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

Yijun Dong

Courant Institute of Mathematical Sciences, New York University

John Hopkins University Postdoc Seminar, Apr 2, 2025

NYU | COURANT INSTITUTE OF MATHEMATICAL SCIENCES

Joint work with

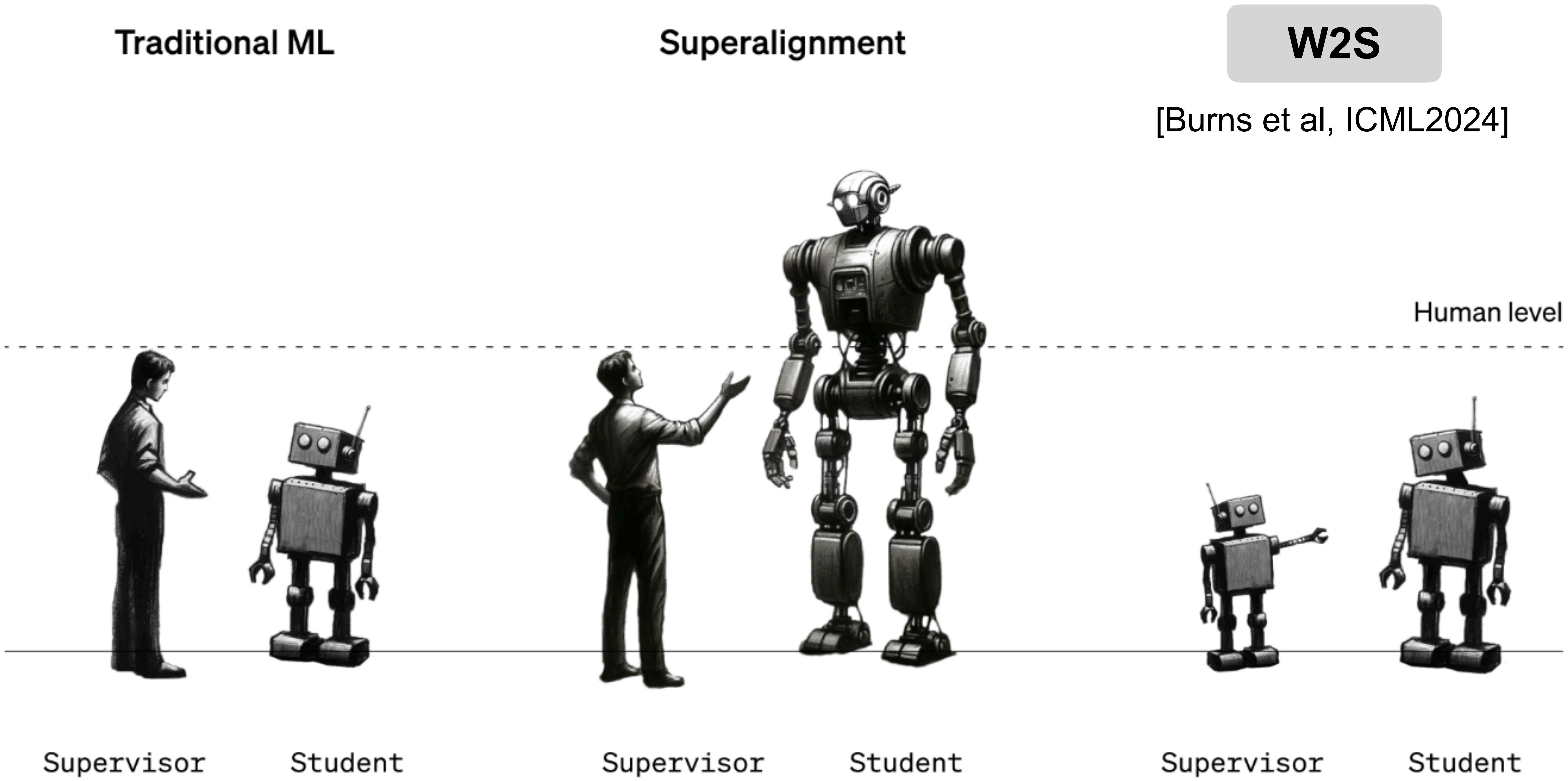
Yicheng Li
NYU

Yunai Li
SJTU

Jason D. Lee
Princeton

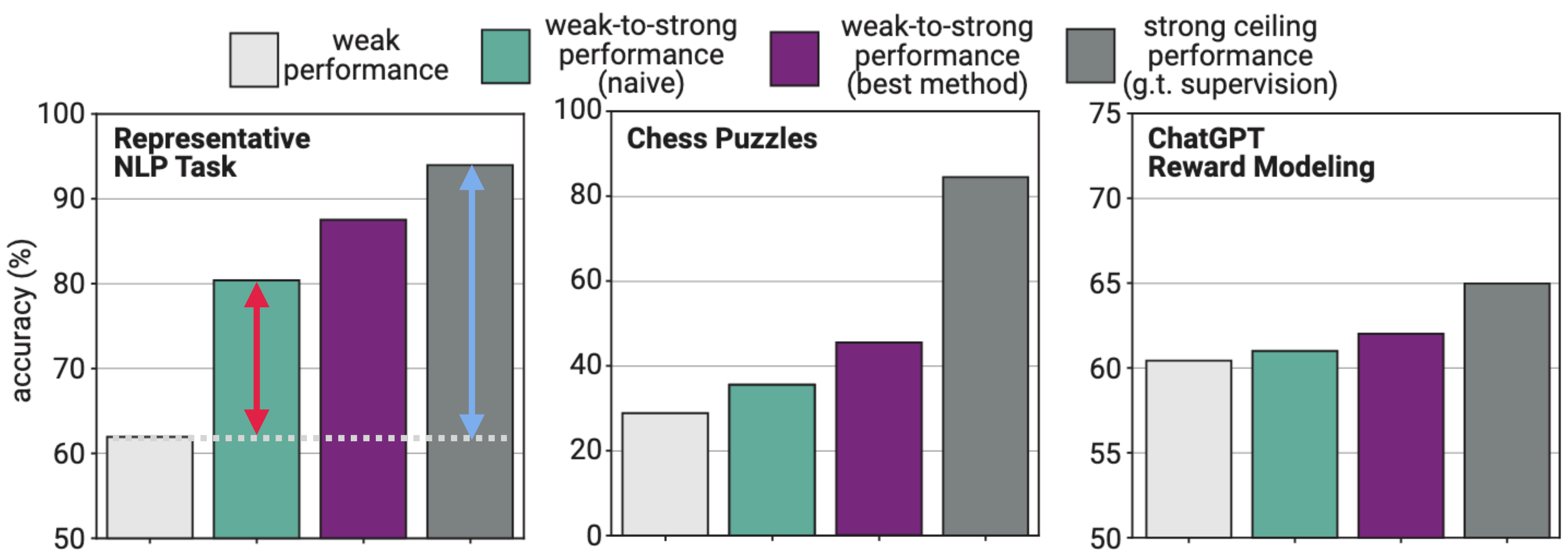
Qi Lei
NYU

Superalignment → weak-to-strong (W2S) generalization



When and how W2S happens?

Better W2S generalization on easier tasks



[Burns et al, ICML2024]

- Difficulty: NLP tasks < Chess puzzles < ChatGPT reward modeling
- Approximation error = error of the model trained over the population
- Better W2S ⇔ **performance gap recovery** closer to 1

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

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

Finetunable parameter of intrinsic dimension $d < D$

$$\boxed{\theta^D} = \boxed{\theta_0^D} + \boxed{\Gamma}\,\boxed{\theta^d}$$

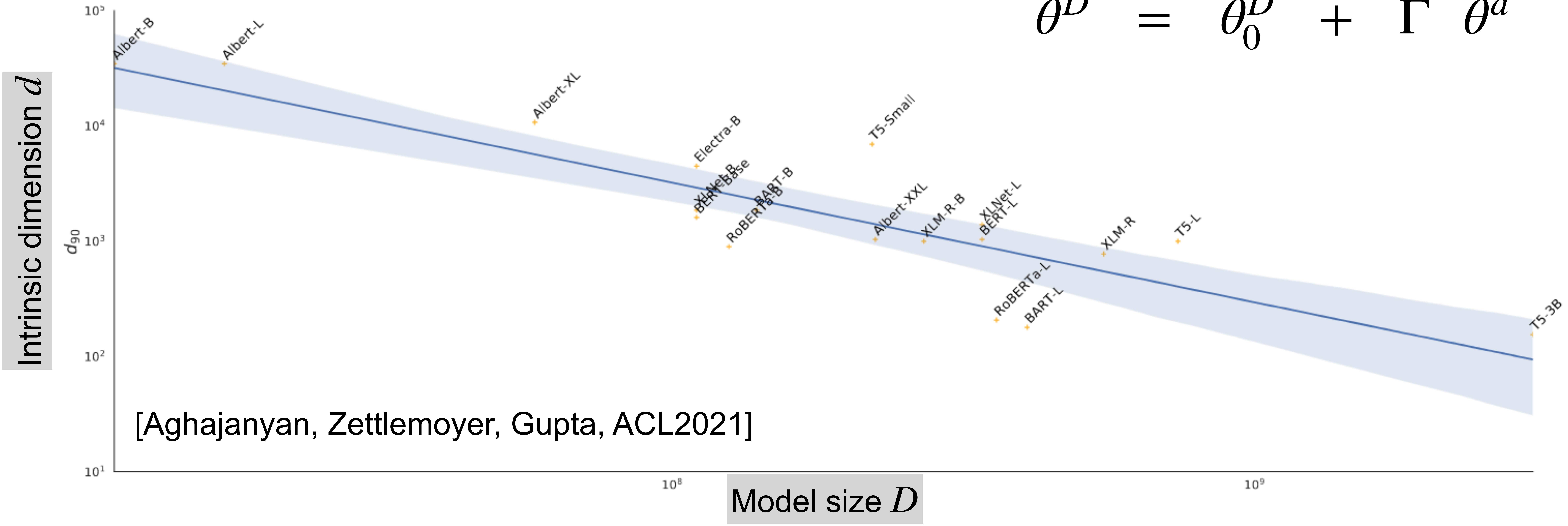
Model parameter of high dimension D

$D \times d$ random projection

Low intrinsic dimension of finetuning

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

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



[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 $f_* : \mathcal{X} \to \mathbb{R}$ given access to two datasets:
  - **Labeled** (small) dataset: $\widetilde{X} \in \mathcal{X}^n$ with noisy labels $\widetilde{y} \in \mathbb{R}^n$
  - **Unlabeled** (large) dataset: $X \in \mathcal{X}^N$ with unknown labels $y \in \mathbb{R}^N$

Finetuning (FT) ≈ linear probing on 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 $\phi_w : \mathcal{X} \to \mathbb{R}^d$ produces $\boxed{\widetilde{\Phi}_w = \phi_w(\widetilde{X}) \in \mathbb{R}^{n\times d},\ \Phi_w = \phi_w(X) \in \mathbb{R}^{N\times d}}$
- Strong model $\phi_s : \mathcal{X} \to \mathbb{R}^d$ produces $\boxed{\widetilde{\Phi}_s = \phi_s(\widetilde{X}) \in \mathbb{R}^{n\times d},\ \Phi_s = \phi_s(X) \in \mathbb{R}^{N\times d}}$

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

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

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

Weak v.s. strong: model capacity + similarity

Representation <u>efficiency</u> — **low intrinsic dimensions**:

$$\mathrm{rank}(\Sigma_w) = d_w \ll d \qquad \mathrm{rank}(\Sigma_s) = d_s \ll d$$

Representation <u>accuracy</u> — **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 <u>similarity</u> — **correlation dimension**: Consider spectral decompositions

$$\Sigma_s = \underset{d \times d_s}{V_s} \; \underset{d_s \times d_s}{\Sigma_s} \; V_s^\top \quad \text{and} \quad \Sigma_w = \underset{d \times d_w}{V_w} \; \underset{d_w \times d_w}{\Sigma_w} \; V_w^\top.$$

The correlation dimension of (ϕ_s, ϕ_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 $\alpha \to 0$

Weak teacher model: $\theta_w = \arg\min_{\theta \in \mathbb{R}^d} \frac{1}{n} \|\widetilde{\Phi}_w \theta - \widetilde{y}\|_2^2 + \alpha \|\theta\|_2^2$

W2S

W2S model: $\theta_{w2s} = \arg\min_{\theta \in \mathbb{R}^d} \frac{1}{N} \|\Phi_s \theta - \Phi_w \theta_w\|_2^2 + \alpha \|\theta\|_2^2$

PGR = $\frac{\text{—}}{\text{—}}$

Strong SFT model: $\theta_s = \arg\min_{\theta \in \mathbb{R}^d} \frac{1}{n} \|\widetilde{\Phi}_s \theta - \widetilde{y}\|_2^2 + \alpha \|\theta\|_2^2$

Strong ceiling model: $\theta_c = \arg\min_{\theta \in \mathbb{R}^d} \frac{1}{n+N} \left\| \begin{bmatrix} \widetilde{\Phi}_s \\ \Phi_s \end{bmatrix} \theta - \begin{bmatrix} \widetilde{y} \\ y \end{bmatrix} \right\|_2^2 + \alpha \|\theta\|_2^2$

W2S v.s. SFT

Is the additional compute of W2S worthwhile?

W2S generalization error: ridgeless regression

With randomness in f from training data:

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

$$\mathrm{Var}(f) = \mathbb{E}_x[\mathbb{E}_f[(f(x) - \mathbb{E}_f[f(x)])^2]]$$

$$\mathrm{Bias}(f) = \mathbb{E}_x[(\mathbb{E}_f[f(x)] - f_*(x))^2]$$

Theorem [**D**LLLL25]. 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 $n > d_w + 1$:

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

$$\mathrm{Bias}(f_{w2s}) \le \rho_w + \rho_s$$

Proposition [**D**LLLL25].

$$\mathrm{Var}(f_w) = \sigma^2\frac{d_w}{n}, \quad \mathrm{Bias}(f_w) \le \rho_w$$

$$\mathrm{Var}(f_s) = \sigma^2\frac{d_s}{n}, \quad \mathrm{Bias}(f_s) \le \rho_s$$

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

$$\mathscr{V}_s = \mathrm{Range}(\Sigma_s),\ \mathscr{V}_w = \mathrm{Range}(\Sigma_w)$$

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

Variance in $\mathscr{V}_w \cap \mathscr{V}_s$

W2S

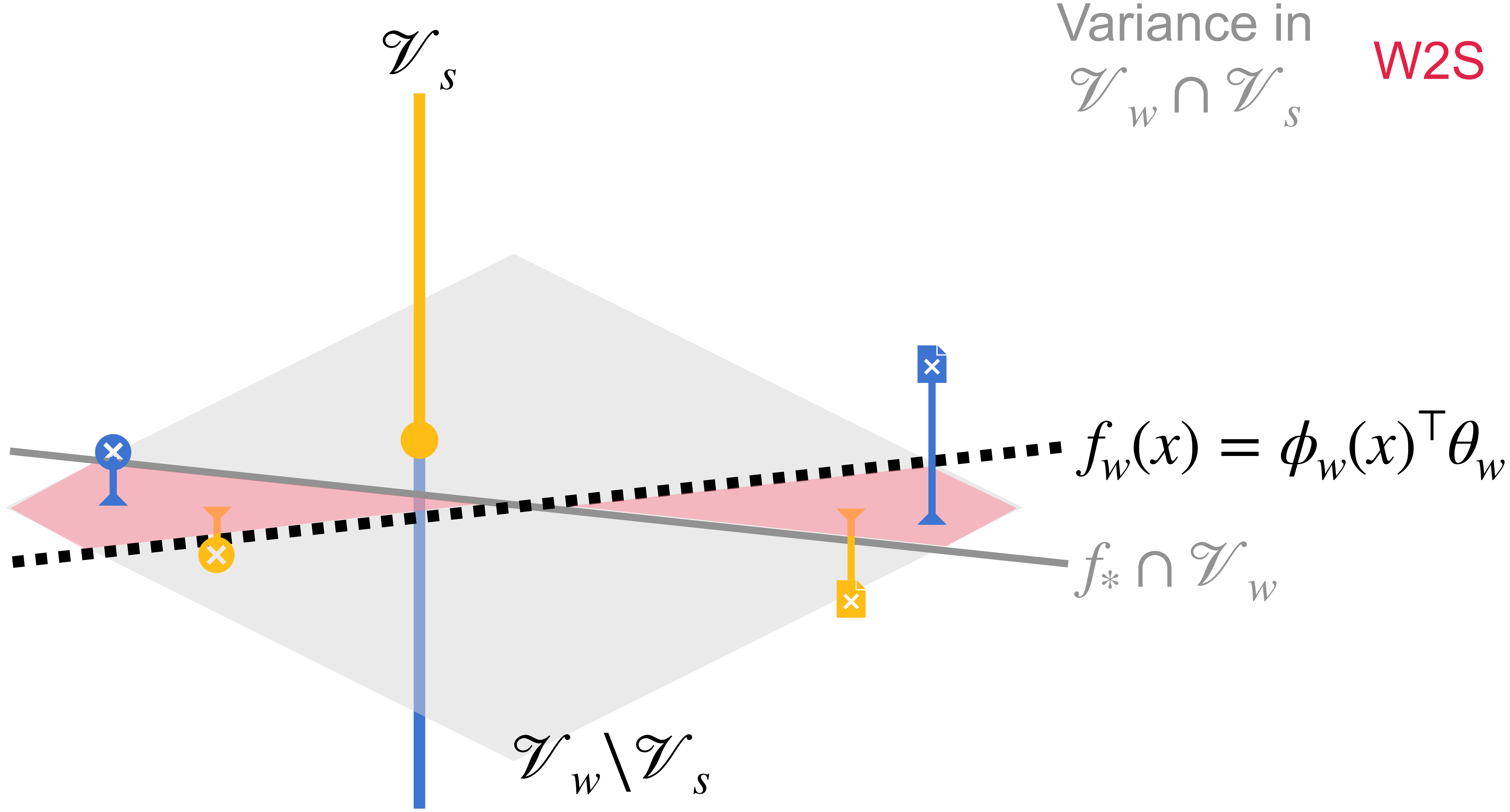
Variance in $\mathscr{V}_w \backslash \mathscr{V}_s$

Intuition: how does variance reduction in W2S happen?

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

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

Variance in $\mathcal{V}_w \cap \mathcal{V}_s$ | W2S | Variance in $\mathcal{V}_w \backslash \mathcal{V}_s$



Psuedolabel error in $\mathcal{V}_w \backslash \mathcal{V}_s$ 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.

W2S generalization error: ridge regression

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

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

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

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

Theorem [**D**LLLL25]. Let $\varrho_w = \|\Sigma_w^{-1/2}\Sigma_*^{1/2}\theta_*\|_2^2$, $\varrho_s = \|\Sigma_s^{-1/2}\Sigma_*^{1/2}\theta_*\|_2^2$. For ridge parameters $\alpha_w = \frac{\sigma^2 \mathrm{tr}(\Sigma_s\Sigma_w)}{4nN}\frac{\varrho_s}{\varrho_w^2}$ and $\alpha_{w2s} = \frac{\sigma^2 \mathrm{tr}(\Sigma_s\Sigma_w)}{4nN}\frac{\varrho_w}{\varrho_s^2}$,

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

- **<u>Multiplicative</u> sample complexity:**

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

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

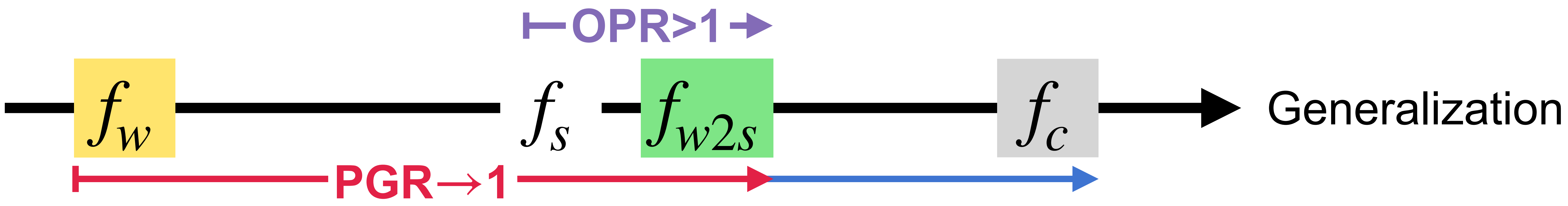
$$\mathrm{tr}(\Sigma_s\Sigma_w) \lesssim \min\{\mathrm{tr}(\Sigma_s), \mathrm{tr}(\Sigma_w)\}$$

- **Coverage ("FT approximation error")**: ϱ_w, ϱ_s are small if the dominating eigenspaces of Σ_w, Σ_s cover that of Σ_*

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

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

Outperformance ratio: $\mathrm{OPR} = \dfrac{\mathrm{ER}(f_s)}{\mathrm{ER}(f_{w2s})}$

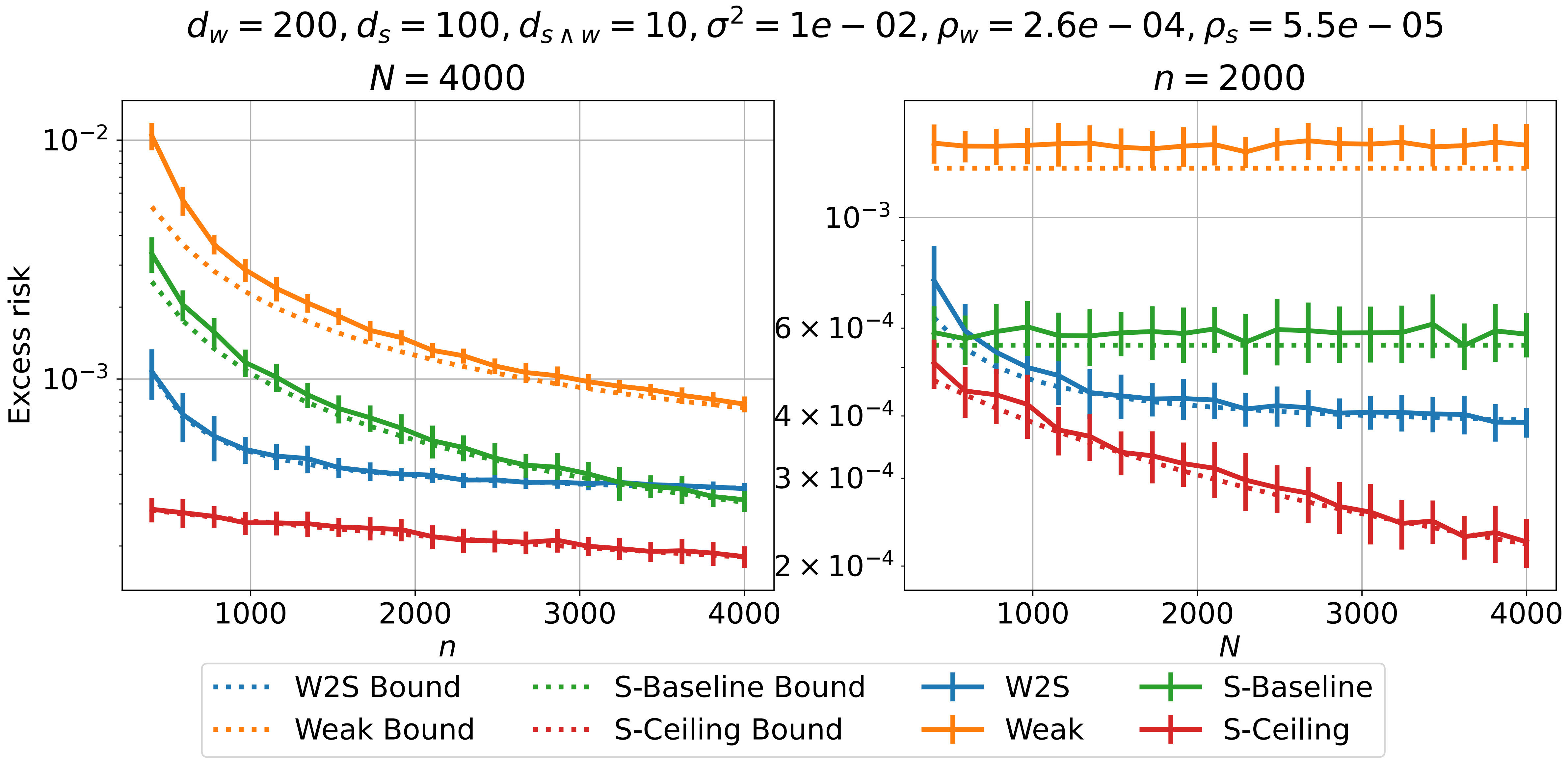


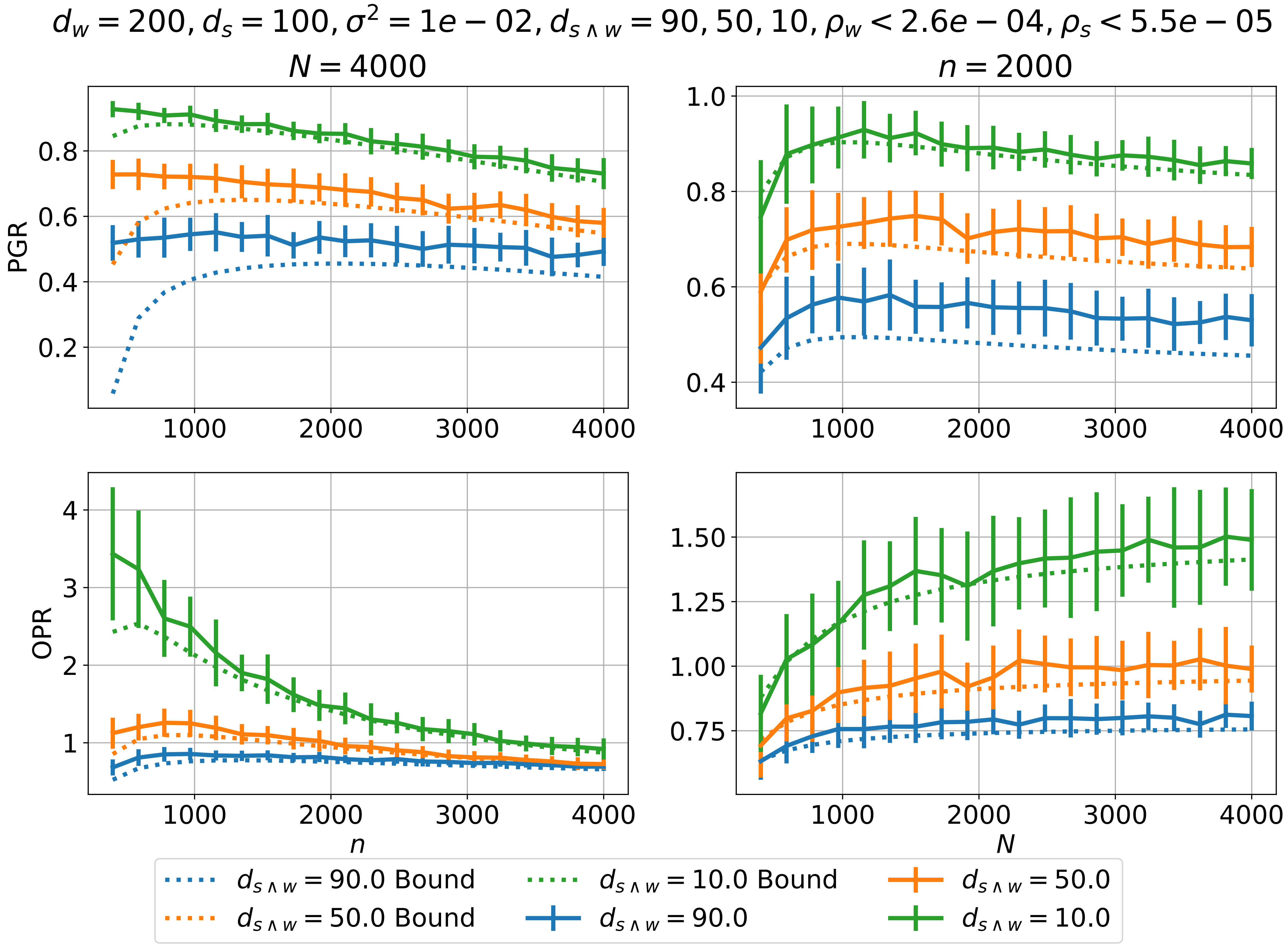
With negligible FT approximation error $(\rho_w + \rho_s)/\sigma^2 \rightarrow 0$, when $n \gtrsim d_w$ and $N \gtrsim d_s(d_w/d_{s \wedge w} - 1)$, we have

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

Synthetic experiments

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



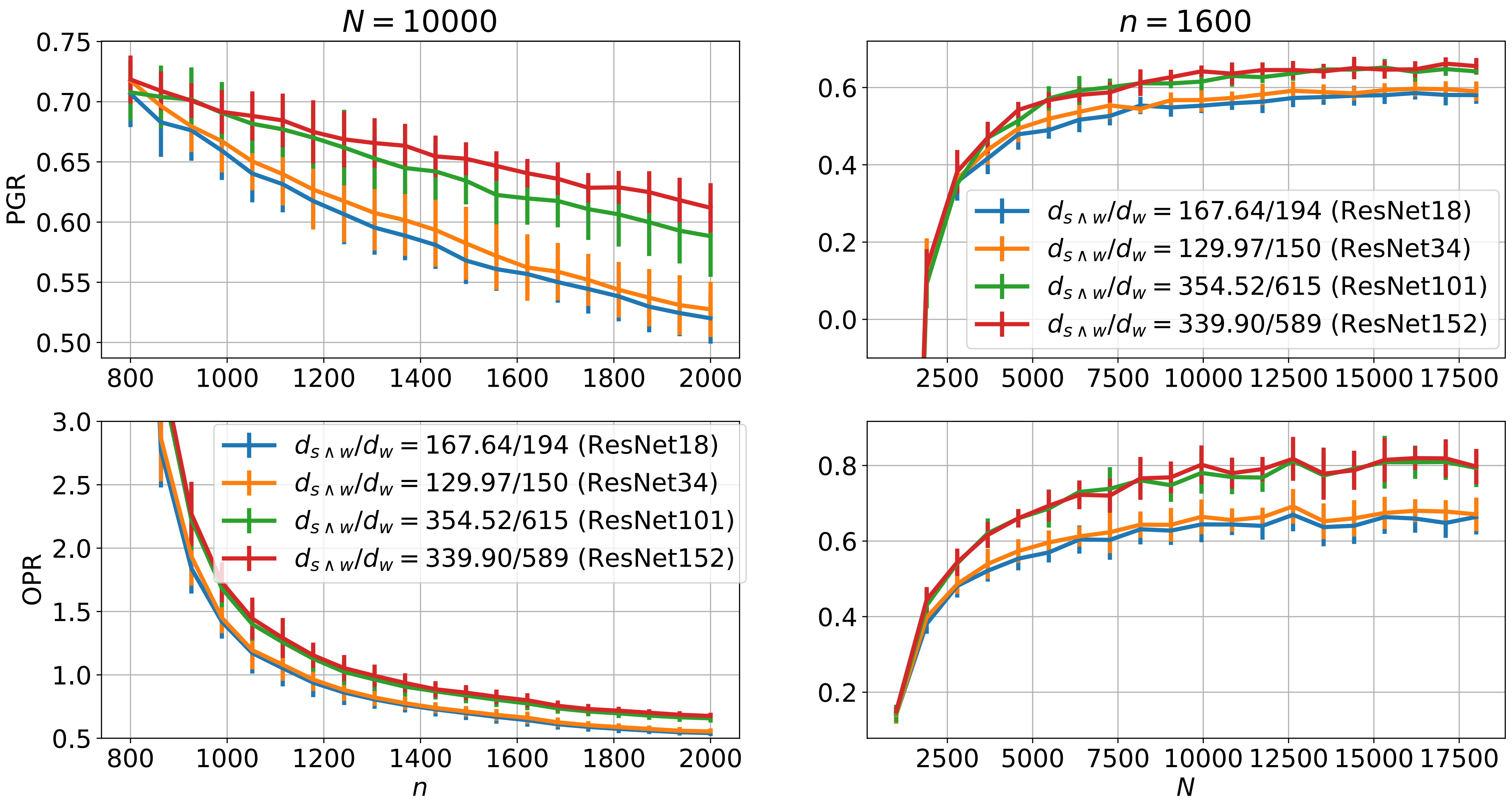


- Our bounds provide reasonably tight characterization for the generalization error, PGR, and OPR.
- W2S is more beneficial with limited label data n — PGR and OPR decrease as n increases!

UTKFace regression

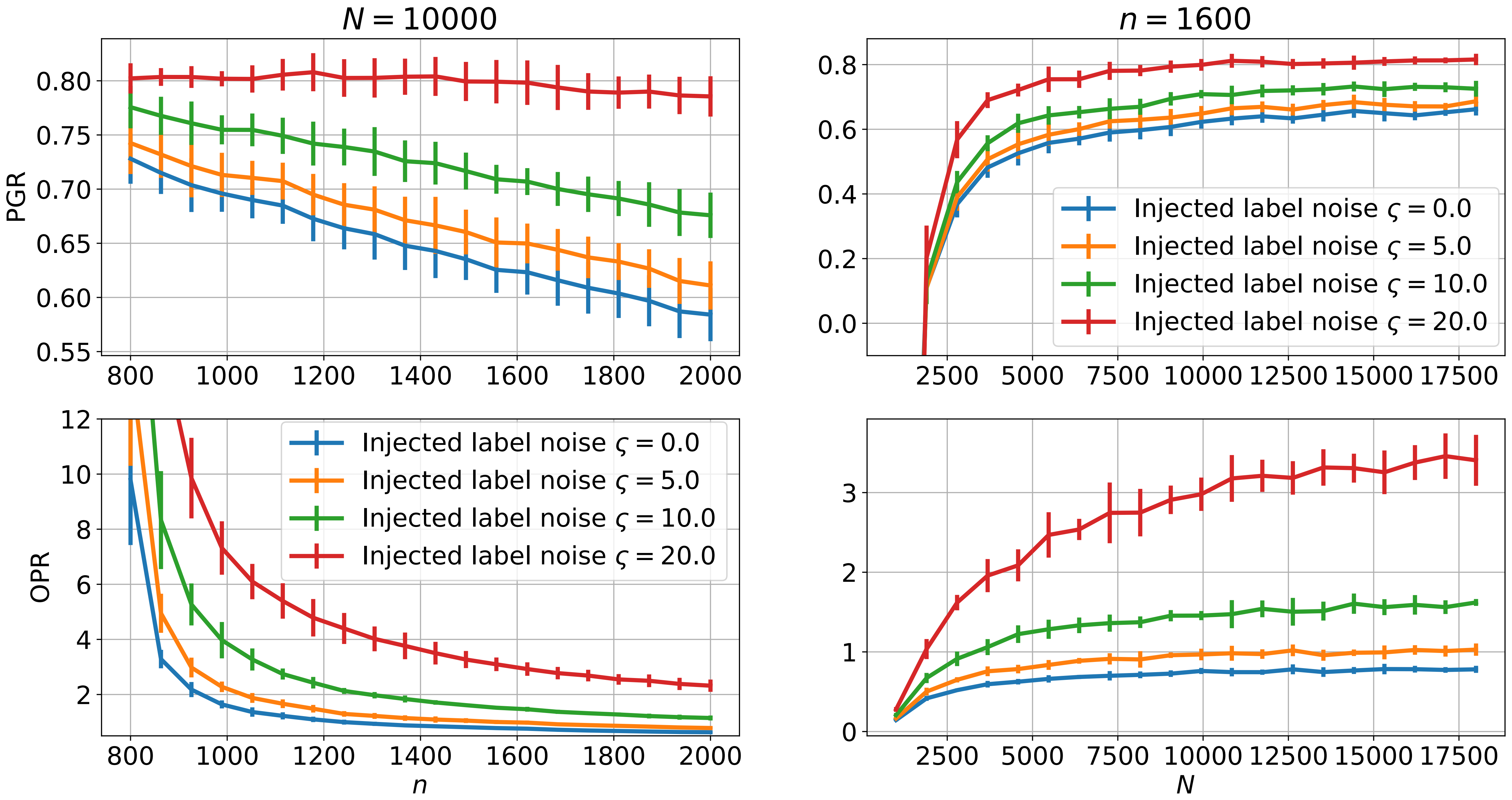
Lower $d_{s\wedge w}/d_w$ → better W2S

$d_w = 194$ (ResNet18), 150 (ResNet34), 615 (ResNet101), 589 (ResNet152), $d_s = 443$ (CLIP-B32)



Larger variance → more pronounced W2S

$d_w = 522$ (ResNet50), $d_s = 443$ (CLIP-B32), $d_{s\wedge w} = 301.06$



- UTKFace: age prediction (0-116) based on images, i.e., image regression.
- Lower $d_{s\wedge w}/d_w$ (larger discrepancy between ϕ_w, ϕ_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**: $\mathrm{rank}(\Sigma_s) = d_s,\ \mathrm{rank}(\Sigma_w) = d_w \ll d$
- Representation **similarity**: correlation dimension $d_{s\wedge w} = \|V_s^\top V_w\|_F^2 \in [0, \min\{d_s, d_w\}]$

$$\mathrm{Var}(f_{w2s}) \asymp \frac{d_{s\wedge w}}{n} + \underbrace{\frac{d_s}{N}}_{\text{W2S}} \frac{d_w - d_{s\wedge w}}{n}$$

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

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

Thank you! Happy to take any questions

Dong, Yijun, Yicheng Li, Yunai Li, Jason D. Lee, and Qi Lei. "Discrepancies are Virtue: Weak-to-Strong Generalization through Lens of Intrinsic Dimension." *arXiv preprint* arXiv:2502.05075 (2025).

References

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).

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