

# **Discrepancies are Virtue: Weak-to-Strong Generalization through Lens of Intrinsic Dimension**

Yijun Dong

Courant Institute of Mathematical Sciences, New York University

Flatiron CCM ML Seminar, May 30, 2025



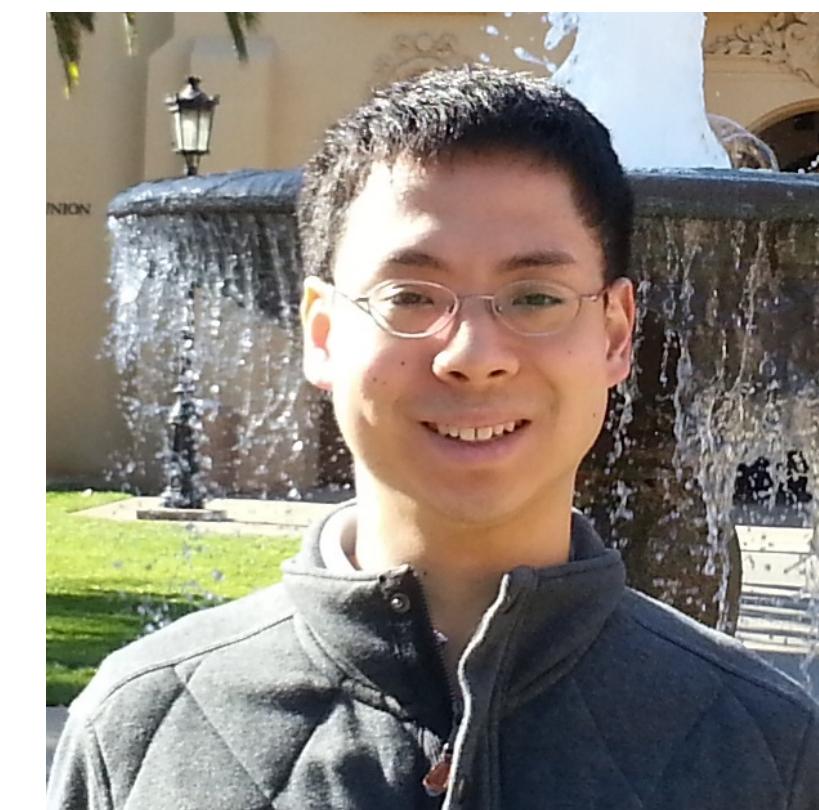
# Joint work with



Yicheng Li  
NYU



Yunai Li  
SJTU

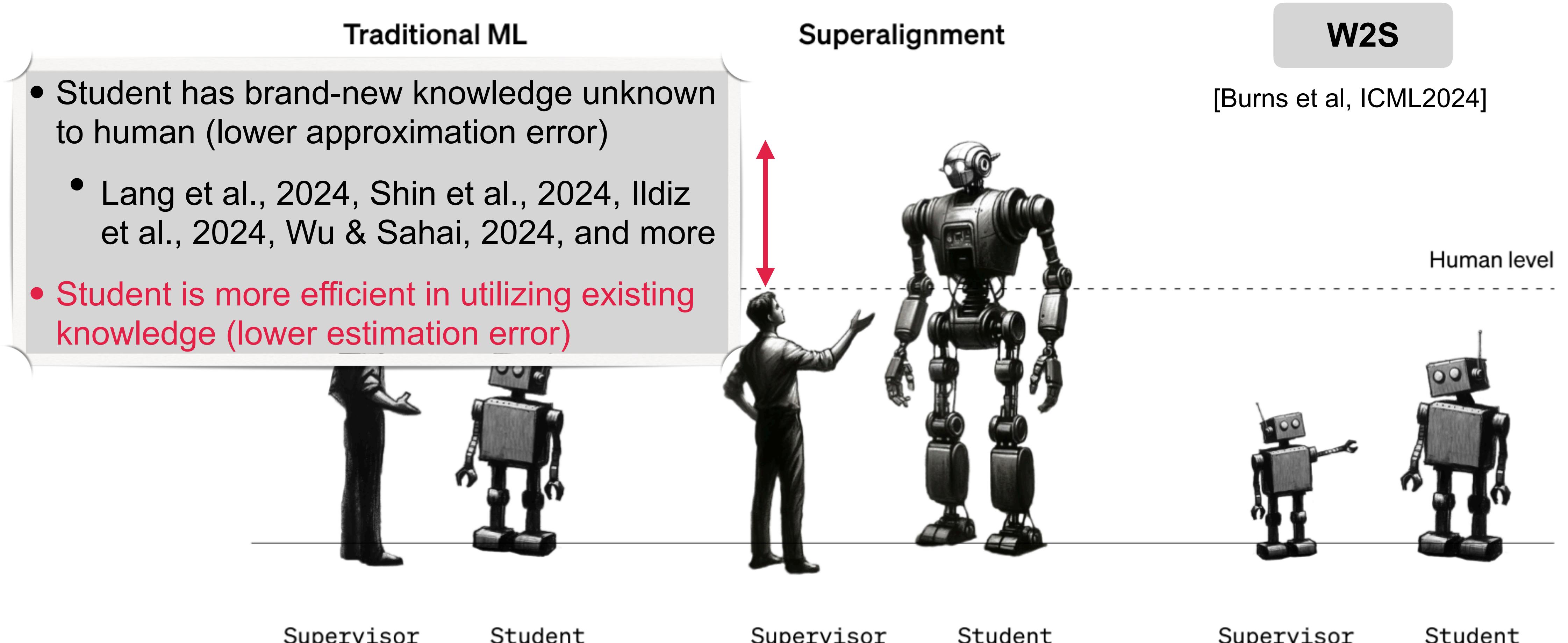


Jason D. Lee  
Princeton



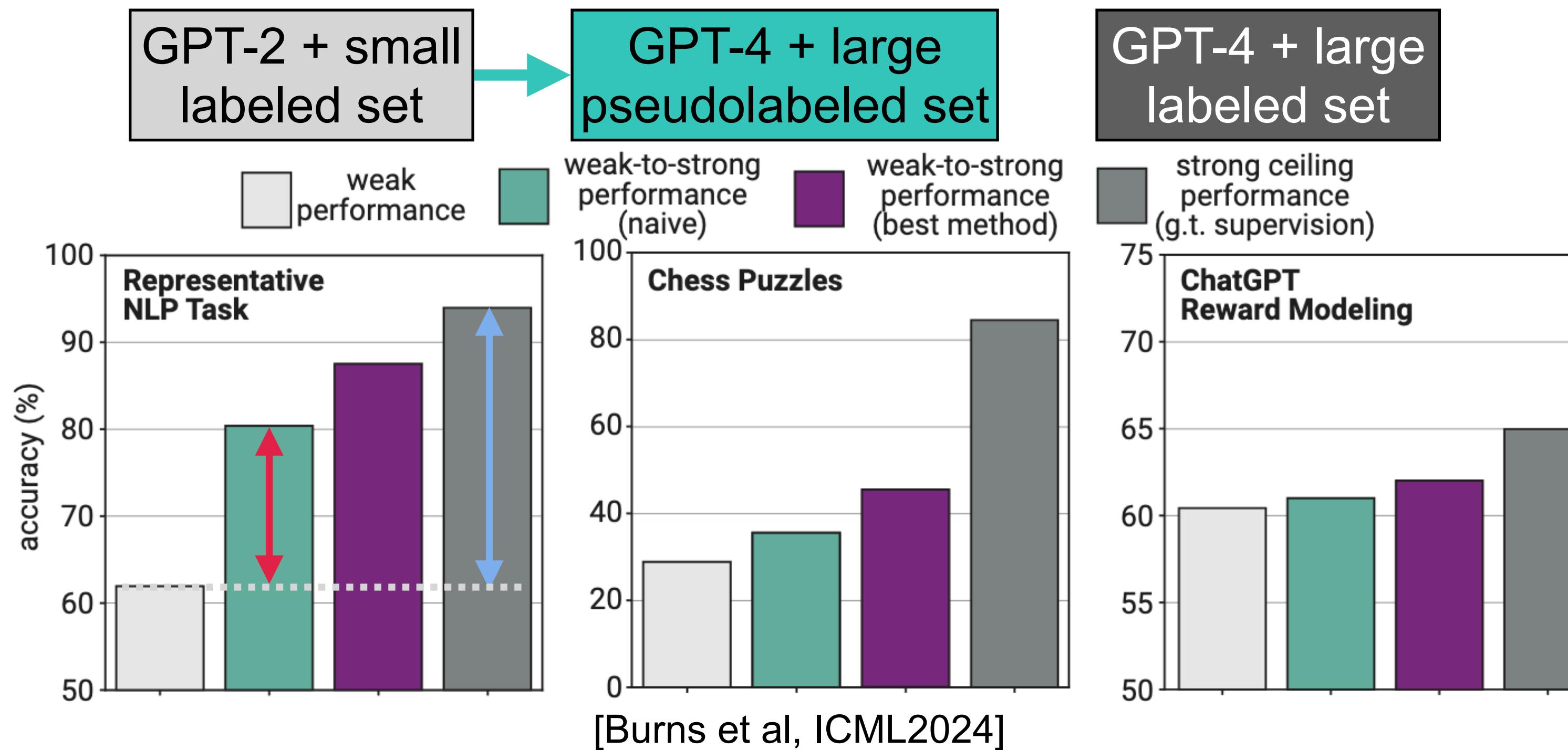
Qi Lei  
NYU

# Superalignment → weak-to-strong (W2S) generalization



When and how does weak-to-strong generalization happen?

# Better W2S generalization on easier tasks



- Difficulty (by **strong ceiling performance**): NLP < Chess < ChatGPT reward model
- Approximation error = error of the model trained over the population
- Better W2S  $\Leftrightarrow$  **performance gap recovery** closer to 1

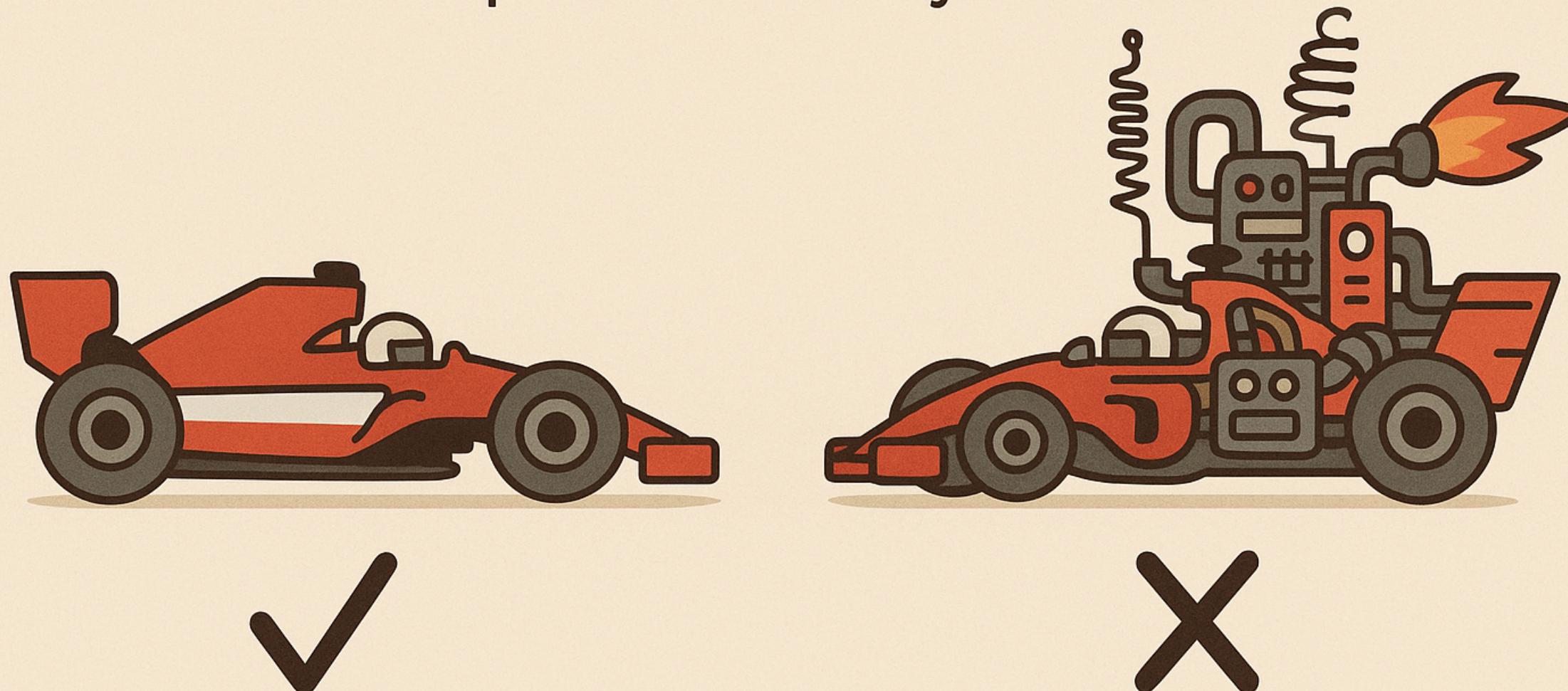
$$\text{PGR} = \frac{\text{W2S-weak gap}}{\text{ceiling-weak gap}}$$

How does W2S happen on easy tasks where weak and strong models both have low approximation errors?

# Intrinsic dimension

## OCCAM'S RAZOR

When faced with multiple hypotheses,  
the simplest is usually the best



**Intrinsic dimension** = the minimal number of model parameters needed to achieve (nearly) optimal performance on a specific task

Pretrained  
initialization

$$\theta^D = \theta_0^D + \Gamma \theta^d$$

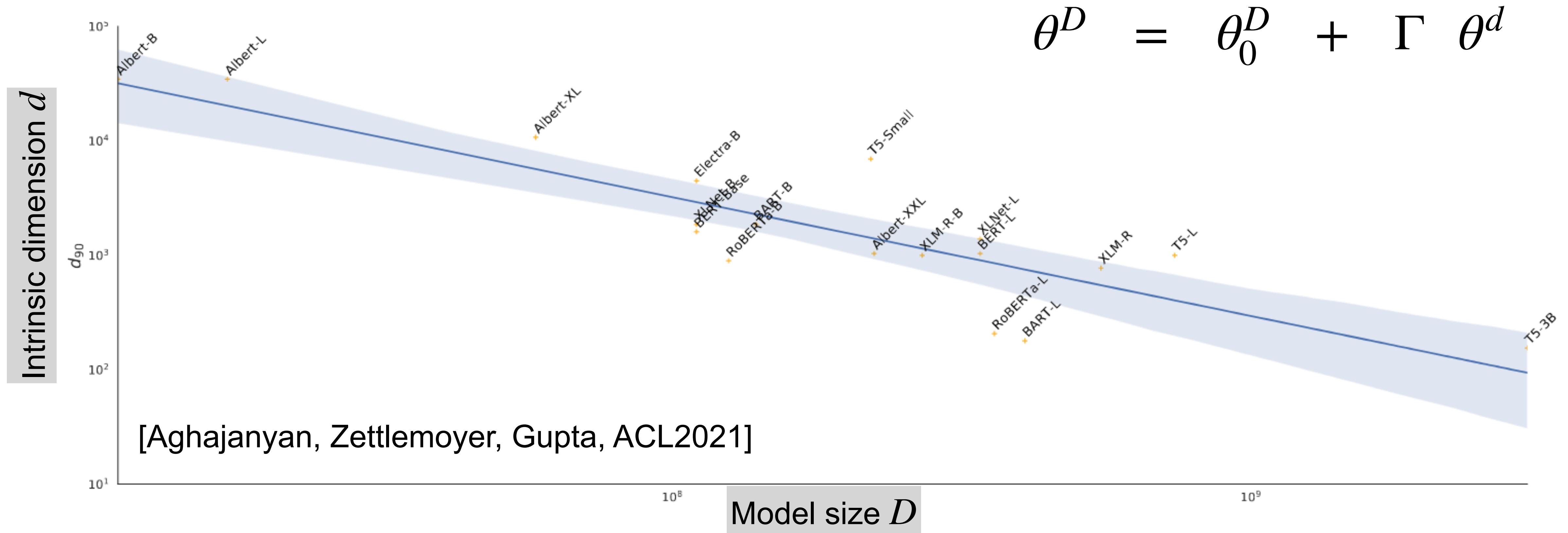
Model parameter of  
high dimension  $D$

Finetunable parameter of  
intrinsic dimension  $d < D$

$$D \times d \text{ random projection}$$

# Low intrinsic dimension of finetuning

Learning over a well-pretrained model (e.g. finetuning) usually exhibits **low intrinsic dimensions**.



[Aghajanyan, Zettlemoyer, Gupta, ACL2021]

Larger pretrained language models have lower intrinsic dimensions on downstream tasks!

# Finetuning with low intrinsic dimensions

## Downstream task

- $(x, y) \sim \mathcal{D}(f_*)$  s.t.  $y = f_*(x) + z$  with i.i.d. noise  $z \sim \mathcal{N}(0, \sigma^2)$  and  $|f_*(x)| < 1$  a.s.
- Want to learn the ground truth function  $f_* : \mathcal{X} \rightarrow \mathbb{R}$  given access to two datasets:
  - **Labeled** (small) dataset:  $\tilde{\mathbf{X}} \in \mathcal{X}^n$  with noisy labels  $\tilde{\mathbf{y}} \in \mathbb{R}^n$
  - **Unlabeled** (large) dataset:  $\mathbf{X} \in \mathcal{X}^N$  with unknown labels  $\mathbf{y} \in \mathbb{R}^N$

## Finetuning (FT) $\approx$ linear probing on low-rank gradient features

- FT fall in **kernel regime**:  $f(x | \theta) = \phi(x)^\top \theta$  with finetunable parameter  $\theta \in \mathbb{R}^d$
- Nonlinear case:  $\phi(x) = \nabla_\theta f(x | \theta_0)$  = gradient at pretrained initialization  $\theta_0 \in \mathbb{R}^d$
- **Weak** model  $\phi_w : \mathcal{X} \rightarrow \mathbb{R}^d$  produces  $\tilde{\Phi}_w = \phi_w(\tilde{\mathbf{X}}) \in \mathbb{R}^{n \times d}$ ,  $\Phi_w = \phi_w(\mathbf{X}) \in \mathbb{R}^{N \times d}$
- **Strong** model  $\phi_s : \mathcal{X} \rightarrow \mathbb{R}^d$  produces  $\tilde{\Phi}_s = \phi_s(\tilde{\mathbf{X}}) \in \mathbb{R}^{n \times d}$ ,  $\Phi_s = \phi_s(\mathbf{X}) \in \mathbb{R}^{N \times d}$

$$\text{rank}(\Sigma_w) = d_w \ll d \quad \text{rank}(\Sigma_s) = d_s \ll d$$

$$\Sigma_w = \mathbb{E}[\phi_w(x)\phi_w(x)^\top]$$

$$\Sigma_s = \mathbb{E}[\phi_s(x)\phi_s(x)^\top]$$

# Weak v.s. strong: model capacity + similarity

Representation efficiency — **low intrinsic dimensions**:

$$\text{rank}(\Sigma_w) = d_w \ll d \quad \text{rank}(\Sigma_s) = d_s \ll d$$

Representation accuracy — **FT approximation error**:  $0 \leq \rho_s \leq \rho_w \leq 1$  where

$$\rho_s := \min_{\theta \in \mathbb{R}^d} \mathbb{E}[(\phi_s(x)^\top \theta - f_*(x))^2] \quad \text{and} \quad \rho_w := \min_{\theta \in \mathbb{R}^d} \mathbb{E}[(\phi_w(x)^\top \theta - f_*(x))^2].$$

We are interested in the **variance-dominated regime**  $\rho_s + \rho_w \ll \sigma^2$ .

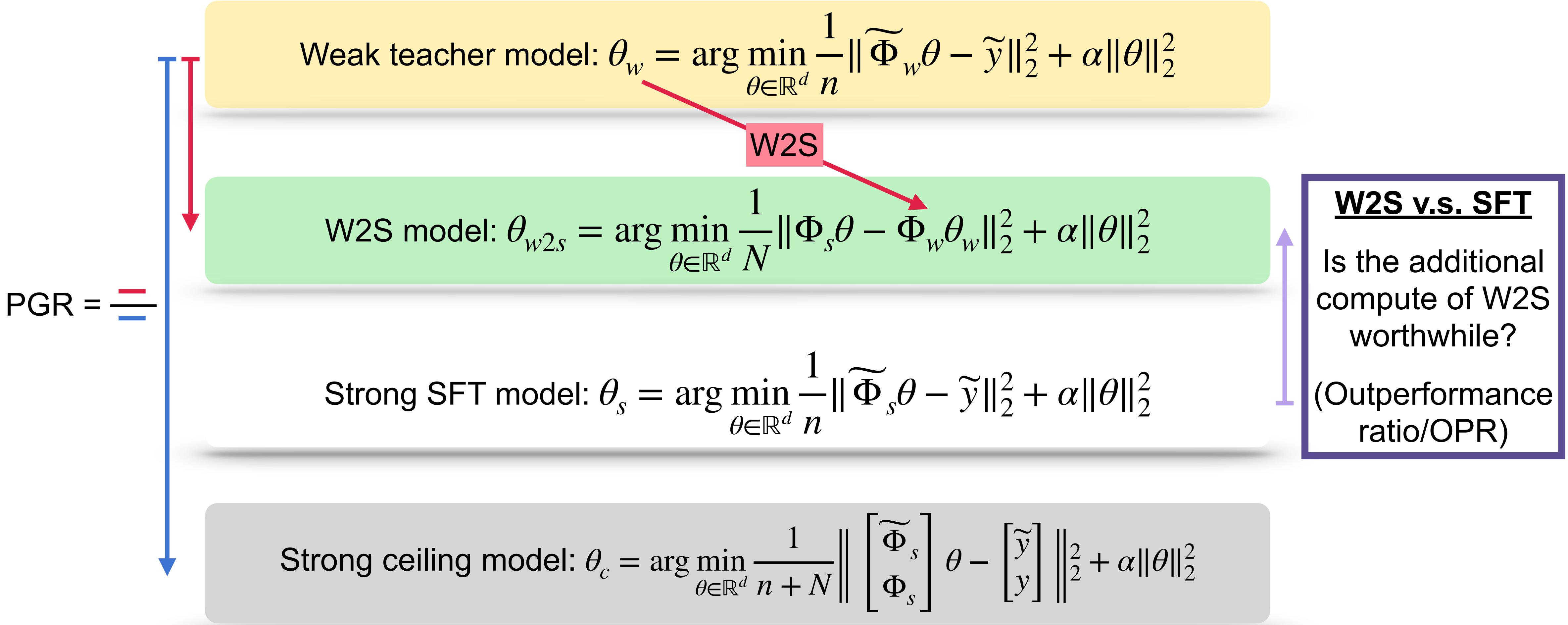
Representation similarity — **correlation dimension**: Consider spectral decompositions

$$\Sigma_s = \begin{matrix} V_s & \Lambda_s & V_s^\top \\ d \times d_s & d_s \times d_s & \end{matrix} \quad \text{and} \quad \Sigma_w = \begin{matrix} V_w & \Lambda_w & V_w^\top \\ d \times d_w & d_w \times d_w & \end{matrix}.$$

The **correlation dimension** of  $(\phi_s, \phi_w)$  is  $d_{s \wedge w} = \|V_s^\top V_w\|_F^2$  s.t.  $0 \leq d_{s \wedge w} \leq \min\{d_s, d_w\}$ .

# W2S finetuning as ridgeless regression

Ridgeless regression: with all  $\alpha \rightarrow 0$



# W2S generalization error: ridgeless regression

With randomness in  $f$  from training data:

$$\text{ER}(f) = \text{Var}(f) + \text{Bias}(f) \text{ where}$$

$$\text{Var}(f) = \mathbb{E}_{X,y} \left[ \frac{1}{N} \|f(X) - \mathbb{E}_{y|X}[f(X)]\|_2^2 \right]$$

$$\text{Bias}(f) = \mathbb{E}_X \left[ \frac{1}{N} \|\mathbb{E}_{y|X}[f(X)] - f_*(X)\|_2^2 \right]$$

Proposition [DLLLL25].

$$\text{Var}(f_w) = \frac{\sigma^2 d_w}{n - d_w - 1}, \quad \text{Bias}(f_w) \lesssim \rho_w$$

$$\text{Var}(f_s) = \frac{\sigma^2 d_s}{n - d_s - 1}, \quad \text{Bias}(f_s) \lesssim \rho_s$$

$$\text{Var}(f_c) = \sigma^2 \frac{d_s}{n + N}, \quad \text{Bias}(f_c) \leq \rho_s$$

Theorem [DLLLL25]. Assume  $\phi_s(x)$  is zero-mean subgaussian and  $\phi_w(x) \sim \mathcal{N}(0_d, \Sigma_w)$  (can be relaxed to subgaussian), for  $n > d_w + 1$ :

$$\text{Var}(f_{w2s}) = \frac{\sigma^2}{n - d_w - 1} \left( d_{s \wedge w} + \frac{d_s}{N} (d_w - d_{s \wedge w}) \right)$$

$$\text{Bias}(f_{w2s}) \leq \text{Bias}(f_w) + \rho_s \leq O(\rho_w) + \rho_s$$

$$\mathcal{V}_s = \text{Range}(\Sigma_s), \quad \mathcal{V}_w = \text{Range}(\Sigma_w)$$

$$\text{Var}(f_{w2s}) \asymp \boxed{\frac{d_{s \wedge w}}{n}} + \boxed{\frac{d_s}{N}} \boxed{\frac{d_w - d_{s \wedge w}}{n}}$$

Var. in  $\mathcal{V}_w \cap \mathcal{V}_s$     W2S    Var. in  $\mathcal{V}_w \setminus \mathcal{V}_s$

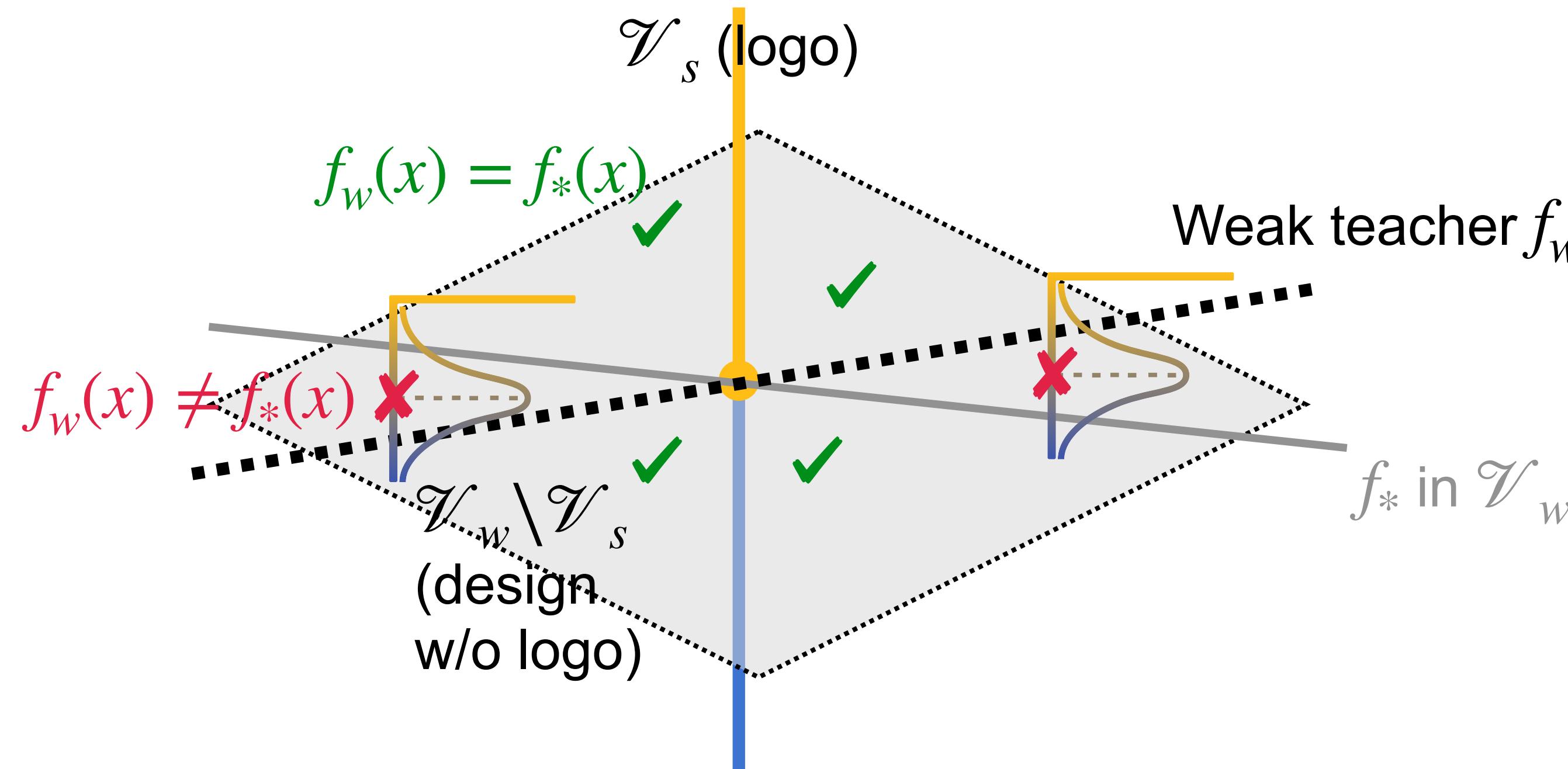
# Intuition: How does variance reduction in W2S happen?

$$\mathcal{V}_s = \text{Range}(\Sigma_s), \mathcal{V}_w = \text{Range}(\Sigma_w)$$

$$\text{Var}(f_{w2s}) \asymp \frac{d_{s \wedge w}}{n} + \frac{d_s}{N} \frac{d_w - d_{s \wedge w}}{n}$$

Var. in  $\mathcal{V}_w \cap \mathcal{V}_s$     W2S    Var. in  $\mathcal{V}_w \setminus \mathcal{V}_s$

Task: Determine the make of a car



Pseudolabel error in  $\mathcal{V}_w \setminus \mathcal{V}_s$  can be viewed as **independent label noise** w.r.t. the orthogonal strong features  $\mathcal{V}_s$ , variance from which reduces proportionally to  $d_s/N$ .

# Suitable regularization is essential for W2S: ridge regression

- Positive-definite covariances:  $\Sigma_w, \Sigma_s, \Sigma_* \succ 0$
- $f_*(x) = \phi_*(x)^\top \theta_*$ ,  $\theta_* \in \mathbb{R}^d$ ,  $\mathbb{E}[\phi_*(x)\phi_*(x)^\top] = \Sigma_*$
- Normalized features:  $\|\Sigma_w\|_2 \asymp \|\Sigma_s\|_2 \asymp \|\Sigma_*\|_2 \asymp 1$
- Intrinsic dimensions:  $\text{tr}(\Sigma_w) \lesssim d_w$ ,  $\text{tr}(\Sigma_s) \lesssim d_s$

Choose some suitable  $\alpha_w, \alpha_{w2s} > 0$  s.t.

$$\theta_w = \arg \min_{\theta \in \mathbb{R}^d} \frac{1}{n} \|\widetilde{\Phi}_w \theta - \tilde{y}\|_2^2 + \alpha_w \|\theta\|_2^2$$

$$\theta_{w2s} = \arg \min_{\theta \in \mathbb{R}^d} \frac{1}{N} \|\Phi_s \theta - \Phi_w \theta_w\|_2^2 + \alpha_{w2s} \|\theta\|_2^2$$

Theorem [DLLL25]. Let  $\varrho_w = \|\Sigma_w^{-1/2} \Sigma_*^{1/2} \theta_*\|_2^2$ ,  $\varrho_s = \|\Sigma_s^{-1/2} \Sigma_*^{1/2} \theta_*\|_2^2$ . Set  $\alpha_w = \frac{\sigma^2 \text{tr}(\Sigma_s \Sigma_w)}{4nN} \frac{\varrho_s}{\varrho_w^2}$  and  $\alpha_{w2s} = \frac{\sigma^2 \text{tr}(\Sigma_s \Sigma_w)}{4nN} \frac{\varrho_w}{\varrho_s^2}$ , when  $N \geq \frac{\text{tr}(\Sigma_s) \text{tr}(\Sigma_w)}{\text{tr}(\Sigma_s \Sigma_w)}$ ,

$$\text{ER}(f_{w2s}) \leq 3 \left( \frac{3\sigma^2}{4nN} \varrho_s \varrho_w \text{tr}(\Sigma_s \Sigma_w) \right)^{1/3}.$$

- **Multiplicative sample complexity:**

$$nN \asymp \sigma^2 \text{tr}(\Sigma_s \Sigma_w) \varrho_s \varrho_w$$

- **Weak-strong similarity (“correlation dimension  $d_{s \wedge w}$ ”):**

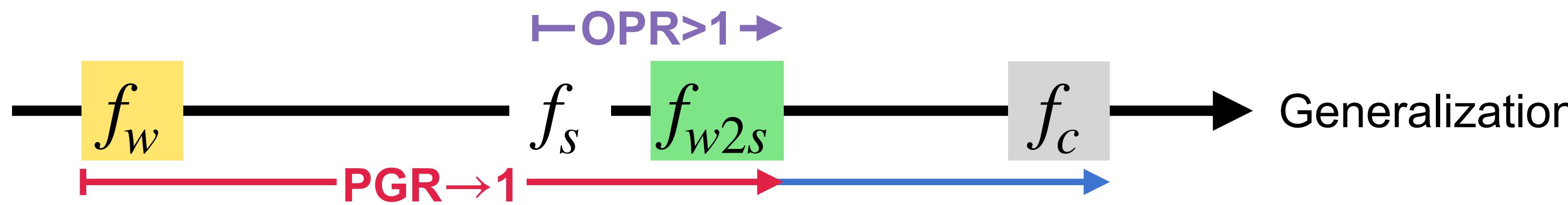
$$\text{tr}(\Sigma_s \Sigma_w) \lesssim \min\{\text{tr}(\Sigma_s), \text{tr}(\Sigma_w)\}$$

- **Coverage (“FT approximation error”):**  $\varrho_w, \varrho_s$  are small if the dominating eigenspaces of  $\Sigma_w, \Sigma_s$  cover that of  $\Sigma_*$

# Larger discrepancy (lower $d_{s \wedge w}$ ) $\rightarrow$ better W2S

**Performance gap recovery:**  $PGR = \frac{ER(f_w) - ER(f_{w2s})}{ER(f_w) - ER(f_c)}$

**Outperformance ratio:**  $OPR = \frac{ER(f_s)}{ER(f_{w2s})}$

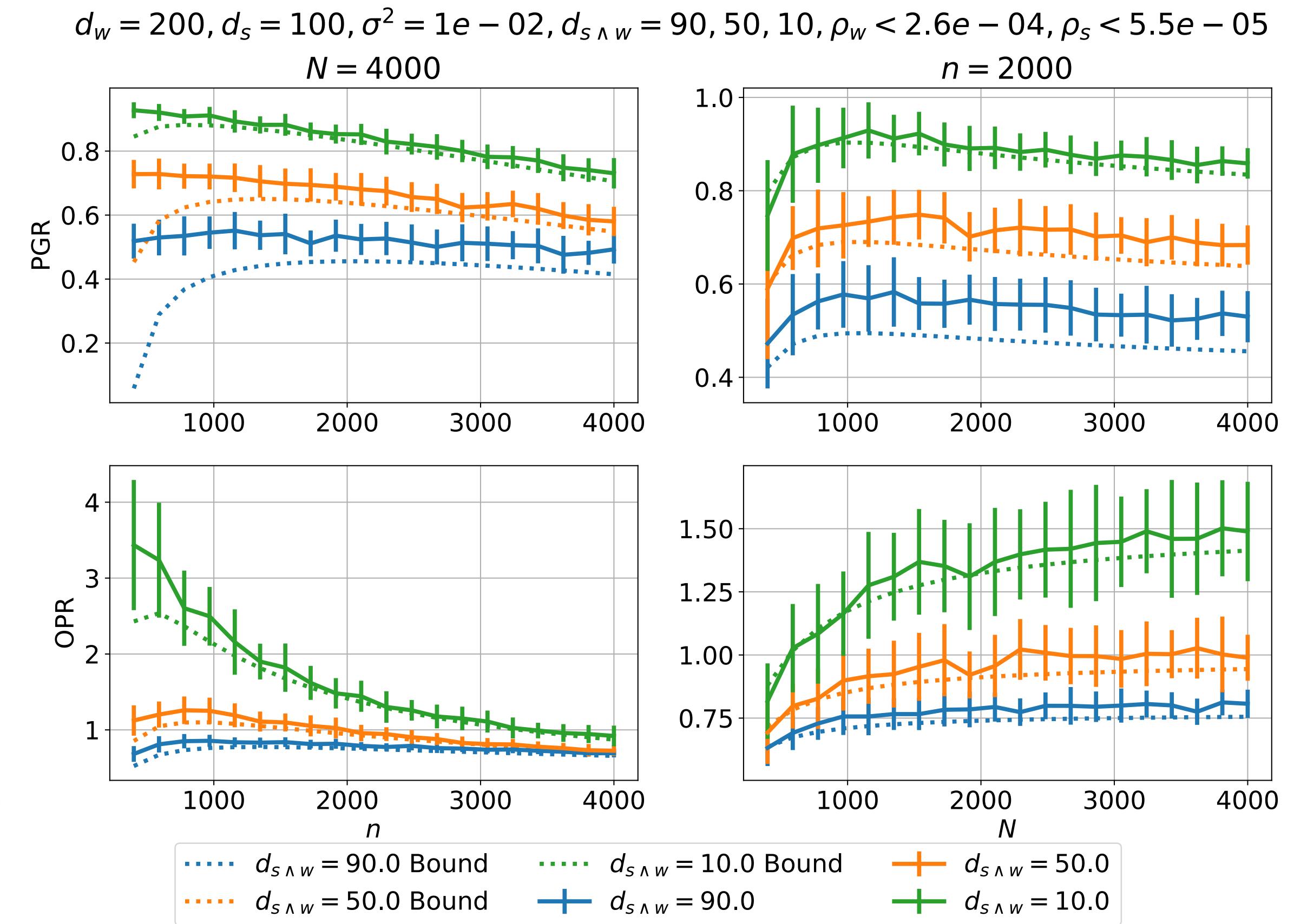
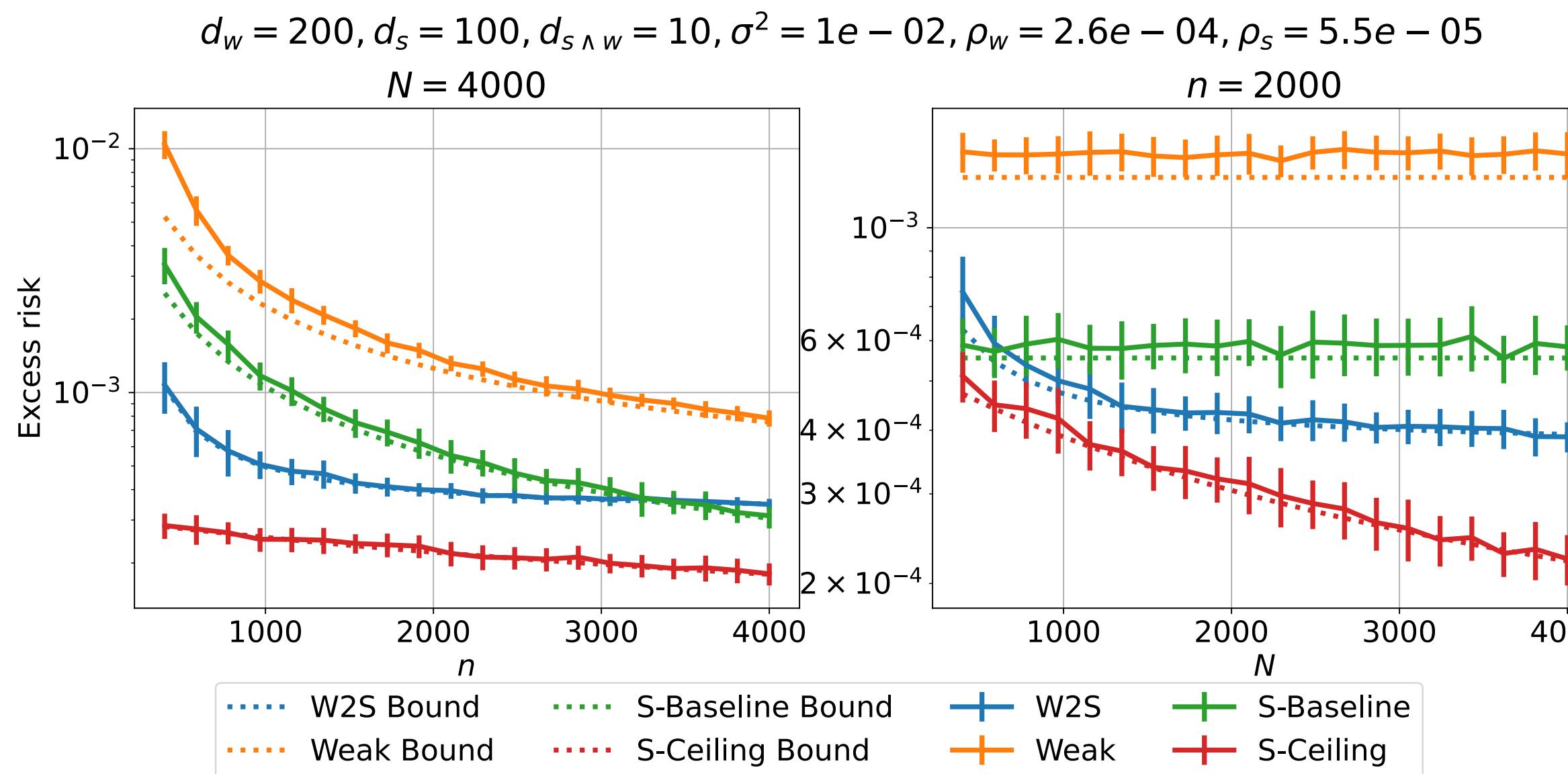


With negligible FT approximation error  $(\rho_w + \rho_s)/\sigma^2 \rightarrow 0$ , when  $n \gtrsim d_w$  and  $N \gtrsim d_s(d_w/d_{s \wedge w} - 1)$ , we have

$$PGR \geq 1 - O(d_{s \wedge w}/d_w) \quad \text{and} \quad OPR \geq \Omega(d_s/d_{s \wedge w})$$

# Synthetic experiments

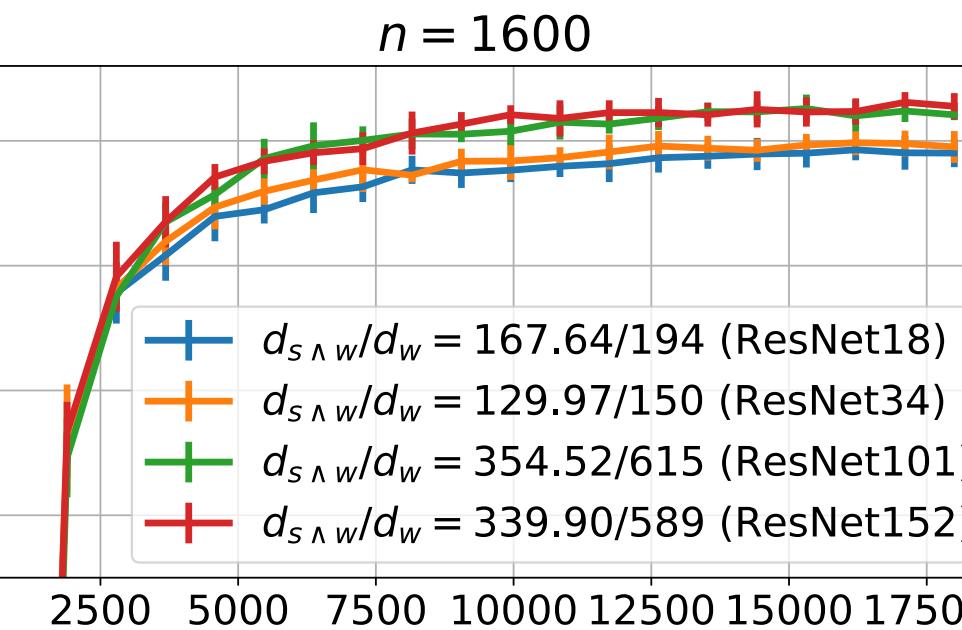
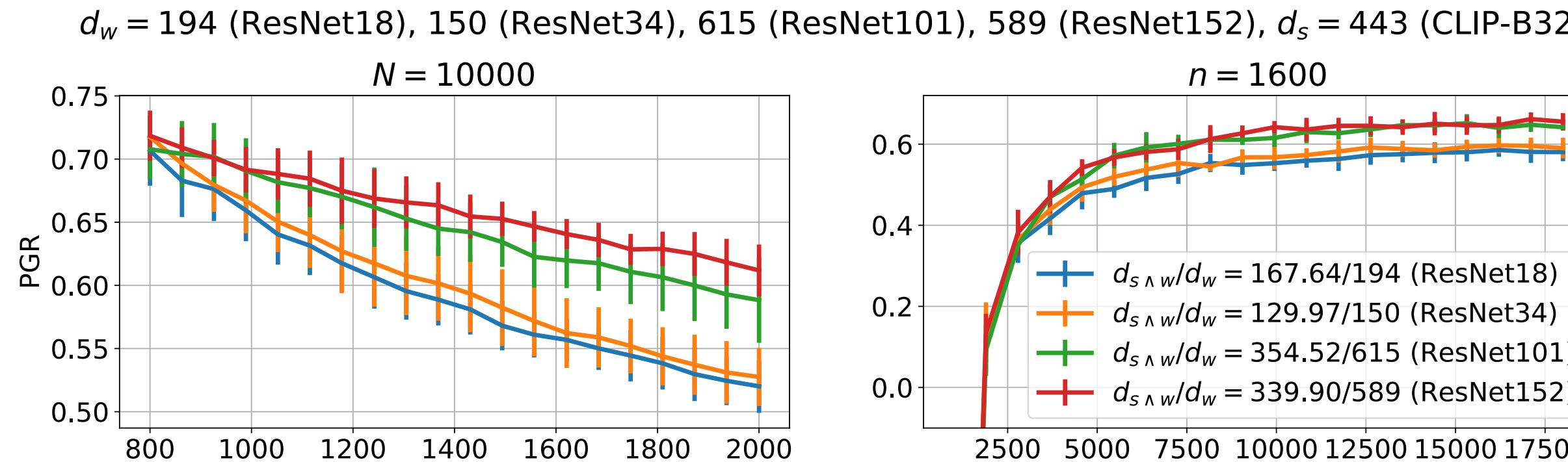
- High-dimensional Gaussian features:  $d = 20000$
- $f_*(x) = x^\top \Lambda_*^{1/2} \theta_*$  where  $\Lambda_* = \text{diag}(\lambda_1^*, \dots, \lambda_d^*)$
- $\lambda_i^* = i^{-1}$  for  $1 \leq i \leq 300$ ,  $\lambda_i^* = 0$  for  $i > 300$



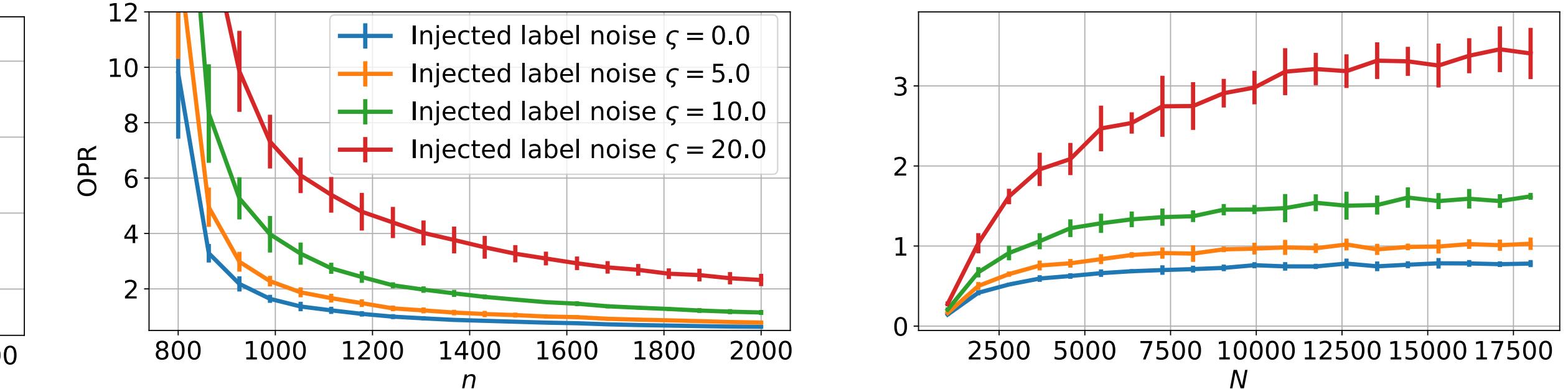
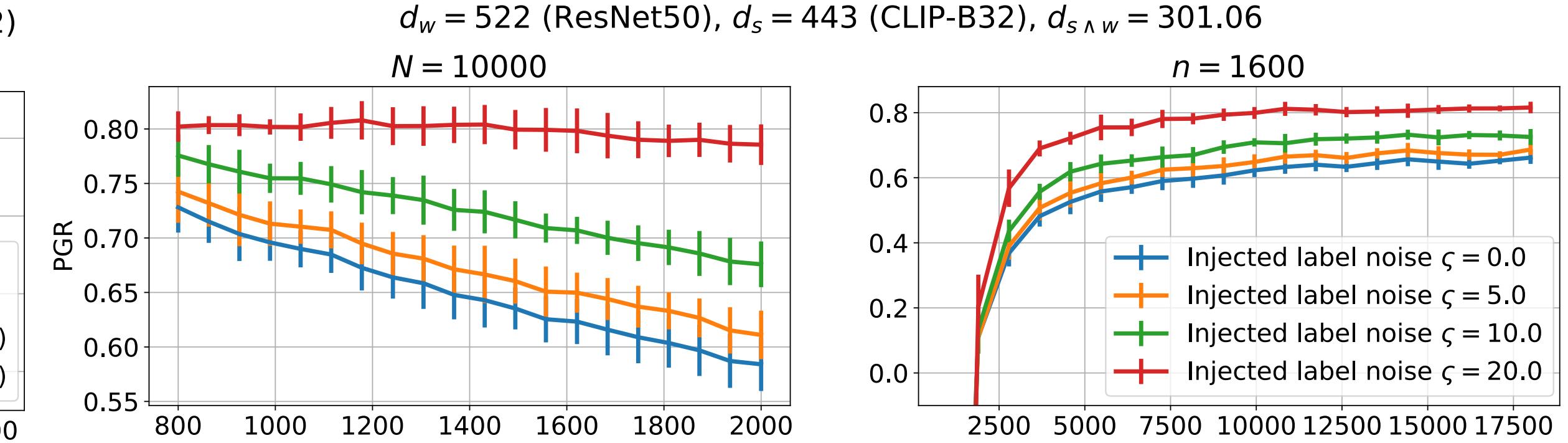
- Our bounds provide reasonably tight characterization for the generalization error, PGR, and OPR.
- W2S is more beneficial with limited label data  $n$  — PGR and OPR decrease as  $n$  increases!

# UTKFace regression

Lower  $d_{s \wedge w}/d_w \rightarrow$  better W2S



Larger variance  $\rightarrow$  more pronounced W2S



- UTKFace: age prediction (0-116) based on images, i.e., image regression.
- Lower  $d_{s \wedge w}/d_w$  (larger discrepancy between  $\phi_w, \phi_s$ ) brings higher PGR and OPR.
- Benefit of W2S is more pronounced on problems with larger variance.

# Takeaway: teacher-student discrepancy → better W2S

How does W2S happen on easy tasks where weak and strong models both have low approximation errors?

Through lens of low intrinsic dimension:

- Representation **efficiency**:  $\text{rank}(\Sigma_s) = d_s, \text{rank}(\Sigma_w) = d_w \ll d$
- Representation **similarity**: correlation dimension  $d_{s \wedge w} = \|V_s^\top V_w\|_F^2 \in [0, \min\{d_s, d_w\}]$

$$\text{Var}(f_{w2s}) \asymp \frac{d_{s \wedge w}}{n} + \frac{d_s}{N} \frac{d_w - d_{s \wedge w}}{n}$$

Var. in  $\mathcal{V}_w \cap \mathcal{V}_s$     W2S    Var. in  $\mathcal{V}_w \setminus \mathcal{V}_s$

With negligible FT approximation error, when  $n \gtrsim d_w$  and  $N \gtrsim d_s(d_w/d_{s \wedge w} - 1)$ ,

$$\text{PGR} \geq 1 - O(d_{s \wedge w}/d_w) \quad \text{and} \quad \text{OPR} \geq \Omega(d_s/d_{s \wedge w})$$

# Thank you! Happy to take any questions



Discrepancies are Virtue: Weak-to-Strong Generalization through Lens of Intrinsic Dimension.  
Yijun Dong, Yicheng Li, Yunai Li, Jason D. Lee, and Qi Lei. ICML 2025.

## References

Aghajanyan, Armen, Luke Zettlemoyer, and Sonal Gupta. "Intrinsic dimensionality explains the effectiveness of language model fine-tuning." *arXiv preprint arXiv:2012.13255* (2020).

Burns, Collin, Pavel Izmailov, Jan Hendrik Kirchner, Bowen Baker, Leo Gao, Leopold Aschenbrenner, Yining Chen et al. "Weak-to-strong generalization: Eliciting strong capabilities with weak supervision." *arXiv preprint arXiv:2312.09390* (2023).

Ildiz, M. Emrullah, Halil Alperen Gozeten, Ege Onur Taga, Marco Mondelli, and Samet Oymak. "High-dimensional analysis of knowledge distillation: Weak-to-strong generalization and scaling laws." *arXiv preprint arXiv:2410.18837* (2024).

Wu, David X., and Anant Sahai. "Provable weak-to-strong generalization via benign overfitting." *arXiv preprint arXiv:2410.04638* (2024).