circ h_s, f_{w}^{(i)} \circ h_w)$, y-axis: $d_{\mathcal{P}}(f_{w}^{(i)} \circ h_w, f^{(i)} \circ h^*) - d_{\mathcal{P}}(f_{sw}^{(i)} \circ h_s, f^{(i)} \circ h^*)$

**Gain of the weak-to-strong supervision $\approx$ misfit error of the weak-to-strong model!**

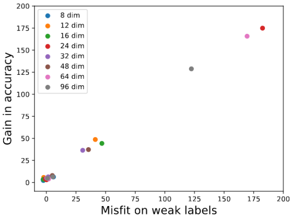
### Experimental Setup

- Three regression datasets: ESOL, FreeSolv and Liplop from MolBERT[1].
- Strong representation $h_s$: Pretrained BERT (hidden dimension 768, 12 layers, 12 attention heads) on GuacaMol dataset.
- Weak representation $h_w$: Transformers with 2 layers and 2 attention heads with hidden dimension $\{8, 12, 16, 32, 48, 64, 96\}$.

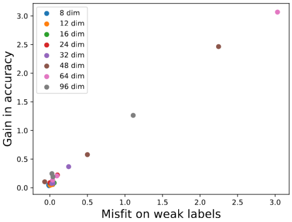
### Results




(d) MolBERT on ESOL




(e) MolBERT on FreeSolv




(f) MolBERT on Lipop

**Gain of the weak-to-strong supervision $\approx$ misfit error of the weak-to-strong model!**

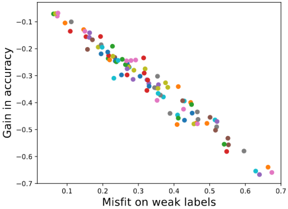
# Real Experiments

| Hidden dimension | Weak error - Misfit | True error of weakly-supervised strong model |
|---|---|---|
| 96 | $\mathbf{0.8969 \pm 0.0327}$ | $\mathbf{1.0713 \pm 0.0489}$ |
| 48 | $0.9731 \pm 0.0707$ | $1.1293 \pm 0.0418$ |
| 24 | $1.0331 \pm 0.0449$ | $1.1204 \pm 0.0261$ |
| 64 | $1.0619 \pm 0.0441$ | $1.1436 \pm 0.0124$ |
| 32 | $1.0624 \pm 0.0527$ | $1.1302 \pm 0.0220$ |
| 16 | $1.1456 \pm 0.0276$ | $1.1950 \pm 0.0484$ |
| 12 | $1.1499 \pm 0.0177$ | $1.1869 \pm 0.0297$ |
| 8 | $1.1958 \pm 0.0194$ | $1.2396 \pm 0.0310$ |

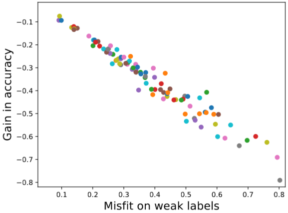
(Weak error - Misfit) $\downarrow$

$\Rightarrow$ (True error of weak-to-strong model) $\downarrow$.

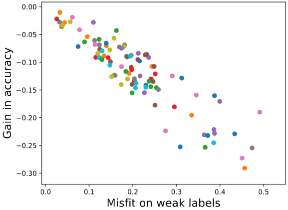
## Strong-to-Weak Generalization




(a) Non-realizable (pretraining). $h^\star$: 5-layer MLP, $h_w$: 8-layer MLP, $h_s$: 2-layer MLP.




(b) Non-realizable (perturbation). $h_w = h^\star + \mathcal{N}(0, 0.01^2)$, $h_s = h^\star + \mathcal{N}(0, 0.05^2)$. $h^\star$ is 5-layer MLP.




(c) Non-realizable (pretraining). $h^\star$: 5-layer MLP, $h_w$: 2-layer MLP, $h_s$: 8-layer MLP. $T = 5$, $N_r = 250$.

- Reverse the weak and strong models $\rightarrow$ flipped figures in (a) and (b).
- Low sample regime $\rightarrow$ weak model learns a better representation $\rightarrow$ flipped figure in (c).



KAIST

# Summary

**Goal**: Quantify the gain of the *weak-to-strong model*

**Theorem 1 and Theorem 2**

- Gain of the weak-to-strong supervision $\geq$ misfit error of the weak-to-strong model

**Synthetic Experiment**

- Representation function: $h^*$: 5 Layer, $h_w$: 2 layer, $h_s$: 8 layer MLP
- Function space $\mathcal{F}_s$: set of linear function
- Three stages of experiment
  1. Representation learning: pretraining (true target), perturbations
  2. Weak model finetuning (true target)
  3. Weak-to-strong supervision (weakly labeled target)

**Real Experiment**

- Real Dataset: ESOL, FreeSolv and Lipop
- Model: Pretrained MolBERT (strong model), 2 layer transformer (weak model)
- Result: Gain of the weak-to-strong supervision $\approx$ misfit error of the weak-to-strong model
- Swapped $h_w$ and $h_s$ or low sample regime $\rightarrow$ flipped figures

# References

[1] B. Fabian, T. Edlich, H. Gaspar, M. Segler, J. Meyers, M. Fiscato, and M. Ahmed, "Molecular representation learning with language models and domain-relevant auxiliary tasks," *arXiv preprint arXiv:2011.13230*, 2020.