Optimal Self-Distillation in Ridge Regression: Precise Asymptotics and One-Shot Tuning

Alessandro Rinaldo

Department of Statistics and Data Sciences

TEXAS
The University of Texas at Austin

DSSV 2026
Università di Trento
June 29, 2026

Collaborators and Papers

Hien Dang
PhD Student, UT Austin

Pratik Patil
Assistant Professor, UT Austin

- Optimal Unconstrained Self-Distillation in Ridge Regression: Strict Improvements, Precise Asymptotics, and One-Shot Tuning, ICML 2026
- Prediction-only Distillation with Optimal Mixing in Linear and Logistic Regression, forthcoming.

Motivation and Setup: Ridge Self-Distillation

Motivation.
Self-distilled ridge regression: the teacher, pure student and optimal self-distilled student.

Knowledge Distillation (Hinton, Vinyals, Dean, 2015)

- A fundamental ML/AI technique in regression/classification that is essential for deployment of large generative models.
- A **large, strong, complex, expensive, general** teacher model is trained.
- A **different and much smaller** student model is then trained on **both ground-truth labels and the teacher's (soft) predictions.**
- Rationale: the teacher's soft labels carry more refined information than the hard labels, making learning easier and more effective.

Knowledge Distillation (Hinton, Vinyals, Dean, 2015)

- A fundamental ML/AI technique in regression/classification that is essential for deployment of large generative models.
- A **large, strong, complex, expensive, general** teacher model is trained.
- A **different and much smaller** student model is then trained on **both ground-truth labels and the teacher's (soft) predictions.**
- Rationale: the teacher's soft labels carry more refined information than the hard labels, making learning easier and more effective.

Goal

(Nearly) Reproduce the performance of the teacher model with a **smaller, cheaper, faster, more specialized** student model.

Knowledge Distillation Example: Self-Driving Cars

- Austin is a hub for self-driving car research and experimentation.
- You can buy a Tesla car and have it drive itself to your home.
- Waymo operated a fully driverless robotaxi fleet in Austin, which can be requested with the Uber app, with a nearly spotless safety track record.

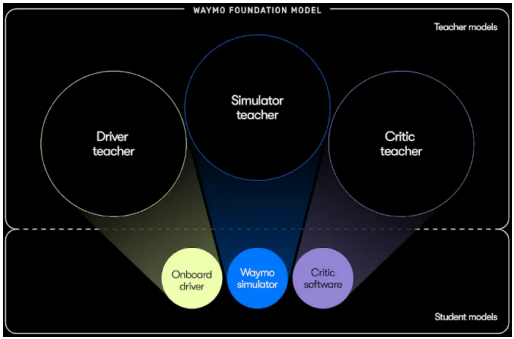


Image credit: Official Waymo blog, December 2025

Self-Distillation

- Self-distillation is the special case in which the teacher and student are the **same model/architecture**.
- The surprising empirical finding: the student often has lower generalization error than the teacher. What is more, repeated rounds of self-distillation typically boost performance.
- Self-distillation does not have a clear rationale.

The Question

If no new data or model is introduced, how could this possibly work?

This talk: exact answers for ridge regression in general settings.

Self-Distillation in Ridge Regression: Set-Up

- We consider a standard regression setting, in which we observe n pairs of data points in $\mathbb{R}^p \times \mathbb{R}$

$$(x_1, y_1), \ldots, (x_n, y_n) \sim P_{\text{data}},$$

written as $\mathcal{D} = (X, y)$, with $X = [x_1\ x_2\ \ldots\ x_n]^\top \in \mathbb{R}^{n \times p}$ the matrix of features and $y = (y_1, \ldots, y_n) \in \mathbb{R}^n$ the vector of responses.

- Assumption-lean setting: We place no assumptions on P_{data}, beyond the existence of two moments, nor on the sampling. **We do not assume a linear model**.

Self-Distillation in Ridge Regression: Set-Up

- We consider a standard regression setting, in which we observe n pairs of data points in $\mathbb{R}^p \times \mathbb{R}$

$$(x_1, y_1), \ldots, (x_n, y_n) \sim P_{\text{data}},$$

written as $\mathcal{D} = (X, y)$, with $X = [x_1\ x_2\ \ldots\ x_n]^\top \in \mathbb{R}^{n \times p}$ the matrix of features and $y = (y_1, \ldots, y_n) \in \mathbb{R}^n$ the vector of responses.

- Assumption-lean setting: We place no assumptions on P_{data}, beyond the existence of two moments, nor on the sampling. **We do not assume a linear model**.

- We are interested in prediction: deploy a $\mathcal{D}$-dependent predictor $\widehat{f}\colon \mathbb{R}^p \to \mathbb{R}$ with small conditional squared prediction risk

$$R(\widehat{f}) = \mathbb{E}_{(x_{\text{new}}, y_{\text{new}})}[(y_{\text{new}} - \widehat{f}(x_{\text{new}}))^2 | \mathcal{D}],$$

where $(x_{\text{new}}, y_{\text{new}})$ is an independent draw from P_{data} or possibly a different distribution (OOD risk).

The Teacher, the PD Student and the SD Student

- The ridge regression predictor, or teacher, is

$$\widehat{f}_\lambda(x) = x^\top \widehat{\beta}_\lambda, \quad x \in \mathbb{R}^p, \quad \lambda > 0,$$

where

$$\widehat{\beta}_\lambda = \underset{\beta \in \mathbb{R}^p}{\operatorname{argmin}} \left\{ \|y - X\beta\|_2^2/n + \lambda \|\beta\|_2^2 \right\} = (X^\top X/n + \lambda I)^{-1} X^\top y/n.$$

The Teacher, the PD Student and the SD Student

- The ridge regression predictor, or teacher, is

$$\widehat{f}_\lambda(x) = x^\top \widehat{\beta}_\lambda, \quad x \in \mathbb{R}^p, \quad \lambda > 0,$$

where

$$\widehat{\beta}_\lambda = \underset{\beta \in \mathbb{R}^p}{\operatorname{argmin}} \left\{ \|y - X\beta\|_2^2/n + \lambda \|\beta\|_2^2 \right\} = (X^\top X/n + \lambda I)^{-1} X^\top y/n.$$

- The pure-distilled (PD) student is the predictor

$$\widehat{f}_{\mathrm{pd},\lambda}(x) = x^\top \widehat{\beta}_{\mathrm{pd},\lambda}, \quad x \in \mathbb{R}^p,$$

where $\widehat{\beta}_{\mathrm{pd},\lambda} = \operatorname{argmin}_{\beta \in \mathbb{R}^p} \left\{ \|\hat{y}_\lambda - X\beta\|_2^2/n + \lambda \|\beta\|_2^2 \right\}$, with $\hat{y}_\lambda = \widehat{f}_\lambda(X)$.

The Teacher, the PD Student and the SD Student

- The ridge regression predictor, or teacher, is

$$\widehat{f}_\lambda(x) = x^\top \widehat{\beta}_\lambda, \quad x \in \mathbb{R}^p, \quad \lambda > 0,$$

where

$$\widehat{\beta}_\lambda = \underset{\beta \in \mathbb{R}^p}{\operatorname{argmin}} \left\{ \|y - X\beta\|_2^2/n + \lambda \|\beta\|_2^2 \right\} = (X^\top X/n + \lambda I)^{-1} X^\top y/n.$$

- The pure-distilled (PD) student is the predictor

$$\widehat{f}_{\mathrm{pd},\lambda}(x) = x^\top \widehat{\beta}_{\mathrm{pd},\lambda}, \quad x \in \mathbb{R}^p,$$

where $\widehat{\beta}_{\mathrm{pd},\lambda} = \operatorname{argmin}_{\beta \in \mathbb{R}^p} \left\{ \|\hat{y}_\lambda - X\beta\|_2^2/n + \lambda \|\beta\|_2^2 \right\}$, with $\hat{y}_\lambda = \widehat{f}_\lambda(X)$.

- The self-distilled (SD) student is the predictor

$$\widehat{f}_{\mathrm{sd},\lambda,\xi}(x) = x^\top \widehat{\beta}_{\mathrm{sd},\lambda,\xi}, \quad x \in \mathbb{R}^p,$$

where

$$\widehat{\beta}_{\mathrm{sd},\lambda,\xi} = \underset{\beta \in \mathbb{R}^p}{\operatorname{argmin}} \left\{ (1-\xi) \|y - X\beta\|_2^2/n + \xi \|\hat{y}_\lambda - X\beta\|_2^2/n + \lambda \|\beta\|_2^2 \right\}$$

and $\xi \in \mathbb{R}$ is the mixing weight.

Affine Mixing

- The SD predictor is an affine mixing of the teacher and PD predictors:

$$\widehat{f}_{\mathrm{sd},\lambda,\xi} = (1-\xi)\widehat{f}_{\lambda} + \xi\widehat{f}_{\mathrm{pd},\lambda}.$$

- Equivalently, it is just the ridge predictor trained on the mixed labels

$$(1-\xi)\, y + \xi\, \widehat{y}_{\lambda}.$$

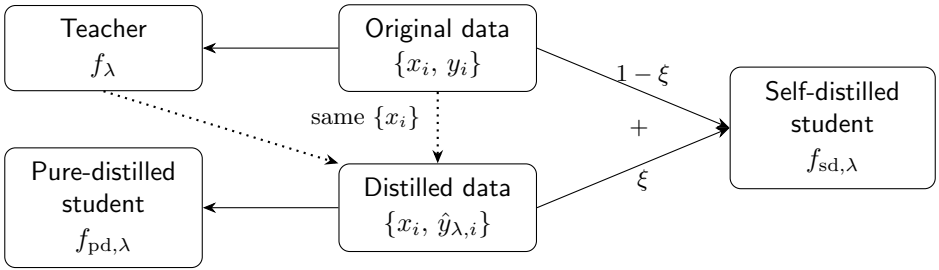
Affine Mixing

- The SD predictor is an affine mixing of the teacher and PD predictors:

$$\widehat{f}_{\mathrm{sd},\lambda,\xi} = (1-\xi)\widehat{f}_{\lambda} + \xi\widehat{f}_{\mathrm{pd},\lambda}.$$

- Equivalently, it is just the ridge predictor trained on the mixed labels

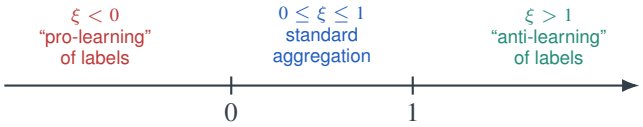
$$(1-\xi)\, y + \xi\, \widehat{y}_{\lambda}.$$



What Does Mixing Do?

The tunable mixing weight ξ controls the trade-off between the original data and the distilled data:

- $\xi = 0$: the teacher;
- $\xi = 1$: the PD student;
- $\xi \in [0, 1]$: standard aggregation, common in practice.
- Here, we consider unconstrained $\xi \in \mathbb{R}$ – this is key!



Optimal-SD

- The choice of the mixing weight ξ plays a crucial role in explaining how SD ridge outperforms standard ridge. Let

$$\xi^\star = \xi^\star(\lambda) = \operatorname*{argmin}_{\xi \in \mathbb{R}} R(\widehat{f}_{\mathrm{sd},\lambda,\xi})$$

  be the oracle optimal mixing weight – a function of λ and the data.

Optimal-SD (not computable): $\widehat{f}_{\mathrm{sd},\lambda,\xi^\star} = \widehat{f}^\star_{\mathrm{sd},\lambda}$.

Optimal-SD

- The choice of the mixing weight ξ plays a crucial role in explaining how SD ridge outperforms standard ridge. Let

$$\xi^\star = \xi^\star(\lambda) = \operatorname*{argmin}_{\xi \in \mathbb{R}} R(\widehat{f}_{\mathrm{sd},\lambda,\xi})$$

  be the oracle optimal mixing weight – a function of λ and the data.

Optimal-SD (not computable): $\widehat{f}_{\mathrm{sd},\lambda,\xi^\star} = \widehat{f}^\star_{\mathrm{sd},\lambda}$.

In the rest of the talk, we will answer the following questions.

Basic Questions

- When does optimal-SD **strictly improve** the teacher?
- Can optimal-SD from a suboptimal teacher rival an **optimally tuned** teacher?
- Can we efficiently tune $\xi^\star$ **without grid search, refitting or data splitting**?

High-Level Answers

We will compare the predictive risk profiles of optimal-SD and the teacher:

$$\lambda \in (0, \infty) \mapsto R(\widehat{f}^{\star}_{\mathrm{sd},\lambda}) \quad \text{vs} \quad \lambda \in (0, \infty) \mapsto R(\widehat{f}_{\lambda}).$$

Pointwise gain

Optimal SD strictly outperforms the teacher at any **non-stationary** point of the teacher risk curve.

Global gain

A student from a suboptimal teacher **can beat** an optimally tuned teacher.

Tuning

Optimal mixing can be tuned **one-shot**: no grid search, no split, no refits.

Punchlines

- The optimal SD mixing weight $\xi^{\star}(\lambda)$ records the **slope of the teacher's risk curve** at λ and makes adjustments akin to a better bias/variance tradeoff.
- Optimal-SD improves on ridge regression via a simple re-tuning.

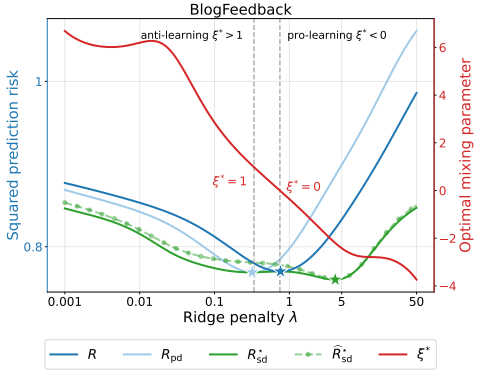
Datasets

We have used some simulated data and the following real datasets in our experiments to validate our results:

- the UCI Blog feedback dataset (Buza, 2014). $n = 2,619$ and $p = 280$;
- the UCI Communities and Crime dataset (Redmond, 2002). $n = 398$ and $p = 99$;
- the UCI Air Quality dataset (Vito, 2008). $n = 5,175$ and $p = 8$;
- the CIFAR10 and CIFAR100 datasets, from ResNet-18 and ResNet-34 (He et al., 2016). $n = 2000/20000$ and $p = 512$;
- the Caltech-101 classification dataset (Griffin et al., 2007), with 102 classes and 2000 samples as the test set.

The prediction risk is evaluated on a large fraction (70% - 95%) of held-out data.

A Preview: Strict Improvement Across Real Datasets



- Blue: teacher risk $R(\lambda)$.
- Light blue: PD risk $R_{\text{pd}}(\lambda)$.
- Green: Optimal-SD risk $R^\star_{\text{sd}}(\lambda)$ — **always below the teacher**.
- Red: Optimal mixing weight $\xi^\star(\lambda)$ — changes sign at the stationary point of $R(\lambda)$.
- Green dashed: One-shot GCV estimate $\widehat{R}^\star_{\text{sd}}(\lambda)$ (from training data only).

Related Work

- A recent and active area of research in statistics/ML.
- A partial list of recent and relevant contributions includes Mobahi et al. (2020), Das and Sanghavi (2023), Pareek et al. (2024), Emrullah Ildiz et al. (2025) , Garg et al. (2025), Lecoiu et al. (2026).

Paper	Focus		Nonasymptotic setting			Proportional asymptotics			Weights		Multi-rounds
	Type	Family	Design	Response	Risk	Features	Signal	Model	Range	Tuning	
DS23	SD	Ridge	Fixed	Linear	Estimation				$\mathbb{R}$		
PDO24	SD	Ridge	Fixed	Linear	Fixed-X prediction				$\mathbb{R}$	Split CV (three refits)	Yes
IGT+25	PD	Ridgeless	Random	Linear	In-distribution prediction				$\{1\}$		
MH25	PD	Ridge (generalized)				Isotropic	Isotropic	Linear	$\{1\}$		
GBS25	SD	Ridge (infinite rounds)				Anisotropic	Deterministic	Linear	$[0, 1]$		Infinite rounds only
Ours	SD	Ridge	Random	Any	Any squared prediction	Anisotropic	Deterministic	Any	$\mathbb{R}$	GCV (one-shot)	Yes

Structural Non-asymptotic Results

Conditionally on the observed training data.
Qualitative Results.
Allow any squared prediction risk.

Closed-Form Expressions for the Optimal-SD Risk and Mixing

Denote the (random, conditional, possibly OOD) prediction risks for the teacher (ridge), PD student and optimal-SD student as

$$R(\lambda) = R(\widehat{f}_\lambda), \quad R_{\mathrm{pd}}(\lambda) = R(\widehat{f}_{\mathrm{pd},\lambda}), \quad \text{and} \quad R^\star_{\mathrm{sd}}(\lambda) = R(\widehat{f}_{\mathrm{sd},\lambda,\xi^\star}).$$

Closed-Form Expressions for the Optimal-SD Risk and Mixing

Denote the (random, conditional, possibly OOD) prediction risks for the teacher (ridge), PD student and optimal-SD student as

$$R(\lambda) = R(\widehat{f}_\lambda), \quad R_{\mathrm{pd}}(\lambda) = R(\widehat{f}_{\mathrm{pd},\lambda}), \quad \text{and} \quad R^\star_{\mathrm{sd}}(\lambda) = R(\widehat{f}_{\mathrm{sd},\lambda,\xi^\star}).$$

Structural Identities

For each $\lambda > 0$, define the correlation and gap terms

$$C(\lambda) = \mathbb{E}\big[(y_{\mathrm{new}} - \widehat{f}_\lambda(x_{\mathrm{new}}))(y_{\mathrm{new}} - \widehat{f}_{\mathrm{pd},\lambda}(x_{\mathrm{new}})) \mid \mathcal{D}\big]$$
$$D(\lambda) = \mathbb{E}\big[(\widehat{f}_\lambda(x_{\mathrm{new}}) - \widehat{f}_{\mathrm{pd},\lambda}(x_{\mathrm{new}}))^2 \mid \mathcal{D}\big].$$

If $D(\lambda) > 0$, then, almost surely,

$$\xi^\star(\lambda) = \frac{R(\lambda) - C(\lambda)}{D(\lambda)} \quad \text{and} \quad R^\star_{\mathrm{sd}}(\lambda) = R(\lambda) - \frac{(R(\lambda) - C(\lambda))^2}{D(\lambda)}.$$

That is, almost surely, $R^\star_{\mathrm{sd}}(\lambda) \leq R(\lambda)$, for all λ!

Pointwise Strict Improvement and Sign Rule

A stronger, more interpretable conclusion follows from relating $\xi^\star(\lambda)$ to $R'(\lambda)$.

Theorem (Sign Rule)

If $D(\lambda) > 0$, then, almost surely,

$$\xi^\star(\lambda) = -\frac{\lambda}{2}\frac{R'(\lambda)}{D(\lambda)}, \quad \textit{and} \quad R^\star_{\mathrm{sd}}(\lambda) = R(\lambda) - \frac{\lambda^2}{4}\frac{(R'(\lambda))^2}{D(\lambda)}.$$

If $R'(\lambda) \neq 0$, then $R^\star_{\mathrm{sd}}(\lambda) < R(\lambda)$ and $\operatorname{sign}(\xi^\star(\lambda)) = -\operatorname{sign}(R'(\lambda))$.

Pointwise Strict Improvement and Sign Rule

A stronger, more interpretable conclusion follows from relating $\xi^\star(\lambda)$ to $R'(\lambda)$.

Theorem (Sign Rule)

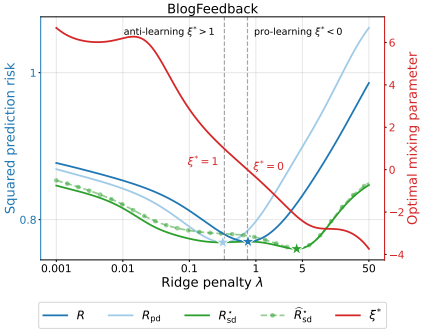
If $D(\lambda) > 0$, then, almost surely,

$$\xi^\star(\lambda) = -\frac{\lambda}{2}\frac{R'(\lambda)}{D(\lambda)}, \quad \textit{and} \quad R^\star_{\mathrm{sd}}(\lambda) = R(\lambda) - \frac{\lambda^2}{4}\frac{(R'(\lambda))^2}{D(\lambda)}.$$

If $R'(\lambda) \neq 0$, then $R^\star_{\mathrm{sd}}(\lambda) < R(\lambda)$ and $\operatorname{sign}(\xi^\star(\lambda)) = -\operatorname{sign}(R'(\lambda))$.

- The optimal-SD mixing re-tunes the ridge predictor based on both the sign and the magnitude of the derivative of $R(\lambda)$.
- To be clear, while the optimal-SD student is itself a ridge predictor (trained on the mixed labels), it is not derived from the teacher just by changing λ.

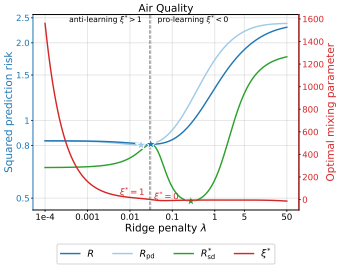
Examples and Commentary



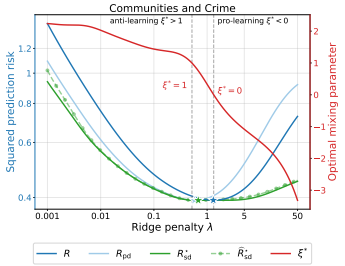
$$n = 2619, p = 280$$

- *Under-regularized case = anti-learner*: When $R'(\lambda) < 0$ (a larger λ decreases the risk), then $\xi^\star(\lambda) > 0$, favoring $\widehat{y}_\lambda$ (corrects for insufficient shrinkage).
- *Over-regularized case = pro-learner*: When $R'(\lambda) > 0$ (a smaller λ decreases the risk), then $\xi^\star(\lambda) < 0$, favoring y (corrects for excessive shrinkage).
- At ridge-optimal $\lambda^\star = \text{argmin}_{\lambda \geq 0} R(\lambda)$, $R^\star_{\text{sd}}(\lambda^\star) = R(\lambda^\star)$, no improvement possible.

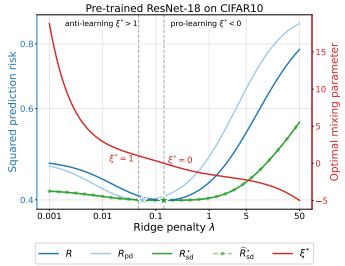
More Examples



(a) $n = 5175, p = 8$



(b) $n = 398, p = 99$



(c) $n = 2000, p = 512$

Note: in general the ridge risk profile need not be quadratic-like, with possible multiple stationary points.

When the Student Beats the Teacher!

- Recall that $\lambda^\star = \text{argmin}_{\lambda>0} R(\lambda)$ is the oracle optimal λ for the teacher. Can we do better than $R(\lambda^\star)$? Yes!

When the Student Beats the Teacher!

- Recall that $\lambda^\star = \operatorname{argmin}_{\lambda>0} R(\lambda)$ is the oracle optimal λ for the teacher. Can we do better than $R(\lambda^\star)$? Yes!

Curvature Test

If

$$D(\lambda^\star) < \frac{(\lambda^\star)^2}{2} R''(\lambda^\star),$$

then

$$\min_{\lambda>0} R^\star_{\mathrm{sd}}(\lambda) < R(\lambda^\star).$$

The best optimal SD-student beats the best teacher!

When the Student Beats the Teacher!

- Recall that $\lambda^\star = \operatorname{argmin}_{\lambda>0} R(\lambda)$ is the oracle optimal λ for the teacher. Can we do better than $R(\lambda^\star)$? Yes!

Curvature Test

If

$$D(\lambda^\star) < \frac{(\lambda^\star)^2}{2} R''(\lambda^\star),$$

then

$$\min_{\lambda>0} R^\star_{\mathrm{sd}}(\lambda) < R(\lambda^\star).$$

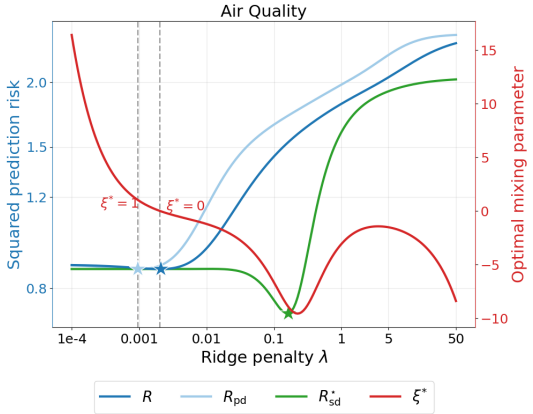
The best optimal SD-student beats the best teacher!

- We empirically verified the curvature test on real-life datasets:

Dataset	Curvature test?	Global gain?
BlogFeedback	✓	✓
Comm. & Crime	✗	✗
CIFAR-10	✗	✗
Air Quality	✓	✓

The Curvature Test is Only Sufficient

With a different split for training/testing in the Air Quality dataset, we verify that the curvature test fails yet optimal-SD is strictly better than optimal ridge.



Proportional Asymptotic Results

Quantify optimal SD gains as a function of the aspect ratio $p/n \to \gamma$, the signal/covariance alignment, SNR, covariance spectrum.

Proportional Asymptotics: Model-Free Random Design

We characterize the limiting optimal-SD risk profile $R^{\star}_{\mathrm{sd}}(\lambda)$ and mixing weight $\xi^{\star}(\lambda)$ in the proportional asymptotics regime as $p/n \to \gamma \in (0, \infty)$.

Assumption (Data assumptions)

- ***Covariates***: *Each feature vector x_i can be decomposed as $x_i = \Sigma^{1/2} z_i$, where $z_i \in \mathbb{R}^p$ contains centered, unit-variance i.i.d. entries having bounded $4 + \mu$ moments for some $\mu > 0$. (RMT feature structure and bounded moments)*
- ***Response***: *$y \sim P_y$ has mean 0 and uniformly bounded $(4 + \nu)$-th moment, for some $\nu > 0$. (model-free, bounded moments response)*

- Limits are parameterized by $(\Sigma, \beta, \sigma^2)$, where $\beta := \Sigma^{-1}\mathbb{E}[xy]$ and $\sigma^2 := \mathrm{Var}(y - x^{\top}\beta)$.
- We allow for an *anisotropic covariance* Σ and a *deterministic signal* β.

Risk Asymptotics

Theorem

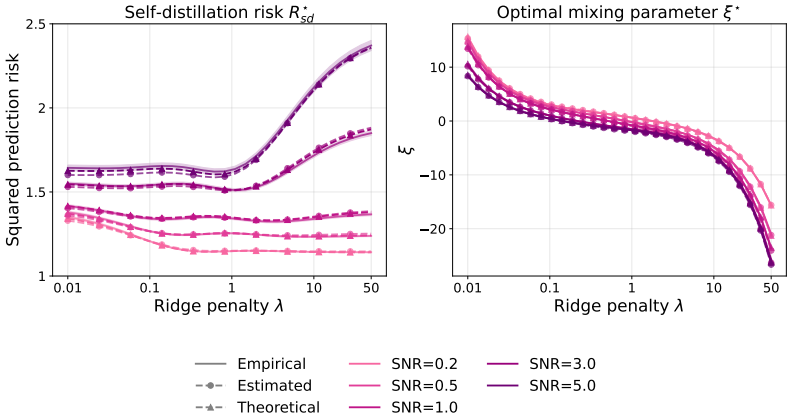
As $p/n \to \gamma \in (0, \infty)$, for each fixed $\lambda > 0$, almost surely,

$$\xi^\star(\lambda) \to \frac{\mathscr{R}(\lambda) - \mathscr{C}(\lambda)}{\mathscr{D}(\lambda)}, \quad R^\star_{\mathrm{sd}}(\lambda) \to \mathscr{R}(\lambda) - \frac{(\mathscr{R}(\lambda) - \mathscr{C}(\lambda))^2}{\mathscr{D}(\lambda)},$$

where $\mathscr{R}(\lambda), \mathscr{C}(\lambda), \mathscr{D}(\lambda)$ are explicit functions of $(\Sigma, \beta, \sigma^2)$.

- The formulae track the spectrum of Σ, the signal-covariance alignment, aspect ratio, and noise.
- A technical challenge in this analysis is the fact that the teacher and the pure distilled student have dependent resolvents. This can be tackled via higher-order deterministic equivalents.

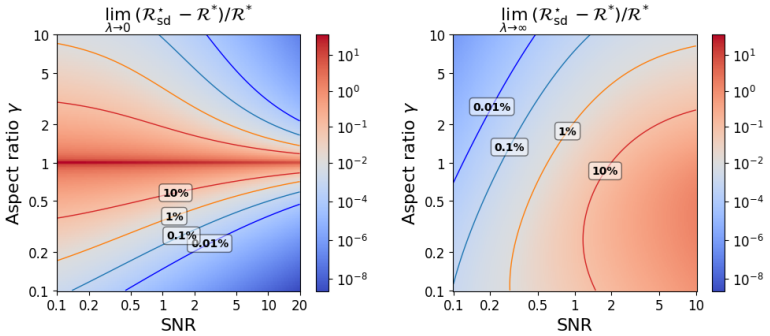
Theory vs Simulations



AR1 design and signal aligned with top eigenvalues, $n = 400, p = 200$.

How Close Can SD Get to Optimal Ridge?

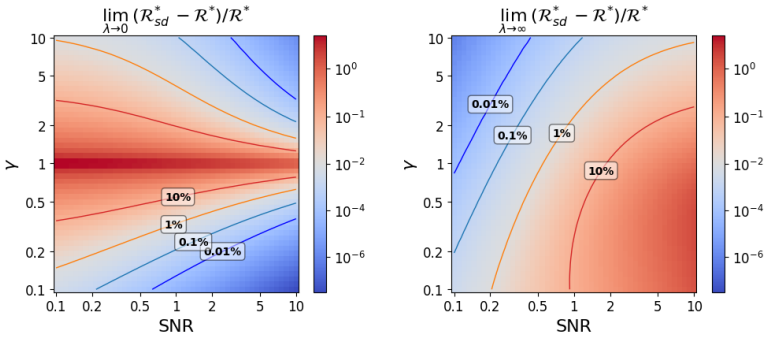
- With an extremely under- or over-regularized teacher ($\lambda \to 0$ or $\lambda \to \infty$), we can quantify how far $\mathscr{R}^\star_{\rm sd}$ is from the Bayes-optimal risk $\mathscr{R}^\star$ (for a Gaussian isotropic signal).
- Optimal SD can be *remarkably close* to the Bayes-optimal ridge even under extreme mis-regularization.
- With low SNR, over-regularization is preferable, and, with high SNR, the opposite.



Isotropic covariance and Gaussian isotropic signal

How Close Can SD Get to Optimal Ridge?

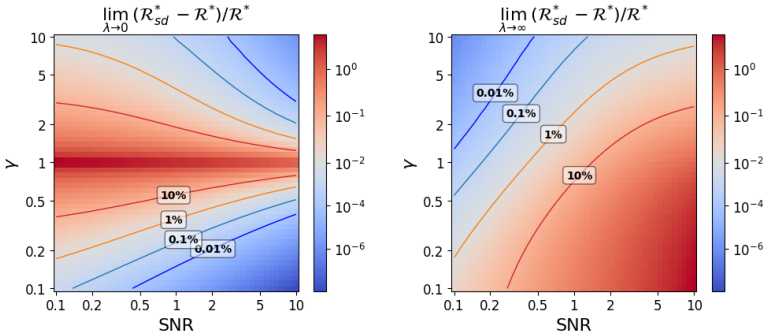
- With an extremely under- or over-regularized teacher ($\lambda \to 0$ or $\lambda \to \infty$), we can quantify how far $\mathscr{R}^{\star}_{\mathrm{sd}}$ is from the Bayes-optimal risk $\mathscr{R}^{\star}$ (for a Gaussian isotropic signal).
- Optimal SD can be *remarkably close* to the Bayes-optimal ridge even under extreme mis-regularization.
- With low SNR, over-regularization is preferable, and, with high SNR, the opposite.



AR1 covariance and Gaussian isotropic signal

How Close Can SD Get to Optimal Ridge?

- With an extremely under- or over-regularized teacher ($\lambda \to 0$ or $\lambda \to \infty$), we can quantify how far $\mathscr{R}^{\star}_{\mathrm{sd}}$ is from the Bayes-optimal risk $\mathscr{R}^{\star}$ (for a Gaussian isotropic signal).
- Optimal SD can be *remarkably close* to the Bayes-optimal ridge even under extreme mis-regularization.
- With low SNR, over-regularization is preferable, and, with high SNR, the opposite.



Spiked covariance and Gaussian isotropic signal

One-shot Tuning

Estimate optimal mixing $\xi^\star(\lambda)$ from training data:
no grid search, no split, no refits.

How to Compute $\xi^\star(\lambda)$ - Only for In-Distribution Risk

- Issue: The oracle optimal mixing weight $\xi^\star(\lambda)$ depends on the quantities ($R(\lambda)$, $C(\lambda)$, $D(\lambda)$), which in turn depend on the unknown data distribution.

- Standard approach: grid search over ξ with split (or leave-one-out) CV. This is problematic:
  - computationally expensive (re-training for each candidate ξ);
  - statistically inefficient and induces high bias when p and n are comparable.

- Our approach: One-shot tuning via Generalized Cross-Validation (GCV).
  - Fit the teacher and the PD student **once**.
  - Estimate $R(\lambda)$, $R_{\mathrm{pd}}(\lambda)$, and $C(\lambda)$.
  - Estimate $\xi^\star(\lambda)$ with a plug-in estimator.

Plug-In Estimator of $\xi^\star(\lambda)$

- Let $\hat{y}_\lambda = f_\lambda(X)$ and $\hat{y}_{\mathrm{pd},\lambda} = f_{\mathrm{pd},\lambda}(X)$ and consider the *effective degrees of freedom* of the teacher and the PD student

$$\mathsf{df}_\lambda = \mathrm{tr}(S_\lambda) \quad \text{and} \quad \mathsf{df}_{\mathrm{pd},\lambda} = \mathrm{tr}(S_{\mathrm{pd},\lambda}),$$

where $S_\lambda = X(X^\top X/n + \lambda I)^{-1} X^\top/n$ is the hat matrix. See Patil et al. (2021).

Plug-In Estimator of $\xi^\star(\lambda)$

- Let $\hat{y}_\lambda = f_\lambda(X)$ and $\hat{y}_{\mathrm{pd},\lambda} = f_{\mathrm{pd},\lambda}(X)$ and consider the *effective degrees of freedom* of the teacher and the PD student

$$\mathsf{df}_\lambda = \mathrm{tr}(S_\lambda) \quad \text{and} \quad \mathsf{df}_{\mathrm{pd},\lambda} = \mathrm{tr}(S_{\mathrm{pd},\lambda}),$$

  where $S_\lambda = X(X^\top X/n + \lambda I)^{-1} X^\top/n$ is the hat matrix. See Patil et al. (2021).
- Evaluate the GCV-corrected residuals

$$\hat{r}_\lambda = \frac{y - \hat{y}_\lambda}{1 - \mathsf{df}_\lambda/n}, \quad \text{and} \quad \hat{r}_{\mathrm{pd},\lambda} = \frac{y - \hat{y}_{\mathrm{pd},\lambda}}{1 - \mathsf{df}_{\mathrm{pd},\lambda}/n}.$$

- This yields estimates of the teacher and PD student prediction risks and their residual correlation (see Patil et al., 2021):

$$\widehat{R}(\lambda) = \frac{\|\hat{r}_\lambda\|^2}{n}, \quad \widehat{R}_{\mathrm{pd}}(\lambda) = \frac{\|\hat{r}_{\mathrm{pd},\lambda}\|^2}{n}, \quad \text{and} \quad \widehat{C}(\lambda) = \frac{\langle \hat{r}_\lambda, \hat{r}_{\mathrm{pd},\lambda} \rangle}{n}.$$

Plug-In Estimator of $\xi^\star(\lambda)$

- Let $\hat{y}_\lambda = f_\lambda(X)$ and $\hat{y}_{\mathrm{pd},\lambda} = f_{\mathrm{pd},\lambda}(X)$ and consider the *effective degrees of freedom* of the teacher and the PD student

$$\mathrm{df}_\lambda = \mathrm{tr}(S_\lambda) \quad \text{and} \quad \mathrm{df}_{\mathrm{pd},\lambda} = \mathrm{tr}(S_{\mathrm{pd},\lambda}),$$

  where $S_\lambda = X(X^\top X/n + \lambda I)^{-1} X^\top/n$ is the hat matrix. See Patil et al. (2021).
- Evaluate the GCV-corrected residuals

$$\hat{r}_\lambda = \frac{y - \hat{y}_\lambda}{1 - \mathrm{df}_\lambda/n}, \quad \text{and} \quad \hat{r}_{\mathrm{pd},\lambda} = \frac{y - \hat{y}_{\mathrm{pd},\lambda}}{1 - \mathrm{df}_{\mathrm{pd},\lambda}/n}.$$

- This yields estimates of the teacher and PD student prediction risks and their residual correlation (see Patil et al., 2021):

$$\widehat{R}(\lambda) = \frac{\|\hat{r}_\lambda\|^2}{n}, \quad \widehat{R}_{\mathrm{pd}}(\lambda) = \frac{\|\hat{r}_{\mathrm{pd},\lambda}\|^2}{n}, \quad \text{and} \quad \widehat{C}(\lambda) = \frac{\langle \hat{r}_\lambda, \hat{r}_{\mathrm{pd},\lambda} \rangle}{n}.$$

- Compute the one-shot estimators for the optimal mixing weight and prediction risk:

$$\hat{\xi}^\star(\lambda) = \frac{\widehat{R}(\lambda) - \widehat{C}(\lambda)}{\widehat{R}(\lambda) + \widehat{R}_{\mathrm{pd}}(\lambda) - 2\widehat{C}(\lambda)} \quad \text{and} \quad \widehat{R}^\star_{\mathrm{sd}}(\lambda) = \widehat{R}(\lambda) - \frac{(\widehat{R}(\lambda) - \widehat{C}(\lambda))^2}{\widehat{R}(\lambda) + \widehat{R}_{\mathrm{pd}}(\lambda) - 2\widehat{C}(\lambda)}.$$

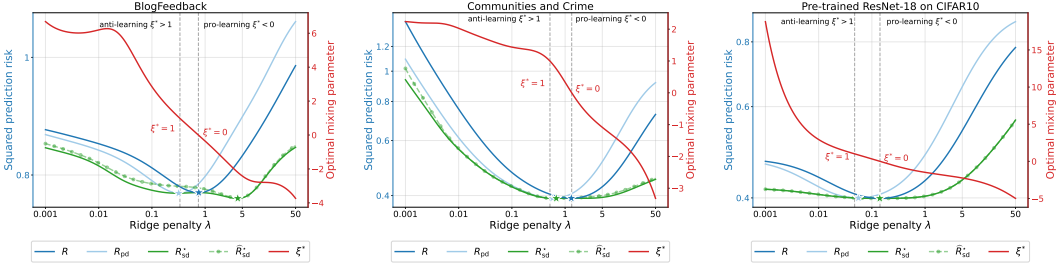
Consistency of One-Shot Tuning

Theorem

For each fixed $\lambda > 0$, almost surely as $p/n \to \gamma$,

$$\hat{\xi}^\star(\lambda) - \xi^\star(\lambda) \to 0, \quad \widehat{R}^\star_{\mathrm{sd}}(\lambda) - R^\star_{\mathrm{sd}}(\lambda) \to 0.$$

Consistency of One-Shot Tuning



(a) $n = 2619, p = 280$ (b) $n = 398, p = 99$ (c) $n = 2000, p = 512$

- $\widehat{R}^\star_{\text{sd}}$ tracks the true test risk closely, especially where train/test discrepancy is small (CIFAR benchmarks).
- One-shot correctly selects *negative* $\hat{\xi}^\star(\lambda)$ in the over-regularized regime.

Variants and Extensions

Multi-round SD

Recursive optimal SD gives monotone risk decrease:

$$R_{k+1}(\lambda) \le R_k(\lambda).$$

Ridge variants

Generalized ridge and kernel ridge inherit the sign rule through:

$$f_\lambda - f_{\mathrm{pd},\lambda} = -\lambda \partial_\lambda f_\lambda.$$

Fresh-X variants

Same-X coupling is crucial; in an isotropic setting, same-X optimal SD dominates affine fresh-X SD.

Structural results are not just a ridge-regression calculation;
they reflect a resolvent-path geometry shared by several ridge-type smoothers.

Multi-Round Optimal-SD

- We consider a **recursive** variant of optimal SD:
  - Set $y^{(0)} := y$ and
  - after iteration $k \geq 0$, define a vector of new response variables

$$y^{(k+1)} = f^{(k)}_{\mathrm{sd},\lambda,\xi_k^\star}(X),$$

    where $f^{(k)}_{\mathrm{sd},\lambda,\xi_k^\star}$ is the round-k optimal SD predictor.
  - Re-run optimal-SD on $(X, y^{(k+1)})$ to get $f^{(k+1)}_{\mathrm{sd},\lambda,\xi_{k+1}^\star}$.

Multi-Round Optimal-SD

- We consider a **recursive** variant of optimal SD:
  - Set $y^{(0)} := y$ and
  - after iteration $k \geq 0$, define a vector of new response variables

$$y^{(k+1)} = f^{(k)}_{\mathrm{sd},\lambda,\xi_k^\star}(X),$$

    where $f^{(k)}_{\mathrm{sd},\lambda,\xi_k^\star}$ is the round-k optimal SD predictor.
  - Re-run optimal-SD on $(X, y^{(k+1)})$ to get $f^{(k+1)}_{\mathrm{sd},\lambda,\xi_{k+1}^\star}$.

- At each round, the properties of multi-round optimal-SD are as above. Thus, the pointwise prediction risk is non-increasing in k.

Monotonicity of Multi-Round Optimal-SD

For any fixed $\lambda > 0$, almost surely,

$$R(f^{(k+1)}_{\mathrm{sd},\lambda,\xi_{k+1}^\star}) \leq R(f^{(k)}_{\mathrm{sd},\lambda,\xi_k^\star}) \quad \text{for all } k \geq 0.$$

Extensions to Ridge (Resolvent-Based) Smoothers

- Let f_λ be a teacher predictor such that there exists a vector-valued map $s_\lambda(\cdot) \in \mathbb{R}^n$ (depending on X, λ but not on y) satisfying $f_\lambda(x) = s_\lambda(x)^\top y$. Define the PD $f_{\mathrm{pd},\lambda}$ and SD predictor $f_{\mathrm{sd},\lambda,\xi}$ similarly as in the ridge setting.
- Assume that this ridge-smoother family satisfies the derivative identity

$$f_\lambda(x) - f_{\mathrm{pd},\lambda}(x) = -\lambda\, \partial_\lambda f_\lambda(x), \quad \text{for all } x \in \mathbb{R}^p.$$

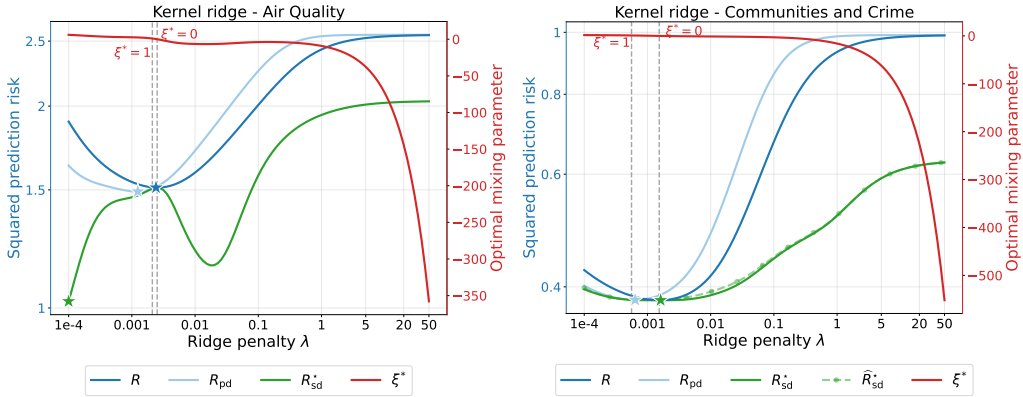
Under the derivative identity, for any $\lambda > 0$, almost surely,

$$\xi^\star(\lambda) = -\frac{\lambda}{2}\frac{R'(\lambda)}{D(\lambda)}, \quad \text{and} \quad R^\star_{\mathrm{sd}}(\lambda) = R(\lambda) - \frac{\lambda^2}{4}\frac{(R'(\lambda))^2}{D(\lambda)}.$$

If $R'(\lambda) \neq 0$, then $R^\star_{\mathrm{sd}}(\lambda) < R(\lambda)$ and $\operatorname{sign}(\xi^\star(\lambda)) = -\operatorname{sign}(R'(\lambda))$.

- The derivative identity is satisfied by *generalized ridge* and *kernel ridge*.

Extensions to Other Ridge Families



Gaussian kernel ridge and kernel SD ridge regression.

Extension to Fresh-X

- So far we have assumed that the teacher and the PD student use the same design matrix X.
- A variant of SD is to train the PD student using fresh independent **unlabeled** data $\tilde{X} \in \mathbb{R}^{m \times p}$ with the teacher's labels $\tilde{y}_\lambda = f_\lambda(\tilde{X}) \in \mathbb{R}^m$.

Extension to Fresh-X

- So far we have assumed that the teacher and the PD student use the same design matrix X.
- A variant of SD is to train the PD student using fresh independent **unlabeled** data $\tilde{X} \in \mathbb{R}^{m \times p}$ with the teacher's labels $\tilde{y}_\lambda = f_\lambda(\tilde{X}) \in \mathbb{R}^m$.
- Let $f^{\mathrm{fr}}_{\mathrm{pd},\lambda}$ denote the ridge predictor trained on the fresh covariates and the pseudo-labeled dataset $(\tilde{X}, \tilde{y}_\lambda)$. The fresh-$X$ SD predictor with mixing ξ is

$$f^{\mathrm{fraff}}_{\mathrm{sd},\lambda,\xi} = (1-\xi) f_\lambda + \xi f^{\mathrm{fr}}_{\mathrm{pd},\lambda}.$$

The optimal-SD risk is $R^{\star,\mathrm{fraff}}_{\mathrm{sd}}(\lambda) = \min_{\xi \in \mathbb{R}} R(f^{\mathrm{fraff}}_{\mathrm{sd},\lambda,\xi})$.

Fresh-X Optimal-SD Improves on the Teacher

For all $\lambda > 0$, almost surely,

$$R^{\star,\mathrm{fraff}}_{\mathrm{sd}}(\lambda) \leq R(\lambda)$$

and the optimal mixing weight obeys the sign rule.

Same-X versus Fresh-X

- It is not clear how to quantify the gain in the fresh-X setting, since the coupling between the teacher and the PD student is lost.
- In simulations, the fresh-X SD predictor is often worse than the same-X SD predictor, especially when the teacher is over-regularized.

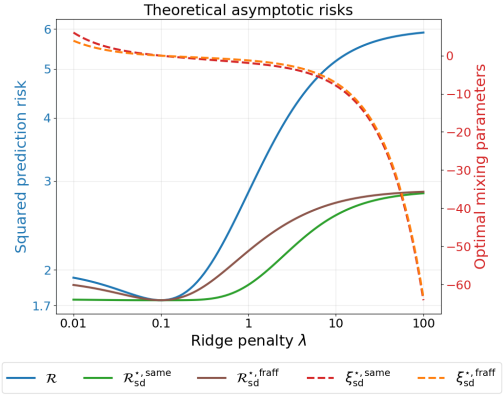
- In the special case of isotropic covariates, we can make this precise.

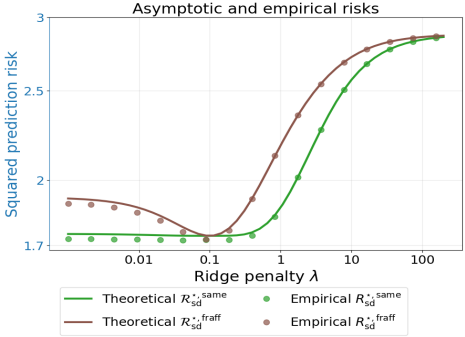
Theorem (Same-X Dominates Fresh-X in Isotropic Setting)

Assume $\Sigma = I_p$ and $\beta \sim \mathcal{N}(0, (r^2/p)I_p)$. Then, for any λ, as $p/m, p/n \to \gamma \in (0, \infty)$, the limiting risks satisfy

$$\mathscr{R}_{\mathrm{sd}}^{\star,\mathrm{same}}(\lambda) \;\leq\; \mathscr{R}_{\mathrm{sd}}^{\star,\mathrm{fraff}}(\lambda).$$

Same-X versus Fresh-X





Theoretical asymptotic and empirical risks (averaged over 20 simulations) in isotropic setting with $n = 400$, $p = 200$, $r^2 = 5$, and $\sigma^2 = 1$.

Prediction-Only SD (Ongoing Work)

SD can help in the PPI-like settings in which we only have access to fresh, unlabeled data and the teacher predictions on them.

Prediction-Only SD

Inspired by common practices in AI modeling, we are studying the scenario in which only the trained teacher predictions (as a black box) and fresh unlabeled data are available. These are precisely the settings considered in the prediction-powered inference (PPI) literature.

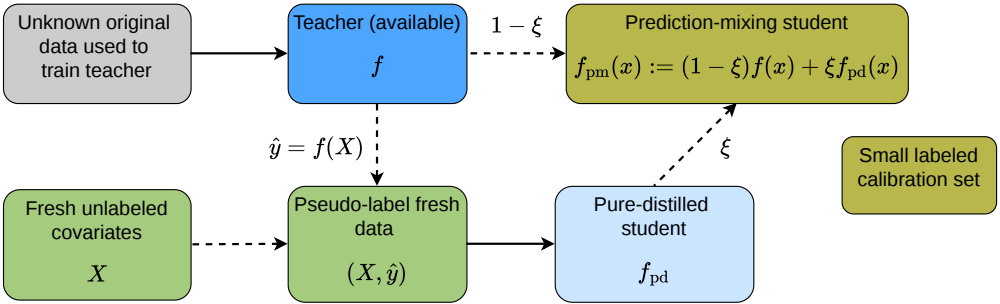
Prediction-Only SD Setting

- The ridge teacher f_{λ_t} is trained on data $\mathcal{D}_{\text{lab}} = (X, y)$. We have no knowledge of $\mathcal{D}_{\text{lab}}$ and $\lambda_t > 0$.
- We have access to a fresh unlabeled dataset $\mathcal{D}_{\text{unlab}} = \tilde{X}$ and the corresponding teacher's predictions $\tilde{y}_{\lambda_t} = f_{\lambda_t}(\tilde{X})$.
- For a tunable $\lambda_s > 0$, we train a PD student f_{pd,λ_s} on $(\tilde{X}, \tilde{y}_{\lambda_t})$ and consider the prediction-mixing (PM) student

$$f_{\text{pm},\lambda_t,\lambda_s,\xi} = (1-\xi) f_{\lambda_t} + \xi f_{\text{pd},\lambda_s}, \quad \xi \in \mathbb{R}.$$

- We may also have a small labeled calibration set for tuning.

Prediction-Only SD



Caveat: As the teacher and PD student use different datasets, the analysis is more challenging. The derivative identity and the resolvent path geometry are lost, and so is the coupling between the teacher and the PD student.

Optimal-PM Student

- Now consider the conditional prediction risk

$$R(f) := \mathbb{E}\big[(y_{\text{new}} - f(x_{\text{new}}))^2 \mid \mathcal{D}_{\text{lab}}, \mathcal{D}_{\text{unlab}}\big],$$

for some fresh test point $(x_{\text{new}}, y_{\text{new}})$ independent of the training data.

- Let $R(\lambda_t) = R(f_{\lambda_t})$ and $R_{\text{pm}}(\lambda_t, \lambda_s, \xi) = R(f_{\text{pm},\lambda_t,\lambda_s,\xi})$ be the teacher's and PM student's risks. The optimal mixing weight

$$\xi^\star = \xi^\star(\lambda_t, \lambda_s) = \operatorname*{argmin}_{\xi \in \mathbb{R}} R_{\text{pm}}(\lambda_t, \lambda_s, \xi)$$

yields the optimal risk $R^\star_{\text{pm}}(\lambda_t, \lambda_s) = R_{\text{pm}}(\lambda_t, \lambda_s, \xi^\star)$ for the optimal-PM student.

Optimal-PM Improves Upon the Teacher

Closed-Form Expressions for Optimal-PM Risk and Mixing

Just like in the SD case, we define the correlation and gap quantities

$$C(\lambda_t, \lambda_s) = \mathbb{E}\big[(y_{\text{new}} - f_{\lambda_t}(x_{\text{new}}))(y_{\text{new}} - f_{\text{pd},\lambda_s}(x_{\text{new}})) \mid \mathcal{D}_{\text{lab}}, \mathcal{D}_{\text{unlab}}\big] \quad \text{and}$$
$$D(\lambda_t, \lambda_s) = \mathbb{E}\big[(f_{\lambda_t}(x_{\text{new}}) - f_{\text{pd},\lambda_s}(x_{\text{new}}))^2 \mid \mathcal{D}_{\text{lab}}, \mathcal{D}_{\text{unlab}}\big].$$

If $D(\lambda_t, \lambda_s) > 0$, then

$$\xi^\star = \frac{R(\lambda_t) - C(\lambda_t, \lambda_s)}{D(\lambda_t, \lambda_s)} \quad \text{and} \quad R^\star_{\text{pm}}(\lambda_t, \lambda_s) = R(\lambda_t) - \underbrace{\frac{(R(\lambda_t) - C(\lambda_t, \lambda_s))^2}{D(\lambda_t, \lambda_s)}}_{\geq 0}.$$

Optimal-PM strictly improves upon the teacher if $R(\lambda_t) \neq C(\lambda_t, \lambda_s)$!

Tuning Optimal-PM

$\xi^\star$ is not estimable!

Due to the lack of labeled data, the optimal mixing parameter $\xi^\star$ is not identifiable.

Tuning Optimal-PM

$\xi^\star$ is not estimable!

Due to the lack of labeled data, the optimal mixing parameter $\xi^\star$ is not identifiable.

- Still, $\xi^\star$ can be estimated with an **independent labeled calibration set** $\mathcal{D}_{\text{cal}} := (x_i^{\text{cal}}, y_i^{\text{cal}})_{i=1}^{n_{\text{cal}}}$ (possibly OOD).
- Define the calibration residuals

$$\hat{e}_i^{\text{t}} := y_i^{\text{cal}} - f_{\lambda_t}(x_i^{\text{cal}}) \quad \text{and} \quad \hat{e}_i^{\text{pd}} := y_i^{\text{cal}} - f_{\text{pd},\lambda_s}(x_i^{\text{cal}}), i = 1, \ldots, n_{\text{cal}},$$

and estimate the three oracle quantities by

$$\widehat{R}_{\lambda_t}^{\text{cal}} := \frac{1}{n_{\text{cal}}} \sum_{i=1}^{n_{\text{cal}}} (\hat{e}_i^{\text{t}})^2, \quad \widehat{R}_{\text{pd},\lambda_s}^{\text{cal}} := \frac{1}{n_{\text{cal}}} \sum_{i=1}^{n_{\text{cal}}} (\hat{e}_i^{\text{pd}})^2, \quad \widehat{C}_{\lambda_t,\lambda_s}^{\text{cal}} := \frac{1}{n_{\text{cal}}} \sum_{i=1}^{n_{\text{cal}}} \hat{e}_i^{\text{t}} \hat{e}_i^{\text{pd}}.$$

The plug-in estimators of the optimal mixing parameter and the optimal PM risk are

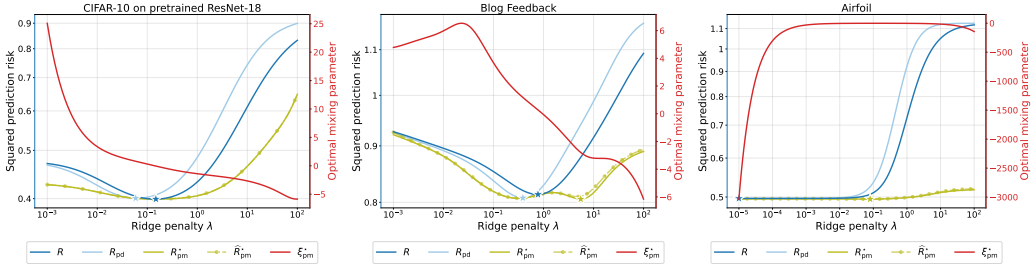
$$\widehat{\xi}_{\text{cal}}^\star(\lambda_t, \lambda_s) = \frac{\widehat{R}_{\lambda_t}^{\text{cal}} - \widehat{C}_{\lambda_t,\lambda_s}^{\text{cal}}}{\widehat{R}_{\lambda_t}^{\text{cal}} + \widehat{R}_{\text{pd},\lambda_s}^{\text{cal}} - 2\widehat{C}_{\lambda_t,\lambda_s}^{\text{cal}}} \quad \text{and} \quad \widehat{R}_{\text{pm,cal}}^\star(\lambda_t, \lambda_s) = \widehat{R}_{\lambda_t}^{\text{cal}} - \frac{(\widehat{R}_{\lambda_t}^{\text{cal}} - \widehat{C}_{\lambda_t,\lambda_s}^{\text{cal}})^2}{\widehat{R}_{\lambda_t}^{\text{cal}} + \widehat{R}_{\text{pd},\lambda_s}^{\text{cal}} - 2\widehat{C}_{\lambda_t,\lambda_s}^{\text{cal}}}.$$

Tuning Optimal-PM

As $n_{\text{cal}} \to \infty$ and assuming that the calibration residuals are square integrable conditionally on $\mathcal{D}_{\text{lab}}, \mathcal{D}_{\text{unlab}}$, almost surely,

$$\widehat{\xi}^{\star}_{\text{cal}}(\lambda_t, \lambda_s) - \xi^{\star}(\lambda_t, \lambda_s) \to 0 \quad \text{and} \quad \widehat{R}^{\star}_{\text{pm,cal}}(\lambda_t, \lambda_s) - R^{\star}_{\text{pm}}(\lambda_t, \lambda_s) \to 0,$$

for all (even data dependent) choices of λ_t and λ_s.



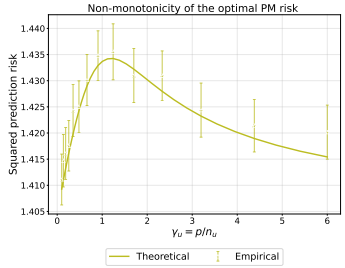
(a) $p = 512, n_l = 2000, n_u = 4000, n_{\text{cal}} = 200$

(b) $p = 280, n_l = 2619, n_u = 15719, n_{\text{cal}} = 524$

(c) $p = 5, n_l = 150, n_u = 150, n_{\text{cal}} = 76$

More Unlabeled Data Does Not Always Help!

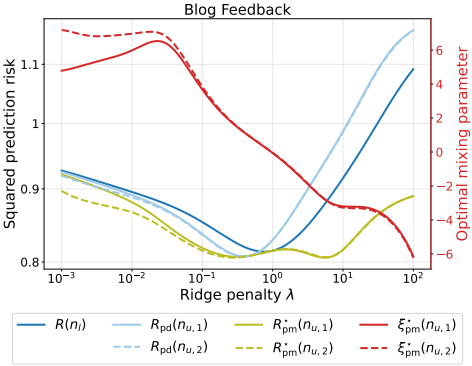
- Intuitively, more unlabeled data should help. Surprisingly, this is not always the case!
- Assume that $\lambda = \lambda_t = \lambda_s$ and that $p/n_{\text{unlab}} \to \gamma_{\text{u}}$, while $p/n_{\text{lab}} \to \gamma_{\text{l}}$. Assume that the feature distribution is isotropic, a linear model and Gaussian signal and noise. In this setting, we can derive an exact asymptotic characterization of the optimal PM risk $R^{\star}_{\text{pm}}(\lambda, \lambda)$ as a function of γ_{l} and γ_{u}.
- For any λ different from the optimal $\lambda^{\star}$ minimizing the asymptotic teacher risk, there are combinations of $(\gamma_{\text{l}}, \gamma_{\text{u}})$ for which the optimal PM risk is non-monotone in γ_{u}.



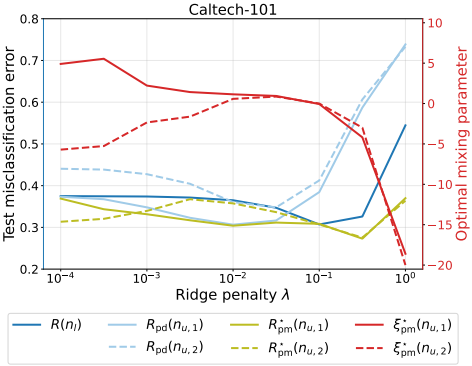
Non-monotonicity in γ_u at $\lambda = 0.2$, isotropic setting.

Non-monotonicity in γ_u

- Experimentally, we have observed this non-monotonicity phenomenon in real datasets, including the Blog Feedback and the Caltech-101 dataset. Below, the solid lines correspond to a larger unlabeled sample size and the dashed lines to the smaller unlabeled sample size. The two types of lines cross.



(a) Ridge regression on Blog Feedback with $p = 280, n_l = 2619, n_{u,1} = 15719, n_{u,2} = 5240$.



(b) Regularized linear probing on Caltech-101 on pre-trained ResNet-34 features with $p = 512, n_l = 1500,$

Summary

We have characterized SD in ridge regression and obtained new results.

- Structural: optimal-SD strictly improves every nonstationary point of the teacher risk curve. Sign rule: $\operatorname{sign}(\xi^\star) = -\operatorname{sign}(R')$.
- Asymptotic: exact deterministic equivalents quantify the SD gain under anisotropic covariance and deterministic signal.
- Algorithmic: one-shot GCV tuning is consistent and avoids grid search, data splitting, and refitting over ξ.
- Extensions: multi-round optimal-SD, kernel and generalized ridge regression, fresh-X and prediction-only SD.

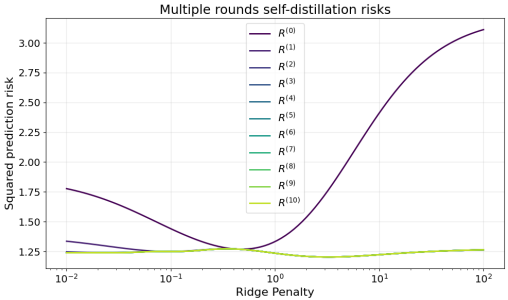
Future Directions

- SD in classification. Preliminary results (by others and us) indicate that SD can deliver significant gains.
- Multi-round optimal-SD. Simulation results seem to indicate that just a few rounds of SD bring the risk profile down significantly.
- Allow the student to optimize its own λ over multiple rounds.

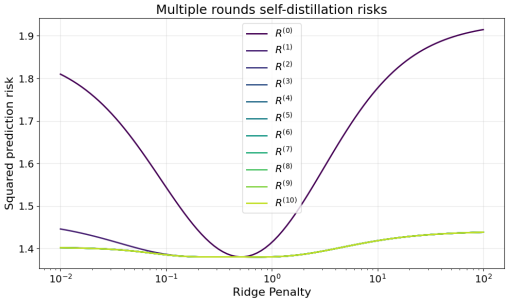
Thank you

Multi-Round Optimal-SD: Does it Converge to What?

- We have noticed in simulations that multi-round Optimal-SD converges quickly to a flat risk profile. However, we do not understand it.



AR1 features/aligned signal



isotropic features/signal

Multi-Round Optimal-SD vs. Anchored SD

- In our iterative scheme

$$f^{(k+1)}_{\mathrm{sd},\lambda,\xi^\star_{k+1}} = (1-\xi^\star_{k+1}) f^{(k)}_{\lambda} + \xi^\star_{k+1} f^{(k)}_{\mathrm{pd},\lambda},$$

  with $f^{(k)}_{\lambda}$ and $f^{(k)}_{\mathrm{pd},\lambda}$ being the teacher and PD student trained on $(X, y^{(k)})$.
- In the literature (Garg et al., 2025; Alemohammad et al., 2023), a common practice is to remain **anchored** to the original teacher and mix it with the PD student trained on the previous round's SD pseudo-labels. This corresponds to a different family of SD predictors:

$$f^{(k+1)}_{\mathrm{sd},\lambda,\xi_{k+1}} = (1-\xi_{k+1}) f^{(0)}_{\lambda} + \xi_{k+1} f^{(k)}_{\mathrm{pd},\lambda},$$

  in which ξ_{k+1} may or may not be optimized.

Multi-Round Optimal-SD vs. Anchored SD

- In our iterative scheme

$$f_{\mathrm{sd},\lambda,\xi_{k+1}^{\star}}^{(k+1)} = (1-\xi_{k+1}^{\star}) f_{\lambda}^{(k)} + \xi_{k+1}^{\star} f_{\mathrm{pd},\lambda}^{(k)},$$

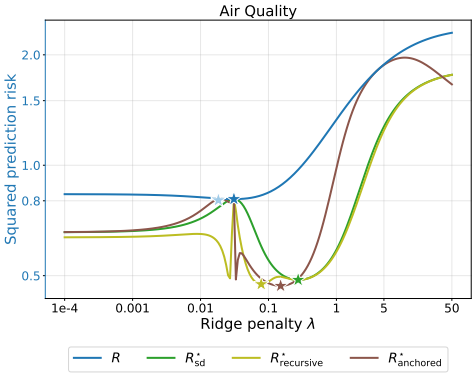
  with $f_{\lambda}^{(k)}$ and $f_{\mathrm{pd},\lambda}^{(k)}$ being the teacher and PD student trained on $(X, y^{(k)})$.
- In the literature (Garg et al., 2025; Alemohammad et al., 2023), a common practice is to remain **anchored** to the original teacher and mix it with the PD student trained on the previous round's SD pseudo-labels. This corresponds to a different family of SD predictors:

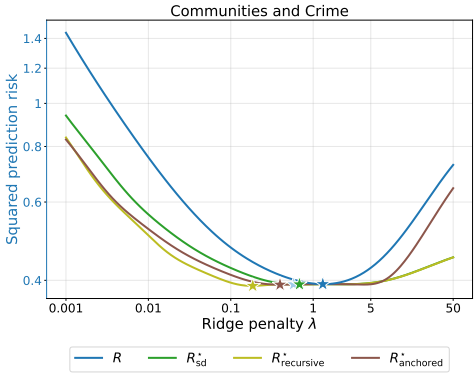
$$f_{\mathrm{sd},\lambda,\xi_{k+1}}^{(k+1)} = (1-\xi_{k+1}) f_{\lambda}^{(0)} + \xi_{k+1} f_{\mathrm{pd},\lambda}^{(k)},$$

  in which ξ_{k+1} may or may not be optimized.

The anchored scheme **does not yield a pointwise monotonically decreasing risk.**

Multi-Round Optimal-SD vs. Anchored SD

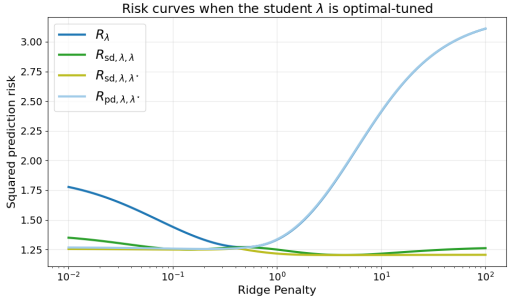




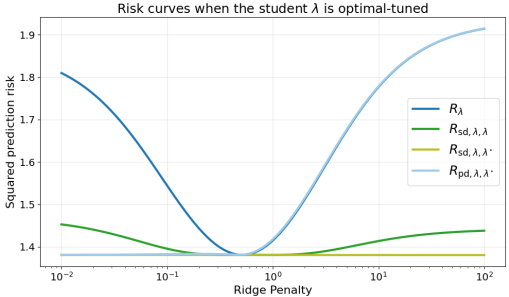
Round-2 recursive ($R^\star_{\text{recursive}}$, in olive) vs. Round-2 anchored ($R^\star_{\text{anchored}}$, in brown). Anchored mixing can be non-monotone (i.e., above the green curve).

What If the SD Student Uses Their Own, Optimal λ?

In the same-X setting, the PD student could use a different value for λ than the teacher. If this value is chosen optimally, then the risk of the resulting optimal-SD student is significantly smaller, according to simulations.



AR1 features/aligned signal



isotropic features/signal

Comments About Prediction-Only SD

- As the sign rule (and curvature test, for that matter) no longer holds, it is hard to relate the gains to mis-regularization.
- We have preliminary results about the form of the asymptotic risks of the teacher, PD and PM students under RMT conditions.
- We find that, in some special cases, the gains over both the teacher and the PD student are robust: the optimal PM risk is *strictly* smaller over all choices of λ_t and for all but finitely many choices of λ_s.