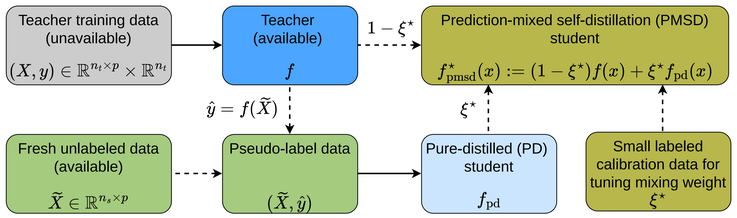
Prediction-only Distillation in Linear and Logistic Regression

Hien Dang[1] Pratik Patil[1] Alessandro Rinaldo[1]

[1]Department of Statistics and Data Sciences, UT Austin

Distillation in prediction-only regime



Motivation. Standard self-distillation retrains the student on the teacher's original training data. After deployment, however, these data may be unavailable because of privacy, storage or data-sharing constraints.

Can fresh, potentially OOD, unlabeled data improve a fixed predictor when its original training data are unavailable?

Our approach. We train a pure-distilled student (PD) on fresh covariates pseudo-labeled by the teacher, then combine the teacher and student predictions affinely. We call the resulting predictor the *prediction-mixed self-distillation* (PMSD) student.

Setup for ridge regression.

Teacher and PD student.

Let β be ridge estimator trained on $\mathcal{D} = (X, y)$ with penalty λ_t, and let β_{pd} be ridge regression trained on $(\widetilde{X}, \widetilde{X}\widehat{\beta})$ with penalty λ_s. Thus,

$$f(x) = x^\top \widehat{\beta}, \qquad f_{\mathsf{pd}}(x) = x^\top \widehat{\beta}_{\mathsf{pd}}.$$

PMSD student.

$$f_{\mathsf{pmsd},\xi}(x) = (1-\xi) f(x) + \xi f_{\mathsf{pd}}(x), \qquad \xi \in \mathbb{R}.$$

Out-of-sample prediction risk:

$$R(f) = \mathbb{E}\big[(y_0 - f(x_0))^2 \mid \mathcal{D}, \widetilde{X}\big].$$

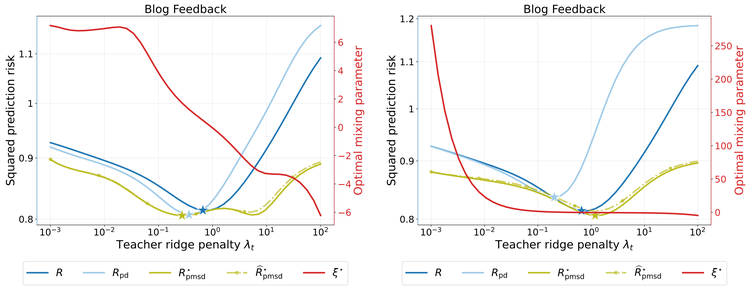
Denote $R := R(f)$, $R_{\mathsf{pd}} := R(f_{\mathsf{pd}})$ and $R^\star_{\mathsf{pmsd}} := \min_\xi R(f_{\mathsf{pmsd},\xi})$.

Correlation and separation terms.

$$C = \mathbb{E}[(y_0 - f(x_0))(y_0 - f_{\mathsf{pd}}(x_0)) \mid \mathcal{D}, \widetilde{X}], \qquad D = R + R_{\mathsf{pd}} - 2C \geq 0.$$

Proposition 1 (Optimal mixing decomposition). Assume $D > 0$, we have,

$$\xi^\star = \frac{R - C}{D}, \qquad R^\star_{\mathsf{pmsd}} = R - \frac{(R-C)^2}{D} \leq \min\{R, R_{\mathsf{pd}}\}.$$



(a) X, $\widetilde{X}$ sampled from the same pool (b) $\widetilde{X}$ sampled i.i.d. from $\mathcal{N}(0, I_p)$.

Figure 1: **Strict improvement over the teacher.** UCI Blog Feedback dataset with $n_t = 2619$, $n_s = 5240$, $n_{\mathsf{cal}} = 524$, $p = 280$, $\lambda_t = \lambda_s$.

Risk characterization in high-dimensional ridge

Data distribution. Each training label y_i has mean 0, and bounded 4^+ moments, and

- $X = Z\Sigma_t^{1/2}$ where $Z \in \mathbb{R}^{n_t \times p}$ has i.i.d. zero-mean, unit-variance entries with bounded 4^+ moment.
- $\widetilde{X} = \widetilde{Z}\Sigma_s^{1/2}$ where $\widetilde{Z} \in \mathbb{R}^{n_s \times p}$ has i.i.d. zero-mean, unit-variance entries with bounded 4^+ moment.
- Σ_s and Σ_t are jointly diagonalizable.

Define the linear projection and variance as $\beta := \Sigma_t^{-1}\mathbb{E}[xy]$ and $\sigma^2 := \mathrm{Var}(y - x^\top\beta)$. We can express $y = x^\top\beta + \varepsilon$ where ε is uncorrelated with x and $E(\varepsilon^2) = \sigma^2$.

Theorem 2 (General anisotropic risk asymptotics). For fixed $\lambda_t, \lambda_s > 0$, as $n_t, n_s, p \to \infty$ with bounded data aspect ratios $\gamma_t := p/n_t$, $\gamma_s := p/n_s$, we have

$$R \xrightarrow{\mathrm{p}} \mathcal{R}, \qquad R_{\mathsf{pd}} \xrightarrow{\mathrm{p}} \mathcal{R}_{\mathsf{pd}}, \qquad C \xrightarrow{\mathrm{p}} \mathcal{C}, \qquad R^\star_{\mathsf{pmsd}} \xrightarrow{\mathrm{p}} \mathcal{R}^\star_{\mathsf{pmsd}} := \mathcal{R} - \frac{\big(\mathcal{R} - \mathcal{C}\big)^2}{\mathcal{R} + \mathcal{R}_{\mathsf{pd}} - 2\mathcal{C}},$$

where $\mathcal{R}, \mathcal{R}_{\mathsf{pd}}$ and $\mathcal{C}$ are computable functions of $(\Sigma_t, \Sigma_s, \beta, \sigma^2, \gamma_t, \gamma_s)$.

Strict improvement and risk monotonicity

Strict improvement over the teacher. We characterize when strict improve $\mathrm{Gain}(\lambda_t, \lambda_s) := \mathcal{R}(\lambda_t) - \mathcal{R}^\star_{\mathsf{pmsd}}(\lambda_t, \lambda_s) > 0$ happen w.r.t. the pairs (λ_t, λ_s). Define the "teacher-degeneracy" and the "student-tie" set as

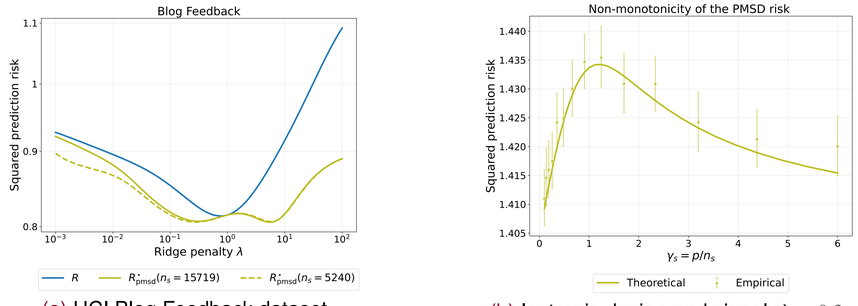
$$\Lambda_t = \{\lambda_t : \mathrm{Gain}(\lambda_t, \lambda_s) = 0 \; \forall \lambda_s > 0\}, \qquad \Lambda_s(\lambda_t) = \{\lambda_s : \mathrm{Gain}(\lambda_t, \lambda_s) = 0\}.$$

Proposition 3 (Crude bound on bad sets). Let m and $\widetilde{m}$ count the distinct eigenvalues of Σ_t and Σ_s, respectively. Then $|\Lambda_t| \leq 4m - 1$, and for $\lambda_t \notin \Lambda_t$, $|\Lambda_s(\lambda_t)| \leq \widetilde{m} - 1$.

Setting	Number of "tie" cases
Isotropic fresh data $\Sigma_s = I_p$, general Σ_t	$\lvert\Lambda_s(\lambda_t)\rvert = 0$
In-distribution isotropic $\Sigma_t = \Sigma_s = I_p$	$\lvert\Lambda_t\rvert = 1$, $\lvert\Lambda_s(\lambda_t)\rvert = 0$
In-distribution s-spike model	$\lvert\Lambda_s(\lambda_t)\rvert \leq s$

Strict improvement over the PD student. In the in-distribution isotropic setting, for every λ_t, $\mathcal{R}_{\mathsf{pd}}(\lambda_s) - \mathcal{R}^\star_{\mathsf{pmsd}}(\lambda_t, \lambda_s) > 0$ for all but at most two values of λ_s.

More unlabeled data need not help monotonically if λ_s is not retuned. In the in-distribution isotropic setting and $\lambda_t = \lambda_s$, the PMSD risk is *not* globally monotone in the unlabeled aspect ratio $\gamma_s = p/n_s$.



(a) UCI Blog Feedback dataset (b) Isotropic design and signal, $\lambda = 0.2$.

Figure 2: Non-monotonicity of the PMSD risk with respect to the amount of unlabeled data at fixed λ_s.

More unlabeled data is helpful when λ_s is optimally retuned. However, if λ_s is optimally retuned, the theory recommends using all available unlabeled data because $\inf_{\lambda_s > 0} \mathcal{R}^\star_{\mathsf{pmsd}}(\gamma_{s,1}, \lambda_s) \leq \inf_{\lambda_s > 0} \mathcal{R}^\star_{\mathsf{pmsd}}(\gamma_{s,2}, \lambda_s)$ for $\gamma_{s,1} < \gamma_{s,2}$. This holds for general anisotropic covariances Σ_t, Σ_s and deterministic signal β.

Data-dependent tuning

Define the calibration residuals using an independent labeled sample

$$\widehat{e}_i := y_i^{\mathrm{cal}} - f(x_i^{\mathrm{cal}}) \quad \text{and} \quad \widehat{e}_i^{\,\mathsf{pd}} := y_i^{\mathrm{cal}} - f_{\mathsf{pd}}(x_i^{\mathrm{cal}}),$$

and estimate the three oracle quantities by

$$\widehat{R}^{\mathrm{cal}} := \frac{1}{n_{\mathrm{cal}}}\sum_{i=1}^{n_{\mathrm{cal}}} (\widehat{e}_i)^2, \qquad \widehat{R}^{\mathrm{cal}}_{\mathsf{pd}} := \frac{1}{n_{\mathrm{cal}}}\sum_{i=1}^{n_{\mathrm{cal}}} (\widehat{e}_i^{\,\mathsf{pd}})^2, \qquad \widehat{C}^{\mathrm{cal}} := \frac{1}{n_{\mathrm{cal}}}\sum_{i=1}^{n_{\mathrm{cal}}} \widehat{e}_i \widehat{e}_i^{\,\mathsf{pd}}.$$

Define the one-shot calibration estimators of $\xi^\star$ and $R^\star_{\mathsf{pmsd}}$ by

$$\widehat{\xi}^\star_{\mathrm{cal}} := \frac{\widehat{R}^{\mathrm{cal}} - \widehat{C}^{\mathrm{cal}}}{\widehat{R}^{\mathrm{cal}} + \widehat{R}^{\mathrm{cal}}_{\mathsf{pd}} - 2\widehat{C}^{\mathrm{cal}}}, \quad \text{and} \quad \widehat{R}^\star_{\mathsf{pmsd},\mathrm{cal}} := \widehat{R}^{\mathrm{cal}} - \frac{(\widehat{R}^{\mathrm{cal}} - \widehat{C}^{\mathrm{cal}})^2}{\widehat{R}^{\mathrm{cal}} + \widehat{R}^{\mathrm{cal}}_{\mathsf{pd}} - 2\widehat{C}^{\mathrm{cal}}}.$$

Theorem 4 (Calibration consistency). For an independent calibration sample, under mild moment and nondegeneracy conditions, as $n_{\mathrm{cal}} \to \infty$,

$$\widehat{\xi}^\star_{\mathrm{cal}} - \xi^\star \xrightarrow{\mathrm{p}} 0, \qquad \widehat{R}^\star_{\mathsf{pmsd},\mathrm{cal}} - R^\star_{\mathsf{pmsd}} \xrightarrow{\mathrm{p}} 0.$$

We target the regime where $n_{\mathrm{cal}}/p \to 0$, thus calibration set is statistically effective for tuning the scalar mixing weight, with estimation rate $O(n_{\mathrm{cal}}^{-1/2})$, but asymptotically insubstantial as training sample.

Remark. Generally, teacher predictions and unlabeled covariates alone cannot identify the optimal mixing weight.

Logistic regression

Dataset. We have access to an *unlabeled*, class-balanced dataset $\mathcal{D}_{\mathrm{unlab}} = \{x_i\}_{i=1}^{2n}$ consisting of $2n$ i.i.d. draws, with unobserved ground-truth labels $y_i \in \{0, 1\}$.

Teacher. A classifier that returns a hard label $\widehat{y} \in \{0, 1\}$ for any x. Suppose its population classification error is an unknown $\rho \in (0, 1)$ and is equal for both classes.

PD student. Using a fixed feature map ϕ, we fit a regularized logistic classifier to the teacher's pseudo-labels $\{\widehat{y}_i\}$ on unlabeled dataset,

$$\theta_{\mathsf{pd}} := \operatorname*{argmin}_\theta \frac{1}{2n}\sum_{i=1}^{2n} \mathrm{CE}\big(\widehat{y}_i, \sigma(\langle\theta, \phi(x_i)\rangle)\big) + \frac{\lambda}{2}\|\theta\|^2,$$
$$y_{\mathsf{pd}}(x) := \sigma(\langle\theta_{\mathsf{pd}}, \phi(x)\rangle).$$

PMSD student.

$$y_{\mathsf{pmsd},\xi}(x) = \mathbb{I}\big\{(1-\xi)\widehat{y}(x) + \xi y_{\mathsf{pd}}(x) \geq 0.5\big\}, \quad \xi \in \mathbb{R}.$$

Assumptions. We analyze a stylized feature geometry in which features have unit norm. Feature vectors are orthogonal across classes, and have constant correlation $c \in (0, 1)$ within each class.

Theorem 5 (Strict improvement). Let $\bar{\lambda}_n := 2n\lambda$ and suppose that $\bar{\lambda}_n \to \bar{\lambda} \in (0, \infty)$ as $n \to \infty$. There exists a threshold $\rho_0 \in (0, 0.5)$ such that, for every $\rho \in (\rho_0, 0.5)$, the PD student y_{pd} has $100(1-\rho)\%$ asymptotic accuracy. For every $\rho < 0.5$, there exists ξ such that the PMSD student $y_{\mathsf{pmsd},\xi}$ achieves 100% asymptotic accuracy.

Table 1: Test classification accuracy on Caltech-101 with $\rho = 40\%$.

Regularization level λ	10^{-4}	$10^{-3.5}$	10^{-3}	$10^{-2.5}$	10^{-2}	$10^{-1.5}$	10^{-1}	$10^{-0.5}$	10^{0}
Teacher (%)	62.5	62.6	62.6	62.9	63.5	65.3	**69.3**	67.4	45.6
PD student (%)	62.7	63.3	65.2	67.7	**69.4**	68.4	61.6	41.3	26.2
Optimal PMSD student (%)	63.1	65.6	66.8	68.3	69.6	68.9	69.2	**72.7**	63.0
Estimated mixing weight	4.9	5.5	2.2	1.4	1.1	0.9	−0.1	**-4.2**	−18.6