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

Yijun Dong[1], Yicheng Li[1], Yunai Li[2], Jason D. Lee[3], Qi Lei[1]
[1]New York University, [2]Shanghai Jiaotong University, [3]Princeton University

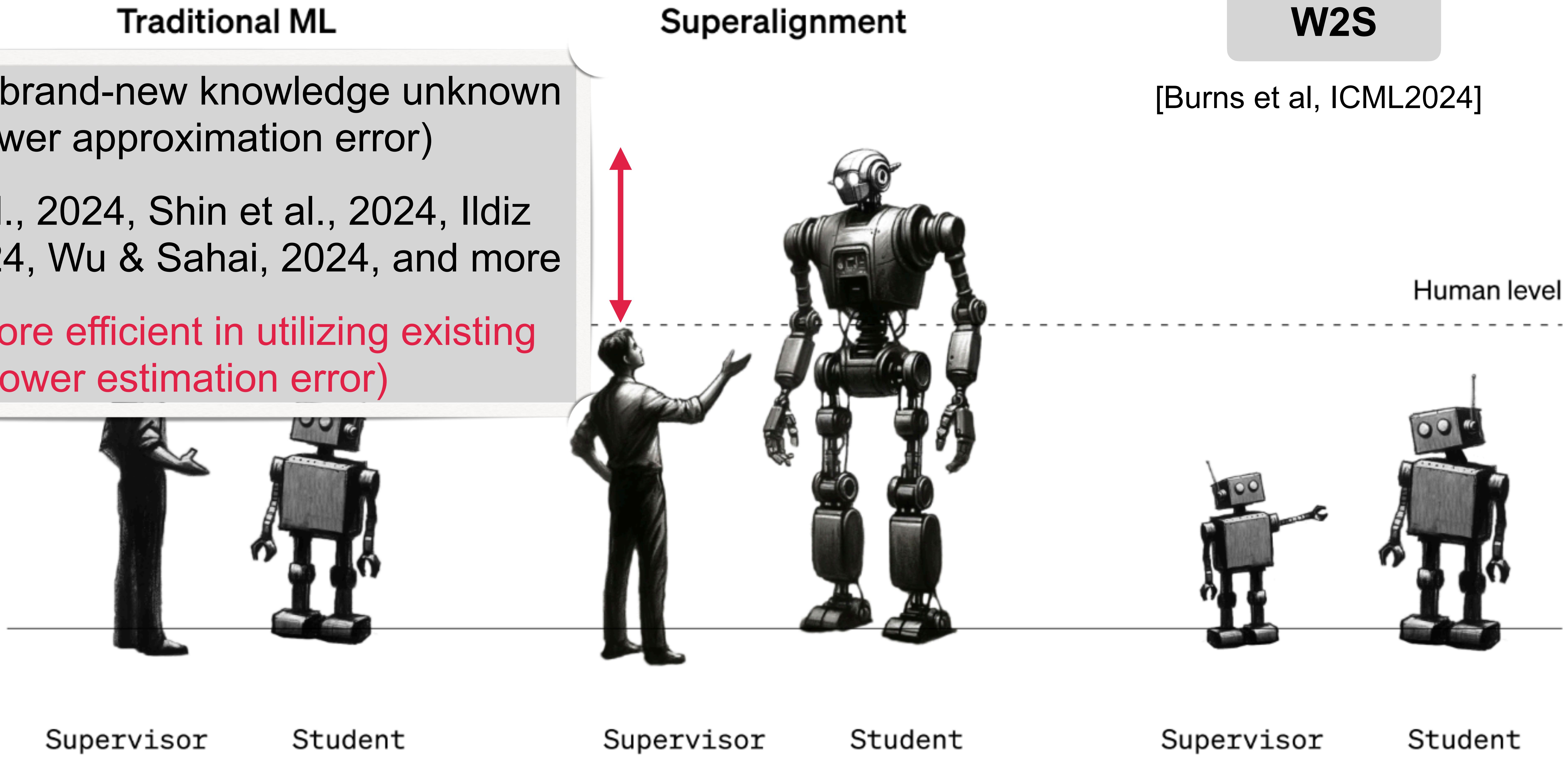
ICML 2025

Superalignment → weak-to-strong (W2S) generalization

W2S

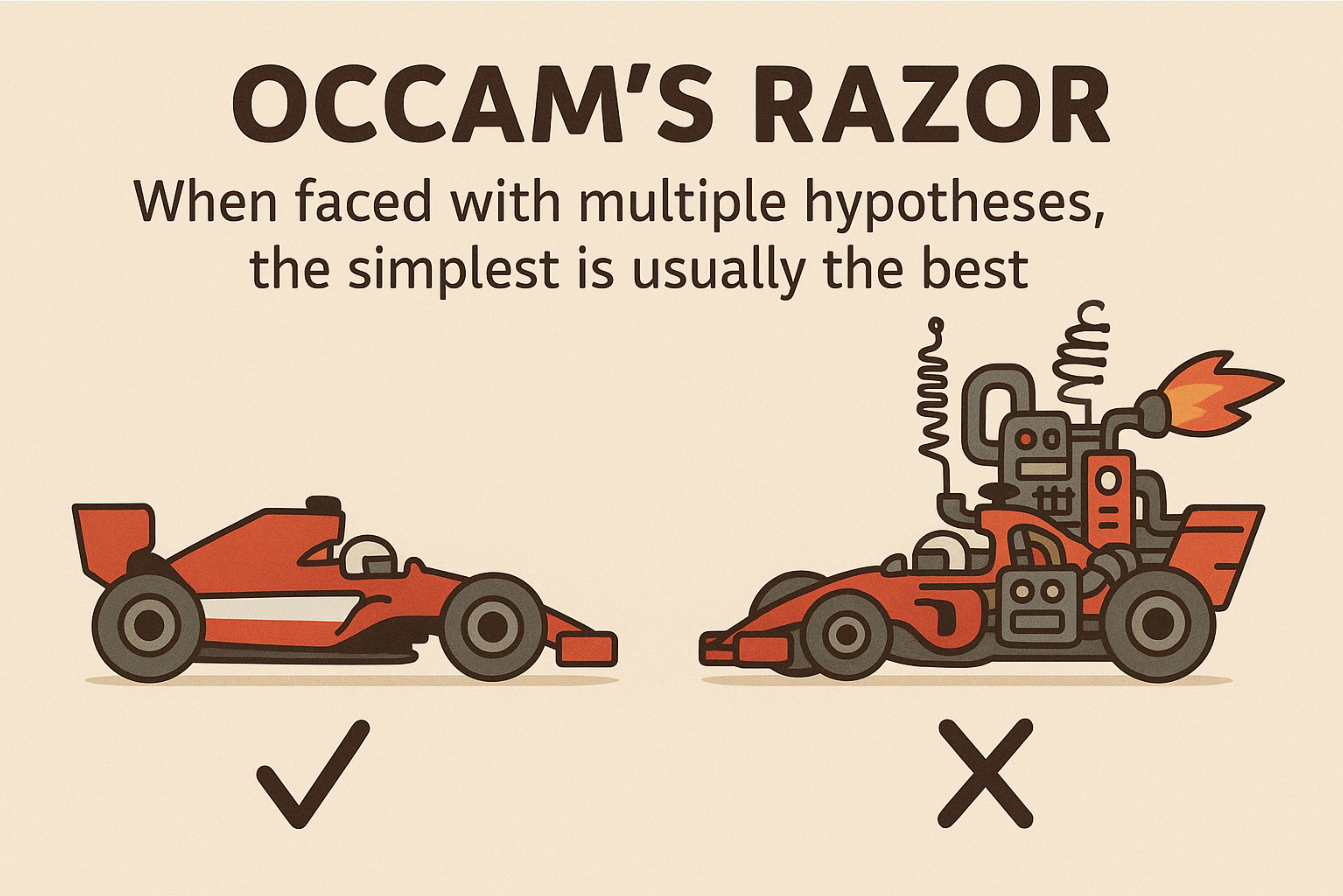
[Burns et al, ICML2024]

- Student has brand-new knowledge unknown to human (lower approximation error)
  - Lang et al., 2024, Shin et al., 2024, Ildiz et al., 2024, Wu & Sahai, 2024, and more
- Student is more efficient in utilizing existing knowledge (lower estimation error)



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

Intrinsic dimension



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

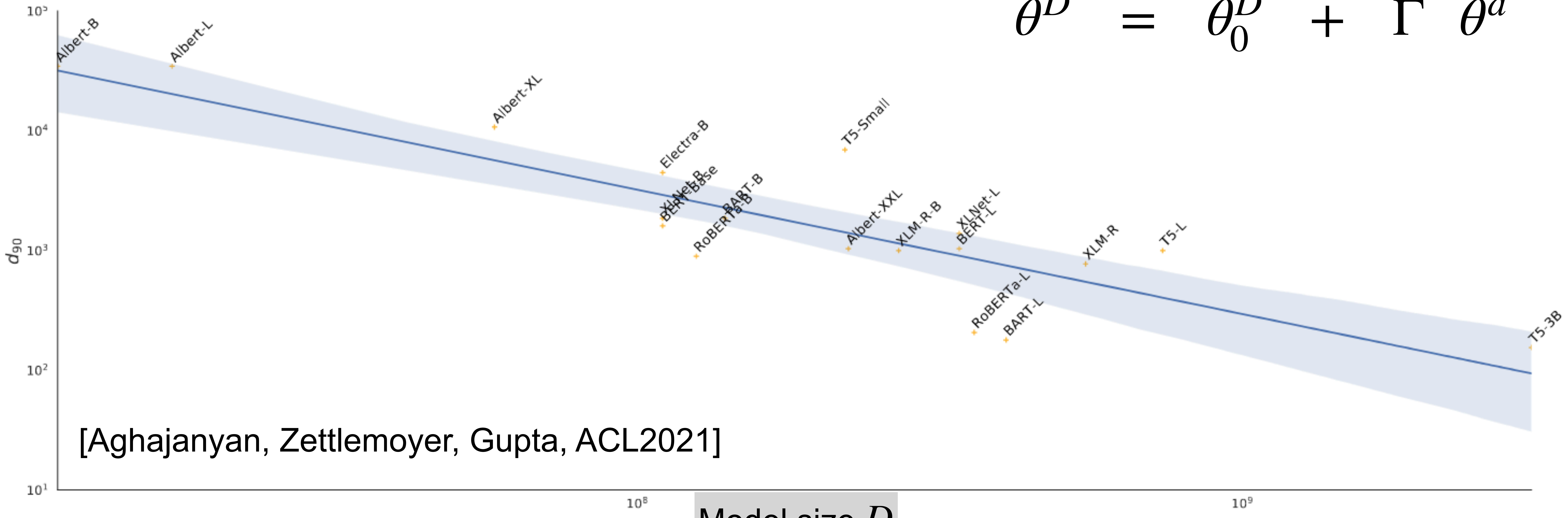
$D \times d$ random projection

Low intrinsic dimension of finetuning

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

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

Intrinsic dimension d



[Aghajanyan, Zettlemoyer, Gupta, ACL2021]

Model size D

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

Finetuning with low intrinsic dimensions

Downstream task

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

Finetuning (FT) ≈ linear probing on low-rank gradient features

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

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

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

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

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**: $0 \le \rho_s \le \rho_w \le 1$ 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

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

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

W2S finetuning as regression

Weak teacher $f_w(x) = \phi_w(x)^\top \theta_w$: $\theta_w = \arg\min_{\theta \in \mathbb{R}^d} \frac{1}{n} \| \widetilde{\Phi}_w \theta - \widetilde{y} \|_2^2 + \alpha_w \| \theta \|_2^2$

W2S

W2S $f_{w2s}(x) = \phi_s(x)^\top \theta_{w2s}$: $\theta_{w2s} = \arg\min_{\theta \in \mathbb{R}^d} \frac{1}{N} \| \Phi_s \theta - \Phi_w \theta_w \|_2^2 + \alpha_{w2s} \| \theta \|_2^2$

PGR =

Strong SFT $f_s(x) = \phi_s(x)^\top \theta_s$: $\theta_s = \arg\min_{\theta \in \mathbb{R}^d} \frac{1}{n} \| \widetilde{\Phi}_s \theta - \widetilde{y} \|_2^2 + \alpha_s \| \theta \|_2^2$

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

W2S v.s. SFT

Is the additional compute of W2S worthwhile?

(Outperformance ratio/OPR)

W2S generalization: ridgeless regression ($\alpha \to 0$)

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

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

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

Proposition [DLLLL25].

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

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

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

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 $n > d_w + 1$:

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

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

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

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

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

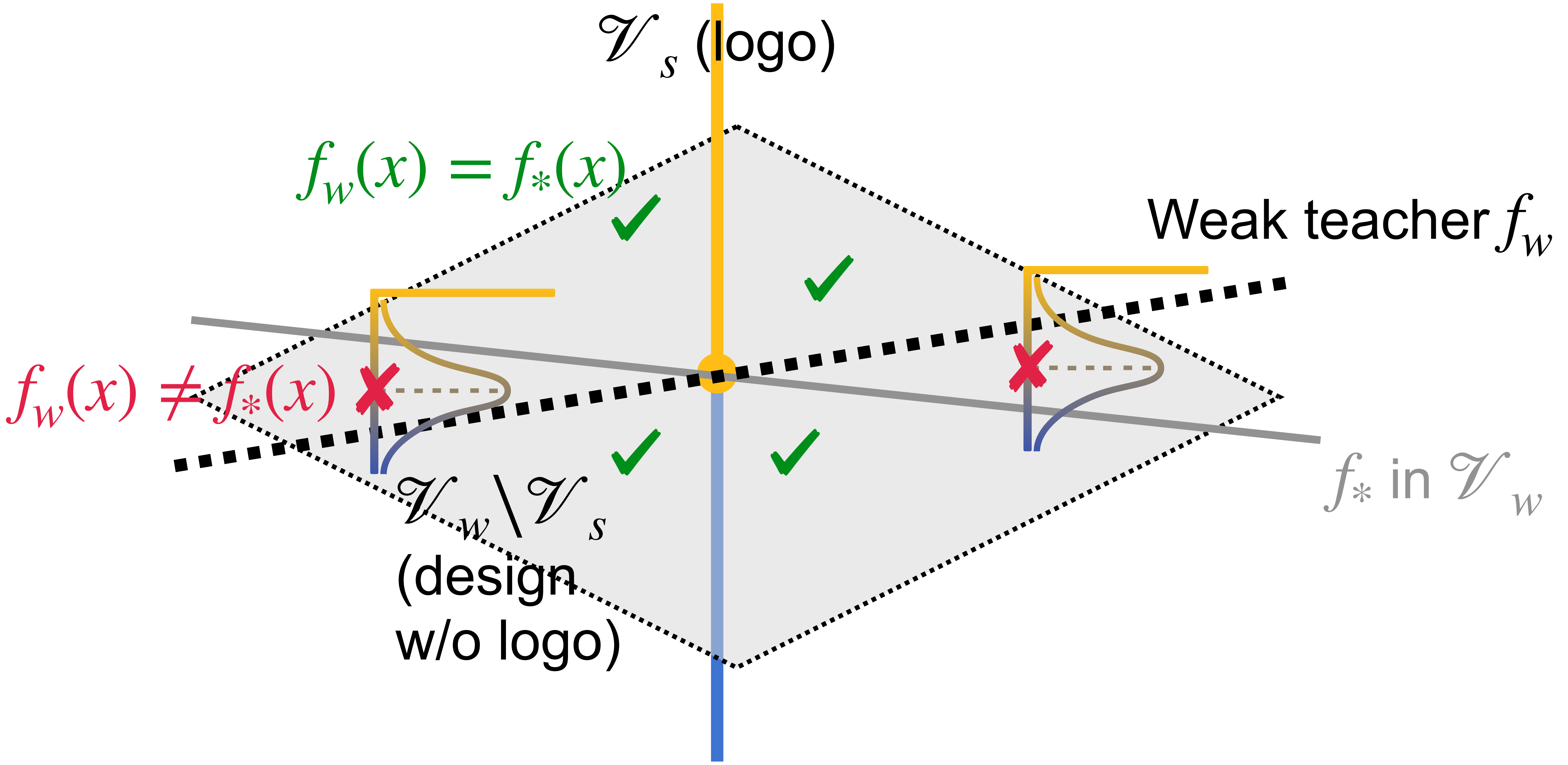
Intuition: How does variance reduction in W2S happen?

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

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

Var. in $\mathscr{V}_w \cap \mathscr{V}_s$ W2S Var. in $\mathscr{V}_w \backslash \mathscr{V}_s$

Task: Determine the make of a car



Pseudolabel error in $\mathscr{V}_w \backslash \mathscr{V}_s$ can be viewed as independent label noise w.r.t. the orthogonal strong features $\mathscr{V}_s$, 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: $\mathrm{tr}(\Sigma_w) \asymp d_w$, $\mathrm{tr}(\Sigma_s) \asymp d_s \ll d$

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

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

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

Theorem [DLLLL25]. 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 $\alpha_w = \frac{\sigma^2\mathrm{tr}(\Sigma_s\Sigma_w)}{4nN}\frac{\varrho_s}{\varrho_w^2}$ and $\alpha_{w2s} = \frac{\sigma^2\mathrm{tr}(\Sigma_s\Sigma_w)}{4nN}\frac{\varrho_w}{\varrho_s^2}$. When $N \geq \frac{\mathrm{tr}(\Sigma_s)\mathrm{tr}(\Sigma_w)}{\mathrm{tr}(\Sigma_s\Sigma_w)}$,

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

- **Multiplicative sample complexity:**

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

- **Representation similarity (“$d_{s\wedge w}$”):**

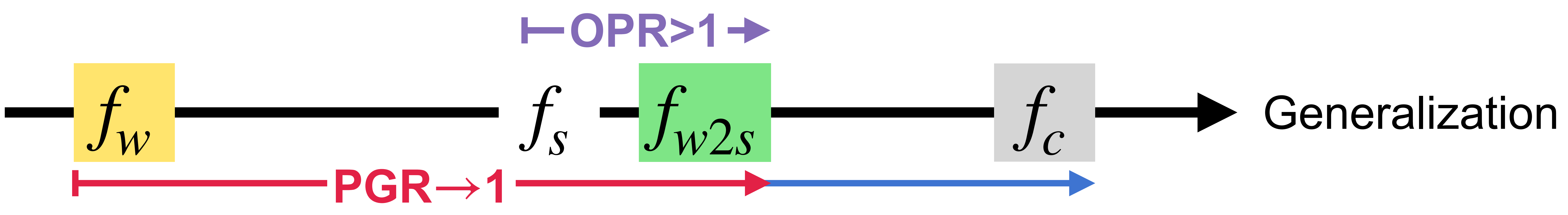
$$\mathrm{tr}(\Sigma_s\Sigma_w) \lesssim \min\{\mathrm{tr}(\Sigma_s), \mathrm{tr}(\Sigma_w)\}$$

- **Representation accuracy**: ϱ_w, ϱ_s are small if the dominating eigenspaces of Σ_w, Σ_s cover that of Σ_*

Larger discrepancy (lower $d_{s\wedge w}$) → 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: $\mathrm{OPR} = \dfrac{\mathrm{ER}(f_s)}{\mathrm{ER}(f_{w2s})}$

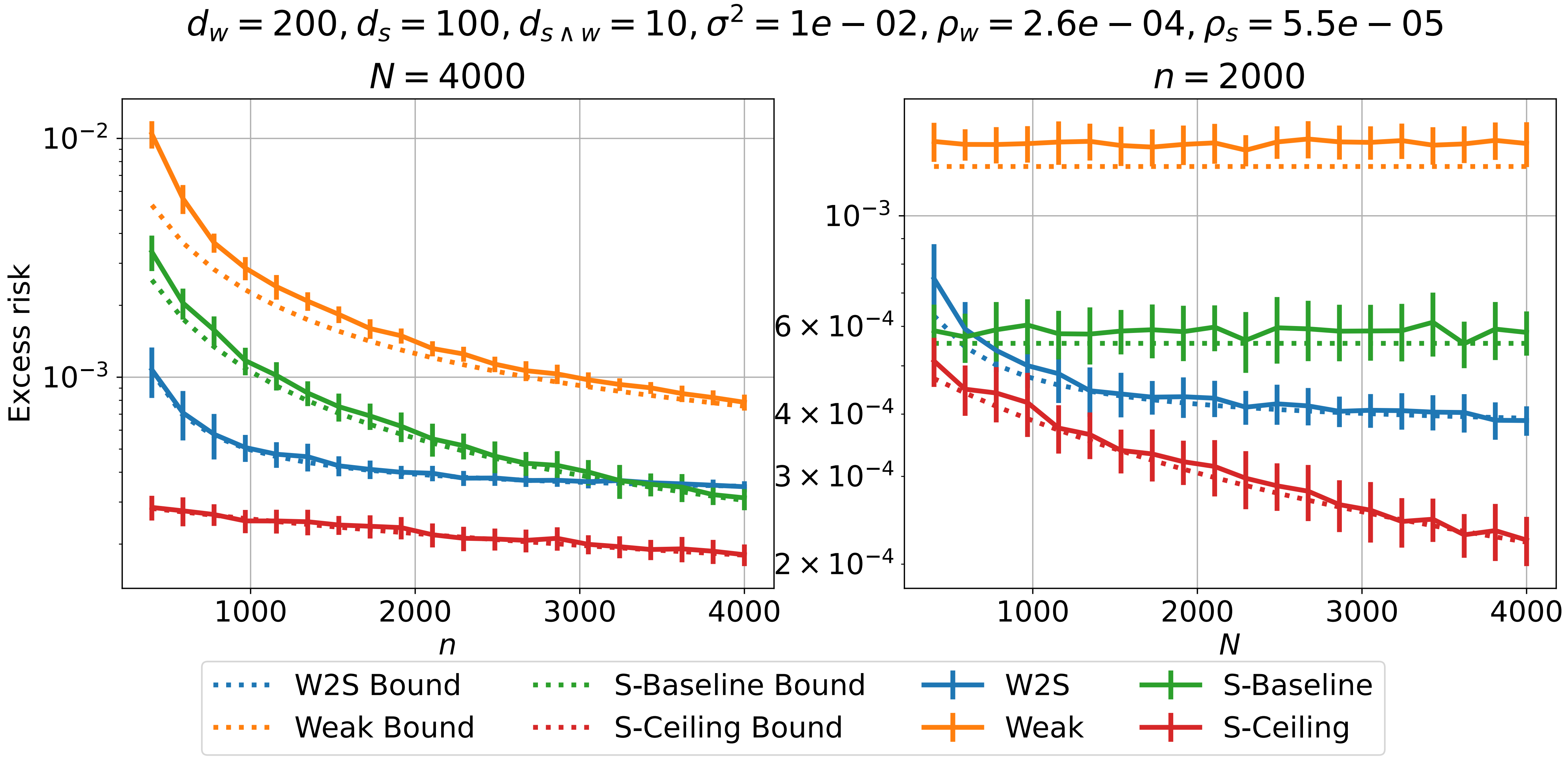


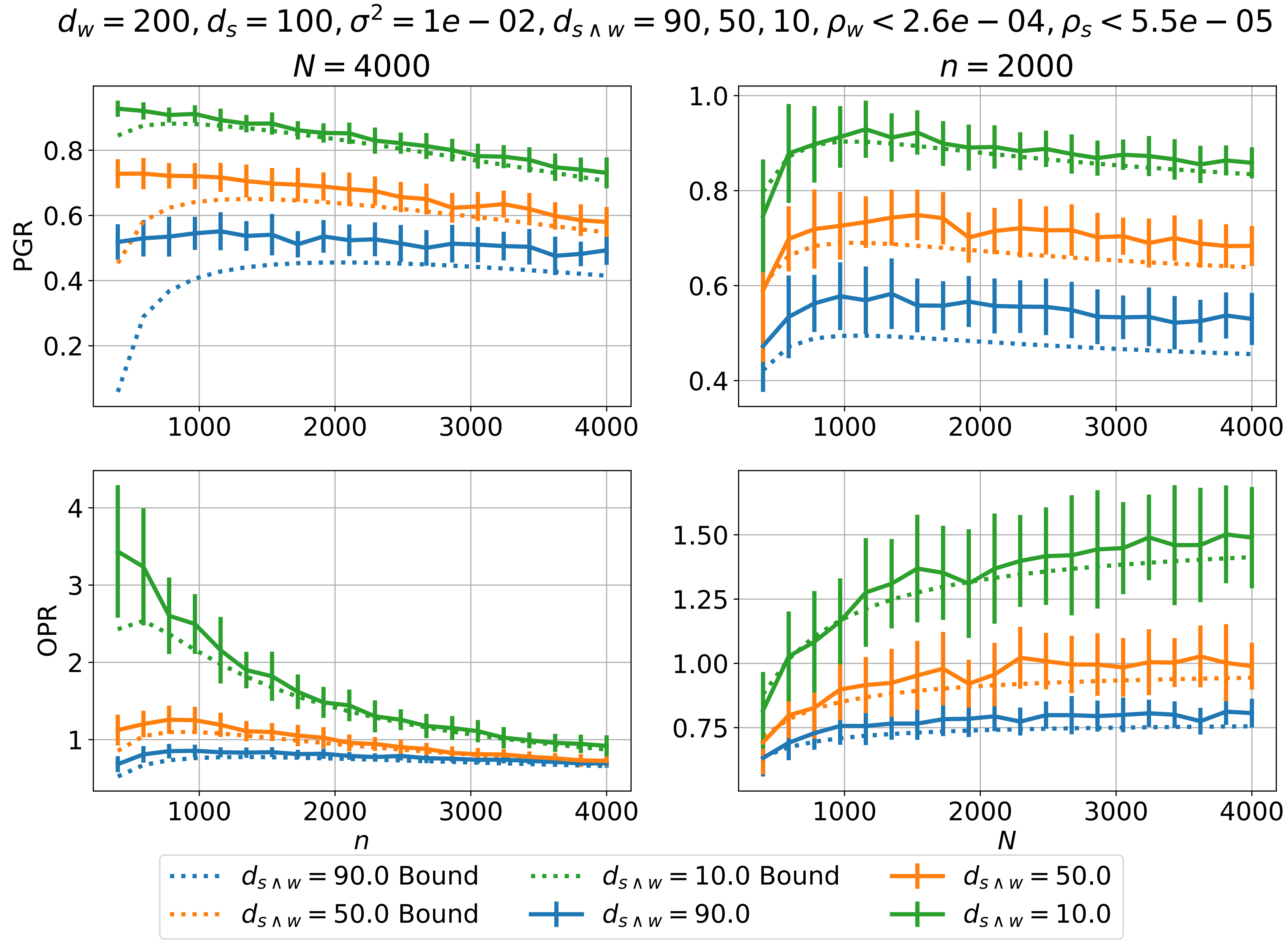
With negligible FT approximation error $(\rho_w + \rho_s)/\sigma^2 \to 0$, when $n \gtrsim d_w$ and $N \gtrsim d_s(d_w/d_{s\wedge w} - 1)$, we have

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

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 $i > 300$

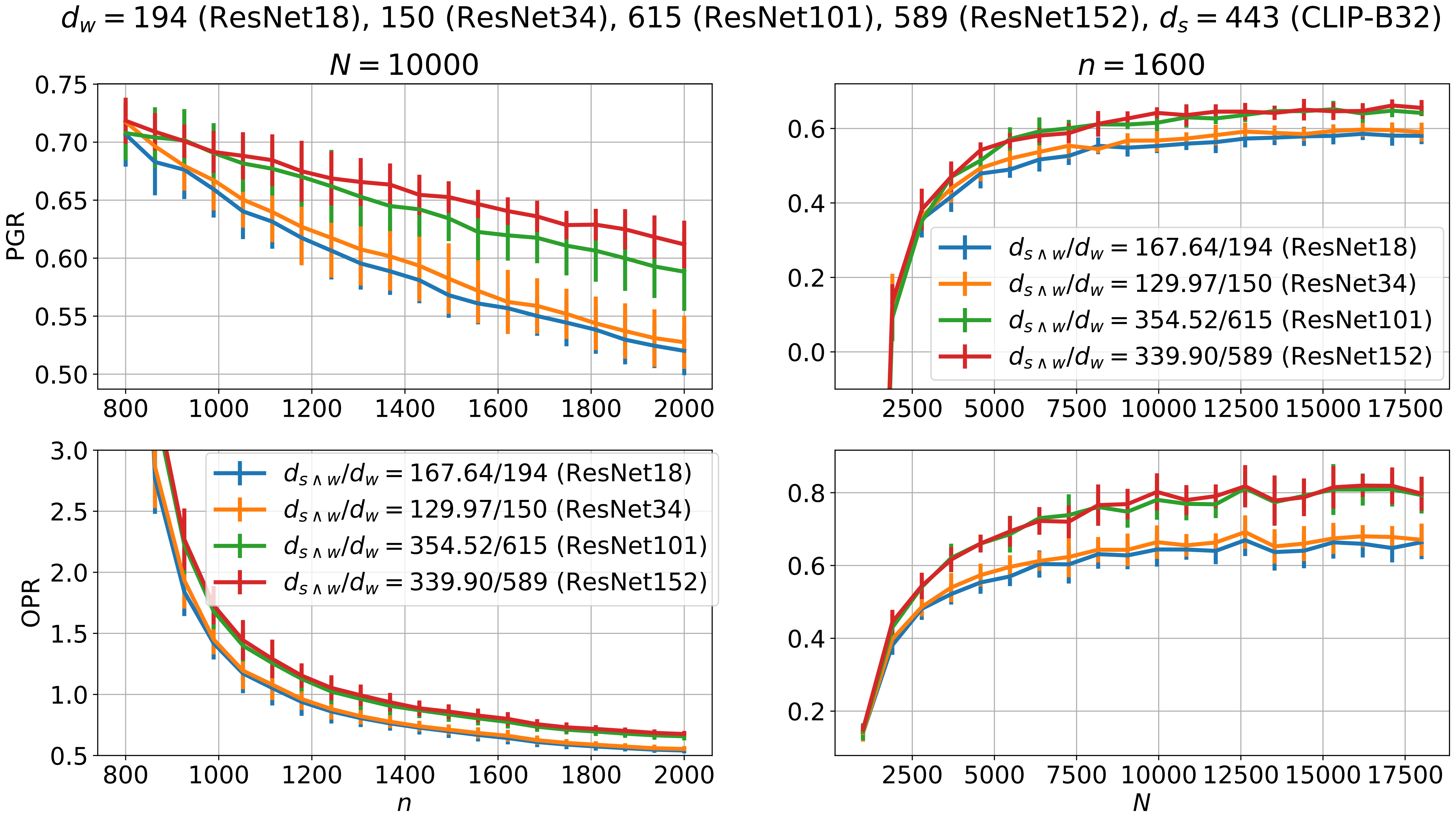




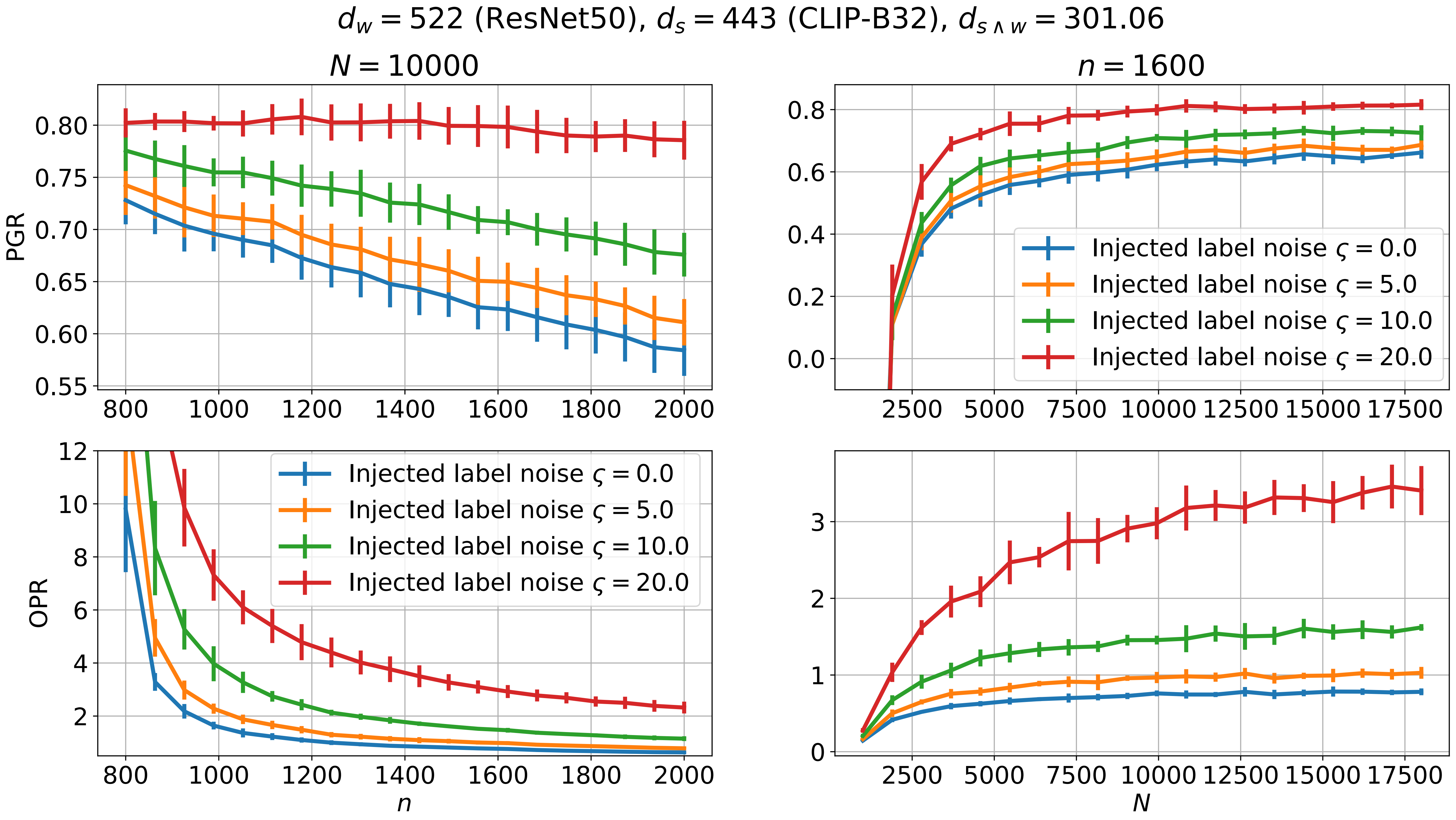
- Our bounds provide reasonably tight characterization for the generalization error, PGR, and OPR.
- W2S is more beneficial with limited label data n — PGR and OPR decrease as n increases!

UTKFace regression

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



Larger variance → 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 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 $d_{s\wedge w} = \|V_s^\top V_w\|_F^2 \in [0, \min\{d_s, d_w\}]$

$$\mathrm{Var}(f_{w2s}) \asymp \underbrace{\boxed{\frac{d_{s\wedge w}}{n}}}_{\text{Var. in } \mathscr{V}_w \cap \mathscr{V}_s} + \underbrace{\boxed{\frac{d_s}{N}}}_{\text{W2S}} \underbrace{\boxed{\frac{d_w - d_{s\wedge w}}{n}}}_{\text{Var. in } \mathscr{V}_w \setminus \mathscr{V}_s}$$

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

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

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).