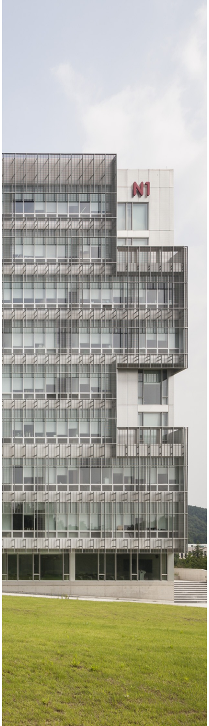


High-dimensional Analysis of Knowledge Distillation: Weak-to-Strong Generalization and Scaling Laws

Arxiv 2024

M. Emrullah Ildiz, Halil Alperen Gozeten, Ege Onur Taga, Marco Mondelli, Samet Oymak

Problem Setup

Stage 1: Surrogate Model.

- Data distribution $(\tilde{\boldsymbol{x}}, \tilde{y}) \sim \mathcal{D}_s$
  - $\tilde{y} = \tilde{\boldsymbol{x}}^\top \boldsymbol{\beta}_\star + \tilde{z}$
  - $\boldsymbol{\beta}_\star \in \mathbb{R}^p$, p: Feature dimension
  - $\tilde{\boldsymbol{x}} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}_s)$, $\boldsymbol{\Sigma}_s$: Covariance matrix of distribution $\mathcal{D}_s$
  - $\tilde{z} \sim \mathcal{N}(0, \sigma_s^2)$: independent of $\tilde{\boldsymbol{x}}$
- Surrogate dataset $\{(\tilde{\boldsymbol{x}}_i, \tilde{y}_i)_{i=1}^m\}$: i.i.d. from $\mathcal{D}_s$
- Estimator $\boldsymbol{\beta}^s$: both under- and over-parameterized settings
  - $\tilde{\boldsymbol{X}} = [\tilde{\boldsymbol{x}}_1^\top, \ldots, \tilde{\boldsymbol{x}}_m^\top]^\top \in \mathbb{R}^{m \times p}$, $\tilde{\boldsymbol{y}} = [\tilde{y}_1, \ldots, \tilde{y}_m]^\top \in \mathbb{R}^m$
  - Under-parametrized ($m \geq p$): Quadratic loss
  - Over-parametrized ($m < p$): Minimum norm interpolator

$$\boldsymbol{\beta}^s = \mathrm{Est}(\tilde{\boldsymbol{X}}, \boldsymbol{y}) := \begin{cases} \arg\min_{\boldsymbol{\beta}} \|\tilde{\boldsymbol{y}} - \boldsymbol{X}\boldsymbol{\beta}\|_2^2, & \text{if } m \geq p, \\ \arg\min_{\boldsymbol{\beta}} \{\|\boldsymbol{\beta}\|_2^2 : \tilde{\boldsymbol{X}}\boldsymbol{\beta} = \tilde{\boldsymbol{y}}\} & \text{if } m < p, \end{cases}$$

KAIST

Problem Setup

Stage 2: Target Model.

- Data distribution $(\boldsymbol{x}, y^s) \sim \mathcal{D}_t(\boldsymbol{\beta}^s)$
  - $y^s = \boldsymbol{x}^\top \boldsymbol{\beta}^s + z$
  - $\boldsymbol{x} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}_t)$, $\boldsymbol{\Sigma}_t$: Covariance matrix of distribution $\mathcal{D}_t(\boldsymbol{\beta}^s)$
  - $z \sim \mathcal{N}(0, \sigma_t^2)$: independent of $\boldsymbol{x}$
- Target Dataset $\{(\boldsymbol{x}_i, y_i^s)_{i=1}^n\}$: i.i.d. from $\mathcal{D}_t(\boldsymbol{\beta}^s)$.
- Estimator $\boldsymbol{\beta}^{s2t}$

$$\boldsymbol{\beta}^{s2t} = \mathrm{Est}(\boldsymbol{X}, \boldsymbol{y}^s),$$

where $\boldsymbol{X} = [\boldsymbol{x}_1^\top, \ldots, \boldsymbol{x}_n^\top]^\top \in \mathbb{R}^{n \times p}$, $\boldsymbol{y}^s = [y_1^s, \ldots, y_n^s]^\top \in \mathbb{R}^n$

Excess (population) Risk for any estimator $\hat{\boldsymbol{\beta}}$

$$\mathcal{R}(\hat{\boldsymbol{\beta}}) := \mathbb{E}_{(\boldsymbol{x},y)\sim\mathcal{D}_t(\boldsymbol{\beta}_\star)}[(y - \boldsymbol{x}^\top \hat{\boldsymbol{\beta}})^2] - \sigma_t^2 = \|\boldsymbol{\Sigma}_t^{1/2}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_\star)\|_2^2.$$

$\Rightarrow$ How close the estimator $\hat{\boldsymbol{\beta}}$ is to $\boldsymbol{\beta}_\star$?

Two Reference Models

Reference 1: Standard Target Model. (No Distillation)

- Access to the ground-truth parameter $\boldsymbol{\beta}_\star$ through labeling
- Target Dataset $\{(\boldsymbol{x}_i, y_i)_{i=1}^n\}$: i.i.d. from $\mathcal{D}_t(\boldsymbol{\beta}_\star)$.
- Estimator $\boldsymbol{\beta}^t$:

$$\boldsymbol{\beta}^t := \mathrm{Est}(\boldsymbol{X}, \boldsymbol{y}),$$

where $\boldsymbol{X} = [\boldsymbol{x}_1^\top, \ldots, \boldsymbol{x}_n^\top]^\top \in \mathbb{R}^{n\times p}$, $\boldsymbol{y} = [y_1, \ldots, y_n]^\top \in \mathbb{R}^n$.

Reference 2: Covariance Shift model. [1, 2]

- Data distribution $(\boldsymbol{x}, y) \sim \mathcal{D}_s^{cs}$
  - $y = \boldsymbol{x}^\top \boldsymbol{\beta}_\star + z$
  - $\boldsymbol{x} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}_s)$, $z \sim \mathcal{N}(0, \sigma_t^2)$
- Target Dataset $\{(\boldsymbol{x}_i, y_i)_{i=1}^n\}$: i.i.d. from $\mathcal{D}_s^{cs}$
- Estimator $\hat{\boldsymbol{\beta}}^{cs}$:

$$\hat{\boldsymbol{\beta}}^{cs} := \mathrm{Est}(\boldsymbol{X}, \boldsymbol{y}),$$

where $\boldsymbol{X} = [\boldsymbol{x}_1^\top, \ldots, \boldsymbol{x}_n^\top]^\top \in \mathbb{R}^{n\times p}$, $\boldsymbol{y} = [y_1, \ldots, y_n]^\top \in \mathbb{R}^n$.

KAIST

Distribution of Ridge(less) Estimator from [3]

Setup: Linear Regression Model

- $y = \boldsymbol{x}^\top \boldsymbol{\beta}_0 + z$, $\boldsymbol{x}, \boldsymbol{\beta} \in \mathbb{R}^p$, $y, z \in \mathbb{R}$
- $\mathbb{E}[\boldsymbol{x}] = 0$, $Cov(\boldsymbol{x}) = \boldsymbol{\Sigma} \in \mathbb{R}^{p \times p}$, $\mathbb{E}[z] = 0$, $Var(z) = \sigma^2$.
- $\boldsymbol{X} = [\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n]^\top \in \mathbb{R}^{n \times p}$, $\boldsymbol{y} = [y_1, \ldots, y_n]^\top = \mathbb{R}^n$.

Estimator: Ridge estimator $\hat{\boldsymbol{\beta}}$ with regularization $\eta > 0$ is defined as

$$\hat{\boldsymbol{\beta}} = \operatorname*{arg\,min}_{\boldsymbol{\beta} \in \mathbb{R}^p} \left\{ \frac{1}{2p} \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\|^2 + \frac{\eta}{2} \|\boldsymbol{\beta}\|^2 \right\} = \frac{1}{n} \left(\frac{1}{n} \boldsymbol{X}^\top \boldsymbol{X} + \eta \boldsymbol{I}_n \right)^{-1} \boldsymbol{X}^\top \boldsymbol{y}$$

and the Ridgeless estimator (also known as minimum-norm interpolator) is defined as

$$\hat{\boldsymbol{\beta}} = \operatorname*{arg\,min}_{\boldsymbol{\beta} \in \mathbb{R}^p} \{\|\boldsymbol{\beta}\|^2 : \boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta}\} = (\boldsymbol{X}^\top \boldsymbol{X})^- \boldsymbol{X}^\top \boldsymbol{y}, \quad \boldsymbol{A}^- : \text{ pesudo-inverse of } \boldsymbol{A}$$

Asymptotic Regime: <u>Fixed</u> $\kappa_t = p/n > 1$ with $n, p \to \infty$.

Gaussian Sequence Model: For a given pair of $(\boldsymbol{\Sigma}, \boldsymbol{\beta}_0)$ and noise level $\gamma > 0$,

$$y^{seq}_{(\boldsymbol{\Sigma}, \boldsymbol{\beta}_0)}(\gamma) := \boldsymbol{\Sigma}^{1/2} \boldsymbol{\beta}_0 + \frac{\gamma \boldsymbol{g}}{\sqrt{p}}, \quad \boldsymbol{g} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_p).$$

KAIST

Distribution of Ridge(less) Estimator from [3]

Ridge estimator with regularization $\tau \geq 0$ in the Gaussian sequence model:

$$\hat{\boldsymbol{\beta}}^{seq}_{(\boldsymbol{\Sigma},\boldsymbol{\beta}_0)}(\gamma;\tau) := \underset{\boldsymbol{\beta}\in\mathbb{R}^p}{\arg\min}\left\{\frac{1}{2}\left\|\boldsymbol{\Sigma}^{1/2}\boldsymbol{\beta} - y^{seq}_{(\boldsymbol{\Sigma},\boldsymbol{\beta}_0)}(\gamma)\right\|^2 + \frac{\tau}{2}\|\boldsymbol{\beta}\|^2\right\}$$
$$= (\boldsymbol{\Sigma} + \tau\boldsymbol{I}_p)^{-1}\boldsymbol{\Sigma}^{1/2}\left(\boldsymbol{\Sigma}^{1/2}\boldsymbol{\beta}_0 + \frac{\gamma\boldsymbol{g}}{\sqrt{p}}\right)$$

Distributional Characterization (Informal).

For any $\eta \geq 0$, there exists a unique pair $(\gamma_{\eta,*}, \tau_{\eta,*}) \in (0,\infty)^2$ determined via a fixed point equation, such that the distribution $\hat{\boldsymbol{\beta}}$ is about the same as that of $\hat{\boldsymbol{\beta}}^{seq}_{(\boldsymbol{\Sigma},\boldsymbol{\beta})}(\gamma_{\eta,*};\tau_{\eta,*})$.

Formally, for any 1-Lipschitz function $g : \mathbb{R}^n \to \mathbb{R}$ and any $K > 0$, with high probability,

$$\sup_{\eta\in[0,K]}\left|g(\hat{\boldsymbol{\beta}}) - \mathbb{E}\left[g(\hat{\boldsymbol{\beta}}^{seq}_{(\boldsymbol{\Sigma},\boldsymbol{\beta})}(\gamma_{\eta,*};\tau_{\eta,*}))\right]\right| \approx 0.$$

Main Idea from [3]: Analyze the behavior of Ridge(less) estimator $\hat{\boldsymbol{\beta}}$ using $\hat{\boldsymbol{\beta}}^{seq}_{(\boldsymbol{\Sigma},\boldsymbol{\beta})}(\gamma_{\eta,*};\tau_{\eta,*})$!

Fixed Point Equation from [3]

For $\gamma, \tau > 0$, define the error $\mathsf{err}_{(\boldsymbol{\Sigma},\boldsymbol{\beta}_0)}(\gamma;\tau)$ and the degrees-of-freedom $\mathsf{dof}_{(\boldsymbol{\Sigma},\boldsymbol{\beta}_0)}(\gamma;\tau)$ as

$$\mathsf{err}_{(\boldsymbol{\Sigma},\boldsymbol{\beta}_0)}(\gamma;\tau) := \left\|\boldsymbol{\Sigma}^{1/2}\left(\hat{\boldsymbol{\beta}}^{seq}_{(\boldsymbol{\Sigma},\boldsymbol{\beta})}(\gamma;\tau) - \boldsymbol{\beta}_0\right)\right\|^2$$
$$\mathsf{dof}_{(\boldsymbol{\Sigma},\boldsymbol{\beta}_0)}(\gamma;\tau) := \left\langle \frac{\gamma \boldsymbol{g}}{\sqrt{p}}, \left(\hat{\boldsymbol{\beta}}^{seq}_{(\boldsymbol{\Sigma},\boldsymbol{\beta})}(\gamma;\tau) - \boldsymbol{\beta}_0\right)\right\rangle$$

Fixed Point Equation: For $\eta \geq 0$,

$$\begin{cases} \kappa_t^{-1}\gamma^2 = \sigma^2 + \mathbb{E}[\mathsf{err}_{(\boldsymbol{\Sigma},\boldsymbol{\beta}_0)}(\gamma;\tau)], \\ \kappa_t^{-1} - \frac{\eta}{\tau} = \frac{1}{p}\mathtt{tr}((\boldsymbol{\Sigma} + \tau \boldsymbol{I}_p)^{-1}\boldsymbol{\Sigma}) = \frac{1}{\gamma^2}\mathbb{E}[\mathsf{dof}_{(\boldsymbol{\Sigma},\boldsymbol{\beta}_0)}(\gamma;\tau)]. \end{cases}$$

For any $\eta \geq 0$, the fixed equation has a unique solution $(\gamma_{\eta,*}, \tau_{\eta,*}) \in (0,\infty)^2$.

Returning to the Original Target Model

Recall the setup of the target model:

- Data distribution $(\boldsymbol{x}, y^s) \sim \mathcal{D}_t(\boldsymbol{\beta}^s)$
  - $y^s = \boldsymbol{x}^\top \boldsymbol{\beta}^s + z$
  - $\boldsymbol{x} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}_t)$, $\boldsymbol{\Sigma}_t$: Covariance matrix of distribution $\mathcal{D}_t(\boldsymbol{\beta}^s)$
  - $z \sim \mathcal{N}(0, \sigma_t^2)$: independent of $\boldsymbol{x}$
- Target Dataset $\{(\boldsymbol{x}_i, y_i^s)_{i=1}^n\}$: i.i.d. from $\mathcal{D}_t(\boldsymbol{\beta}^s)$.
- Estimator $\boldsymbol{\beta}^{s2t}$

$$\boldsymbol{\beta}^{s2t} = \mathrm{Est}(\boldsymbol{X}, \boldsymbol{y}^s),$$

where $\boldsymbol{X} = [\boldsymbol{x}_1^\top, \ldots, \boldsymbol{x}_n^\top]^\top \in \mathbb{R}^{n \times p}$, $\boldsymbol{y}^s = [y_1^s, \ldots, y_n^s]^\top \in \mathbb{R}^n$

Main Idea: $\boldsymbol{\beta}^{s2t} \approx \hat{\boldsymbol{\beta}}^{seq}_{(\boldsymbol{\Sigma}_t, \boldsymbol{\beta}^s)}(\gamma_t; \tau_t)$, where γ_t, τ_t satisfies the fixed point equation

$$\begin{cases} \gamma_t^2 = \kappa_t \left(\sigma_t^2 + \mathbb{E}[\mathsf{err}_{(\boldsymbol{\Sigma}_t, \boldsymbol{\beta}^s)}(\gamma_t; \tau_t)]\right), \\ \kappa_t^{-1} = \frac{1}{p}\mathtt{tr}((\boldsymbol{\Sigma}_t + \tau_t \boldsymbol{I}_p)^{-1}\boldsymbol{\Sigma}_t) \end{cases}$$

and

$$\hat{\boldsymbol{\beta}}^{seq}_{(\boldsymbol{\Sigma}_t, \boldsymbol{\beta}^s)}(\gamma_t; \tau_t) = (\boldsymbol{\Sigma}_t + \tau_t \boldsymbol{I}_p)^{-1}\boldsymbol{\Sigma}_t^{1/2}\left(\boldsymbol{\Sigma}_t^{1/2}\boldsymbol{\beta}^s + \frac{\gamma_t \boldsymbol{g}_t}{\sqrt{p}}\right), \quad \boldsymbol{g}_t \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I}_p).$$

KAIST

Asymptotic Risk Estimate

Excess Risk $\mathcal{R}(\hat{\boldsymbol{\beta}})$ for any estimator $\hat{\boldsymbol{\beta}} \in \mathbb{R}^p$:

$$\mathcal{R}(\hat{\boldsymbol{\beta}}) := \mathbb{E}_{(\boldsymbol{x},y)\sim\mathcal{D}_t(\boldsymbol{\beta}_\star)}[(y - \boldsymbol{x}^\top \hat{\boldsymbol{\beta}})^2] - \sigma_t^2 = \|\boldsymbol{\Sigma}_t^{1/2}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_\star)\|_2^2.$$

Risk of the surrogate-to-target problem: $\mathcal{R}(\boldsymbol{\beta}^{s2t})$:

$$\mathcal{R}(\boldsymbol{\beta}^{s2t}) = \|\boldsymbol{\Sigma}_t^{1/2}(\boldsymbol{\beta}^{s2t} - \boldsymbol{\beta}_\star)\|_2^2$$

Asymptotic Risk Estimate $\bar{\mathcal{R}}^{s2t}_{\kappa_t,\sigma_t}(\boldsymbol{\Sigma}_t, \boldsymbol{\beta}_\star, \boldsymbol{\beta}^s)$:

$$\bar{\mathcal{R}}^{s2t}_{\kappa_t,\sigma_t}(\boldsymbol{\Sigma}_t, \boldsymbol{\beta}_\star, \boldsymbol{\beta}^s) = \mathbb{E}_{\boldsymbol{g}_t}[\mathcal{R}(\hat{\boldsymbol{\beta}}^{seq}_{(\boldsymbol{\Sigma}_t,\boldsymbol{\beta}^s)}(\gamma_t;\tau_t))] = \mathbb{E}_{\boldsymbol{g}_t}\left[\left\|\boldsymbol{\Sigma}_t^{1/2}\left(\hat{\boldsymbol{\beta}}^{seq}_{(\boldsymbol{\Sigma}_t,\boldsymbol{\beta}^s)}(\gamma_t;\tau_t) - \boldsymbol{\beta}_\star\right)\right\|_2^2\right]$$

Recall the definition of $\hat{\boldsymbol{\beta}}^{seq}_{(\boldsymbol{\Sigma}_t,\boldsymbol{\beta}^s)}(\gamma_t;\tau_t)$:

$$\begin{aligned}
\hat{\boldsymbol{\beta}}^{seq}_{(\boldsymbol{\Sigma}_t,\boldsymbol{\beta}^s)}(\gamma_t;\tau_t) &= (\boldsymbol{\Sigma}_t + \tau_t \boldsymbol{I}_p)^{-1}\boldsymbol{\Sigma}_t^{1/2}\left(\boldsymbol{\Sigma}^{1/2}\boldsymbol{\beta}^s + \frac{\gamma_t \boldsymbol{g}_t}{\sqrt{p}}\right) \\
&= \underbrace{(\boldsymbol{\Sigma}_t + \tau_t \boldsymbol{I}_p)^{-1}\boldsymbol{\Sigma}_t}_{:=\boldsymbol{\theta}_1}\boldsymbol{\beta}^s + \gamma_t \underbrace{(\boldsymbol{\Sigma}_t + \tau_t \boldsymbol{I}_p)^{-1}\boldsymbol{\Sigma}_t^{1/2}\frac{\boldsymbol{g}_t}{\sqrt{p}}}_{:=\boldsymbol{\theta}_2} \\
&= \boldsymbol{\theta}_1\boldsymbol{\beta}^s + \gamma_t\boldsymbol{\theta}_2
\end{aligned}$$

Note that $\mathbb{E}_{\boldsymbol{g}_t}[\boldsymbol{\theta}_2] = 0$ and $\boldsymbol{\theta}_1$ is not a function of $\boldsymbol{g}_t$.

$$\begin{aligned}\bar{\mathcal{R}}^{s2t}_{\kappa_t,\sigma_t}(\boldsymbol{\Sigma}_t, \boldsymbol{\beta}_\star, \boldsymbol{\beta}^s) &= \mathbb{E}_{\boldsymbol{g}_t}\left[\left(\boldsymbol{\theta}_1\boldsymbol{\beta}^s + \gamma_t\boldsymbol{\theta}_2 - \boldsymbol{\beta}_\star\right)^\top \boldsymbol{\Sigma}_t \left(\boldsymbol{\theta}_1\boldsymbol{\beta}^s + \gamma_t\boldsymbol{\theta}_2 - \boldsymbol{\beta}_\star\right)\right] \\ &= \left(\boldsymbol{\theta}_1\boldsymbol{\beta}^s - \boldsymbol{\beta}_\star\right)^\top \boldsymbol{\Sigma}_t \left(\boldsymbol{\theta}_1\boldsymbol{\beta}^s - \boldsymbol{\beta}_\star\right) + \gamma_t^2\mathbb{E}_{\boldsymbol{g}_t}[\boldsymbol{\theta}_2^\top\boldsymbol{\Sigma}_t\boldsymbol{\theta}_2]\end{aligned}$$

The former term can be expressed as

$$\begin{aligned}&(\boldsymbol{\theta}_1(\boldsymbol{\beta}^s - \boldsymbol{\beta}_\star) - (\boldsymbol{I} - \boldsymbol{\theta}_1)\boldsymbol{\beta}_\star)^\top\boldsymbol{\Sigma}_t(\boldsymbol{\theta}_1(\boldsymbol{\beta}^s - \boldsymbol{\beta}_\star) - (\boldsymbol{I} - \boldsymbol{\theta}_1)\boldsymbol{\beta}_\star) \\ &\quad = (\boldsymbol{\beta}^s - \boldsymbol{\beta}_\star)^\top\boldsymbol{\theta}_1^\top\boldsymbol{\Sigma}_t\boldsymbol{\theta}_1(\boldsymbol{\beta}^s - \boldsymbol{\beta}_\star) + \boldsymbol{\beta}_\star^\top(\boldsymbol{I} - \boldsymbol{\theta}_1)^\top\boldsymbol{\Sigma}_t(\boldsymbol{I} - \boldsymbol{\theta}_1)\boldsymbol{\beta}_\star \\ &\quad\quad - 2\boldsymbol{\beta}_\star^\top(\boldsymbol{I} - \boldsymbol{\theta}_1)^\top\boldsymbol{\Sigma}_t(\boldsymbol{\beta}^s - \boldsymbol{\beta}_\star)\end{aligned}$$

Hence, we can formulate Asymptotic risk estimate as

$$\begin{aligned}\bar{\mathcal{R}}^{s2t}_{\kappa_t,\sigma_t}(\boldsymbol{\Sigma}_t, \boldsymbol{\beta}_\star, \boldsymbol{\beta}^s) &:= (\boldsymbol{\beta}^s - \boldsymbol{\beta}_\star)^\top\boldsymbol{\theta}_1^\top\boldsymbol{\Sigma}_t\boldsymbol{\theta}_1(\boldsymbol{\beta}^s - \boldsymbol{\beta}_\star) + \gamma_t^2\mathbb{E}_{\boldsymbol{g}_t}[\boldsymbol{\theta}_2^\top\boldsymbol{\Sigma}_t\boldsymbol{\theta}_2] \\ &\quad + \boldsymbol{\beta}_\star^\top(\boldsymbol{I} - \boldsymbol{\theta}_1)^\top\boldsymbol{\Sigma}_t(\boldsymbol{I} - \boldsymbol{\theta}_1)\boldsymbol{\beta}_\star - 2\boldsymbol{\beta}_\star^\top(\boldsymbol{I} - \boldsymbol{\theta}_1)^\top\boldsymbol{\Sigma}_t(\boldsymbol{\beta}^s - \boldsymbol{\beta}_\star)\end{aligned}$$

Recall the fixed point equation:

$$\begin{cases}\gamma_t^2 = \kappa_t\left(\sigma_t^2 + \mathbb{E}[\mathsf{err}_{(\boldsymbol{\Sigma}_t,\boldsymbol{\beta}^s)}(\gamma_t;\tau_t)]\right) = \kappa_t\left(\sigma_t^2 + \bar{\mathcal{R}}^{s2t}_{\kappa_t,\sigma_t}(\boldsymbol{\Sigma}_t, \boldsymbol{\beta}^s, \boldsymbol{\beta}^s)\right) \\ \kappa_t^{-1} = \frac{1}{p}\mathtt{tr}((\boldsymbol{\Sigma}_t + \tau_t\boldsymbol{I}_p)^{-1}\boldsymbol{\Sigma}_t)\end{cases}$$

Non-Asymptotic Characterization of the Risk

Theorem 1 (Theorem 2.3 in [3])

Suppose that some constant $M_t > 1$, we have $1/M_t \le \kappa_t, \sigma_t^2 \le M_t$ and $\|\boldsymbol{\Sigma}_t\|_{op}$, $\|\boldsymbol{\Sigma}_t^{-1}\|_{op} \le M_t$. Let $\boldsymbol{B}_p(R) := \{\boldsymbol{x} \in \mathbb{R}^p : \|\boldsymbol{x}\|_2 \le R\}$. Then, there exists a constant $C = C(M_t)$ such that, for any $\epsilon \in (0, 1/2]$, the following holds with $R + 1 < M_t$:

$$\sup_{\boldsymbol{\beta}_\star, \boldsymbol{\beta}^s \in \boldsymbol{B}_p(R)} \boldsymbol{P}\left(|\mathcal{R}(\boldsymbol{\beta}^{s2t}) - \bar{\mathcal{R}}^{s2t}_{\kappa_t,\sigma_t}(\boldsymbol{\Sigma}_t, \boldsymbol{\beta}_\star, \boldsymbol{\beta}^s)| \ge \epsilon\right) \le Cpe^{-p\epsilon^4/C}.$$

Meaning of Theorem 1: In the asymptotic regime, $\mathcal{R}(\boldsymbol{\beta}^{s2t}) \xrightarrow{p} \bar{\mathcal{R}}^{s2t}_{\kappa_t,\sigma_t}(\boldsymbol{\Sigma}_t, \boldsymbol{\beta}_\star, \boldsymbol{\beta}^s)$

Proposition 1

Let $\Omega = \frac{\mathrm{tr}(\boldsymbol{\Sigma}_t^2(\boldsymbol{\Sigma}_t + \tau_t \boldsymbol{I}_p)^{-2})}{n}$. The optimal surrogate $\boldsymbol{\beta}^s$ minimizing the asymptotic risk $\bar{\mathcal{R}}^{s2t}_{\kappa_t,\sigma_t}(\boldsymbol{\Sigma}_t, \boldsymbol{\beta}_\star, \boldsymbol{\beta}^s)$ is

$$\boldsymbol{\beta}^{s*} = \left((\boldsymbol{\Sigma}_t + \tau_t \boldsymbol{I}_p)^{-1}\boldsymbol{\Sigma}_t + \frac{\Omega\tau_t^2}{1-\Omega}\boldsymbol{\Sigma}_t^{-1}(\boldsymbol{\Sigma}_t + \tau_t \boldsymbol{I}_p)^{-1}\right)\boldsymbol{\beta}_\star.$$

Proof: Differentiate $\bar{\mathcal{R}}^{s2t}_{\kappa_t,\sigma_t}(\boldsymbol{\Sigma}_t, \boldsymbol{\beta}_\star, \boldsymbol{\beta}^s)$ with respect to $\boldsymbol{\beta}^s$.

Optimal Surrogate Parameter is not $\beta_\star$

Corollary 1

Without loss of generality, suppose that $\mathbf{\Sigma}_t$ is diagonal. Let $(\lambda_i)_{i=1}^p$ be the eigenvalues of $\mathbf{\Sigma}_t$ in non-increasing order and let $\xi_i = \frac{\tau_t}{\lambda_i + \tau_t}$ for $i \in [p]$. Then, the following results hold:

1. $(\beta^{s*})_i = (\beta_\star)_i \left((1 - \xi_i) + \xi_i \frac{\Omega}{1-\Omega} \frac{\xi_i}{1-\xi_i} \right)^{-1}$ for every $i \in [p]$.
2. $|(\beta^{s*})_i| > |(\beta_\star)_i|$ if and only if $1 - \xi_i > \Omega = \frac{\sum_{j=1}^p (1-\xi_j)^2}{\sum_{j=1}^p (1-\xi_j)}$ for every $i \in [p]$.
3. $\boldsymbol{\beta}^{s*} = \boldsymbol{\beta}_\star$ if and only if the covariance matrix $\mathbf{\Sigma}_t = c\boldsymbol{I}_p$ for some $c \in \mathbb{R}$.

Meaning of Corollary 1

1. Optimal surrogate parameter $\boldsymbol{\beta}^{s*}$ only depends on the covariance spectrum λ_i.
2. For regions where λ_i is small, $1 - \xi_i$ is also small, leading to further amplification through the ground truth parameter $\boldsymbol{\beta}_\star$. Conversely, for regions with large λ_i, $\boldsymbol{\beta}_\star$ induces shrinkage.
3. The threshold Ω corresponds to the ratio of the sample second moment to the sample first moment, arising from the trade-off between bias and variance terms of $\bar{\mathcal{R}}^{s2t}_{\kappa_t,\sigma_t}(\mathbf{\Sigma}_t, \boldsymbol{\beta}_\star, \boldsymbol{\beta}^s)$.
4. Unless the eigenvalues of the $\mathbf{\Sigma}_t$ are constant, there is potential for improvement by using the surrogate parameter $\boldsymbol{\beta}^s$ rather than using $\boldsymbol{\beta}_\star$.

Weak-to-Strong Generalization

How to design *weak model*?

- Surrogate model uses <u>fewer features</u>, $p_s < p$, by a mask operation $\mathcal{M}(\boldsymbol{x})$.
- **Masked Surrogate Model**: $(\mathcal{M}(\tilde{\boldsymbol{x}}), \tilde{y}) \sim \mathcal{D}_s^{p_s}$: Masked distribution
  - $\tilde{y} = \mathcal{M}(\tilde{\boldsymbol{x}})^\top \mathcal{M}(\boldsymbol{\beta}_\star) + \tilde{z}$
  - $\boldsymbol{\beta}_\star \in \mathbb{R}^p$, $\tilde{\boldsymbol{x}} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}_s)$, $\tilde{z} \sim \mathcal{N}(0, \sigma_s^2)$.

  $\Rightarrow$ Estimator $\boldsymbol{\beta}^s = \mathrm{Est}(\tilde{\boldsymbol{X}}, \tilde{\boldsymbol{y}})$
- **Masked Target Model**: $(\boldsymbol{x}, y^s) \sim \mathcal{D}_t^{p_s}(\boldsymbol{\beta}^s)$
  - $y^s = \mathcal{M}(\boldsymbol{x})^\top \boldsymbol{\beta}^s + z$
  - $\boldsymbol{\beta}^s \in \mathbb{R}^{p_s}$, $\boldsymbol{x} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}_t)$, $z \sim \mathcal{N}(0, \sigma_t^2)$.

  $\Rightarrow$ Estimator $\boldsymbol{\beta}^{s2t} = \mathrm{Est}(\boldsymbol{X}, \boldsymbol{y})$
- **Standard Target Model**: $(\boldsymbol{x}, y^s) \sim \mathcal{D}_t^{p}(\boldsymbol{\beta}_\star)$
  - $y = \boldsymbol{x}^\top \boldsymbol{\beta}_\star + z$
  - $\boldsymbol{\beta}_\star \in \mathbb{R}^p$, $\boldsymbol{x} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}_t)$, $z \sim \mathcal{N}(0, \sigma_t^2)$.

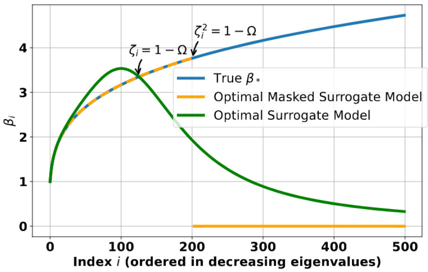
  $\Rightarrow$ Estimator $\boldsymbol{\beta}^t = \mathrm{Est}(\boldsymbol{X}, \boldsymbol{y})$

Leveraging Information from a Weak Teacher

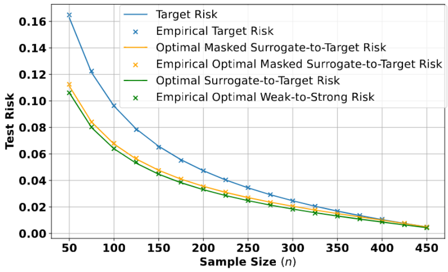
Proposition 2

Assume that $\boldsymbol{\Sigma}_s = \boldsymbol{\Sigma}_t$ and $\boldsymbol{\Sigma}_t$ is diagonal. In the absence of model shift ($\mathcal{M}(\boldsymbol{\beta}_\star) = \boldsymbol{\beta}^s$), the following results hold:

1. If the mask operation $\mathcal{M}$ select all the features that satisfy $1 - \xi_i^2 > \Omega$, then the surrogate-to-target model outperforms the standard target model in the asymptotic risk.
2. Let $\boldsymbol{M}$ represent the all possible $\mathcal{M}$, where $|\boldsymbol{M}| = 2^p$. The optimal $\mathcal{M}^*$ for the asymptotic risk within $\boldsymbol{M}$ is the one that selects all features satisfying $1 - \xi^2 > \Omega$.



(a) Ground-truth and surrogate model weights



(b) Test risks as a function of sample size

Fundamental Limits and Scaling Laws

Observation 2. For any covariance matrix $\boldsymbol{\Sigma} \in \mathbb{R}^{p\times p}$, there exists an orthonormal matrix $\boldsymbol{U} \in \mathbb{R}^{p\times p}$ such that the transformation of $\boldsymbol{x} \to \boldsymbol{U}^\top \boldsymbol{x}$ and $\boldsymbol{\beta} \to \boldsymbol{u}^\top \boldsymbol{\beta}$ does not affect the labels $\boldsymbol{y}$ but ensures that the covariance matrix is diagonal. $\Rightarrow$ Consider only the diagonal covariance!

Omniscient Test Risk Estimate

Fix $p > n \geq 1$. Let $\boldsymbol{\Sigma} = \boldsymbol{U}\mathrm{diag}(\boldsymbol{\lambda})\boldsymbol{U}^\top$ and $\bar{\boldsymbol{\beta}} = \boldsymbol{U}^\top \boldsymbol{\beta}_\star$. Then, the excess test risk estimate is the following:

$$\mathcal{R}(\hat{\boldsymbol{\beta}}) \approx \mathbb{E}_{\hat{\boldsymbol{\beta}}\sim D(\boldsymbol{\beta}_\star)}\left[(y - \boldsymbol{x}^\top \hat{\boldsymbol{\beta}})^2\right] - \sigma^2 = \frac{\sigma^2\Omega + \mathcal{B}(\bar{\boldsymbol{\beta}})}{1-\Omega},$$

where $\quad n = \sum_{i=1}^{p} \frac{1}{\lambda_i + \tau}, \quad \xi_i = \frac{\tau}{\lambda_i + \tau}, \quad \Omega = \frac{1}{n}\sum_{i=1}^{p}(1-\xi_i)^2, \quad \mathcal{B}(\bar{\boldsymbol{\beta}}) = \sum_{i=1}^{p} \lambda_i \xi_i^2 \bar{\beta}_i^2.$

Asymptotic behavior of omniscient risk: As $n, p \to \infty$ with a fixed ratio $p/n = \kappa$, the approximation becomes an equality. (Same as Theorem 1)

What is $D(\boldsymbol{\beta}_\star)$?

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{\Sigma} + \tau \boldsymbol{I}_p)^{-1}\boldsymbol{\Sigma}^{1/2}\left(\boldsymbol{\Sigma}^{1/2}\boldsymbol{\beta}_\star + \frac{\gamma \boldsymbol{g}}{\sqrt{p}}\right), \quad (\gamma, \tau) : \text{sol. of fixed point eq.}, \boldsymbol{g} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_p).$$

Framework for Degree Analysis

Setup

- $p, n \to \infty$ with a fixed ratio $\kappa = p/n > 1$. (Over-parametrized regime)
- Covariance $\mathbf{\Sigma}$: diagonal, $\mathbf{\Sigma}_{i,i} = \lambda_i = i^{-\alpha}$ for $\alpha > 1$.

Equations for the τ_t and Ω:

$$\sum_{i=1}^{\infty} \frac{\tau_t}{\lambda_i + \tau_t} = n, \quad n\Omega = \sum_{i=1}^{\infty} \left(\frac{i^{-\alpha}}{i^{-\alpha} + \tau_t} \right)^2$$

Proposition 3. When $\lambda_i = i^{-\alpha}$,

$$\tau_t = cn^{-\alpha}(1 + O(n^{-1})), \Omega = \frac{\alpha - 1}{\alpha} - O(n^{-1}) \text{ for } c = \left(\frac{\pi}{\alpha \sin(\pi/\alpha)} \right)^{\alpha}.$$

Proposition 4. When $C_1 := \frac{\alpha \sin(\pi/\alpha)}{n(\alpha-1)^{1/\alpha}}$ and $C_2 := \frac{\alpha \sin(\pi/\alpha)}{n(\sqrt{\alpha}-1)^{1/\alpha}}$. Then, the indices i for $\xi_i < 1 - \Omega$ are $i < nC_1 + O(1)$; while the indices for $\xi_i^2 < 1 - \Omega$ are $i < nC_2 + O(1)$.

Meaning of Proposition 4: As sample size n increases, the cut-off indices of $\boldsymbol{\beta}^{s*}$ and the optimal mask $\mathcal{M}^*$ increase linearly in n.

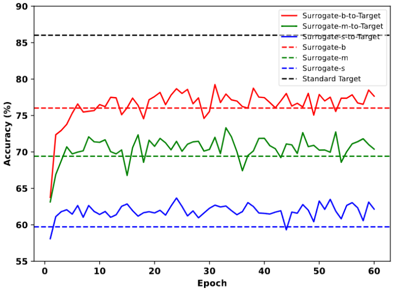
Scaling Law for S2T Model

Scaling Law

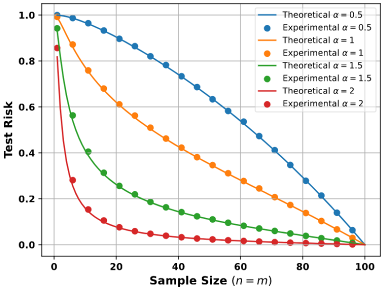
Assume $\lambda_i = i^{-\alpha}$ and $\lambda_i \beta_i^2 = i^{-\beta}$ for $\alpha, \beta > 1$. In the asymptotic regime ($p \to \infty$), the excess risk of the surrogate-to-target model with an *optimal* surrogate parameter scales the same as the standard target model. Specifically, we have

$$\mathcal{R}^*(\boldsymbol{\beta}^{s2t}) = \Theta(n^{-\min(2\alpha, \beta-1)}) = \mathcal{R}(\boldsymbol{\beta}^t),$$

where the minimum surrogate-to-target risk $\mathcal{R}^*(\boldsymbol{\beta}^{s2t})$ is attained by $\boldsymbol{\beta}^{s*}$.



(a) Weak-to-strong on CIFAR-10



(b) Comparison of theoretical and experimental risks

Asymptotic Risk for the Two-Stage Model

Definition 3. Recall the definition of τ_t and γ_t in Theorem 1. Let $\kappa_s = p/m > 1$ and define $\tau_s \in \mathbb{R}$ similarly to τ_t. We define the random variable $X^s_{(\kappa_s, \sigma_s^2)}$ based on $\boldsymbol{g}_s \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I})$ and the function $\gamma_s : \mathbb{R}^p \to \mathbb{R}$ as follows:

$$X^s_{(\kappa_s,\sigma_s^2)}(\boldsymbol{\Sigma}_s, \boldsymbol{\beta}_\star, \boldsymbol{g}_s) := (\boldsymbol{\Sigma}_s + \tau_s \boldsymbol{I}_p)^{-1} \boldsymbol{\Sigma}_s^{1/2} \left(\boldsymbol{\Sigma}_s^{1/2} \boldsymbol{\beta}_\star + \frac{\gamma_s(\boldsymbol{\beta}_\star) \boldsymbol{g}_s}{\sqrt{p}} \right)$$

$$\gamma_s^2(\boldsymbol{\beta}_\star) := \kappa_s \left(\sigma_s^2 + \mathbb{E}_{\boldsymbol{g}_s} \left[\boldsymbol{\Sigma}_s^{1/2} \left(X^s_{(\kappa_s,\sigma_s^2)}(\boldsymbol{\Sigma}_s, \boldsymbol{\beta}_\star, \boldsymbol{g}_s) - \boldsymbol{\beta}_\star \right)^2 \right] \right)$$

Let $\dot{\kappa} = (\kappa_s, \kappa_t)$, $\dot{\boldsymbol{\Sigma}} = (\boldsymbol{\Sigma}_s, \boldsymbol{\Sigma}_t)$ and $\dot{\sigma} = (\sigma_s, \sigma_t)$. Then, the asymptotic risk estimate is

$$\begin{aligned}
\bar{\mathcal{R}}_{\dot{\kappa},\dot{\sigma}}(\dot{\boldsymbol{\Sigma}}, \boldsymbol{\beta}_\star) &= \|\sigma_t^{1/2} \left(\boldsymbol{I} - (\boldsymbol{\Sigma}_t + \tau_t \boldsymbol{I})^{-1} \boldsymbol{\Sigma}_t (\boldsymbol{\Sigma}_s + \tau_s \boldsymbol{I}) \boldsymbol{\Sigma}_s \right) \boldsymbol{\beta}_\star\|_2^2 \\
&+ \frac{\mathbb{E}_{\boldsymbol{\beta}^s \sim X^s}[\gamma_t^2(\boldsymbol{\beta}^s)]}{p} \operatorname{tr}(\boldsymbol{\Sigma}_t^2 (\boldsymbol{\Sigma}_t + \tau_t \boldsymbol{I})^{-2}) \\
&+ \frac{\gamma_s^2(\boldsymbol{\beta}_\star)}{p} \operatorname{tr} \left(\boldsymbol{\Sigma}_s^{1/2} (\boldsymbol{\Sigma}_s + \tau_s \boldsymbol{I}) \boldsymbol{\Sigma}_t (\boldsymbol{\Sigma}_t + \tau_t \boldsymbol{I})^{-1} \boldsymbol{\Sigma}_t (\boldsymbol{\Sigma}_t + \tau_t \boldsymbol{I})^{-1} \boldsymbol{\Sigma}_t (\boldsymbol{\Sigma}_s + \tau_s \boldsymbol{I})^{-1} \boldsymbol{\Sigma}_s^{1/2} \right).
\end{aligned}$$

KAIST

Non-Asymptotic Risk of Two-Stage Model

Theorem 2

Suppose that some constant $M_t > 1$, we have $1/M_t \leq \kappa_s, \sigma_s^2, \kappa_t, \sigma_t^2 \leq M_t$ and $\|\boldsymbol{\Sigma}_s\|_{op}$, $\|\boldsymbol{\Sigma}_s^{-1}\|_{op}, \|\boldsymbol{\Sigma}_t\|_{op}, \|\boldsymbol{\Sigma}_t^{-1}\|_{op} \leq M_t$. Let $\boldsymbol{B}_p(R) := \{\boldsymbol{x} \in \mathbb{R}^p : \|\boldsymbol{x}\|_2 \leq R\}$. Then, there exists a constant $C = C(M_t)$ such that, for any $\epsilon \in (0, 1/2]$, the following holds with $R + 1 < M_t$:

$$\sup_{\boldsymbol{\beta}_\star \in \boldsymbol{B}_p(R)} \boldsymbol{P}\left(|\mathcal{R}(\boldsymbol{\beta}^{s2t}) - \bar{\mathcal{R}}_{\dot{\kappa},\dot{\sigma}}(\dot{\boldsymbol{\Sigma}}, \boldsymbol{\beta}_\star)| \geq \epsilon \right) \leq Cpe^{-p\epsilon^4/C}.$$

Future research direction

- Extend the two-stage process to multiple stages.
- Apply the two-stage learning to data pruning, using the surrogate model to decide whether keep of discard each data during the training of the target model.

Summary

Problem Setup

- Surrogate model: $\tilde{y} = \tilde{\boldsymbol{x}}^\top \boldsymbol{\beta}_\star + \tilde{z}$, $\tilde{\boldsymbol{x}} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}_s)$, $z \sim \mathcal{N}(0, \sigma_s^2) \rightarrow$ estimate: $\boldsymbol{\beta}^s$
- Target model: $y^s = \boldsymbol{x}^\top \boldsymbol{\beta}^s + z$, $\boldsymbol{x} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}_t)$, $z \sim \mathcal{N}(0, \sigma_t^2) \rightarrow$ estimate: $\boldsymbol{\beta}^{s2t}$
- Standard target model: $y = \boldsymbol{x}^\top \boldsymbol{\beta}_\star + z$, $\boldsymbol{x} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}_t)$, $z \sim \mathcal{N}(0, \sigma_t^2) \rightarrow$ estimate: $\boldsymbol{\beta}^t$
- Excess Risk: $\mathcal{R}(\hat{\boldsymbol{\beta}}) = \|\boldsymbol{\Sigma}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_\star)\|^2$

Asymptotic Risk[3]

- Asymptotic regime: $n, p \rightarrow \infty$, $\kappa = p/n > 1$
- Theorem 1, Theorem 2: non-asymptotic risk $\rightarrow 0$, excess risk $\rightarrow$ asymptotic risk
- Only consider the asymptotic risk for further analysis

Optimal Surrogate Parameter

- $\boldsymbol{\beta}_\star$ is not a optimal surrogate parameter! $\rightarrow$ room for improvement from the surrogate model
- Optimal surrogate parameter only depends on the eigenvalues of the covariance matrix.
- Threshold: $1 - \xi > \Omega \rightarrow$ originates from the trade-off between bias and variance.

KAIST

Summary

Weak-to-Strong Generalization

- Masked surrogate model: Only uses fewer features by a mask operation $\mathcal{M}$
- Select all features with $1-\xi^2 > \Omega$, the target model outperforms the standard target model.
  $\Rightarrow$ Weak-to-strong model outperforms strong model!

Scaling law

- Assumption: $\boldsymbol{\Sigma}$: diagonal, $\boldsymbol{\Sigma}_{i,i} = \lambda_i = i^{-\alpha}$, $\lambda_i \beta_i^2 = i^{-\beta}$
- $\mathcal{R}^*(\boldsymbol{\beta}^{s2t}) = \Theta(n^{-\min(2\alpha, \beta-1)}) = \mathcal{R}(\boldsymbol{\beta}^t)$

Experiment

- CIFAR-10: Surrogate model < Target model < Standard target model
- Why Target model < Standard target model? $\rightarrow$ Due to the lack of a feature selection process

Future direction

- Multi-stage process (Multi-round distillation)
- Data pruning $\rightarrow$ Utilize the surrogate model for data selection

References

[1] N. Mallinar, A. Zane, S. Frei, and B. Yu, "Minimum-norm interpolation under covariate shift," *arXiv preprint arXiv:2404.00522*, 2024.

[2] P. Patil, J.-H. Du, and R. J. Tibshirani, "Optimal ridge regularization for out-of-distribution prediction," *arXiv preprint arXiv:2404.01233*, 2024.

[3] Q. Han and X. Xu, "The distribution of ridgeless least squares interpolators," *arXiv preprint arXiv:2307.02044*, 2023.