

Learning from Data of All Sorts (Synthetic, Private, Surrogate and Mixture): Lessons from High-Dimensional Regression

Marco Mondelli

Institute of Science and Technology Austria (ISTA)

WU Vienna, October 14, 2026



Credits



Credits



Learning from data of all sorts

Modern ML systems are trained on

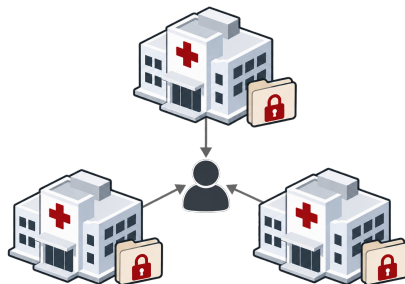
(#1) data obtained from generative models
(**synthetic**)



Learning from data of all sorts

Modern ML systems are trained on

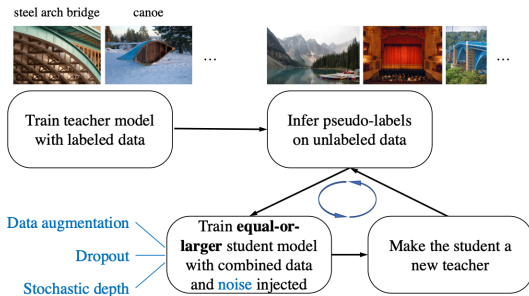
- (#1) data obtained from generative models
(**synthetic**)
- (#2) sensitive data held by external
providers (**private**)



Learning from data of all sorts

Modern ML systems are trained on

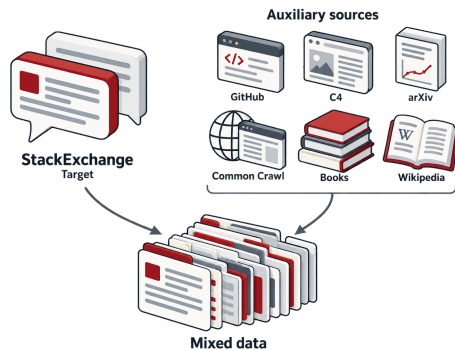
- (#1) data obtained from generative models (**synthetic**)
- (#2) sensitive data held by external providers (**private**)
- (#3) data labeled by a pre-trained teacher (**surrogate**)



Learning from data of all sorts

Modern ML systems are trained on

- (#1) data obtained from generative models (**synthetic**)
- (#2) sensitive data held by external providers (**private**)
- (#3) data labeled by a pre-trained teacher (**surrogate**)
- (#4) data obtained by mixing different sources (**mixture**)



Lessons from high-dimensional regression

Methodology based on ridge regression:

- Asymptotic risk characterization with number of features proportional to sample size
- Non-asymptotic deterministic equivalent for general growth rates, leading to scaling laws

Lessons from high-dimensional regression

Methodology based on ridge regression:

- Asymptotic risk characterization with number of features proportional to sample size
- Non-asymptotic deterministic equivalent for general growth rates, leading to scaling laws

Insights into practice:

- How to select synthetic data?
- When to collaborate with an external provider holding private data?
- To what extent can data labeled by an imperfect teacher help?
- When does data mixing produce synergy?

Learning from synthetic data

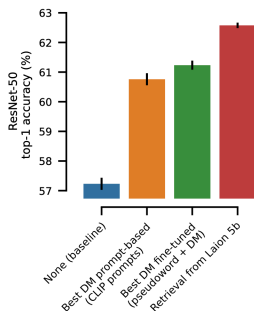
Progress in generative models provides high-quality synthetic data [Trabucco et al., 2024; He et al., 2023; Azizi et al., 2023]



Learning from synthetic data

Progress in generative models provides high-quality synthetic data, but the **usefulness of synthetic data is unclear**:

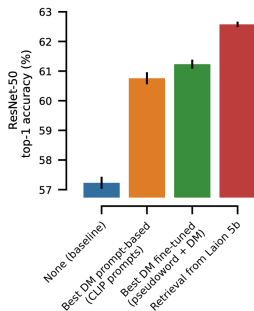
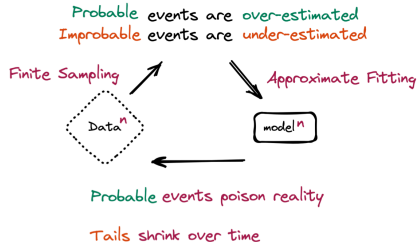
- directly selecting data used to train the generative model is better [Burg et al., 2023; Geng et al., 2024]



Learning from synthetic data

Progress in generative models provides high-quality synthetic data, but the **usefulness of synthetic data is unclear**:

- directly selecting data used to train the generative model is better [Burg et al., 2023; Geng et al., 2024]
- model collapse [Shumailov et al., 2024], bias [Wyllie et al., 2024]



Formalization of the problem

How to select synthetic data?

Formalization of the problem

How to select synthetic data?

- Training and synthetic data are Gaussian mixtures
- Augmentation occurs per class

Formalization of the problem

How to select synthetic data?

- Training and synthetic data are Gaussian mixtures
- Augmentation occurs per class

We study data selection for a single class, neglecting class interactions



Strong assumption, but insights translate to practice!

Formalization of the problem

Training data: $\{(\mathbf{x}_i^t, y_i^t)\}_{i=1}^{n_t} \sim_{\text{i.i.d.}} \mathcal{D}(\boldsymbol{\mu}_t, \boldsymbol{\Sigma}_t), \mathbf{x}_i^t \in \mathbb{R}^p$

$$y_i^t = (\mathbf{x}_i^t)^\top \boldsymbol{\beta}^* + z_i^t, \quad \mathbb{E}[\mathbf{x}_i^t] = \boldsymbol{\mu}_t, \quad \mathbb{E}[\mathbf{x}_i^t (\mathbf{x}_i^t)^\top] = \boldsymbol{\Sigma}_t, \quad \mathbb{E}[z_i^t] = 0, \quad \mathbb{E}[(z_i^t)^2] = \sigma^2$$

Formalization of the problem

Training data: $\{(\mathbf{x}_i^t, y_i^t)\}_{i=1}^{n_t} \sim \text{i.i.d. } \mathcal{D}(\boldsymbol{\mu}_t, \boldsymbol{\Sigma}_t), \mathbf{x}_i^t \in \mathbb{R}^p$

$$y_i^t = (\mathbf{x}_i^t)^\top \boldsymbol{\beta}^* + z_i^t, \quad \mathbb{E}[\mathbf{x}_i^t] = \boldsymbol{\mu}_t, \quad \mathbb{E}[\mathbf{x}_i^t (\mathbf{x}_i^t)^\top] = \boldsymbol{\Sigma}_t, \quad \mathbb{E}[z_i^t] = 0, \quad \mathbb{E}[(z_i^t)^2] = \sigma^2$$

Synthetic data: $\{(\mathbf{x}_i^s, y_i^s)\}_{i=1}^{n_s} \sim \text{i.i.d. } \mathcal{D}(\boldsymbol{\mu}_s, \boldsymbol{\Sigma}_s), \mathbf{x}_i^s \in \mathbb{R}^p$

$$y_i^s = (\mathbf{x}_i^s)^\top \boldsymbol{\beta}^* + z_i^s, \quad \mathbb{E}[\mathbf{x}_i^s] = \boldsymbol{\mu}_s, \quad \mathbb{E}[\mathbf{x}_i^s (\mathbf{x}_i^s)^\top] = \boldsymbol{\Sigma}_s, \quad \mathbb{E}[z_i^s] = 0, \quad \mathbb{E}[(z_i^s)^2] = \sigma^2$$

- $\hat{\boldsymbol{\beta}}$ = minimum-norm interpolator of $n := n_t + n_s$ samples
 $(\mathbf{x}_1^t, y_1^t), \dots, (\mathbf{x}_{n_t}^t, y_{n_t}^t), (\mathbf{x}_1^s, y_1^s), \dots, (\mathbf{x}_{n_s}^s, y_{n_s}^s)$

Formalization of the problem

Training data: $\{(\mathbf{x}_i^t, y_i^t)\}_{i=1}^{n_t} \sim \text{i.i.d. } \mathcal{D}(\boldsymbol{\mu}_t, \boldsymbol{\Sigma}_t), \mathbf{x}_i^t \in \mathbb{R}^p$

$$y_i^t = (\mathbf{x}_i^t)^\top \boldsymbol{\beta}^* + z_i^t, \quad \mathbb{E}[\mathbf{x}_i^t] = \boldsymbol{\mu}_t, \quad \mathbb{E}[\mathbf{x}_i^t (\mathbf{x}_i^t)^\top] = \boldsymbol{\Sigma}_t, \quad \mathbb{E}[z_i^t] = 0, \quad \mathbb{E}[(z_i^t)^2] = \sigma^2$$

Synthetic data: $\{(\mathbf{x}_i^s, y_i^s)\}_{i=1}^{n_s} \sim \text{i.i.d. } \mathcal{D}(\boldsymbol{\mu}_s, \boldsymbol{\Sigma}_s), \mathbf{x}_i^s \in \mathbb{R}^p$

$$y_i^s = (\mathbf{x}_i^s)^\top \boldsymbol{\beta}^* + z_i^s, \quad \mathbb{E}[\mathbf{x}_i^s] = \boldsymbol{\mu}_s, \quad \mathbb{E}[\mathbf{x}_i^s (\mathbf{x}_i^s)^\top] = \boldsymbol{\Sigma}_s, \quad \mathbb{E}[z_i^s] = 0, \quad \mathbb{E}[(z_i^s)^2] = \sigma^2$$

- $\hat{\boldsymbol{\beta}}$ = minimum-norm interpolator of $n := n_t + n_s$ samples
 $(\mathbf{x}_1^t, y_1^t), \dots, (\mathbf{x}_{n_t}^t, y_{n_t}^t), (\mathbf{x}_1^s, y_1^s), \dots, (\mathbf{x}_{n_s}^s, y_{n_s}^s)$

Excess population risk:

$$\mathcal{R}(\hat{\boldsymbol{\beta}}) = \mathbb{E}_{(\mathbf{x}, y) \sim \mathcal{D}(\boldsymbol{\mu}_t, \boldsymbol{\Sigma}_t)} \left[(y - \mathbf{x}^\top \hat{\boldsymbol{\beta}})^2 \right] - \sigma^2$$

Analysis in high dimensions

Proportional regime: $p, n \rightarrow \infty$ with fixed ratio p/n

- Train on two distributions and test on one of them [Fan et al., 2025; Song et al., 2024]: same setup, except $\mu_t = \mu_s = 0$

Analysis in high dimensions

Proportional regime: $p, n \rightarrow \infty$ with fixed ratio p/n

- Train on two distributions and test on one of them [Fan et al., 2025; Song et al., 2024]: same setup, except $\mu_t = \mu_s = 0$

Cannot set $\mu_t = \mu_s = 0$ for synthetic data selection

- Centering the data requires the mean of the test sample, which corresponds to the unknown label.

Asymptotic risk characterization

Theorem [Rezaei, Kovacevic, Locatello and M., 2026]

Let $p, n \rightarrow \infty$, $p/n \in (0, 1)$ ($p < n$) and $M = \Sigma_t^{-1/2} \Sigma_s \Sigma_t^{-1/2}$. Then, the random quantity $\mathcal{R}(\hat{\beta})$ (dependent on data, noise) concentrates to the **deterministic quantity**

$$\bar{\mathcal{R}}(M) := \frac{\sigma^2}{n} \text{Tr} \left[(\alpha_1 M + \alpha_2 I_p)^{-1} \right],$$

with α_1, α_2 given by a fixed point equation.

P. Rezaei, F. Kovacevic, F. Locatello, and M. Mondelli, "High-dimensional Analysis of Synthetic Data Selection", *ICLR'26 (oral)*.

Asymptotic risk characterization

Theorem [Rezaei, Kovacevic, Locatello and M., 2026]

Let $p, n \rightarrow \infty$, $p/n \in (0, 1)$ ($p < n$) and $M = \Sigma_t^{-1/2} \Sigma_s \Sigma_t^{-1/2}$. Then, the random quantity $\mathcal{R}(\hat{\beta})$ (dependent on data, noise) concentrates to the **deterministic quantity**

$$\bar{\mathcal{R}}(M) := \frac{\sigma^2}{n} \text{Tr} \left[(\alpha_1 M + \alpha_2 I_p)^{-1} \right],$$

with α_1, α_2 given by a fixed point equation.

Idea: Factor out means μ_t, μ_s as rank-2 perturbation and use anisotropic local laws



Asymptotic risk characterization

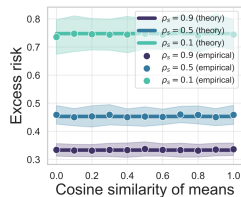
Theorem [Rezaei, Kovacevic, Locatello and M., 2026]

Let $p, n \rightarrow \infty$, $p/n \in (0, 1)$ ($p < n$) and $M = \Sigma_t^{-1/2} \Sigma_s \Sigma_t^{-1/2}$. Then, the random quantity $\mathcal{R}(\hat{\beta})$ (dependent on data, noise) concentrates to the **deterministic quantity**

$$\bar{\mathcal{R}}(M) := \frac{\sigma^2}{n} \text{Tr} \left[(\alpha_1 M + \alpha_2 I_p)^{-1} \right],$$

with α_1, α_2 given by a fixed point equation.

Covariances Σ_t, Σ_s matter, but **means μ_t, μ_s do not!**



Asymptotic risk characterization

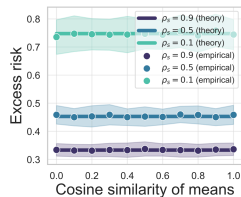
Theorem [Rezaei, Kovacevic, Locatello and M., 2026]

Let $p, n \rightarrow \infty$, $p/n \in (0, 1)$ ($p < n$) and $M = \Sigma_t^{-1/2} \Sigma_s \Sigma_t^{-1/2}$. Then, the random quantity $\mathcal{R}(\hat{\beta})$ (dependent on data, noise) concentrates to the **deterministic quantity**

$$\bar{\mathcal{R}}(M) := \frac{\sigma^2}{n} \text{Tr} \left[(\alpha_1 M + \alpha_2 I_p)^{-1} \right],$$

with α_1, α_2 given by a fixed point equation.

Covariances Σ_t, Σ_s matter, but **means μ_t, μ_s do not!**



Means μ_t, μ_s do matter when training only on synthetic data ($n_t = 0$)

Synthetic data selection

How to select synthetic data?

Synthetic data selection

Given Σ_t , what is the **optimal Σ_s that minimizes the risk?**

Synthetic data selection

Given Σ_t , what is the **optimal Σ_s that minimizes the risk?**

Theorem

The matrix M minimizing $\overline{\mathcal{R}}(M)$ among matrices of fixed trace is a multiple of identity.

Synthetic data selection

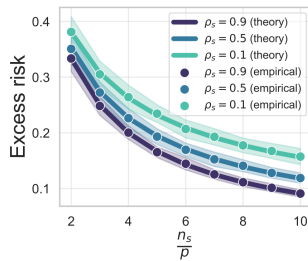
Given Σ_t , what is the **optimal Σ_s that minimizes the risk?**

Theorem

The matrix M minimizing $\bar{\mathcal{R}}(M)$ among matrices of fixed trace is a multiple of identity.

$$M \propto I \implies \Sigma_s \propto \Sigma_t$$

Matching covariances is optimal



Synthetic data selection

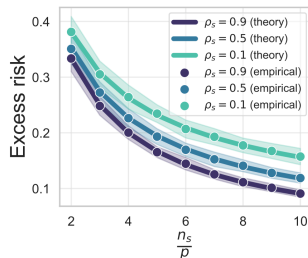
Given Σ_t , what is the **optimal Σ_s that minimizes the risk?**

Theorem

The matrix M minimizing $\overline{\mathcal{R}}(M)$ among matrices of fixed trace is a multiple of identity.

$$M \propto I \implies \Sigma_s \propto \Sigma_t$$

Matching covariances is optimal



Same high-level picture for overparameterized regime $p/n \in (1, \infty)$ ($p > n$)

Experimental evidence

- Three paradigms to train ResNets on CIFAR-10
- Synthetic data from five truncated StyleGAN2-Ada models

Experimental evidence

- Three paradigms to train ResNets on CIFAR-10
- Synthetic data from five truncated StyleGAN2-Ada models

Method	Scratch	Distillation	Pretrained
No synthetic	44.36 ± 1.51	47.33 ± 0.57	63.40 ± 1.33
Center matching (He et al., 2023)	50.04 ± 2.84	53.83 ± 0.59	67.01 ± 0.89
Center sampling (Lin et al., 2023)	50.48 ± 2.03	54.91 ± 1.07	67.71 ± 0.90
DS3 (Hulkund et al., 2025)	52.83 ± 2.19	58.32 ± 0.43	68.21 ± 0.66
K-means (Lin et al., 2023)	50.74 ± 1.77	56.06 ± 0.68	66.50 ± 1.11
Random	49.38 ± 2.43	54.89 ± 0.91	67.65 ± 0.77
Text matching (Lin et al., 2023)	50.94 ± 1.40	55.17 ± 0.57	67.81 ± 0.76
Text sampling (Lin et al., 2023)	50.28 ± 1.18	54.82 ± 0.72	67.45 ± 1.02
Covariance matching (ours)	54.00 ± 1.89	59.77 ± 0.61	69.20 ± 0.56
Real upper bound	61.08 ± 2.54	65.38 ± 0.51	74.35 ± 0.56

Covariance matching outperforms all baselines

More experimental evidence

- ResNet-18 trained from scratch on ImageNet-100; synthetic images from StyleGAN-XL and text-to-image (T2I) models

Method	Truncated models	T2I models
No synthetic	40.78 ± 1.29	
Center matching (He et al., 2023)	53.39 ± 0.37	53.96 ± 1.06
DS3 (Hulkund et al., 2025)	57.47 ± 0.87	53.51 ± 0.31
Random	54.14 ± 0.82	49.84 ± 1.32
Text matching (Lin et al., 2023)	53.39 ± 0.99	53.37 ± 0.72
Covariance matching (ours)	57.52 ± 0.36	53.07 ± 0.89
Real upper bound	62.67 ± 0.65	

Similar results on a broader dataset from computer vision

More experimental evidence

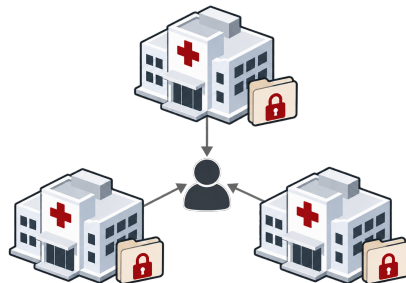
- ResNet-18 trained from scratch on ImageNet-100; synthetic images from StyleGAN-XL and text-to-image (T2I) models
- Linear model (on top of ImageNet-pretrained ResNet) trained on RxRx1; synthetic images from MorphGen

Method	MorphGen
No synthetic	86.83 ± 2.44
Center matching (He et al., 2023)	88.17 ± 2.35
Random	87.33 ± 2.03
K-means (Lin et al., 2023)	89.00 ± 1.70
DS3 (Hulkund et al., 2025)	89.67 ± 1.45
Center sampling (Lin et al., 2023)	88.75 ± 2.27
Covariance matching (ours)	90.00 ± 1.86

Similar results on a broader dataset from computer vision and a specialized one from fluorescence microscopy

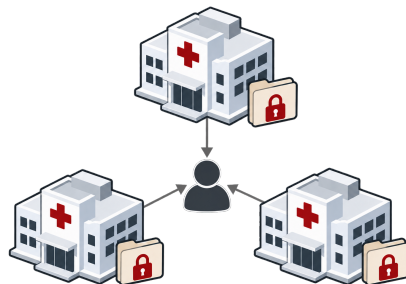
Learning from private data

- Only summary statistics available $\rightarrow$ ridge regression



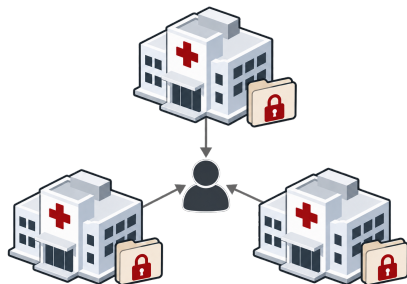
Learning from private data

- Only summary statistics available $\rightarrow$ ridge regression
- (Differential) privacy guarantee $\rightarrow$ add noise to labels (and possibly features)



Learning from private data

- Only summary statistics available $\rightarrow$ ridge regression
- (Differential) privacy guarantee $\rightarrow$ add noise to labels (and possibly features)
- Apply for access before seeing data itself $\rightarrow$ theory prediction depends only on population-level quantities



Formalization of the problem

When to collaborate with an external provider holding private data?

Formalization of the problem

Training data from C clients: $\{(\mathbf{x}_i^c, y_i^c)\}_{i=1}^{n_c} \sim_{\text{i.i.d.}} \mathcal{D}^{(c)}, \mathbf{x}_i^c \in \mathbb{R}^p, c \in \{1, \dots, C\}$

$$y_i^c = (\mathbf{x}_i^c)^\top \boldsymbol{\beta}_c^* + z_i^c, \quad \mathbb{E}[\mathbf{x}_i^c] = 0, \quad \mathbb{E}[\mathbf{x}_i^c (\mathbf{x}_i^c)^\top] = \boldsymbol{\Sigma}_c, \quad \mathbb{E}[z_i^c] = 0, \quad \mathbb{E}[(z_i^c)^2] = \sigma_c^2$$

- Both model and covariance shift

Formalization of the problem

Training data from C clients: $\{(\mathbf{x}_i^c, y_i^c)\}_{i=1}^{n_c} \sim_{\text{i.i.d.}} \mathcal{D}^{(c)}, \mathbf{x}_i^c \in \mathbb{R}^p, c \in \{1, \dots, C\}$

$$y_i^c = (\mathbf{x}_i^c)^\top \boldsymbol{\beta}_c^* + z_i^c, \quad \mathbb{E}[\mathbf{x}_i^c] = 0, \quad \mathbb{E}[\mathbf{x}_i^c (\mathbf{x}_i^c)^\top] = \boldsymbol{\Sigma}_c, \quad \mathbb{E}[z_i^c] = 0, \quad \mathbb{E}[(z_i^c)^2] = \sigma_c^2$$

Private release of labels:

$$\tilde{y}_i^c = y_i^c + \xi_i^c, \quad \xi_i^c \sim \mathcal{N}(0, \sigma_{\text{DP},c}^2), \quad c \in \{2, \dots, C\}$$

- Guarantees in terms of ρ zero-concentrated differential privacy
- Features can be privatized as well

Formalization of the problem

Training data from C clients: $\{(\mathbf{x}_i^c, y_i^c)\}_{i=1}^{n_c} \sim \text{i.i.d. } \mathcal{D}^{(c)}, \mathbf{x}_i^c \in \mathbb{R}^p, c \in \{1, \dots, C\}$

$$y_i^c = (\mathbf{x}_i^c)^\top \boldsymbol{\beta}_c^* + z_i^c, \quad \mathbb{E}[\mathbf{x}_i^c] = 0, \quad \mathbb{E}[\mathbf{x}_i^c (\mathbf{x}_i^c)^\top] = \boldsymbol{\Sigma}_c, \quad \mathbb{E}[z_i^c] = 0, \quad \mathbb{E}[(z_i^c)^2] = \sigma_c^2$$

Private release of labels:

$$\tilde{y}_i^c = y_i^c + \xi_i^c, \quad \xi_i^c \sim \mathcal{N}(0, \sigma_{\text{DP},c}^2), \quad c \in \{2, \dots, C\}$$

Weighted ridge regression:

$$(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\delta}}_2, \dots, \hat{\boldsymbol{\delta}}_C) = \arg \min_{\hat{\mathbf{b}}, \hat{\mathbf{h}}_2, \dots, \hat{\mathbf{h}}_C} \left\{ \frac{\|\mathbf{y}^1 - \mathbf{X}^1 \hat{\mathbf{b}}\|_2^2}{n} + \sum_{c=2}^C w_c \frac{\|\tilde{\mathbf{y}}^c - \mathbf{X}^c (\hat{\mathbf{b}} + \hat{\mathbf{h}}_c)\|_2^2}{n} + \lambda_{\beta} \|\hat{\mathbf{b}}\|_2^2 + \sum_{c=2}^C \lambda_{\delta_c} \|\hat{\mathbf{h}}_c\|_2^2 \right\}$$

- $n = \sum_{c=1}^C n_c$

Formalization of the problem

Training data from C clients: $\{(\mathbf{x}_i^c, y_i^c)\}_{i=1}^{n_c} \sim_{\text{i.i.d.}} \mathcal{D}^{(c)}, \mathbf{x}_i^c \in \mathbb{R}^p, c \in \{1, \dots, C\}$

$$y_i^c = (\mathbf{x}_i^c)^\top \boldsymbol{\beta}_c^* + z_i^c, \quad \mathbb{E}[\mathbf{x}_i^c] = 0, \quad \mathbb{E}[\mathbf{x}_i^c (\mathbf{x}_i^c)^\top] = \boldsymbol{\Sigma}_c, \quad \mathbb{E}[z_i^c] = 0, \quad \mathbb{E}[(z_i^c)^2] = \sigma_c^2$$

Private release of labels:

$$\tilde{y}_i^c = y_i^c + \xi_i^c, \quad \xi_i^c \sim \mathcal{N}(0, \sigma_{\text{DP},c}^2), \quad c \in \{2, \dots, C\}$$

Weighted ridge regression:

$$(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\delta}}_2, \dots, \hat{\boldsymbol{\delta}}_C) = \arg \min_{\hat{\mathbf{b}}, \hat{\mathbf{h}}_2, \dots, \hat{\mathbf{h}}_C} \left\{ \frac{\|\mathbf{y}^1 - \mathbf{X}^1 \hat{\mathbf{b}}\|_2^2}{n} + \sum_{c=2}^C w_c \frac{\|\tilde{\mathbf{y}}^c - \mathbf{X}^c (\hat{\mathbf{b}} + \hat{\mathbf{h}}_c)\|_2^2}{n} + \lambda_\beta \|\hat{\mathbf{b}}\|_2^2 + \sum_{c=2}^C \lambda_{\delta_c} \|\hat{\mathbf{h}}_c\|_2^2 \right\}$$

Excess population risk:

$$\mathcal{R}(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\delta}}_2, \dots, \hat{\boldsymbol{\delta}}_C) = \mathbb{E}_{(\mathbf{x}, y) \sim \mathcal{D}^{(1)}} \left[(y - \mathbf{x}^\top \hat{\boldsymbol{\beta}})^2 \right] - \sigma_1^2$$

Asymptotic risk characterization

Theorem [Kovacevic, Cyffers, Mannelli and M., 2026]

Let $p, n \rightarrow \infty$, $p/n \in (0, 1)$, $n_c/n \rightarrow \bar{\pi}_c$, and assume $\{\beta_c\}_{c=1}^C$ have independent, zero-mean entries. Then, the random quantity $\mathcal{R}(\hat{\beta}, \hat{\delta}_2, \dots, \hat{\delta}_C)$ (dependent on data, noise) concentrates to the **deterministic quantity** $\bar{\mathcal{R}}$ (dependent only on $\{\bar{\pi}_c, \Sigma_c, \sigma_c\}_{c=1}^C$, $\{w_c, \sigma_{\text{DP},c}^2, \lambda_{\delta_c}\}_{c=2}^C$, λ_β).

- Explicit expression for $\bar{\mathcal{R}}$ (in terms of fixed point equations) depending on **population-level quantities**

F. Kovacevic, E. Cyffers, S. S. Mannelli, and M. Mondelli, "High-dimensional Asymptotics and Dataset Selection for Private Transfer Learning", arXiv:2610.02578, 2026.

Asymptotic risk characterization

Theorem [Kovacevic, Cyffers, Mannelli and M., 2026]

Let $p, n \rightarrow \infty$, $p/n \in (0, 1)$, $n_c/n \rightarrow \bar{\pi}_c$, and assume $\{\beta_c\}_{c=1}^C$ have independent, zero-mean entries. Then, the random quantity $\mathcal{R}(\hat{\beta}, \hat{\delta}_2, \dots, \hat{\delta}_C)$ (dependent on data, noise) concentrates to the **deterministic quantity** $\bar{\mathcal{R}}$ (dependent only on $\{\bar{\pi}_c, \Sigma_c, \sigma_c\}_{c=1}^C, \{w_c, \sigma_{\text{DP},c}^2, \lambda_{\delta_c}\}_{c=2}^C, \lambda_\beta$).

- Explicit expression for $\bar{\mathcal{R}}$ (in terms of fixed point equations) depending on **population-level quantities**
- Independence of entries of $\{\beta_c\}_{c=1}^C$ can be lifted $\rightarrow \bar{\mathcal{R}}$ depends on alignment of β_c 's weighted by the covariance

Asymptotic risk characterization

Theorem [Kovacevic, Cyffers, Mannelli and M., 2026]

Let $p, n \rightarrow \infty$, $p/n \in (0, 1)$, $n_c/n \rightarrow \bar{\pi}_c$, and assume $\{\beta_c\}_{c=1}^C$ have independent, zero-mean entries. Then, the random quantity $\mathcal{R}(\hat{\beta}, \hat{\delta}_2, \dots, \hat{\delta}_C)$ (dependent on data, noise) concentrates to the **deterministic quantity** $\bar{\mathcal{R}}$ (dependent only on $\{\bar{\pi}_c, \Sigma_c, \sigma_c\}_{c=1}^C, \{w_c, \sigma_{\text{DP},c}^2, \lambda_{\delta_c}\}_{c=2}^C, \lambda_\beta$).

- Explicit expression for $\bar{\mathcal{R}}$ (in terms of fixed point equations) depending on **population-level quantities**
- Independence of entries of $\{\beta_c\}_{c=1}^C$ can be lifted $\rightarrow \bar{\mathcal{R}}$ depends on alignment of β_c 's weighted by the covariance
- Regularization and weights can be tuned (Bayes-optimally)

Asymptotic risk characterization

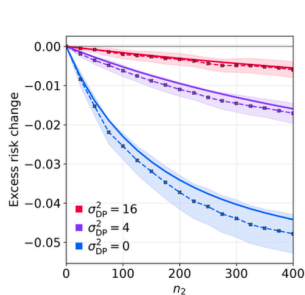
Theorem [Kovacevic, Cyffers, Mannelli and M., 2026]

Let $p, n \rightarrow \infty$, $p/n \in (0, 1)$, $n_c/n \rightarrow \bar{\pi}_c$, and assume $\{\beta_c\}_{c=1}^C$ have independent, zero-mean entries. Then, the random quantity $\mathcal{R}(\hat{\beta}, \hat{\delta}_2, \dots, \hat{\delta}_C)$ (dependent on data, noise) concentrates to the **deterministic quantity** $\bar{\mathcal{R}}$ (dependent only on $\{\bar{\pi}_c, \Sigma_c, \sigma_c\}_{c=1}^C, \{w_c, \sigma_{\text{DP},c}^2, \lambda_{\delta_c}\}_{c=2}^C, \lambda_\beta$).

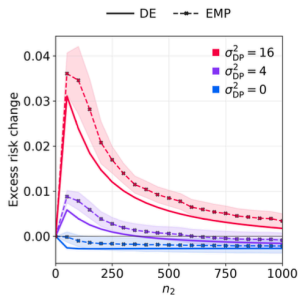
- Explicit expression for $\bar{\mathcal{R}}$ (in terms of fixed point equations) depending on **population-level quantities**
- Independence of entries of $\{\beta_c\}_{c=1}^C$ can be lifted $\rightarrow \bar{\mathcal{R}}$ depends on alignment of β_c 's weighted by the covariance
- Regularization and weights can be tuned (Bayes-optimally)
- Extension to privatization of the features

Experimental evidence

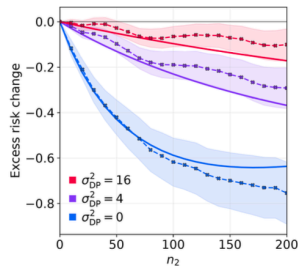
- Real-world regression datasets (likely not linear models)



(a) LSAC



(b) Lianjia

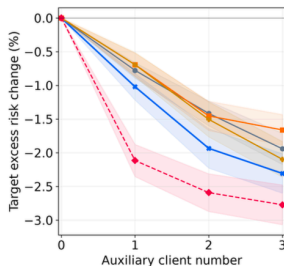


(c) Student Performance

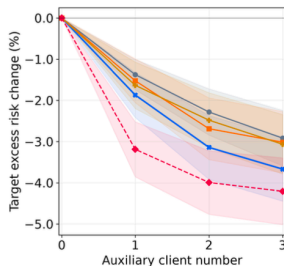
Theoretical predictions match well empirical performance

Experimental evidence

- Real-world regression datasets (likely not linear models)
- Dataset must be selected with **access only to population-level quantities**



(a) LSAC



(b) Lianjia

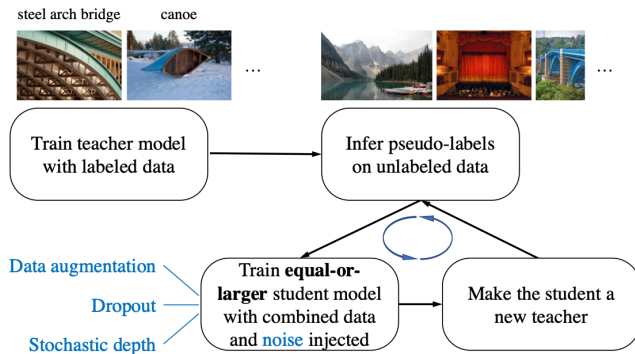


Dataset selection based on theoretical predictions improves on baselines

Learning from surrogate data

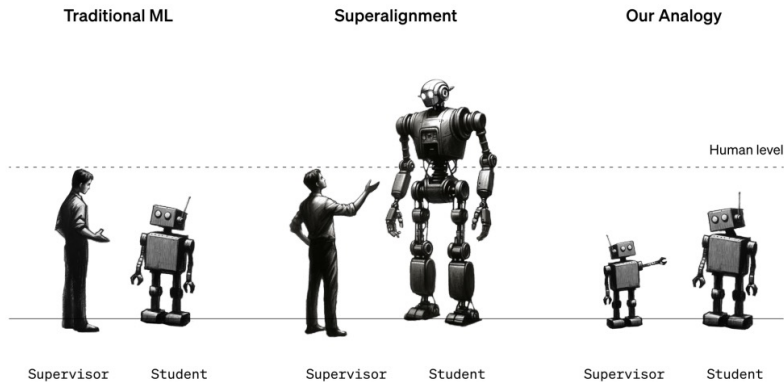
Pre-trained models label data:

- Knowledge distillation [Hinton et al., 2015]
- Self-training [Xie et al., 2020]



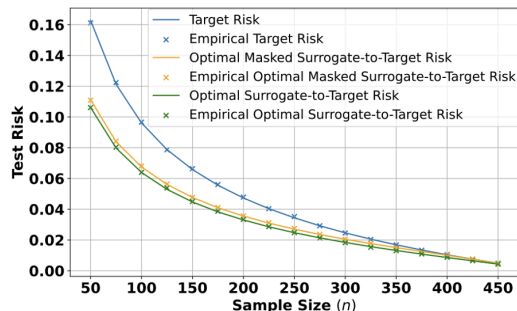
Weak-to-strong generalization

Strong student outperforms weak teacher [Burns et al., 2023]



A surrogate may be better than the ground truth

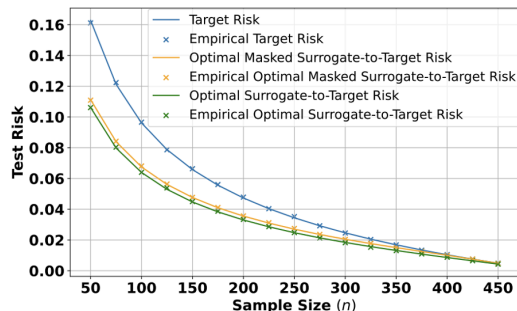
For ridgeless regression in proportional regime, optimal surrogate has lower risk than ground truth



M. E. Ildiz, H. A. Gozeten, E. O. Taga, M. Mondelli, and S. Oymak, "High-dimensional Analysis of Knowledge Distillation: Weak-to-Strong Generalization and Scaling Laws", *ICLR'25 (spotlight)*.

A surrogate may be better than the ground truth, but not by much

For ridgeless regression in proportional regime, optimal surrogate has lower risk than ground truth, but it cannot improve scaling laws!



M. E. Ildiz, H. A. Gozeten, E. O. Taga, M. Mondelli, and S. Oymak, "High-dimensional Analysis of Knowledge Distillation: Weak-to-Strong Generalization and Scaling Laws", *ICLR'25 (spotlight)*.

Formalization of the problem

To what extent can data labeled by an imperfect teacher help?

Random feature ridge regression (RFRR)

Random feature model:

$$\hat{f}(\mathbf{x}; \mathbf{a}) = \frac{1}{\sqrt{p}} \sum_{j=1}^p a_j \varphi(\mathbf{x}; \mathbf{w}_j)$$

Random feature ridge regression (RFRR)

Random feature model: $\{(\mathbf{x}_i, y_i)\}_{i=1}^n \sim_{\text{i.i.d.}} \mathcal{D}, \mathbf{x}_i \in \mathbb{R}^d$

$$\hat{f}(\mathbf{x}; \mathbf{a}) = \frac{1}{\sqrt{p}} \sum_{j=1}^p a_j \varphi(\mathbf{x}; \mathbf{w}_j), \quad y_i = f^*(\mathbf{x}_i) + \varepsilon_i, \quad \mathbf{x}_i \sim \mu_{\mathbf{x}}, \quad \varepsilon_i \sim \mathcal{N}(0, \sigma^2)$$

Random feature ridge regression (RFRR)

Random feature model: $\{(\mathbf{x}_i, y_i)\}_{i=1}^n \sim_{\text{i.i.d.}} \mathcal{D}, \mathbf{x}_i \in \mathbb{R}^d$

$$\hat{f}(\mathbf{x}; \mathbf{a}) = \frac{1}{\sqrt{p}} \sum_{j=1}^p a_j \varphi(\mathbf{x}; \mathbf{w}_j), \quad y_i = f^*(\mathbf{x}_i) + \varepsilon_i, \quad \mathbf{x}_i \sim \mu_{\mathbf{x}}, \quad \varepsilon_i \sim \mathcal{N}(0, \sigma^2)$$

Ridge regression:

$$\hat{\mathbf{a}} = (\mathbf{Z}^\top \mathbf{Z} + \lambda I_p)^{-1} \mathbf{Z}^\top \mathbf{y}$$

- $(\mathbf{Z})_{i,j} = \frac{1}{\sqrt{p}} \varphi(\mathbf{x}_i; \mathbf{w}_j)$ ($\mathbf{Z} \in \mathbb{R}^{n \times p}$), $\mathbf{y} = [y_1, \dots, y_n]^\top \in \mathbb{R}^n$

Formalization of the problem

Teacher: $\{(\mathbf{x}_i^{(t)}, y_i^{(t)})\}_{i=1}^{n_t} \sim_{\text{i.i.d.}} \mathcal{D}^{(t)}$

$$y_i^{(t)} = f^*(\mathbf{x}_i^{(t)}) + \varepsilon_i^{(t)}, \quad \mathbf{x}_i^{(t)} \sim \mu_{\mathbf{x}}, \quad \varepsilon_i^{(t)} \sim \mathcal{N}(0, \sigma_t^2)$$

Formalization of the problem

Teacher: $\{(\mathbf{x}_i^{(t)}, y_i^{(t)})\}_{i=1}^{n_t} \sim_{\text{i.i.d.}} \mathcal{D}^{(t)}$

$$y_i^{(t)} = f^*(\mathbf{x}_i^{(t)}) + \varepsilon_i^{(t)}, \quad \mathbf{x}_i^{(t)} \sim \mu_{\mathbf{x}}, \quad \varepsilon_i^{(t)} \sim \mathcal{N}(0, \sigma_t^2)$$

- $\hat{\mathbf{a}}^{(t)}$ = ridge regression solution, $\hat{f}^{(t)}(\mathbf{x})$ teacher predictor

Formalization of the problem

Teacher: $\{(\mathbf{x}_i^{(t)}, y_i^{(t)})\}_{i=1}^{n_t} \sim \text{i.i.d. } \mathcal{D}^{(t)}$

$$y_i^{(t)} = f^*(\mathbf{x}_i^{(t)}) + \varepsilon_i^{(t)}, \quad \mathbf{x}_i^{(t)} \sim \mu_{\mathbf{x}}, \quad \varepsilon_i^{(t)} \sim \mathcal{N}(0, \sigma_t^2)$$

- $\hat{\mathbf{a}}^{(t)}$ = ridge regression solution, $\hat{f}^{(t)}(\mathbf{x})$ teacher predictor

Student: $\{(\mathbf{x}_i^{(s)}, y_i^{(s)})\}_{i=1}^{n_s} \sim \text{i.i.d. } \mathcal{D}^{(s)}$

$$y_i^{(s)} = \hat{f}^{(t)}(\mathbf{x}_i^{(s)}), \quad \mathbf{x}_i^{(s)} \sim \mu_{\mathbf{x}}$$

Formalization of the problem

Teacher: $\{(\mathbf{x}_i^{(t)}, y_i^{(t)})\}_{i=1}^{n_t} \sim \text{i.i.d. } \mathcal{D}^{(t)}$

$$y_i^{(t)} = f^*(\mathbf{x}_i^{(t)}) + \varepsilon_i^{(t)}, \quad \mathbf{x}_i^{(t)} \sim \mu_{\mathbf{x}}, \quad \varepsilon_i^{(t)} \sim \mathcal{N}(0, \sigma_t^2)$$

- $\hat{\mathbf{a}}^{(t)}$ = ridge regression solution, $\hat{f}^{(t)}(\mathbf{x})$ teacher predictor

Student: $\{(\mathbf{x}_i^{(s)}, y_i^{(s)})\}_{i=1}^{n_s} \sim \text{i.i.d. } \mathcal{D}^{(s)}$

$$y_i^{(s)} = \hat{f}^{(t)}(\mathbf{x}_i^{(s)}), \quad \mathbf{x}_i^{(s)} \sim \mu_{\mathbf{x}}$$

- $\hat{\mathbf{a}}^{(s)}$ = ridge regression solution, $\hat{f}^{(s)}(\mathbf{x})$ teacher predictor

Formalization of the problem

Teacher: $\{(\mathbf{x}_i^{(t)}, y_i^{(t)})\}_{i=1}^{n_t} \sim \text{i.i.d. } \mathcal{D}^{(t)}$

$$y_i^{(t)} = f^*(\mathbf{x}_i^{(t)}) + \varepsilon_i^{(t)}, \quad \mathbf{x}_i^{(t)} \sim \mu_{\mathbf{x}}, \quad \varepsilon_i^{(t)} \sim \mathcal{N}(0, \sigma_t^2)$$

- $\hat{\mathbf{a}}^{(t)}$ = ridge regression solution, $\hat{f}^{(t)}(\mathbf{x})$ teacher predictor

Student: $\{(\mathbf{x}_i^{(s)}, y_i^{(s)})\}_{i=1}^{n_s} \sim \text{i.i.d. } \mathcal{D}^{(s)}$

$$y_i^{(s)} = \hat{f}^{(t)}(\mathbf{x}_i^{(s)}), \quad \mathbf{x}_i^{(s)} \sim \mu_{\mathbf{x}}$$

- $\hat{\mathbf{a}}^{(s)}$ = ridge regression solution, $\hat{f}^{(s)}(\mathbf{x})$ student predictor

Excess population risk:

$$\mathcal{R}(\hat{f}^{(s)}) = \mathbb{E}_{\mathbf{x} \sim \mu_{\mathbf{x}}} \left[(f^*(\mathbf{x}) - \hat{f}^{(s)}(\mathbf{x}))^2 \right]$$

Two-stage learning via random feature ridge regression

- Teacher performs RFRR with n_t samples, p_t features and ridge λ_t

Two-stage learning via random feature ridge regression

- Teacher performs RFRR with n_t samples, p_t features and ridge λ_t
- Student receives fresh samples labeled by trained teacher

Two-stage learning via random feature ridge regression

- Teacher performs RFRR with n_t samples, p_t features and ridge λ_t
- Student receives fresh samples labeled by trained teacher
- Student performs RFRR with n_s samples, p_s features and ridge λ_s

Deterministic equivalent of student risk

Theorem [Wu, Chen, Misiakiewicz and M., 2026]

Under technical assumptions, the random quantity $\mathcal{R}(\hat{f}^{(s)})$ (dependent on data, noise) concentrates to the **deterministic quantity** $\bar{\mathcal{R}}^{(s)}$ (dependent only on model parameters Σ , β^* , teacher parameters n_t, p_t, λ_t , and student parameters n_s, p_s, λ_s):

$$|\mathcal{R}(\hat{f}^{(s)}) - \bar{\mathcal{R}}^{(s)}| \leq \mathcal{E}_{n_s, p_s, n_t, p_t} \tilde{\mathcal{R}}^{(s)}$$

D. Wu, L. Chen, T. Misiakiewicz and M. Mondelli, "Improved Scaling Laws via Weak-to-Strong Generalization in Random Feature Ridge Regression", *ICML*, 2026.

Deterministic equivalent of student risk

Theorem [Wu, Chen, Misiakiewicz and M., 2026]

Under technical assumptions, the random quantity $\mathcal{R}(\hat{f}^{(s)})$ (dependent on data, noise) concentrates to the **deterministic quantity** $\bar{\mathcal{R}}^{(s)}$ (dependent only on model parameters Σ , β^* , teacher parameters n_t, p_t, λ_t , and student parameters n_s, p_s, λ_s):

$$|\mathcal{R}(\hat{f}^{(s)}) - \bar{\mathcal{R}}^{(s)}| \leq \mathcal{E}_{n_s, p_s, n_t, p_t} \tilde{\mathcal{R}}^{(s)}$$

Idea:

1. Deterministic equivalent for student given teacher model
2. Deterministic equivalent for teacher model



D. Wu, L. Chen, T. Misiakiewicz and M. Mondelli, "Improved Scaling Laws via Weak-to-Strong Generalization in Random Feature Ridge Regression", *ICML*, 2026.

Deterministic equivalent of student risk

Theorem [Wu, Chen, Misiakiewicz and M., 2026]

Under technical assumptions, the random quantity $\mathcal{R}(\hat{f}^{(s)})$ (dependent on data, noise) concentrates to the **deterministic quantity** $\bar{\mathcal{R}}^{(s)}$ (dependent only on model parameters Σ , β^* , teacher parameters n_t, p_t, λ_t , and student parameters n_s, p_s, λ_s):

$$|\mathcal{R}(\hat{f}^{(s)}) - \bar{\mathcal{R}}^{(s)}| \leq \mathcal{E}_{n_s, p_s, n_t, p_t} \tilde{\mathcal{R}}^{(s)}$$

- Explicit expression for $\bar{\mathcal{R}}^{(s)}$ (in terms of fixed point equations)

Deterministic equivalent of student risk

Theorem [Wu, Chen, Misiakiewicz and M., 2026]

Under technical assumptions, the random quantity $\mathcal{R}(\hat{f}^{(s)})$ (dependent on data, noise) concentrates to the **deterministic quantity** $\bar{\mathcal{R}}^{(s)}$ (dependent only on model parameters Σ , β^* , teacher parameters n_t, p_t, λ_t , and student parameters n_s, p_s, λ_s):

$$|\mathcal{R}(\hat{f}^{(s)}) - \bar{\mathcal{R}}^{(s)}| \leq \mathcal{E}_{n_s, p_s, n_t, p_t} \tilde{\mathcal{R}}^{(s)}$$

- Explicit expression for $\bar{\mathcal{R}}^{(s)}$ (in terms of fixed point equations)
- Error term $\mathcal{E}_{n_s, p_s, n_t, p_t} = \tilde{O}(n_t^{-1/2} + p_t^{-1/2} + n_s^{-1/2} + p_s^{-1/2})$ in typical cases

Deterministic equivalent of student risk

Theorem [Wu, Chen, Misiakiewicz and M., 2026]

Under technical assumptions, the random quantity $\mathcal{R}(\hat{f}^{(s)})$ (dependent on data, noise) concentrates to the **deterministic quantity** $\bar{\mathcal{R}}^{(s)}$ (dependent only on model parameters Σ , β^* , teacher parameters n_t, p_t, λ_t , and student parameters n_s, p_s, λ_s):

$$|\mathcal{R}(\hat{f}^{(s)}) - \bar{\mathcal{R}}^{(s)}| \leq \mathcal{E}_{n_s, p_s, n_t, p_t} \tilde{\mathcal{R}}^{(s)}$$

- Explicit expression for $\bar{\mathcal{R}}^{(s)}$ (in terms of fixed point equations)
- Error term $\mathcal{E}_{n_s, p_s, n_t, p_t} = \tilde{O}(n_t^{-1/2} + p_t^{-1/2} + n_s^{-1/2} + p_s^{-1/2})$ in typical cases
- Dimension-free (covers $d = \infty$)

Deterministic equivalent of student risk

Theorem [Wu, Chen, Misiakiewicz and M., 2026]

Under technical assumptions, the random quantity $\mathcal{R}(\hat{f}^{(s)})$ (dependent on data, noise) concentrates to the **deterministic quantity** $\bar{\mathcal{R}}^{(s)}$ (dependent only on model parameters Σ , β^* , teacher parameters n_t, p_t, λ_t , and student parameters n_s, p_s, λ_s):

$$|\mathcal{R}(\hat{f}^{(s)}) - \bar{\mathcal{R}}^{(s)}| \leq \mathcal{E}_{n_s, p_s, n_t, p_t} \tilde{\mathcal{R}}^{(s)}$$

- Explicit expression for $\bar{\mathcal{R}}^{(s)}$ (in terms of fixed point equations)
- Error term $\mathcal{E}_{n_s, p_s, n_t, p_t} = \tilde{O}(n_t^{-1/2} + p_t^{-1/2} + n_s^{-1/2} + p_s^{-1/2})$ in typical cases
- Dimension-free (covers $d = \infty$)
- Additive guarantee: $\tilde{\mathcal{R}}^{(s)} = O(1)$ under mild conditions (multiplicative guarantee conjectured)

Student scaling laws

Theorem [Wu, Chen, Misiakiewicz and M., 2026]

Consider the source-capacity conditions $(\Sigma)_{i,i} = i^{-\alpha}$, $(\beta^*)_i = i^{-\frac{(1+2\alpha r)}{2}}$ and the scaling of model width and regularization $p_t = n_t^{\gamma_{p_t}}$, $\lambda_t = n_t^{-\gamma_{\lambda_s}}$, $p_s = n_t^{\gamma_{p_s}}$, $n_s = n_t^{\gamma_{n_s}}$, $\lambda_s = n_t^{-\gamma_{\lambda_s}}$.

Student scaling laws

Theorem [Wu, Chen, Misiakiewicz and M., 2026]

Consider the source-capacity conditions $(\Sigma)_{i,i} = i^{-\alpha}$, $(\beta^*)_i = i^{-\frac{(1+2\alpha r)}{2}}$ and the scaling of model width and regularization $p_t = n_t^{\gamma_{p_t}}$, $\lambda_t = n_t^{-\gamma_{\lambda_t}}$, $p_s = n_t^{\gamma_{p_s}}$, $n_s = n_t^{\gamma_{n_s}}$, $\lambda_s = n_t^{-\gamma_{\lambda_s}}$. Then, the student deterministic equivalent $\bar{\mathcal{R}}^{(s)} = \bar{\mathcal{B}}_{\text{bias}}^{(s)} + \bar{\mathcal{B}}_{\text{var}}^{(s)}$ satisfies

$$\bar{\mathcal{B}}_{\text{bias}}^{(s)} = \Theta \left(n_t^{-\gamma_{\text{bias}}^{(s)}} \right), \quad \bar{\mathcal{B}}_{\text{var}}^{(s)} = \Theta \left(\sigma_t^2 n_t^{-\gamma_{\text{var}}^{(s)}} \right)$$

- Explicit expressions for $\gamma_{\text{bias}}^{(s)} := \gamma_{\text{bias}}^{(s)}(\alpha, r, \gamma_{\lambda_t}, \gamma_{p_t}, \gamma_{\lambda_s}, \gamma_{p_s}, \gamma_{n_s})$,
 $\gamma_{\text{var}}^{(s)} := \gamma_{\text{var}}^{(s)}(\alpha, \gamma_{\lambda_t}, \gamma_{p_t}, \gamma_{\lambda_s}, \gamma_{p_s}, \gamma_{n_s})$

Student scaling laws

Theorem [Wu, Chen, Misiakiewicz and M., 2026]

Consider the source-capacity conditions $(\Sigma)_{i,i} = i^{-\alpha}$, $(\beta^*)_i = i^{-\frac{(1+2\alpha r)}{2}}$ and the scaling of model width and regularization $p_t = n_t^{\gamma_{p_t}}$, $\lambda_t = n_t^{-\gamma_{\lambda_t}}$, $p_s = n_t^{\gamma_{p_s}}$, $n_s = n_t^{\gamma_{n_s}}$, $\lambda_s = n_t^{-\gamma_{\lambda_s}}$. Then, the student deterministic equivalent $\bar{\mathcal{R}}^{(s)} = \bar{\mathcal{B}}_{\text{bias}}^{(s)} + \bar{\mathcal{B}}_{\text{var}}^{(s)}$ satisfies

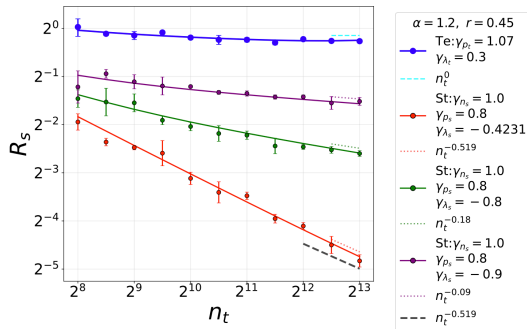
$$\bar{\mathcal{B}}_{\text{bias}}^{(s)} = \Theta \left(n_t^{-\gamma_{\text{bias}}^{(s)}} \right), \quad \bar{\mathcal{B}}_{\text{var}}^{(s)} = \Theta \left(\sigma_t^2 n_t^{-\gamma_{\text{var}}^{(s)}} \right)$$

- Explicit expressions for $\gamma_{\text{bias}}^{(s)} := \gamma_{\text{bias}}^{(s)}(\alpha, r, \gamma_{\lambda_t}, \gamma_{p_t}, \gamma_{\lambda_s}, \gamma_{p_s}, \gamma_{n_s})$,
 $\gamma_{\text{var}}^{(s)} := \gamma_{\text{var}}^{(s)}(\alpha, \gamma_{\lambda_t}, \gamma_{p_t}, \gamma_{\lambda_s}, \gamma_{p_s}, \gamma_{n_s})$
- Can compare with teacher scaling law [Defilippis et al., 2024] to establish when **weak-to-strong generalization** takes place

Weak-to-strong generalization via variance reduction

For variance-dominated teacher, student can **always improve** scaling law via **variance reduction** (choosing properly number of samples, features and ridge).

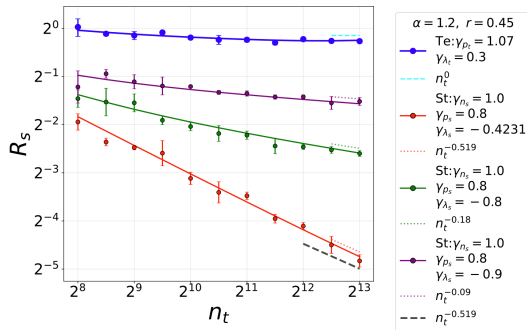
- Three **students** improving the scaling law of their **teacher**



Weak-to-strong generalization via variance reduction

For variance-dominated teacher, student can **always improve** scaling law via **variance reduction** (choosing properly number of samples, features and ridge).

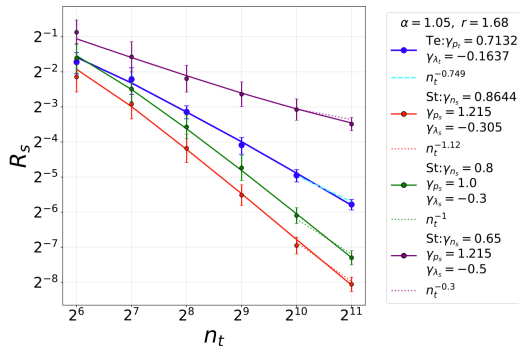
- Three **students** improving the scaling law of their **teacher**
- **Student** achieves **minimax exponent** when **teacher** risk does not even decay with n_t



Weak-to-strong generalization via bias reduction

For bias-dominated teacher, student **may improve** scaling law via **bias reduction**, albeit in more restrictive settings.

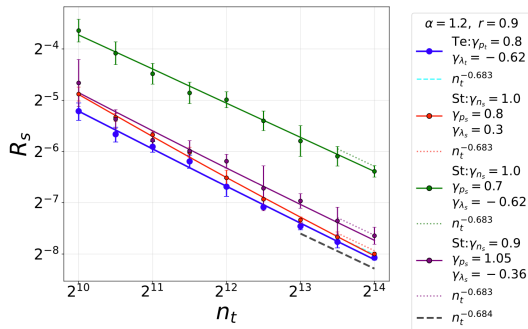
- Earlier work focused on improvement in variance-dominated regimes [Dong et al., 2025]
- Even if **teacher** is trained with clean labels ($\sigma_t = 0$), two **students** still improve scaling law



No weak-to-strong generalization for optimal teachers

For minimax-optimal teacher, **no improvement possible** in scaling law.

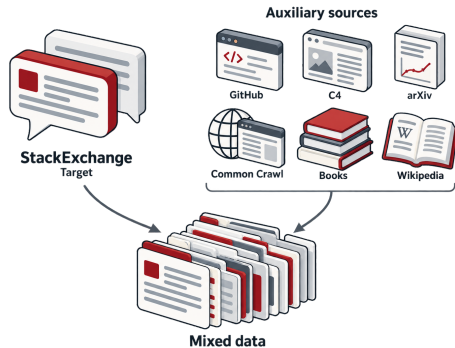
- Three **students** match but do not improve upon **teacher** scaling law
- Scaling law improvement stems from teacher sub-optimality



Learning from mixture data

Lots of empirical work on

- optimizing data mixture [Xie et al., 2023; Fan et al., 2024; Liu et al., 2025; ...]
- empirical scaling laws [Aghajanyan et al., 2023; Shukor et al., 2025; Sedova et al., 2026; ...]



Formalization of the problem

When does data mixing produce synergy?

Here, synergy = scaling law of the mixture better than that of each dataset alone

Formalization of the problem

Training data from K sources: $\{(\mathbf{x}_i^k, y_i^k)\}_{i=1}^{n_k} \sim_{\text{i.i.d.}} \mathcal{D}^{(k)}, \mathbf{x}_i^k \in \mathbb{R}^p, k \in \{1, \dots, K\}$

$$y_i^k = (\mathbf{x}_i^k)^\top \boldsymbol{\beta}^* + z_i^k, \quad \mathbb{E}[\mathbf{x}_i^k] = 0, \quad \mathbb{E}[\mathbf{x}_i^k (\mathbf{x}_i^k)^\top] = \boldsymbol{\Sigma}_k, \quad z_i^k \sim \mathcal{N}(0, \sigma_k^2)$$

- Covariance shift

Formalization of the problem

Training data from K sources: $\{(\mathbf{x}_i^k, y_i^k)\}_{i=1}^{n_k} \sim_{\text{i.i.d.}} \mathcal{D}^{(k)}, \mathbf{x}_i^k \in \mathbb{R}^p, k \in \{1, \dots, K\}$

$$y_i^k = (\mathbf{x}_i^k)^\top \boldsymbol{\beta}^* + z_i^k, \quad \mathbb{E}[\mathbf{x}_i^k] = 0, \quad \mathbb{E}[\mathbf{x}_i^k (\mathbf{x}_i^k)^\top] = \boldsymbol{\Sigma}_k, \quad z_i^k \sim \mathcal{N}(0, \sigma_k^2)$$

- Covariance shift

Mixed ridge regression:

$$\hat{\boldsymbol{\beta}} = \arg \min_{\hat{\mathbf{b}}} \left\{ \sum_{k=1}^K \left\| \mathbf{y}^k - \mathbf{X}^k \hat{\mathbf{b}} \right\|_2^2 + \lambda \left\| \hat{\mathbf{b}} \right\|_2^2 \right\}$$

Formalization of the problem

Training data from K sources: $\{(\mathbf{x}_i^k, y_i^k)\}_{i=1}^{n_k} \sim_{\text{i.i.d.}} \mathcal{D}^{(k)}, \mathbf{x}_i^k \in \mathbb{R}^p, k \in \{1, \dots, K\}$

$$y_i^k = (\mathbf{x}_i^k)^\top \beta^* + z_i^k, \quad \mathbb{E}[\mathbf{x}_i^k] = 0, \quad \mathbb{E}[\mathbf{x}_i^k (\mathbf{x}_i^k)^\top] = \Sigma_k, \quad z_i^k \sim \mathcal{N}(0, \sigma_k^2)$$

- Covariance shift

Mixed ridge regression:

$$\hat{\beta} = \arg \min_{\hat{\mathbf{b}}} \left\{ \sum_{k=1}^K \left\| \mathbf{y}^k - \mathbf{X}^k \hat{\mathbf{b}} \right\|_2^2 + \lambda \left\| \hat{\mathbf{b}} \right\|_2^2 \right\}$$

Excess population risk:

$$\mathcal{R}(\hat{\beta}) = \sum_{k=1}^K \pi_k^* \left(\mathbb{E}_{(\mathbf{x}, y) \sim \mathcal{D}^{(k)}} \left[(y - \mathbf{x}^\top \hat{\beta})^2 \right] - \sigma_k^2 \right)$$

Minimax error

Theorem [Wu, Chen, Misiakiewicz and M., 2026]

1. **Minimax error** under ellipsoid constraint is approximated by variational problem depending only on model parameters $\mathcal{M} = \{\Sigma_k, n_k, \pi_k^*, \sigma_k^2\}_{k=1}^K$:

$$\inf_{\hat{\mathbf{b}}} \sup_{\beta^*} \sum_{k=1}^K \pi_k^* \left(\mathbb{E}_{(\mathbf{x}, y) \sim \mathcal{D}^{(k)}} \left[(y - \mathbf{x}^\top \hat{\beta})^2 \right] - \sigma_k^2 \right) \asymp \sup_{\mathbf{A} \succeq 0} \mathcal{L}(\mathcal{M}, \mathbf{A}).$$

D. Wu, L. Chen, T. Misiakiewicz and M. Mondelli, "When do data mixtures improve scaling laws? Insights from high-dimensional regression", [arXiv:2609.38011](https://arxiv.org/abs/2609.38011), 2026.

Minimax error and deterministic equivalent of the risk

Theorem [Wu, Chen, Misiakiewicz and M., 2026]

1. **Minimax error** under ellipsoid constraint is approximated by variational problem depending only on model parameters $\mathcal{M} = \{\Sigma_k, n_k, \pi_k^*, \sigma_k^2\}_{k=1}^K$:

$$\inf_{\hat{\mathbf{b}}} \sup_{\beta^*} \sum_{k=1}^K \pi_k^* \left(\mathbb{E}_{(\mathbf{x}, y) \sim \mathcal{D}^{(k)}} \left[(y - \mathbf{x}^\top \hat{\beta})^2 \right] - \sigma_k^2 \right) \asymp \sup_{\mathbf{A} \succeq 0} \mathcal{L}(\mathcal{M}, \mathbf{A}).$$

2. When Σ_k commute, random quantity $\mathcal{R}(\hat{\beta})$ (dependent on data, noise) concentrates to **deterministic quantity** $\bar{\mathcal{R}}$ depending only on $\mathcal{M}$:

$$|\mathcal{R}(\hat{\beta}) - \bar{\mathcal{R}}| \leq \mathcal{E}_{n_1, \dots, n_K} \tilde{\mathcal{R}}.$$

D. Wu, L. Chen, T. Misiakiewicz and M. Mondelli, "When do data mixtures improve scaling laws? Insights from high-dimensional regression", arXiv:2609.38011, 2026.

Minimax error

Theorem [Wu, Chen, Misiakiewicz and M., 2026]

1. **Minimax error** under ellipsoid constraint is approximated by variational problem depending only on model parameters $\mathcal{M} = \{\Sigma_k, n_k, \pi_k^*, \sigma_k^2\}_{k=1}^K$:

$$\inf_{\hat{\mathbf{b}}} \sup_{\beta^*} \sum_{k=1}^K \pi_k^* \left(\mathbb{E}_{(\mathbf{x}, y) \sim \mathcal{D}^{(k)}} \left[(y - \mathbf{x}^\top \hat{\beta})^2 \right] - \sigma_k^2 \right) \asymp \sup_{\mathbf{A} \succeq 0} \mathcal{L}(\mathcal{M}, \mathbf{A}).$$

2. When Σ_k commute, random quantity $\mathcal{R}(\hat{\beta})$ (dependent on data, noise) concentrates to **deterministic quantity** $\bar{\mathcal{R}}$ depending only on $\mathcal{M}$:

$$|\mathcal{R}(\hat{\beta}) - \bar{\mathcal{R}}| \leq \mathcal{E}_{n_1, \dots, n_K} \tilde{\mathcal{R}}.$$

- Can derive scaling laws under source-capacity conditions ($K = 2$)

D. Wu, L. Chen, T. Misiakiewicz and M. Mondelli, "When do data mixtures improve scaling laws? Insights from high-dimensional regression", arXiv:2609.38011, 2026.

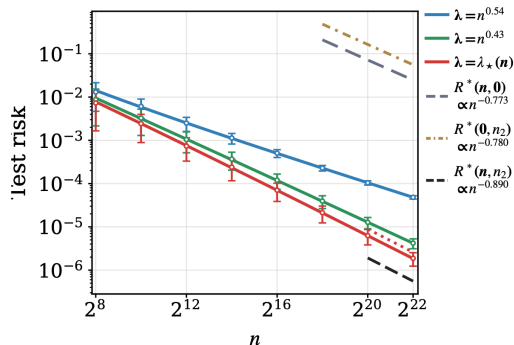
Scaling laws for mixture data

Whenever **mixed minimax rate is strictly better** than minimax rate of either dataset alone (source-only, auxiliary-only), **ridge regression achieves mixed minimax rate**.

Scaling laws for mixture data

Whenever **mixed minimax rate is strictly better** than minimax rate of either dataset alone (source-only, auxiliary-only), **ridge regression achieves mixed minimax rate**.

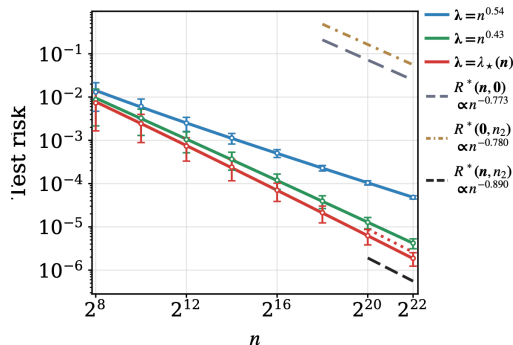
- **Mixed** minimax rate better than **source-only** and **target-only** rates
- **Optimally regularized ridge** achieves mixed minimax rate



Scaling laws for mixture data

Whenever **mixed minimax rate is strictly better** than minimax rate of either dataset alone (source-only, auxiliary-only), **ridge regression achieves mixed minimax rate**.

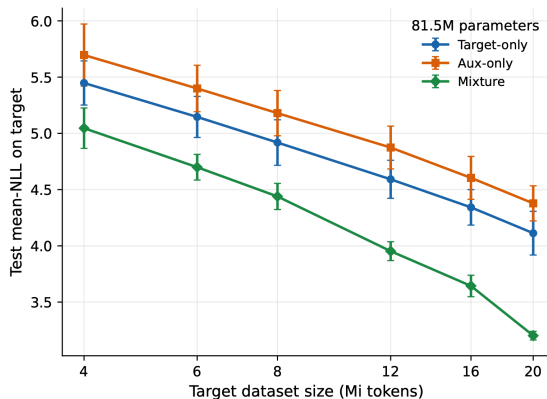
- **Mixed** minimax rate better than **source-only** and **target-only** rates
- **Optimally regularized ridge** achieves mixed minimax rate
- Size of auxiliary dataset needs to be “just about right”
- Auxiliary tails need to be heavier



Experimental evidence

Same qualitative phenomena in language models:

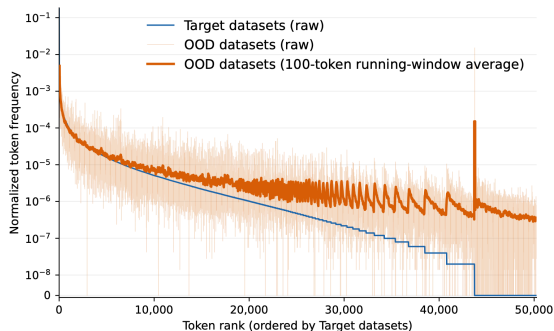
- Size of auxiliary dataset needs to be “just about right”



Experimental evidence

Same qualitative phenomena in language models:

- Size of auxiliary dataset needs to be “just about right”
- Auxiliary tails need to be heavier



Take-home message

Insights from **asymptotic risk characterizations** and **deterministic equivalents** on learning from data of all sorts (**synthetic**):



How to select synthetic data?

Take-home message

Insights from **asymptotic risk characterizations** and **deterministic equivalents** on learning from data of all sorts (**synthetic**, **private**):



How to select synthetic data?



When to collaborate with an external provider holding private data?

Take-home message

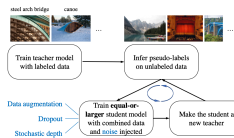
Insights from **asymptotic risk characterizations** and **deterministic equivalents** on learning from data of all sorts (**synthetic**, **private**, **surrogate**):



How to select synthetic data?



When to collaborate with an external provider holding private data?



To what extent can data labeled by an imperfect teacher help?

Take-home message

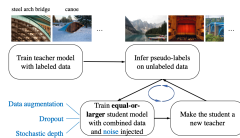
Insights from **asymptotic risk characterizations** and **deterministic equivalents** on learning from data of all sorts (**synthetic**, **private**, **surrogate** and **mixture**):



How to select synthetic data?



When to collaborate with an external provider holding private data?



To what extent can data labeled by an imperfect teacher help?



When does data mixing produce synergy?

Take-home message

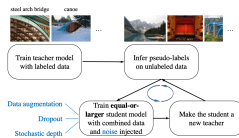
Insights from **asymptotic risk characterizations** and **deterministic equivalents** on learning from data of all sorts (**synthetic**, **private**, **surrogate** and **mixture**):



How to select synthetic data?



When to collaborate with an external provider holding private data?



To what extent can data labeled by an imperfect teacher help?



When does data mixing produce synergy?

Thank you for your attention!