# Prediction-Only Distillation in Linear and Logistic Regression

**Hien Dang** **Pratik Patil** **Alessandro Rinaldo**
Department of Statistics and Data Sciences, University of Texas at Austin
{hiendang, pratikpatil, alessandro.rinaldo}@utexas.edu

## Abstract

Self-distillation (SD) is typically studied when the student is retrained on the teacher's original training inputs. In many practical deployments, however, the labeled training data are no longer available, and one has access only to the trained predictor and fresh unlabeled covariates. We study SD in this *prediction-only* regime through a fresh-$X$ prediction-mixed scheme: a pure-distilled student is trained only on fresh covariates pseudo-labeled by the teacher, and the final predictor is an affine combination of the teacher and student predictions. For ridge regression under proportional asymptotics, we derive deterministic equivalents for the optimally mixed prediction risk. This risk is *strictly smaller* than the teacher risk for almost every pair of teacher and student regularization levels, including when the fresh covariates are out-of-distribution and even when their covariance is isotropic. Increasing the amount of fresh unlabeled data can worsen the optimally mixed risk at fixed student regularization, whereas optimal re-tuning guarantees non-increasing risk. Furthermore, we show that the optimal mixing weight is not identifiable in general from unlabeled covariates and teacher predictions alone, but can be consistently estimated in a single post-training step using a small independent labeled calibration set, without additional model fitting. Finally, as an extension, for binary logistic regression, we demonstrate that prediction mixing can outperform both the teacher and the pure-distilled classifier.

## 1 Introduction

Knowledge distillation (KD) has become a fundamental technique in modern machine learning [8, 3, 30]. In its original formulation, KD transfers knowledge from a high-capacity teacher to a more compact student and allows the student to achieve comparable performance with fewer parameters. This paradigm has since been extended to *self-distillation*, in which a model is retrained using its own predictions while retaining the same architecture [22, 70]. Extensive empirical evidence suggests that self-distillation can improve generalization [1, 14, 13, 25].

Self-distillation (SD) is typically studied when the student is retrained on the teacher's original training data. In many practical deployments, however, those labeled data are no longer available: one can query a trained predictor but cannot access its training data or parameters. We instead assume access to a stream of fresh unlabeled covariates, which may be out-of-distribution (OOD) relative to the teacher's training covariates. We refer to this regime as the *prediction-only* setting. It arises naturally in post-deployment adaptation and model reuse, particularly when the original training data cannot be retained or shared because of privacy, safety, or storage constraints [47, 40]. This setting raises a basic question:

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

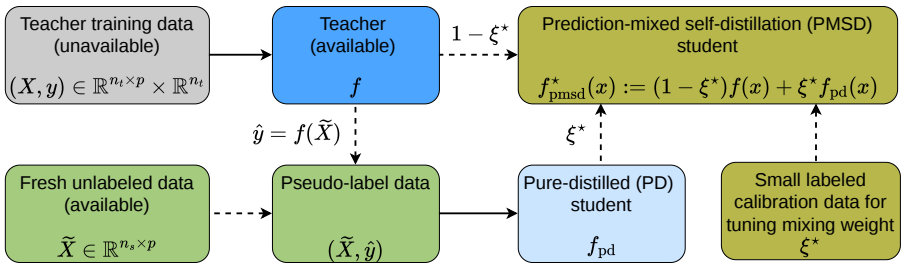



Figure 1: Schematic of the prediction-only setting and fresh-$X$ prediction-mixed SD. Solid arrows indicate training steps; dashed arrows indicate other information flows.

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

A natural approach is to train a student on fresh unlabeled covariates using the teacher's predictions as pseudo-labels. Because no ground-truth labels are used, this procedure yields a *pure-distilled* (PD) student [41]. Pure-distillation has been studied in several settings, including high-dimensional ridgeless regression [31], ridge regression [45], random-feature ridge regression [68], and Gaussian mixture models [12, 66, 60]. Perhaps surprisingly, the PD student can outperform its teacher. Such gains are not guaranteed, however, and the PD student can instead perform substantially worse; see, for example, Figure 2.

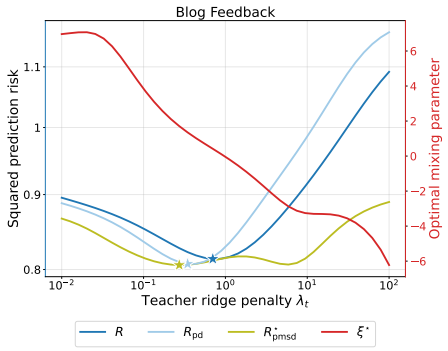
Recent work on same-$X$ ridge SD establishes strict-improvement guarantees [17, 49, 15, 39]. In ridge regression, the optimally mixed student lies on the affine path between the teacher and the PD student; this observation suggests directly combining their predictions as the final output. We call this affine combination predictor the *prediction-mixed self-distillation* (PMSD) student. Its mixing weight is real-valued and need not lie in $[0, 1]$, so the combination need not be convex. The method is illustrated in Figure 1.

This approach is operationally attractive: it requires no access to the original labeled data and reduces adaptation to a one-dimensional post hoc aggregation problem, with no additional model fitting. Establishing an analogous strict-improvement guarantee in the fresh-$X$ setting is theoretically nontrivial. The same-$X$ result relies on geometric identities induced by the shared design, and these identities disappear when the student is trained on a fresh design independent of the teacher's design. Before turning to theory, Figure 2 provides an empirical preview of the fresh-$X$ setting on the UCI Blog Feedback dataset. In both the in-distribution and out-of-distribution settings, the optimally mixed PMSD student outperforms the teacher and the PD student. This empirical preview motivates three theoretical and practical questions:

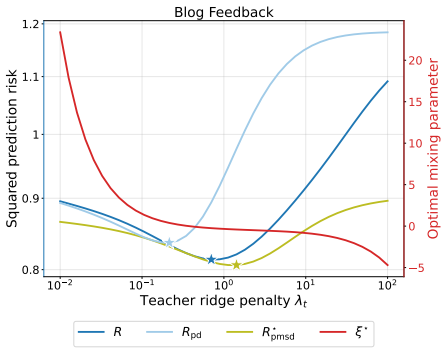
1. Under what conditions can the PMSD student improve upon the teacher and the PD student?
2. How does its performance depend on the amount of fresh unlabeled data?
3. Can the optimal mixing weight be tuned when the teacher's training data are unavailable?

Our main contributions, which also provide an outline of the paper, are as follows:

- **High-dimensional characterization for ridge regression (Section 3).** We first derive exact oracle identities for the optimal PMSD mixing weight (Proposition 3.1). Under proportional asymptotics, we then characterize the PMSD risk and optimal mixing weight under general anisotropic covariance and deterministic signal (Theorem 3.3).
- **Strict improvement and risk monotonicity (Section 4).** Using these deterministic equivalents, we show that the PMSD student strictly improves upon the teacher for almost every pair of regularization levels $(\lambda_t, \lambda_s)$ used by the teacher and the PD student (Theorem 4.1). Thus, strict improvement is generic even when $\lambda_t$ is unknown or $\lambda_s$ is not optimally tuned, provided that the mixing weight is chosen optimally.
  In the same-$\lambda$ isotropic setting, we also highlight a phenomenon absent from same-$X$ self-distillation: at fixed regularization, the PMSD risk is generically *non-monotone* in the unlabeled aspect ratio, so more unlabeled data need not improve and can hurt performance. However, if $\lambda_s$ is optimally re-tuned when adding more unlabeled data, the PMSD risk is non-increasing, so our theory recommends using all available unlabeled data (Section 4.2).




(a) Fresh unlabeled covariates follow the teacher's training distribution.




(b) Fresh unlabeled covariates are sampled i.i.d. from the isotropic Gaussian distribution $\mathcal{N}(0, I_p)$.

Figure 2: **Strict improvement of the PMSD student.** Test squared prediction risks of the ridge teacher ($R$, in blue), the fresh-$X$ PD student ($R_{\text{pd}}$, in light blue), and the optimally mixed PMSD student ($R_{\text{pm}}$, in olive), plotted as functions of the teacher's ridge penalty $\lambda_t$. The oracle optimal mixing weight $\xi^\star$ is in red. Setting: UCI Blog Feedback with $n_t = 2619$, $n_s = 5240$, and $p = 280$. We use the representative choice $\lambda_s = \lambda_t$. Experiments with alternative choices of $\lambda_s$, including values optimally tuned over a grid, are presented in Section H.1.

- **Consistent data-dependent tuning (Section 5).** We show that the teacher's outputs and fresh unlabeled covariates alone generally do not identify the oracle mixing weight. We then prove that a small independent labeled calibration sample enables consistent estimation in a single post-training step, without additional model fitting (Theorem 5.1).
- **Gains in logistic regression (Section 6).** Finally, we extend the analysis beyond regression under squared loss by studying binary logistic regression. We show that prediction mixing can strictly outperform both the teacher and the PD student (Theorems 6.2, 6.3 and G.2).

## 2 Related work

**Same-$X$ self-distillation.** A recent line of work studies SD in regression settings where the original labeled data remain available and the student is retrained on the same design. [44] interpret repeated SD as a regularization mechanism in Hilbert space. [17] and [49] analyze one-step and repeated SD for linear models under fixed-design assumptions. Under random design, [15] establish pointwise strict improvement guarantees, proportional asymptotic results, and a tuning method for ridge regression. [39] further show that, for spiked covariance matrices with $s$ spikes, $s$ rounds of SD attain optimal performance among spectral shrinkage estimators. Their results motivate using different regularization levels across SD rounds. Our paper studies a different regime: the *prediction-only fresh-$X$* setting, where the original labeled sample is unavailable and same-design geometric identities no longer apply. We instead ask whether prediction mixing can improve the teacher when its training data and regularization level $\lambda_t$ are unknown and the student penalty $\lambda_s > 0$ is arbitrary. Such pointwise guarantees are important because finding the optimal ridge penalty can be challenging. For example, [64] show that leave-one-out cross-validation loss is generally neither convex nor quasi-convex in $\lambda$.

**Fresh-$X$ pseudo-labeling and pure-distillation.** Empirical work shows that training on unlabeled data with teacher-generated pseudo-labels can improve performance [69, 55]. On the theoretical side, pure-distillation on fresh unlabeled covariates has been studied in high-dimensional ridgeless regression [31, 20] and ridge regression [45]. [68] analyze scaling laws for the teacher and PD student in random feature ridge regression. These works often consider a student stronger than the teacher, for example because it is more overparameterized or has better regularization; this regime is known as "weak-to-strong" distillation [9]. Although the PD student can outperform the teacher, such gains are not guaranteed. For example, in high-dimensional ridge regression, [45] show that when the teacher is over-regularized, the PD student cannot improve upon it and may perform worse; see Figure 2. These works focus primarily on the benefits of pure-distillation in teacher–student transfer, whereas we ask whether combining teacher and PD-student predictions can guarantee improvement over the teacher. We also study how PMSD risk depends on the fresh unlabeled sample size and how to calibrate the optimal mixing weight post hoc.

**Theory of self-distillation in classification.** A substantial literature studies the benefits of retraining on teacher pseudo-labels in classification. This includes linear probing on frozen feature backbones

with cross-entropy loss [17, 33], as well as self-training and retraining in Gaussian mixture models [48, 21, 65, 12, 32, 16, 66, 60]. For example, in the two-component Gaussian mixture setting, [16] show that retraining on the teacher's hard labels can improve accuracy under conditions on the label noise rate and sample size. For binary and multiclass classification under label noise, using a linear classifier on frozen features followed by a sigmoid or softmax layer (i.e., linear probing), [17, 33] show that the pure-distilled student, trained only on the teacher's soft labels, can tolerate higher noise rates and improve accuracy over the teacher. Our logistic regression analysis and linear probing experiments (see Section 6) complement both lines of work: we study a prediction-mixing mechanism that combines teacher and student predictions and can improve upon both the teacher and pure-distilled student.

# 3 Ridge regression

In this section, we introduce prediction-mixed distillation for ridge regression, derive its structural risk decomposition, and obtain deterministic equivalents under proportional asymptotics.

## 3.1 Setup and preliminaries

Let $f_{\lambda_t} : \mathbb{R}^p \to \mathbb{R}$ be the teacher ridge predictor trained on an *unknown* labeled dataset $\mathcal{D}_{\text{lab}} := (X, y) \in \mathbb{R}^{n_t \times p} \times \mathbb{R}^{n_t}$. Here, $X \in \mathbb{R}^{n_t \times p}$ is the design matrix with i.i.d. $p$-dimensional observations $x_i \in \mathbb{R}^p$ as its rows, and $y \in \mathbb{R}^{n_t}$ is the vector of corresponding scalar ground-truth labels $y_i$, for $i = 1, \ldots, n_t$. Denote the regularization parameter used by the teacher by $\lambda_t > 0$ (potentially depending on $\mathcal{D}_{\text{lab}}$). That is, the teacher ridge predictor is the random function

$$x \in \mathbb{R}^p \mapsto f_{\lambda_t}(x) := x^\top \operatorname*{argmin}_{\beta \in \mathbb{R}^p} \left\{ \|y - X\beta\|_2^2 / n_t + \lambda_t \|\beta\|_2^2 \right\}. \tag{1}$$

In the prediction-only regime considered here, we can query the teacher at any input but need not know the explicit form of $f_{\lambda_t}$ or its regularization parameter $\lambda_t$. Thus, we observe only the teacher's predictions at the queried points.

Given a fresh sample of unlabeled covariates $\mathcal{D}_{\text{unlab}} := \widetilde{X} \in \mathbb{R}^{n_s \times p}$ consisting of $n_s$ i.i.d. observations that may be *out-of-distribution* (OOD) relative to the teacher's training distribution, we query the teacher on this unlabeled set and denote its predictions by $\widehat{y}_{\lambda_t} := f_{\lambda_t}(\widetilde{X}) \in \mathbb{R}^{n_s}$. The pure-distilled (PD) student $f_{\mathsf{pd},\lambda_s}$ is the ridge predictor trained on $(\widetilde{X}, \widehat{y}_{\lambda_t})$ with penalty $\lambda_s > 0$ [1],

$$x \in \mathbb{R}^p \mapsto f_{\mathsf{pd},\lambda_s}(x) := x^\top \operatorname*{argmin}_{\beta \in \mathbb{R}^p} \left\{ \|\widehat{y}_{\lambda_t} - \widetilde{X}\beta\|_2^2 / n_s + \lambda_s \|\beta\|_2^2 \right\}. \tag{2}$$

A natural approach to improving the student's predictive performance is to combine the teacher's queried predictions with the student's own PD predictions. We focus on *affine mixing strategies*, in which the student deploys an affine combination of the available predictions. Specifically, for a mixing parameter $\xi \in \mathbb{R}$, the prediction-mixed self-distillation (PMSD) student predictor is given by

$$x \in \mathbb{R}^p \mapsto f_{\mathsf{pm},\lambda_t,\lambda_s,\xi}(x) := (1 - \xi) f_{\lambda_t}(x) + \xi f_{\mathsf{pd},\lambda_s}(x). \tag{3}$$

We consider affine rather than convex combinations because negative weights or weights exceeding one may be preferable. Indeed, empirical and theoretical work shows that the optimal mixing weight lies outside the unit interval in many settings [17, 49, 15].

Let $(x_0, y_0)$ denote an independent test point drawn from the teacher's training distribution, corresponding to the out-of-sample in-distribution setting. For any predictor $f$, define the conditional squared prediction risk

$$R(f) := \mathbb{E}\big[(y_0 - f(x_0))^2 \mid \mathcal{D}_{\text{lab}}, \mathcal{D}_{\text{unlab}}\big]. \tag{4}$$

For convenience, denote $R(\lambda_t) := R(f_{\lambda_t})$, $R_{\mathsf{pd}}(\lambda_s) := R(f_{\mathsf{pd},\lambda_s})$, and $R_{\mathsf{pm}}(\lambda_t, \lambda_s, \xi) := R(f_{\mathsf{pm},\lambda_t,\lambda_s,\xi})$. When the regularization parameters are clear from context, we use the shorthand $R$, $R_{\mathsf{pd}}$, and $R_{\mathsf{pm}}$.

[1] For brevity, we omit the dependence on $\lambda_t$ in the notation for the PD student when the context is clear.

### 3.2 Optimal prediction-mixed decomposition

We first decompose the optimally mixed PMSD risk in terms of three key quantities: the teacher risk $R(\lambda_t)$, the PD student risk $R_{\mathsf{pd}}(\lambda_s)$, and the teacher-student residual correlation

$$C(\lambda_t, \lambda_s) := \mathbb{E}\big[(y_0 - f_{\lambda_t}(x_0))(y_0 - f_{\mathsf{pd},\lambda_s}(x_0)) \mid \mathcal{D}_{\mathrm{lab}}, \mathcal{D}_{\mathrm{unlab}}\big].$$

Let $\xi^\star(\lambda_t, \lambda_s) \in \operatorname{argmin}_{\xi\in\mathbb{R}} R_{\mathsf{pm}}(\lambda_t, \lambda_s, \xi)$ denote the optimal mixing weight, and let $R^\star_{\mathsf{pm}}(\lambda_t, \lambda_s) := R_{\mathsf{pm}}(\lambda_t, \lambda_s, \xi^\star)$ denote the corresponding optimal PMSD risk.

**Proposition 3.1** (Optimal prediction-mixed decomposition)**.** *Suppose that $\lambda_t$, $\lambda_s$, and $D(\lambda_t, \lambda_s) := R(\lambda_t) + R_{\mathsf{pd}}(\lambda_s) - 2C(\lambda_t, \lambda_s)$ are strictly positive. Then,*

$$\xi^\star(\lambda_t, \lambda_s) = \frac{R(\lambda_t) - C(\lambda_t, \lambda_s)}{D(\lambda_t, \lambda_s)}, \quad \textit{and } R^\star_{\mathsf{pm}}(\lambda_t, \lambda_s) = R(\lambda_t) - \frac{(R(\lambda_t) - C(\lambda_t, \lambda_s))^2}{D(\lambda_t, \lambda_s)}. \tag{5}$$

The decomposition implies that $R^\star_{\mathsf{pm}} \le R$. The gain over the teacher is determined by the magnitude of $R(\lambda_t) - C(\lambda_t, \lambda_s)$, while the sign of this difference determines the sign of $\xi^\star$, which may therefore be negative. The denominator is the mean squared difference between the teacher and the PD student:

$$D(\lambda_t, \lambda_s) = R(\lambda_t) + R_{\mathsf{pd}}(\lambda_s) - 2C(\lambda_t, \lambda_s) = \mathbb{E}\big[(f_{\lambda_t}(x_0) - f_{\mathsf{pd},\lambda_s}(x_0))^2 \mid \mathcal{D}_{\mathrm{lab}}, \mathcal{D}_{\mathrm{unlab}}\big] \ge 0.$$

It is zero if and only if $f_{\lambda_t}$ and $f_{\mathsf{pd},\lambda_s}$ coincide almost surely on the test distribution, in which case $R_{\mathsf{pm}} \equiv R \equiv R_{\mathsf{pd}}$ for all $\xi$, so any $\xi$ is trivially optimal. Hence the assumption $D > 0$ is mild.

### 3.3 Proportional asymptotics under general anisotropic covariance

In this subsection, we derive deterministic equivalents for the optimal mixing weight and conditional PMSD risk in the proportional asymptotic regime, where $n_t$, $n_s$, and $p$ diverge at the same rate.

Throughout, we assume that the teacher's training data $\{(x_i, y_i)\}_{i=1}^{n_t}$ are i.i.d. samples from a distribution satisfying the standard assumptions used in the random matrix analysis of high-dimensional regression [27, 5]. We further assume that the fresh unlabeled covariates $\{\widetilde{x}_i\}_{i=1}^{n_s}$ are i.i.d. samples whose population covariance matrix is *simultaneously diagonalizable* with the population covariance matrix of the teacher's covariates $\{x_i\}_{i=1}^{n_t}$. This assumption of simultaneous diagonalizability across covariate sources has also been considered in the transfer learning literature [43, 63].

**Assumption 3.2** (Data distribution)**.** We assume the following about the data distribution.

(a) The training covariates $X \sim P_x^{\otimes n_t}$ admit the representation $X = Z\Sigma_t^{1/2}$, where $\Sigma_t \in \mathbb{R}^{p\times p}$ is deterministic and positive definite, with eigenvalues bounded away from 0 and $\infty$, and $Z \in \mathbb{R}^{n_t\times p}$ has i.i.d. entries with mean 0, variance 1, and uniformly bounded $(4^+)$-th moment.
(b) The unlabeled covariates $\widetilde{X} \sim \widetilde{P}_x^{\otimes n_s}$ admit the representation $\widetilde{X} = \widetilde{Z}\Sigma_s^{1/2}$, where $\Sigma_s \in \mathbb{R}^{p\times p}$ is deterministic and positive definite, with eigenvalues bounded away from 0 and $\infty$. Moreover, $\widetilde{Z} \in \mathbb{R}^{n_s\times p}$ has i.i.d. entries with mean 0, variance 1, and uniformly bounded $(4^+)$-th moment.
(c) $\Sigma_t$ and $\Sigma_s$ are simultaneously diagonalizable. Equivalently, they are commute.
(d) The response $y \sim P_y$ have mean 0 and uniformly bounded $(4^+)$-th moment.

**Key quantities and notation.** We define the population projection and variance parameters as $\beta := \Sigma_t^{-1}\mathbb{E}[xy]$ and $\sigma^2 := \mathrm{Var}(y - x^\top\beta)$, respectively. The signal energy is $r^2 := \|\beta\|_2^2$ and $\mathrm{SNR} := r^2/\sigma^2$. Denote the labeled and unlabeled data aspect ratios $\gamma_t := p/n_t$ and $\gamma_s := p/n_s$, respectively. We introduce several scalar quantities to characterize the limiting risk. Let $\overline{\mathrm{tr}}(A) := \mathrm{tr}(A)/p$ denote the normalized trace. For any fixed $\lambda_t > 0$ and $\lambda_s > 0$, define $\kappa_t = \kappa_t(\lambda_t) > 0$ and $\kappa_s = \kappa_s(\lambda_s) > 0$ as the unique solutions to the corresponding fixed-point equations,

$$\kappa_t = \lambda_t + \gamma_t\kappa_t\ \overline{\mathrm{tr}}\big(\Sigma_t(\Sigma_t + \kappa_t I_p)^{-1}\big) \quad \text{and} \quad \kappa_s = \lambda_s + \gamma_s\kappa_s\ \overline{\mathrm{tr}}\big(\Sigma_s(\Sigma_s + \kappa_s I_p)^{-1}\big), \tag{6}$$

respectively, and let $G_t = (\Sigma_t + \kappa_t I_p)^{-1}$ and $G_s = (\Sigma_s + \kappa_s I_p)^{-1}$ be the associated resolvents. For fixed $(\Sigma_t, \gamma_t)$ and $(\Sigma_s, \gamma_s)$, the maps $\lambda_t \mapsto \kappa_t$ and $\lambda_s \mapsto \kappa_s$ are both one-to-one on $(0, \infty)$. Finally, for $i, j \ge 0$, define the mixed signal-covariance alignment and trace functionals as

$$q_{i,j} := \beta^\top G_t^i G_s^j \Sigma_t\, \beta, \qquad \widetilde{q}_{i,j} := \beta^\top G_t^i G_s^j \Sigma_s\, \beta, \qquad t_{i,j} := \gamma_t\ \overline{\mathrm{tr}}(\Sigma_t^2 G_t^i G_s^j), \tag{7}$$

and set $b_t := (1 - t_{2,0})^{-1}$ and $b_s := \big(\gamma_s\,\overline{\mathrm{tr}}(\Sigma_s\Sigma_t G_s^2)\big)/\big(1 - \gamma_s\,\overline{\mathrm{tr}}(\Sigma_s^2 G_s^2)\big)$.

Our first result states that the three oracle ingredients in Proposition 3.1 have deterministic equivalents, in the sense defined in Section C.5, under general anisotropic covariance and deterministic signal.

**Theorem 3.3** (General anisotropic risk asymptotics)**.** *Under Assumption 3.2, for fixed $\lambda_t > 0$ and $\lambda_s > 0$, as $n_t, n_s, p \to \infty$ with $\gamma_t$ and $\gamma_s$ are bounded away from 0 and $\infty$, we have*

$$R(\lambda_t) \xrightarrow{\text{p}} \mathcal{R}(\lambda_t), \qquad R_{\mathsf{pd}}(\lambda_s) \xrightarrow{\text{p}} \mathcal{R}_{\mathsf{pd}}(\lambda_s), \qquad C(\lambda_t, \lambda_s) \xrightarrow{\text{p}} \mathcal{C}(\lambda_t, \lambda_s), \tag{8}$$

*where*

$$\begin{aligned}
\mathcal{R}(\lambda_t) &= \sigma^2 + \kappa_t^2 b_t q_{2,0} + \sigma^2 b_t t_{2,0}, \\
\mathcal{C}(\lambda_t, \lambda_s) &= \sigma^2 + \kappa_t^2 b_t q_{2,0} + \sigma^2 b_t t_{2,0} - \kappa_s(-\kappa_t q_{1,1} + \kappa_t^2 q_{2,1} + \kappa_t^2 b_t t_{2,1} q_{2,0} + \sigma^2 b_t t_{2,1}), \\
\mathcal{R}_{\mathsf{pd}}(\lambda_s) &= \kappa_s^2 \Big( q_{0,2} + b_s \widetilde{q}_{0,2} - 2\kappa_t (q_{1,2} + b_s \widetilde{q}_{1,2}) + \kappa_t^2 (q_{2,2} + b_s \widetilde{q}_{2,2}) \\
&\qquad + b_t(\kappa_t^2 q_{2,0} + \sigma^2)\big(t_{2,2} + b_s \gamma_t \, \overline{\text{tr}}(\Sigma_t \Sigma_s G_t^2 G_s^2)\big) \Big) - \mathcal{R}(\lambda_t) + 2\mathcal{C}(\lambda_t, \lambda_s).
\end{aligned}$$

*The denominator $\mathcal{R}(\lambda_t) + \mathcal{R}_{\mathsf{pd}}(\lambda_s) - 2\mathcal{C}(\lambda_t, \lambda_s)$ is strictly larger than 0, and*

$$\xi^\star \xrightarrow{\text{p}} \frac{\mathcal{R} - \mathcal{C}}{\mathcal{R} + \mathcal{R}_{\mathsf{pd}} - 2\mathcal{C}}, \qquad R^\star_{\mathsf{pm}} \xrightarrow{\text{p}} \mathcal{R}^\star_{\mathsf{pm}} := \mathcal{R} - \frac{\big(\mathcal{R} - \mathcal{C}\big)^2}{\mathcal{R} + \mathcal{R}_{\mathsf{pd}} - 2\mathcal{C}}. \tag{9}$$

As shown in the experiments in Section H.3, the theoretical predictions closely align with the empirical risks even at moderate values of $n_t$, $n_s$, and $p$.

# 4 Strict improvement and risk monotonicity

Using the risk characterization from Section 3, we now study two qualitative questions: when the PMSD student strictly improves upon the teacher and the pure-distilled student, and how its optimal risk depends on the amount of fresh unlabeled data.

## 4.1 Strict improvement over the teacher

Leveraging the deterministic equivalents from Theorem 3.3, we study when the PMSD student strictly improves upon the teacher. Equation (9) shows that the PMSD risk is asymptotically strictly less than the teacher risk whenever

$$\mathcal{R}(\lambda_t) - \mathcal{C}(\lambda_t, \lambda_s) \neq 0, \tag{10}$$

and the two risks coincide otherwise. Thus, characterizing when the PMSD student fails to improve upon the teacher amounts to understanding when $\mathcal{R}(\lambda_t) - \mathcal{C}(\lambda_t, \lambda_s)$ vanishes.

For a fixed and unknown $\lambda_t > 0$ used by the teacher, we are interested in the choices of $\lambda_s > 0$ for which the PMSD student strictly improves. Hence, we study the cardinality of the set of $\lambda_s$ for which there is no strict improvement at a given $\lambda_t$, which we call the "student-tie" set:

$$\Lambda_s(\lambda_t) := \{\lambda_s > 0 : \mathcal{R}(\lambda_t) - \mathcal{C}(\lambda_t, \lambda_s) = 0\}.$$

If this equation is not degenerated, it is equivalent to a polynomial equation with degree $\widetilde{m} - 1$ with respect to $\lambda_s$, where $\widetilde{m}$ is the number of distinct eigenvalues of $\Sigma_s$. Therefore, $|\Lambda_s(\lambda_t)| \leq \widetilde{m} - 1$.

On the other hand, if for a given $\lambda_t$, the gap $\mathcal{R}(\lambda_t) - \mathcal{C}(\lambda_t, \lambda_s)$ is identically zero as a function of $\lambda_s$, then the PMSD student cannot improve upon the teacher for any choice of $\lambda_s$. Equivalently, $\Lambda_s(\lambda_t) = \mathbb{R}^+$. We call this the "teacher-degeneracy" set:

$$\Lambda_t := \{\lambda_t > 0 : \mathcal{R}(\lambda_t) - \mathcal{C}(\lambda_t, \lambda_s) = 0 \text{ for all } \lambda_s > 0\}.$$

**Theorem 4.1** (Upper bound the size of no-improvement sets)**.** *Let $m$ and $\widetilde{m}$ denote the number of distinct eigenvalues of $\Sigma_t$ and $\Sigma_s$, respectively. We have,*

$$|\Lambda_t| \leq 4m - 1. \tag{11}$$

*Moreover, for any $\lambda_t \notin \Lambda_t$,*

$$|\Lambda_s(\lambda_t)| \leq \widetilde{m} - 1. \tag{12}$$

Thus, under the nondegeneracy condition $\lambda_t \notin \Lambda_t$, the deterministic equivalent of the optimal PMSD risk is *strictly smaller* than that of the teacher for almost every $\lambda_s$. Equivalently, the two risks can tie only at finitely many values of $\lambda_s$ for each fixed $p$.

In many settings, we may not know $\Sigma_t$ or have access to fresh covariates from the same distribution as the teacher's training data. A simple and practical alternative is to sample the fresh unlabeled covariates from an isotropic distribution, which commutes with any $\Sigma_t$. In this case, we obtain the following strict improvement guarantee.

**Proposition 4.2.** *Assume $\Sigma_s = I_p$. For any $\lambda_t > 0$ such that $\lambda_t \notin \Lambda_t$, we have $\mathcal{R}^\star_{\mathsf{pm}}(\lambda_t, \lambda_s) < \mathcal{R}(\lambda_t)$ for all $\lambda_s > 0$.*

Hence, by using isotropic unlabeled features, the PMSD student strictly improves upon the teacher for all $\lambda_s > 0$ when $\lambda_t$ satisfies the non-degenerate condition.

### 4.2 Risk monotonicity in the amount of unlabeled data

A natural question in this framework is whether more fresh unlabeled data always improves the PMSD student. In the proportional regime, this amounts to fixing the labeled aspect ratio $\gamma_t$ and studying the optimal PMSD risk as a function of $\gamma_s$, where smaller $\gamma_s$ means more unlabeled data. Surprisingly, for fixed $\lambda_t = \lambda_s = \lambda$ in the isotropic setting, the optimal PMSD risk is *not* globally monotone in $\gamma_s$. Non-monotone risk profiles in aspect ratio are well documented in high-dimensional prediction, and subsampling and aggregation can sometimes restore monotonicity [52, 51]. The phenomenon here is different: the labeled sample and teacher remain fixed, while only the amount of fresh unlabeled data varies. Even so, the optimal PMSD risk has a simple one-dimensional structure and is unimodal in $\gamma_s$. See Section G.1 for the formal non-monotonicity statement and Figure 4 for illustrations on real datasets.

This non-monotonicity disappears when the student penalty is re-tuned as the amount of unlabeled data changes. Specifically, if $\lambda_s$ is optimized separately for each $\gamma_s$, then the resulting optimal PMSD risk is non-increasing as more unlabeled data are added. Thus, although additional unlabeled data can hurt at fixed regularization, they cannot hurt under optimal re-tuning of $\lambda_s$. In this sense, our theory recommends using all available unlabeled data when the student regularization can be tuned accordingly. The following result holds for general anisotropic covariance and deterministic signal.

**Theorem 4.3** (Optimally tuned risk monotonicity)**.** *Fix $\gamma_t$, for any $0 < \gamma_{s,1} < \gamma_{s,2}$, we have*

$$\inf_{\lambda_s > 0} \mathcal{R}^\star_{\mathsf{pm}}(\gamma_{s,1}, \lambda_s) \le \inf_{\lambda_s > 0} \mathcal{R}^\star_{\mathsf{pm}}(\gamma_{s,2}, \lambda_s).$$

## 5 Data-dependent tuning

The results in Section 4 demonstrate the benefits of fresh-$X$ distillation with optimal prediction mixing $\xi^\star$. However, the deterministic equivalents in Theorem 3.3 depend on population quantities that are unknown in practice, such as the covariance spectrum and signal-covariance alignment. In this section, we study how to tune the mixing weight $\xi$.

Recall that the optimal *oracle* mixing weight from Proposition 3.1 depends on residual correlations with the unknown response. Thus, in the fully prediction-only regime, unlabeled covariates and teacher outputs alone are insufficient to identify $\xi^\star$. This is an information-theoretic limitation, not a computational one. We give a formal non-identifiability setting in Section G.2. However, $\xi^\star$ can be consistently estimated using an independent labeled calibration set $\mathcal{D}_{\mathrm{cal}} := \{(x_i^{\mathrm{cal}}, y_i^{\mathrm{cal}})\}_{i=1}^{n_{\mathrm{cal}}}$, which is used only for evaluation and not retraining. Conditional on the teacher predictor and fresh unlabeled data, both $f$ and $f_{\mathsf{pd}}$ are fixed predictors. For an independent test point $(x, y)$ drawn from the target distribution, define the residuals $e := y - f(x)$ and $e^{\mathsf{pd}} := y - f_{\mathsf{pd}}(x)$. Define the calibration residuals

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

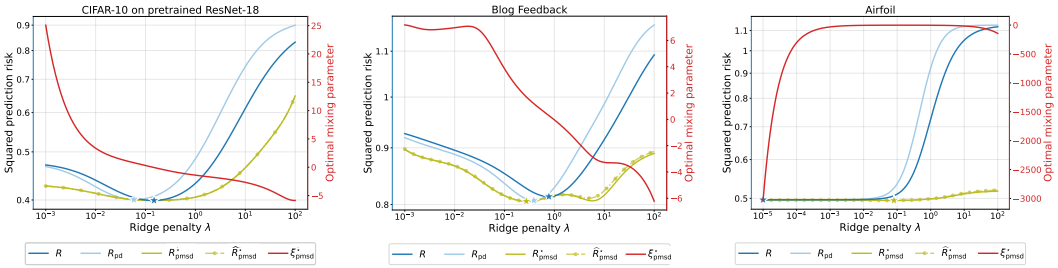
and estimate the three oracle quantities by

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

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

$$\widehat{\xi}^\star_{\mathrm{cal}} := \frac{\widehat{R}^{\mathrm{cal}} - \widehat{C}^{\mathrm{cal}}}{\widehat{R}^{\mathrm{cal}} + \widehat{R}^{\mathrm{cal}}_{\mathsf{pd}} - 2\widehat{C}^{\mathrm{cal}}}, \quad \text{and} \quad \widehat{R}^\star_{\mathsf{pm},\mathrm{cal}} := \widehat{R}^{\mathrm{cal}} - \frac{(\widehat{R}^{\mathrm{cal}} - \widehat{C}^{\mathrm{cal}})^2}{\widehat{R}^{\mathrm{cal}} + \widehat{R}^{\mathrm{cal}}_{\mathsf{pd}} - 2\widehat{C}^{\mathrm{cal}}}. \tag{13}$$

The next result establishes their consistency.




(a) $p = 512$, $n_t = 2000$, $n_s = 4000$, $n_{\text{cal}} = 200$.

(b) $p = 280$, $n_t = 2619$, $n_s = 5240$, $n_{\text{cal}} = 524$.

(c) $p = 5$, $n_t = 150$, $n_s = 150$, $n_{\text{cal}} = 76$.

Figure 3: PMSD risk estimated using a small calibration set of size $n_{\text{cal}}$ on real datasets. We use the same-$\lambda$ setting with $\lambda_t = \lambda_s = \lambda$.

**Theorem 5.1** (Calibration consistency)**.** *Fix the teacher and the fresh unlabeled sample, and suppose the oracle denominator $D := R + R_{\mathsf{pd}} - 2C$ is bounded away from zero. If the calibration sample is independent, $n_{\text{cal}} \to \infty$, and the conditional variances of $(e)^2$, $(e^{\mathsf{pd}})^2$, and $ee^{\mathsf{pd}}$ are uniformly bounded, then*

$$\widehat{\xi}^{\star}_{\text{cal}} - \xi^{\star} \xrightarrow{\text{p}} 0, \qquad \widehat{R}^{\star}_{\mathsf{pm},\text{cal}} - R^{\star}_{\mathsf{pm}} \xrightarrow{\text{p}} 0.$$

*Moreover, the same convergence holds uniformly over any fixed finite collection of candidate students under the corresponding uniform conditions.*

The assumptions required for consistency are minimal. Because the calibration set estimates only the risk and residual correlation of fixed predictors, no asymptotic condition involving $n_{\text{cal}}/p$ is required. This is particularly useful when the calibration sample grows but remains small relative to the model dimension where $n_{\text{cal}}/p \to 0$. In this regime, the calibration set contains enough information to estimate the single mixing weight consistently, while being asymptotically insubstantial as a standalone training sample for a $p$-dimensional predictor. The estimation error of the mixing weight is of order $O_p(n_{\text{cal}}^{-1/2})$. Computationally, tuning $\xi$ is inexpensive: after training the PD student once, (13) requires only evaluating the two predictors on the calibration points and computing three empirical averages. The uniform finite-candidate statement also allows the same calibration set to choose among finitely many student penalties $\lambda_s$ or unlabeled sample sizes $n_s$. Figure 3 illustrates calibration-based tuning with small labeled samples on real datasets.

## 6 Logistic regression

We turn to the fresh-$X$ prediction-only setting for logistic regression with cross-entropy loss. Our goal is to show that the PMSD student can strictly outperform both the teacher and the PD student.

**Setting.** A theoretical analysis of self-distillation for classification is more challenging than for regression. To streamline the presentation, we adopt the setting of [17, 33], which we review here. We study a binary classification problem in which each instance $x \in \mathcal{X}$ has a ground-truth label $y(x) \in \{0, 1\}$. Let $\mathcal{P}$ denote the underlying distribution on $\mathcal{X}$. We have access to an unlabeled dataset $\mathcal{D}_{\text{unlab}} = \{x_i\}_{i=1}^{2n}$ consisting of $2n$ i.i.d. draws from $\mathcal{P}$. The corresponding labels $y_i := y(x_i)$ are deterministic functions of the instances but are unobserved. We assume that the dataset is balanced and index the observations so that $y_i = 1$ for $i \in \mathcal{S}_1 := \{1, \dots, n\}$ and $y_i = 0$ for $i \in \mathcal{S}_0 := \{n + 1, \dots, 2n\}$.

**Teacher.** We have access to a teacher classifier that returns a hard label $\widehat{y}(x) \in \{0, 1\}$ for any $x \in \mathcal{X}$. Although its form is unknown, we can query its predictions. Suppose that its population classification error is the same within each class and equals an unknown $\rho \in (0, 1)$:

$$\mathbb{P}_{x\sim\mathcal{P}}\big(\widehat{y}(x) \neq y(x) \mid y(x) = 0\big) = \mathbb{P}_{x\sim\mathcal{P}}\big(\widehat{y}(x) \neq y(x) \mid y(x) = 1\big) = \rho.$$

Following [17], let $\widehat{n} := \lfloor n\rho \rfloor$ and assume that the teacher makes $\widehat{n}$ errors in each class. After reindexing the observations within each class, write the teacher's labels on $\mathcal{D}_{\text{unlab}}$ as

$$\widehat{y}_i := \widehat{y}(x_i) = \begin{cases} 1 - y_i & \text{for } i \in \underbrace{\{1, \dots, \widehat{n}\}}_{\mathcal{S}_{1,\text{bad}}} \cup \underbrace{\{n + 1, \dots, n + \widehat{n}\}}_{\mathcal{S}_{0,\text{bad}}}, \\ y_i & \text{for } i \in \underbrace{\{\widehat{n} + 1, \dots, n\}}_{\mathcal{S}_{1,\text{good}}} \cup \underbrace{\{n + \widehat{n} + 1, \dots, 2n\}}_{\mathcal{S}_{0,\text{good}}}. \end{cases}$$

Thus, $|\mathcal{S}_{1,\text{bad}}| = |\mathcal{S}_{0,\text{bad}}| = \widehat{n}$, and the fraction of incorrect labels in each class converges to $\rho$.

**Students.** We assume a feature map $\phi : \mathcal{X} \to \mathcal{H}$, where $\mathcal{H}$ is a Hilbert space, and keep $\phi$ fixed throughout. We train a linear classifier in $\mathcal{H}$ on the teacher's hard pseudo-labels $\{\widehat{y}_i\}_{i=1}^{2n}$ using $\ell_2$-regularized logistic loss:

$$\theta_{\mathsf{pd}} := \operatorname*{argmin}_{\theta} \frac{1}{2n} \sum_{i=1}^{2n} \mathrm{CE}\big(\widehat{y}_i, \sigma(\langle \theta, \phi(x_i) \rangle)\big) + \frac{\lambda}{2}\|\theta\|^2, \tag{14}$$

where $\mathrm{CE}(q, \widehat{q}) := -q \log(\widehat{q}) - (1-q)\log(1-\widehat{q})$ is the binary cross-entropy loss and $\sigma(z) := (1+e^{-z})^{-1}$ is the sigmoid function. The resulting PD student's soft prediction is $y_{\mathsf{pd}}(x) := \sigma(\langle \theta_{\mathsf{pd}}, \phi(x) \rangle) \in (0,1)$. This setting corresponds to linear probing in neural networks, where a learnable linear layer and sigmoid activation are placed on top of a frozen feature backbone. Given a mixing weight $\xi \in \mathbb{R}$, the PMSD classifier is

$$y_{\mathsf{pm},\xi}(x) = \mathbb{I}\big\{(1-\xi)\widehat{y}(x) + \xi y_{\mathsf{pd}}(x) \geq 0.5\big\}. \tag{15}$$

We make the following assumption on the feature map $\phi$ and the unlabeled dataset $\mathcal{D}_{\text{unlab}}$.

**Assumption 6.1** (Feature norm and correlation structure)**.** The features have unit norm, $\|\phi(x)\| = 1$ for all $x \in \mathcal{X}$. Let $\Phi \in \mathbb{R}^{2n \times 2n}$ denote the Gram matrix on $\mathcal{D}_{\text{unlab}}$, with $[\Phi]_{ij} := \langle \phi(x_i), \phi(x_j) \rangle$. We assume that $[\Phi]_{ij} = 0$ when $y_i \neq y_j$ and $[\Phi]_{ij} = c \in (0,1)$ when $y_i = y_j$ and $i \neq j$.

This assumption, formulated by [17], is technically convenient because it makes the data separable with respect to their unobserved ground-truth classes, thereby ensuring the existence of a perfect classifier. Moreover, [33] empirically validate this assumption on six real datasets: CIFAR-100, Caltech-101, Caltech-256, Flowers-102, Food-101, and Stanford Cars. They use features from a ResNet-34 pretrained on ImageNet; see their Figure 1 and Appendix D.2.

**Goal.** Given a soft prediction rule $y_\theta : \mathcal{X} \to [0,1]$, an instance $x$ is classified correctly when $y(x) = \mathbb{I}\{y_\theta(x) \geq 0.5\}$. By construction, the teacher's accuracy on $\mathcal{D}_{\text{unlab}}$ converges to $100(1-\rho)\%$. We show in Theorem 6.2 that, when $\rho$ is sufficiently close to $0.5$ from below, the PD student trained by (14) also achieves only $100(1-\rho)\%$ accuracy as $n \to \infty$. Our goal is to construct a PMSD classifier as in (15) that achieves asymptotic $100\%$ accuracy and thus perfectly recovers the true labels in the limit. Prior work [17, 33] characterizes conditions on $\rho$ and $\lambda$ under which the PD student $y_{\mathsf{pd}}$ itself achieves this goal. These works refer to the result as "$100\%$ population accuracy" because the generalization gap vanishes for finite-dimensional linear classifiers as $n \to \infty$ [34, 36].

**Theorem 6.2.** *Let $\bar{\lambda}_n := 2n\lambda$ and suppose that $\bar{\lambda}_n \to \bar{\lambda} \in (0, \infty)$ as $n \to \infty$. Under Assumption 6.1, there exists a threshold $\rho_0 = \rho_0(c, \bar{\lambda}) \in (0, 0.5)$ such that, for every $\rho \in (\rho_0, 0.5)$, the PD student $y_{\mathsf{pd}}$ has $100(1-\rho)\%$ population accuracy. On the other hand, for every $\rho < 0.5$, there exists $\xi \geq 0$ such that the PMSD student $y_{\mathsf{pm},\xi}$ achieves $100\%$ population accuracy.*

Self-distillation has been observed to induce label averaging among highly correlated instances [33]. When $\rho$ is small, the PD student can therefore correct mislabeled samples by aggregating labels from correctly classified neighbors, although incorrect neighbors also reduce its confidence on correctly labeled samples. Relative to the teacher, the student can thus trade confidence for accuracy. As $\rho$ approaches $0.5$ from below, however, incorrectly labeled neighbors increasingly mislead the student, and its predictions no longer cross the classification threshold on the mislabeled samples. Choosing $\xi > 1$ extrapolates beyond the PD student's soft prediction and can amplify this corrective direction, although an excessively large $\xi$ may hurt performance. We illustrate this behavior with synthetic experiments in Figures 14 and 16.

### 6.1 Soft teacher–student prediction mixing

Theorem 6.2 mixes the teacher's hard label with the PD student's soft prediction and strictly improves upon both when $\rho_0 < \rho < 0.5$. We now consider mixing soft predictions from both models. Specifically, we treat the PD student $y_{\mathsf{pd}}(\cdot) = \sigma(\langle \theta_{\mathsf{pd}}, \phi(\cdot) \rangle)$ defined by (14) as a teacher that produces soft labels. Using its predictions on $\mathcal{D}_{\text{unlab}}$, we train a second PD student $y^{(2)}_{\mathsf{pd}}(\cdot) = \sigma(\langle \theta^{(2)}_{\mathsf{pd}}, \phi(\cdot) \rangle)$ with the same regularization parameter $\lambda$:

$$\theta^{(2)}_{\mathsf{pd}} := \operatorname*{argmin}_{\theta} \frac{1}{2n} \sum_{i=1}^{2n} \mathrm{CE}\big(y_{\mathsf{pd}}(x_i), \sigma(\langle \theta, \phi(x_i) \rangle)\big) + \frac{\lambda}{2}\|\theta\|^2. \tag{16}$$

Table 1: Test classification accuracy on CIFAR-100 with hierarchical corruption $\rho = 0.2$. The mixing weight is tuned over the grid $[-20, 20]$ for all linear probing experiments.

| Regularization level $\lambda$ | $10^{-4}$ | $10^{-3.5}$ | $10^{-3}$ | $10^{-2.5}$ | $10^{-2}$ | $10^{-1.5}$ | $10^{-1}$ | $10^{-0.5}$ | $10^{0}$ |
|---|---|---|---|---|---|---|---|---|---|
| Teacher (%) | 52.8 | 52.9 | 52.9 | 53.2 | 54.0 | 56.1 | 57.7 | 50.4 | 35.9 |
| PD student (%) | 53.1 | 53.8 | 55.0 | 56.1 | 57.3 | 56.8 | 47.3 | 20.8 | 2.3 |
| Optimal PMSD student (%) | 54.5 | 54.9 | 55.6 | 56.6 | 57.4 | 57.3 | 59.0 | 54.1 | 43.1 |
| Estimated optimal mixing weight $\widehat{\xi}^\star$ | 10.5 | 4.1 | 2.1 | 1.3 | 0.9 | 0.7 | $-17.8$ | $-15.0$ | $-20.0$ |

For a mixing weight $\xi \in \mathbb{R}$, the resulting PMSD classifier thresholds the affine combination of the two soft predictions:

$$y^{(2)}_{\mathsf{pm},\xi}(x) = \mathbb{I}\big\{(1-\xi)y_{\mathsf{pd}}(x) + \xi y^{(2)}_{\mathsf{pd}}(x) \geq 0.5\big\}. \tag{17}$$

Previous results (Theorem 5 of [17] and Theorem 4.1 of [33]) show that, as $n \to \infty$, the second student $y^{(2)}_{\mathsf{pd}}$ can tolerate a higher label noise rate $\rho$ than its teacher $y_{\mathsf{pd}}$, subject to $\rho < 0.5$. Specifically, there exists a threshold $\rho_1$ with $0 < \rho_0 < \rho_1 < 0.5$ such that (i) for $\rho_0 < \rho < \rho_1$, the student $y^{(2)}_{\mathsf{pd}}$ attains $100\%$ accuracy while the teacher $y_{\mathsf{pd}}$ has $100(1-\rho)\%$ accuracy, and (ii) for $\rho_1 < \rho < 0.5$, both achieve $100(1-\rho)\%$ accuracy. In contrast, we show that, for every $\rho < 0.5$, an appropriate mixing parameter $\xi \geq 0$ allows the PMSD classifier $y^{(2)}_{\mathsf{pm},\xi}$ in (17) to attain $100\%$ accuracy.

**Theorem 6.3.** *Let $\bar{\lambda}_n := 2n\lambda$ and suppose that $\bar{\lambda}_n \to \bar{\lambda} \in (0, \infty)$ as $n \to \infty$. Under Assumption 6.1, for every $\rho < 0.5$, there exists $\xi \geq 0$ such that the PMSD student $y^{(2)}_{\mathsf{pm},\xi}$ achieves $100\%$ accuracy.*

We also demonstrate the improvement of the PMSD performance when $\rho > 0.5$, see Section G.3.

### 6.2 Linear probing experiments on real datasets

To illustrate the performance of PMSD, we consider multiclass classification with $\ell_2$-regularized cross-entropy loss on Caltech-101, Caltech-256 [26], and CIFAR-100 [38]. For each dataset, we train a linear classifier with a softmax output on frozen features from a ResNet-34 pretrained on ImageNet. We use this setup for both the teacher and the pure-distilled student: the teacher is trained on corrupted one-hot labels, whereas the student is trained on the teacher's soft pseudo-labels using the same $\lambda$. Table 1 reports the test accuracies of the teacher, PD student, and PMSD student. We estimate the mixing weight $\xi$ on a small calibration set using a fine grid over $[-20, 20]$. Additional results across datasets and corruption levels appear in Section H.2 and experimental details are provided in Section I.

## 7 Concluding Remarks

In this work, we study how to improve an available predictor through distillation without access to its original training data or parameters. Using teacher queries on fresh unlabeled data, which can be out-of-distribution, we propose to first train a pure-distilled student on the teacher's pseudo-labels and then combine their predictions using an optimal affine combination. For ridge regression, we derive exact characterizations of the optimal mixing weight, the optimally mixed risk and show that strict improvement upon the teacher holds quite generically. Such pointwise gains are important because using only the pure-distilled student does not guarantee them, and tuning its regularization level requires retraining and is generally not easy [64]. We also show that increasing the amount of unlabeled data can worsen the prediction-mixing risk when regularization is held fixed, whereas optimal re-tuning guarantees non-increasing risk. On the implementation side, a small labeled calibration set suffices to tune the optimal mixing weight at low computational cost and without further retraining. We further demonstrate the benefits of prediction mixing for logistic regression.

There are some further questions for this framework that we discuss in the appendix, including the comparison between the PMSD student versus the student distilled using the same training data as the teacher (see Section H.5), and the optimal distribution for fresh unlabeled covariates (see Section G.4). As a limitation, our theoretical analysis assumes the teacher and student feature covariance are commute and extension beyond this assumption is a direction for future work.

## Acknowledgements

We thank Sujay Sanghavi, Adel Javanmard, Peter Müeller, Antonio Linero, Edward George, and participants in the Best of Statistical Science ("BOSS") Workshop for helpful discussions.

**Funding disclosure.** The authors received no third-party funding or support for this work.

**Competing Interests.** The authors declare no competing financial interests.

## Bibliography

[1] Sungsoo Ahn, Shell Xu Hu, Andreas Damianou, Neil D. Lawrence, and Zhenwen Dai. Variational information distillation for knowledge transfer. In *Conference on Computer Vision and Pattern Recognition*, 2019.

[2] Sina Alemohammad, Josue Casco-Rodriguez, Lorenzo Luzi, Ahmed Imtiaz Humayun, Hossein Babaei, Daniel LeJeune, Ali Siahkoohi, and Richard Baraniuk. Self-consuming generative models go MAD. In *International Conference on Learning Representations*, 2023.

[3] Jimmy Ba and Rich Caruana. Do deep nets really need to be deep? In *Advances in Neural Information Processing Systems*, 2014.

[4] Soham Bakshi and Sunrit Chakraborty. From collapse to improvement: Statistical perspectives on the evolutionary dynamics of iterative training on contaminated sources. arXiv: 2602.10531, 2026.

[5] Peter L. Bartlett, Andrea Montanari, and Alexander Rakhlin. Deep learning: A statistical viewpoint. *Acta Numerica*, 30:87–201, 2021.

[6] Pierre C. Bellec, Jin-Hong Du, Takuya Koriyama, Pratik Patil, and Kai Tan. Corrected generalized cross-validation for finite ensembles of penalized estimators. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 87(2):289–318, 2025.

[7] Thomas F. Brooks, D. Stuart Pope, and Michael A. Marcolini. Airfoil Self-Noise. UCI Machine Learning Repository, 1989.

[8] Cristian Buciluǎ, Rich Caruana, and Alexandru Niculescu-Mizil. Model compression. In *International Conference on Knowledge Discovery and Data Mining*, 2006.

[9] Collin Burns, Pavel Izmailov, Jan Hendrik Kirchner, Bowen Baker, Leo Gao, Leopold Aschenbrenner, Yining Chen, Adrien Ecoffet, Manas Joglekar, Jan Leike, Ilya Sutskever, and Jeff Wu. Weak-to-strong generalization: Eliciting strong capabilities with weak supervision. arXiv: 2312.09390, 2023.

[10] Krisztian Buza. BlogFeedback. UCI Machine Learning Repository, 2014.

[11] Richard H. Byrd, Peihuang Lu, Jorge Nocedal, and Ciyou Zhu. A limited-memory algorithm for bound-constrained optimization. *SIAM Journal on Scientific Computing*, 16(5):1190–1208, 1995.

[12] Yair Carmon, Aditi Raghunathan, Ludwig Schmidt, John C. Duchi, and Percy S. Liang. Unlabeled data improves adversarial robustness. In *Advances in Neural Information Processing Systems*, 2019.

[13] Defang Chen, Jian-Ping Mei, Hailin Zhang, Can Wang, Yan Feng, and Chun Chen. Knowledge distillation with the reused teacher classifier. In *Conference on Computer Vision and Pattern Recognition*, 2022.

[14] Guobin Chen, Wongun Choi, Xiang Yu, Tony Han, and Manmohan Chandraker. Learning efficient object detection models with knowledge distillation. In *Advances in Neural Information Processing Systems*, 2017.

[15] Hien Dang, Pratik Patil, and Alessandro Rinaldo. Optimal unconstrained self-distillation in ridge regression: Strict improvements, precise asymptotics, and one-shot tuning. In *International Conference on Machine Learning*, 2026.

[16] Rudrajit Das, Inderjit S. Dhillon, Alessandro Epasto, Adel Javanmard, Jieming Mao, Vahab Mirrokni, Sujay Sanghavi, and Peilin Zhong. Retraining with predicted hard labels provably increases model accuracy. In *International Conference on Machine Learning*, 2025.

[17] Rudrajit Das and Sujay Sanghavi. Understanding self-distillation in the presence of label noise. In *International Conference on Machine Learning*, 2023.

[18] Elvis Dohmatob, Yunzhen Feng, and Julia Kempe. Model collapse demystified: The case of regression. In *Advances in Neural Information Processing Systems*, 2024.

[19] Elvis Dohmatob, Yunzhen Feng, Arjun Subramonian, and Julia Kempe. Strong model collapse. In *International Conference on Learning Representations*, 2025.

[20] Yijun Dong, Yicheng Li, Yunai Li, Jason D Lee, and Qi Lei. Discrepancies are virtue: Weak-to-strong generalization through lens of intrinsic dimension. *arXiv preprint arXiv:2502.05075*, 2025.

[21] Spencer Frei, Difan Zou, Zixiang Chen, and Quanquan Gu. Self-training converts weak learners to strong learners in mixture models. In *International Conference on Artificial Intelligence and Statistics*, pages 8003–8021. PMLR, 2022.

[22] Tommaso Furlanello, Zachary Lipton, Michael Tschannen, Laurent Itti, and Anima Anandkumar. Born again neural networks. In *International Conference on Machine Learning*, 2018.

[23] Anvit Garg, Sohom Bhattacharya, and Pragya Sur. Preventing model collapse under over-parametrization: Optimal mixing ratios for interpolation learning and ridge regression. In *International Conference on Learning Representations*, 2026.

[24] Matthias Gerstgrasser, Rylan Schaeffer, Apratim Dey, Rafael Rafailov, Tomasz Korbak, Henry Sleight, Rajashree Agrawal, John Hughes, Dhruv Bhandarkar Pai, Andrey Gromov, Daniel A. Roberts, Diyi Yang, David L. Donoho, and Sanmi Koyejo. Is model collapse inevitable? Breaking the curse of recursion by accumulating real and synthetic data. In *Conference on Language Modeling*, 2024.

[25] Jianping Gou, Baosheng Yu, Stephen J. Maybank, and Dacheng Tao. Knowledge distillation: A survey. *International Journal of Computer Vision*, 129(6):1789–1819, 2021.

[26] Gregory Griffin, Alex Holub, and Pietro Perona. Caltech-256 object category dataset. CaltechDATA, 2007.

[27] Trevor Hastie, Andrea Montanari, Saharon Rosset, and Ryan J. Tibshirani. Surprises in high-dimensional ridgeless least squares interpolation. *Annals of Statistics*, 50(2):949–986, 2022.

[28] Hengzhi He, Shirong Xu, and Guang Cheng. Recursive learning without collapse: A weighting-based stabilization framework. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, page qkag099, 2026.

[29] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Conference on Computer Vision and Pattern Recognition*, 2016.

[30] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. arXiv: 1503.02531, 2015.

[31] M. Emrullah Ildiz, Halil Alperen Gozeten, Ege Onur Taga, Marco Mondelli, and Samet Oymak. High-dimensional analysis of knowledge distillation: Weak-to-strong generalization and scaling laws. In *International Conference on Learning Representations*, 2025.

[32] Adel Javanmard, Rudrajit Das, Alessandro Epasto, and Vahab Mirrokni. Self-boost via optimal retraining: An analysis via approximate message passing. In *Advances in Neural Information Processing Systems*, 2025.

[33] Hyeonsu Jeong and Hye Won Chung. Rethinking self-distillation: Label averaging and enhanced soft label refinement with partial labels. In *International Conference on Learning Representations*, 2025.

[34] Sham M. Kakade, Karthik Sridharan, and Ambuj Tewari. On the complexity of linear prediction: Risk bounds, margin bounds, and regularization. In *Advances in Neural Information Processing Systems*, 2008.

[35] Antti Knowles and Jun Yin. Anisotropic local laws for random matrices. *Probability Theory and Related Fields*, 169(1–2):257–352, 2017.

[36] Vladimir Koltchinskii and Dmitry Panchenko. Empirical margin distributions and bounding the generalization error of combined classifiers. *Annals of Statistics*, 30(1):1–50, 2002.

[37] Takuya Koriyama, Pratik Patil, Jin-Hong Du, Kai Tan, and Pierre C. Bellec. Precise asymptotics of bagging regularized M-estimators. *Annals of Statistics*, 54(2):796–833, 2026.

[38] Alex Krizhevsky. Learning multiple layers of features from tiny images. Technical report, University of Toronto, 2009.

[39] Radu Lecoiu, Debarghya Mukherjee, and Pragya Sur. Self-distillation is optimal among spectral shrinkage estimators in spiked covariance models. arXiv: 2605.17778, 2026.

[40] Jian Liang, Dapeng Hu, Jiashi Feng, and Ran He. DINE: Domain adaptation from single and multiple black-box predictors. In *Conference on Computer Vision and Pattern Recognition*, 2022.

[41] David Lopez-Paz, Léon Bottou, Bernhard Schölkopf, and Vladimir Vapnik. Unifying distillation and privileged information. arXiv: 1511.03643, 2015.

[42] Michal Lukasik, Srinadh Bhojanapalli, Aditya Menon, and Sanjiv Kumar. Does label smoothing mitigate label noise? In *International conference on machine learning*, pages 6448–6458. PMLR, 2020.

[43] Neil Rohit Mallinar, Austin Zane, Spencer Frei, and Bin Yu. Minimum-norm interpolation under covariate shift. In *International Conference on Machine Learning*, 2024.

[44] Hossein Mobahi, Mehrdad Farajtabar, and Peter L. Bartlett. Self-distillation amplifies regularization in Hilbert space. In *Advances in Neural Information Processing Systems*, 2020.

[45] Behrad Moniri and Hamed Hassani. On the mechanisms of weak-to-strong generalization: A theoretical perspective. In *Advances in Neural Information Processing Systems*, 2025.

[46] Nagarajan Natarajan, Inderjit S Dhillon, Pradeep K Ravikumar, and Ambuj Tewari. Learning with noisy labels. *Advances in neural information processing systems*, 26, 2013.

[47] Gaurav Kumar Nayak, Konda Reddy Mopuri, Vaisakh Shaj, Venkatesh Babu Radhakrishnan, and Anirban Chakraborty. Zero-shot knowledge distillation in deep networks. In *International Conference on Machine Learning*, 2019.

[48] Samet Oymak and Talha Cihad Gulcu. A theoretical characterization of semi-supervised learning with self-training for gaussian mixture models. In *International Conference on Artificial Intelligence and Statistics*, pages 3601–3609. PMLR, 2021.

[49] Divyansh Pareek, Simon S. Du, and Sewoong Oh. Understanding the gains from repeated self-distillation. In *Advances in Neural Information Processing Systems*, 2024.

[50] Pratik Patil and Jin-Hong Du. Generalized equivalences between subsampling and ridge regularization. In *Advances in Neural Information Processing Systems*, 2023.

[51] Pratik Patil, Jin-Hong Du, and Arun Kumar Kuchibhotla. Bagging in overparameterized learning: Risk characterization and risk monotonization. *Journal of Machine Learning Research*, 24(319):1–113, 2023.

[52] Pratik Patil, Arun Kumar Kuchibhotla, Yuting Wei, and Alessandro Rinaldo. Mitigating multiple descents: A model-agnostic framework for risk monotonization. arXiv: 2205.12937, 2022.

[53] Pratik Patil, Alessandro Rinaldo, and Ryan J. Tibshirani. Estimating functionals of the out-of-sample error distribution in high-dimensional ridge regression. In *International Conference on Artificial Intelligence and Statistics*, 2022.

[54] Pratik Patil, Yuting Wei, Alessandro Rinaldo, and Ryan J. Tibshirani. Uniform consistency of cross-validation estimators for high-dimensional ridge regression. In *International Conference on Artificial Intelligence and Statistics*, 2021.

[55] Hieu Pham, Zihang Dai, Qizhe Xie, and Quoc V. Le. Meta pseudo labels. In *Conference on Computer Vision and Pattern Recognition*, 2021.

[56] Kamiar Rahnama Rad and Arian Maleki. A scalable estimate of the out-of-sample prediction error via approximate leave-one-out cross-validation. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 82(4):965–996, 2020.

[57] Michael Redmond. Communities and Crime. UCI Machine Learning Repository, 2002.

[58] Parham Rezaei, Filip Kovačević, Francesco Locatello, and Marco Mondelli. High-dimensional analysis of synthetic data selection. In *International Conference on Learning Representations*, 2025.

[59] Francisco Rubio and Xavier Mestre. Spectral convergence for a general class of random matrices. *Statistics and Probability Letters*, 81(5):592–602, 2011.

[60] Luca Saglietti and Lenka Zdeborová. Solvable model for inheriting the regularization through knowledge distillation. In *Mathematical and Scientific Machine Learning Conference*, 2022.

[61] Rylan Schaeffer, Joshua Kazdan, Alvan Caleb Arulandu, and Sanmi Koyejo. Position: Model collapse does not mean what you think. arXiv: 2503.03150, 2025.

[62] Ilia Shumailov, Zakhar Shumaylov, Yiren Zhao, Nicolas Papernot, Ross Anderson, and Yarin Gal. AI models collapse when trained on recursively generated data. *Nature*, 631(8022):755–759, 2024.

[63] Yanke Song, Sohom Bhattacharya, and Pragya Sur. Generalization error of min-norm interpolators in transfer learning. arXiv: 2406.13944, 2024.

[64] Will Stephenson, Zachary Frangella, Madeleine Udell, and Tamara Broderick. Can we globally optimize cross-validation loss? Quasiconvexity in ridge regression. In *Advances in Neural Information Processing Systems*, 2021.

[65] Takashi Takahashi. The role of pseudo-labels in self-training linear classifiers on high-dimensional gaussian mixture data. *arXiv preprint arXiv:2205.07739*, 2022.

[66] Kaito Takanami, Takashi Takahashi, and Ayaka Sakata. The effect of optimal self-distillation in noisy Gaussian mixture model. In *Advances in Neural Information Processing Systems*, 2025.

[67] Alexander Wei, Wei Hu, and Jacob Steinhardt. More than a toy: Random matrix models predict how real-world neural representations generalize. In *International Conference on Machine Learning*, 2022.

[68] Diyuan Wu, Lehan Chen, Theodor Misiakiewicz, and Marco Mondelli. Improved scaling laws via weak-to-strong generalization in random feature ridge regression. In *International Conference on Machine Learning*, 2026.

[69] Qizhe Xie, Minh-Thang Luong, Eduard Hovy, and Quoc V. Le. Self-training with Noisy Student improves ImageNet classification. In *Conference on Computer Vision and Pattern Recognition*, 2020.

[70] Linfeng Zhang, Chenglong Bao, and Kaisheng Ma. Self-distillation: Towards efficient and compact neural networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(8):4388–4403, 2022.

# Supplement

This supplement accompanies the paper "Prediction-Only Distillation in Linear and Logistic Regression". It contains additional details, illustrations, and proofs of all the theoretical results. The supplement is organized as follows.


**A Notation** **17**

**B Further related work** **17**

**C Proofs in Section 3** **18**

C.1 Preliminaries . . . . . 18

C.2 Proof of Proposition 3.1 . . . . . 18

C.3 Deterministic equivalents for the proof of Theorem 3.3 . . . . . 19

C.4 Proof of Theorem 3.3 . . . . . 21

C.5 Background on deterministic equivalents . . . . . 24

**D Proofs in Section 4** **24**

D.1 Proof of Theorem 4.1 . . . . . 24

D.2 Proof of Proposition 4.2 . . . . . 27

D.3 Proof of Theorem 4.3 . . . . . 27

**E Proofs in Section 5** **28**

E.1 Proof of Theorem 5.1 . . . . . 28

E.2 Vector-valued predictors . . . . . 30

**F Proofs in Section 6** **30**

F.1 Preliminaries . . . . . 30

F.2 Proof of Theorem 6.2 . . . . . 32

F.3 Proof of Theorem 6.3 . . . . . 34

**G Additional theoretical results** **35**

G.1 Non-monotonicity in the fresh unlabeled sample size . . . . . 35

G.2 Non-identifiability of tuning without labeled data . . . . . 38

G.3 Logistic regression under heavy label noise . . . . . 39

G.4 Optimal fresh-X covariance structure . . . . . 41

**H Additional experiments** **43**

H.1 Real-world regression tasks with different student regularization levels . . . . . 43

H.2 Linear probing experiments on real datasets . . . . . 47

H.3 Synthetic experiments for ridge regression . . . . . 50

H.4 Synthetic experiments for logistic regression . . . . . 51

H.5 Fresh-X prediction-mixed versus same-X loss-mixed SD . . . . . 53

**I Experiment details** **54**

I.1 Ridge regression experiments . . . . . 54

I.2 Linear probing experiments . . . . . 55

I.3 Synthetic experiments . . . . . 56

# A Notation

**General notation.** We denote vectors in non-bold lowercase and matrices in non-bold upper case. We use blackboard letters to denote some special sets: $\mathbb{N}$ denotes the set of positive integers, $\mathbb{R}$ denotes the set of real numbers, and $\overline{\mathbb{R}}$ denotes the extended real line.

For a positive integer $n$, we use the shorthand $[n]$ to denote the set $\{1, \dots, n\}$. For a pair of real numbers $x$ and $y$, we use $x \wedge y$ to denote $\min\{x, y\}$, and $x \vee y$ to denote $\max\{x, y\}$. For an event or set $A$, $\mathbb{1}_A$ denotes the indicator random variable associated with $A$.

**Linear algebra notation.** For a vector $x$, $\|x\|_2$ denotes its $\ell_2$ norm (If no subscript is specified for the norm $\|x\|$ of a vector $x$, then it is assumed to be the $\ell_2$ norm of $x$.). For a matrix $X \in \mathbb{R}^{n \times p}$, $X^\top \in \mathbb{R}^{p \times n}$ denotes its transpose, and $X^\dagger \in \mathbb{R}^{p \times n}$ denotes its Moore-Penrose inverse. For a square matrix $A \in \mathbb{R}^{p \times p}$, $\mathrm{tr}(A)$ denotes its trace, $\overline{\mathrm{tr}}(A)$ denotes average trace ($\mathrm{tr}(A)/p$) and $A^{-1} \in \mathbb{R}^{p \times p}$ denotes its inverse, provided it is invertible. For a positive semidefinite matrix $\Sigma$, $\Sigma^{1/2}$ denotes its principal square root. A $p \times p$ identity matrix is denoted $I_p$, or simply by $I$, when it is clear from the context. For a matrix $X$, we denote its operator norm with respect to $\ell_2$ vector norm by $\|X\|_{\mathrm{op}}$ and its Frobenius norm by $\|X\|_F$. For a matrix $M$, $\|M\|_{\mathrm{tr}}$ denotes the trace norm of $M$, which is the sum of all its singular values.

**Asymptotic notation.** For a nonnegative quantity $Y$, we use $X = O_\alpha(Y)$ to denote the deterministic big-O notation that indicates the bound $|X| \le C_\alpha Y$, where $C_\alpha$ is some numerical constant that can depend on the ambient parameter $\alpha$ but otherwise does not depend on other parameters in the context. We denote the probabilistic big-O notation by $O_p$. We use $\Theta(1)$ to define constant that does not depend on the converging parameter. We denote convergence in probability by "$\xrightarrow{\mathrm{p}}$" and almost sure convergence by "$\xrightarrow{\mathrm{a.s.}}$". We sometimes write $X_n \to Y_n$ for two sequences of random variables $\{X_n\}$ and $\{Y_n\}$ to mean $|X_n - Y_n| \to 0$, where the stochastic mode of convergence is almost surely or in probability. We use generic letters $C$, $C'$ to denote any positive absolute constants.

# B Further related work

**Retraining on synthetic labels and model collapse.** Beyond a single round of pure-distillation, recent work studies how recursively training on model-generated data can degrade performance, a phenomenon known as *model collapse* [62, 2, 18, 24, 61, 19]. [28] and [23] study infinitely repeated schemes for ridge and ridgeless regression that mix ground-truth and synthetic labels with a fixed weight $w \in [0, 1]$ to prevent such degradation. [4] studies an extension of this iterative training framework to the next-token prediction setting, using a mixture of real and synthetic data at the population level. [19] use the term "strong model collapse" when a model trained on synthetic data has worse population risk than one trained only on ground-truth data. Our setting differs in two respects: we study a fixed, finite distillation procedure using fresh unlabeled covariates, and the final predictor mixes predictions rather than recursively replacing ground-truth data with synthetic data. Accordingly, our results concern the benefits of post hoc prediction mixing, rather than the long-run stability of repeated synthetic training; see Sections 4 and 6.

**Risk estimation and tuning.** Our calibration analysis in Section 5 connects to the literature on out-of-sample risk estimation for high-dimensional ridge regression and related models, including analyses based on leave-one-out and generalized cross-validation [56, 54, 53, 67]. Risk estimation has also been studied for finite ensembles of penalized estimators and subagged regularized $M$-estimators [6, 37]. Unlike those settings, PMSD combines two predictors linked through pseudo-labeling, and our goal is to estimate their residual correlation and optimal mixing weight using a small independent labeled calibration sample, without retraining either model.

**Robust classification under label noise.** Our analysis of binary classification with logistic loss under noisy supervision (Section 6) is also related to the robust classification literature. Robust classification methods generally allow broader corruption models than the structured label noise considered here in this work, but typically provide excess-risk or estimation-error guarantees that depend on the corruption level. For example, [46] propose a loss-correction approach based on class-conditional noise rates, requiring $\rho_+ + \rho_- < 1$, which reduces to $\rho < 1/2$ under symmetric label noise. [42] connect label smoothing to loss correction and empirically demonstrate its denoising effect, including improvements obtained by smoothing the predictions of a teacher trained on noisy labels. More closely related to our fresh-$X$ setting, [12] study a two-cluster Gaussian mixture model in which a pure-distilled classifier is trained on fresh unlabeled data using the teacher's hard pseudo-labels, with the goal of improving adversarial robustness. In contrast to these approaches, we leave the cross-entropy loss unchanged and perform distillation through prediction mixing, specializing to the setting considered in prior work [17, 33] on the benefits of self-distillation without modifying the training loss.

# C Proofs in Section 3

## C.1 Preliminaries

Define the empirical covariance matrices for the teacher's training covariates and the student's fresh covariates, together with their resolvents, by

$$\widehat{\Sigma} := \frac{1}{n_t} X^\top X, \qquad \widetilde{\Sigma} := \frac{1}{n_s} \widetilde{X}^\top \widetilde{X}, \qquad Q_{\lambda_t} := (\widehat{\Sigma} + \lambda_t I_p)^{-1}, \qquad \widetilde{Q}_{\lambda_s} := (\widetilde{\Sigma} + \lambda_s I_p)^{-1}.$$

The teacher coefficient is $\beta_{\lambda_t} := Q_{\lambda_t} X^\top y / n_t$. The *fresh-$X$ pure-distilled* (PD) coefficient is

$$\beta_{\mathsf{pd},\lambda_s} := \widetilde{Q}_{\lambda_s} \widetilde{\Sigma} \beta_{\lambda_t} = \big(I_p - \lambda_s \widetilde{Q}_{\lambda_s}\big)\beta_{\lambda_t}.$$

The *prediction-mixed* student family is

$$f_{\mathsf{pm},\lambda_t,\lambda_s,\xi}(x) = (1-\xi) f_{\lambda_t}(x) + \xi f_{\mathsf{pd},\lambda_s}(x), \qquad f_{\lambda_t}(x) = x^\top \beta_{\lambda_t}, \;\; f_{\mathsf{pd},\lambda_s}(x) = x^\top \beta_{\mathsf{pd},\lambda_s}.$$

Recall the in-distribution risks and correlation

$$R(\lambda_t) := \mathbb{E}[(y_0 - f_{\lambda_t}(x_0))^2 \mid \mathcal{D}_{\mathrm{lab}}], \qquad R_{\mathsf{pd}}(\lambda_s) := \mathbb{E}[(y_0 - f_{\mathsf{pd},\lambda_s}(x_0))^2 \mid \mathcal{D}_{\mathrm{lab}}, \mathcal{D}_{\mathrm{unlab}}],$$
$$C(\lambda_t, \lambda_s) := \mathbb{E}[(y_0 - f_{\lambda_t}(x_0))(y_0 - f_{\mathsf{pd},\lambda_s}(x_0)) \mid \mathcal{D}_{\mathrm{lab}}, \mathcal{D}_{\mathrm{unlab}}],$$
$$R_{\mathsf{pm}}(\lambda_t, \lambda_s, \xi) = \mathbb{E}[(y_0 - f_{\mathsf{pm},\lambda_t,\lambda_s,\xi}(x_0))^2 \mid \mathcal{D}_{\mathrm{lab}}, \mathcal{D}_{\mathrm{unlab}}],$$

Denote the deterministic equivalents of the first three quantities by $\mathcal{R}(\lambda_t)$, $\mathcal{R}_{\mathsf{pd}}(\lambda_s)$, and $\mathcal{C}(\lambda_t, \lambda_s)$, respectively. Also define $D(\lambda_t, \lambda_s) := R(\lambda_t) + R_{\mathsf{pd}}(\lambda_s) - 2C(\lambda_t, \lambda_s) \geq 0$.

Recall the random-design model in Assumption 3.2. The population $L_2$-projection coefficient, residual, and residual variance are

$$\beta := \mathbb{E}[xx^\top]^{-1}\mathbb{E}[xy] = \Sigma_t^{-1}\mathbb{E}[xy], \qquad \varepsilon := y - x^\top \beta, \qquad \sigma^2 := \mathbb{E}[\varepsilon^2].$$

Then $\mathbb{E}[x\varepsilon] = 0$ (equivalently, $\mathbb{E}[z\,\varepsilon] = 0$) and $\mathbb{E}[\varepsilon] = 0$. Throughout, we consider in-distribution prediction: the test pair $(x_0, y_0)$ is an independent copy of $(x, y)$, so $y_0 = x_0^\top \beta + \varepsilon_0$, with $\mathbb{E}[x_0\varepsilon_0] = 0$ and $\mathbb{E}[\varepsilon_0^2] = \sigma^2$.

## C.2 Proof of Proposition 3.1

*Proof of Proposition 3.1.* Let $e := y_0 - f_{\lambda_t}(x_0)$ and $e_{\mathsf{pd}} := y_0 - f_{\mathsf{pd},\lambda_s}(x_0)$. Since $y_0 - f_{\mathsf{pm},\lambda_t,\lambda_s,\xi}(x_0) = (1-\xi)e + \xi e_{\mathsf{pd}}$, we have

$$\begin{aligned} R_{\mathsf{pm}}(\lambda_t, \lambda_s, \xi) &= (1-\xi)^2 R(\lambda_t) + \xi^2 R_{\mathsf{pd}}(\lambda_s) + 2\xi(1-\xi) C(\lambda_t, \lambda_s) \\ &= R(\lambda_t) - 2\xi(R(\lambda_t) - C(\lambda_t, \lambda_s)) + \xi^2 D(\lambda_t, \lambda_s). \end{aligned} \tag{18}$$

If $D(\lambda_t, \lambda_s) > 0$, this is a strictly convex quadratic in $\xi$ with unique minimizer $\xi^\star(\lambda_t, \lambda_s) = (R(\lambda_t) - C(\lambda_t, \lambda_s))/D(\lambda_t, \lambda_s)$. Substitution gives

$$R^\star_{\mathsf{pm}}(\lambda_t, \lambda_s) = R(\lambda_t) - \frac{(R(\lambda_t) - C(\lambda_t, \lambda_s))^2}{D(\lambda_t, \lambda_s)}. \tag{19}$$

Equivalently, expanding the same quadratic around $\xi = 1$ gives

$$R^{\star}_{\mathsf{pm}}(\lambda_t, \lambda_s) = R_{\mathsf{pd}}(\lambda_s) - \frac{(R_{\mathsf{pd}}(\lambda_s) - C(\lambda_t, \lambda_s))^2}{D(\lambda_t, \lambda_s)},$$

which proves both identities in Proposition 3.1. □

### C.3 Deterministic equivalents for the proof of Theorem 3.3

Fix $\lambda_t > 0$. Let $\kappa_t = \kappa_t(\lambda_t) > 0$ denote the teacher fixed-point solution

$$\kappa_t = \lambda_t + \gamma_t \kappa_t \ \overline{\mathrm{tr}}\big(\Sigma_t(\Sigma_t + \kappa_t I_p)^{-1}\big), \tag{20}$$

and define

$$G_t := (\Sigma_t + \kappa_t I_p)^{-1}.$$

Likewise, let $\kappa_s = \kappa_s(\lambda_s) > 0$ denote the PD student fixed-point solution

$$\kappa_s = \lambda_s + \gamma_s \kappa_s \ \overline{\mathrm{tr}}\big(\Sigma_s(\Sigma_s + \kappa_s I_p)^{-1}\big), \tag{21}$$

and define

$$G_s := (\Sigma_s + \kappa_s I_p)^{-1}.$$

For $i, j \geq 0$, define the mixed signal-covariance alignment and trace functionals by

$$q_{i,j} := \beta^\top G_t^i G_s^j \Sigma_t \, \beta, \qquad \widetilde{q}_{i,j} := \beta^\top G_t^i G_s^j \Sigma_s \, \beta, \qquad t_{i,j} := \gamma_t \ \overline{\mathrm{tr}}(\Sigma_t^2 G_t^i G_s^j), \tag{22}$$

and set

$$b_t := (1 - t_{2,0})^{-1}, \quad b_s := \frac{\gamma_s \, \overline{\mathrm{tr}}(\Sigma_s \Sigma_t G_s^2)}{1 - \gamma_s \, \overline{\mathrm{tr}}(\Sigma_s^2 G_s^2)}.$$

**Lemma C.1** (Deterministic equivalents with a deterministic insertion)**.** *Let $A \in \mathbb{R}^{p \times p}$ be deterministic, symmetric, uniformly bounded in operator norm, and commuting with $\Sigma_t$. Define*

$$\tau_A := \gamma_t \ \overline{\mathrm{tr}}(\Sigma_t G_t A G_t).$$

*Then, for each fixed $\lambda_t > 0$,*

$$\lambda_t Q_{\lambda_t} A \asymp \kappa_t \, G_t A, \tag{23}$$

$$\lambda_t^2 Q_{\lambda_t} A Q_{\lambda_t} \asymp \kappa_t^2 G_t A G_t + \kappa_t^2 b_t \, \tau_A \, G_t \Sigma_t G_t, \tag{24}$$

*and*

$$\frac{1}{n_t} \operatorname{tr}(A \widehat{\Sigma} Q_{\lambda_t}^2) \xrightarrow{\mathrm{p}} b_t \, \tau_A, \tag{25}$$

*where $b_t := (1 - t_{2,0})^{-1}$ and $t_{2,0} = \gamma_t \, \overline{\mathrm{tr}}(\Sigma_t^2 G_t^2)$.*

*Proof of Lemma C.1.* The first display (23) follows immediately from the scaled resolvent deterministic equivalent

$$\lambda_t Q_{\lambda_t} \asymp \kappa_t G_t$$

by right-multiplying with the bounded deterministic matrix $A$.

From here, we will skip the subscript $t$ on the objects for brevity. For a fixed $\lambda$ and a real parameter $h$ near zero, we define,

$$B(h) := \lambda I_p + hA, \qquad Q(h) := (\widehat{\Sigma} + B(h))^{-1}. \tag{26}$$

Since $A$ is uniformly bounded and $\lambda > 0$ is fixed, we may choose $h_0 > 0$ independently of $p$ so that $B(h) \succeq (\lambda/2) I_p$ whenever $|h| \leq h_0$.

Write $\widehat{\Sigma}_h = B(h)^{-1/2} \widehat{\Sigma} B(h)^{-1/2}$ and $\Sigma_h = B(h)^{-1/2} \Sigma B(h)^{-1/2}$. Since $A$ commute with $\Sigma$, we have $\Sigma_h = \Sigma B(h)^{-1}$. Notice that,

$$Q(h) = B(h)^{-1/2} (\widehat{\Sigma}_h + I_p)^{-1} B(h)^{-1/2}. \tag{27}$$

Apply Lemma C.6 with $\lambda = 1$, we have,

$$(\widehat{\Sigma}_h + I_p)^{-1} \asymp \kappa_h(\Sigma_h + \kappa_h I_p)^{-1}, \tag{28}$$

where $\kappa_h > 0$ is the unique solution of

$$\kappa_h = 1 + \gamma\kappa_h \operatorname{tr}\big[\Sigma_h(\Sigma_h + \kappa_h I_p)^{-1}\big]. \tag{29}$$

Multiplying on both sides the bounded deterministic matrix $B(h)^{-1/2}$, we obtain,

$$Q(h) \asymp \kappa_h(\Sigma + \kappa_h B(h))^{-1} = \left(B(h) + \frac{\Sigma}{\kappa_h}\right)^{-1} =: T(h), \tag{30}$$

and equivalently, $\kappa_h$ is the unique solution of,

$$\kappa_h = 1 + \gamma \operatorname{tr}(\Sigma T(h)). \tag{31}$$

At $h = 0$, we have,

$$\kappa_0 = \frac{\kappa}{\lambda}, \qquad T(0) = \kappa_0 G. \tag{32}$$

Differentiating (30) at $h = 0$, we have,

$$Q'(0) \asymp T'(0). \tag{33}$$

Since $B'(h) = A$, from the formula $(M^{-1})' = -M^{-1}M'M^{-1}$,

$$Q'(0) = -QAQ. \tag{34}$$

To compute $T'(0)$, differentiate $T(h) = (B(h) + \Sigma/k(h))^{-1}$,

$$T'(h) = -T(h)AT(h) + \frac{\kappa_h'}{\kappa_h^2} T(h)\Sigma T(h). \tag{35}$$

To derive $\kappa_h'$, we differentiate (31),

$$\kappa_h' = \gamma \operatorname{tr}(\Sigma T'(h)) = -\gamma \operatorname{tr}(\Sigma T(h)AT(h)) + \frac{\kappa_h'}{\kappa_h^2}\gamma \operatorname{tr}(\Sigma T(h)\Sigma T(h)). \tag{36}$$

Therefore,

$$\kappa_h' \left[1 - \frac{\gamma}{\kappa_h^2} \operatorname{tr}(\Sigma T(h)\Sigma T(h))\right] = -\gamma \operatorname{tr}(\Sigma T(h)AT(h)). \tag{37}$$

At $h = 0$, the term inside the bracket equals $1 - t_{2,0}$, while the right-hand side equals $-\kappa_0^2\tau_A$. Thus,

$$\kappa_0' = -\frac{\kappa_0^2\tau_A}{1 - t_{2,0}} = -\kappa_0^2 b\tau_A. \tag{38}$$

Substitute this into (35) yields,

$$T'(0) = -\kappa_0^2 GAG + b\tau_A G\Sigma G. \tag{39}$$

From this, we have (24).

For the remaining trace term, notice that,

$$\widehat{\Sigma}Q_\lambda^2 = \frac{d}{d\lambda}(\lambda Q_\lambda) \asymp \frac{d}{d\lambda}(\kappa G). \tag{40}$$

Lemma C.7 gives $\kappa' = b$ and therefore,

$$\frac{d}{d\lambda}(\kappa G) = b(G - \kappa G^2) = b\Sigma G^2. \tag{41}$$

Since $A$ is deterministic and uniformly bounded, we can multiply $A$ on both sides and take normalized trace,

$$\begin{aligned}\frac{1}{n_t}\operatorname{tr}(A\widehat{\Sigma}Q^2) &= \gamma\,\overline{\operatorname{tr}}(A\widehat{\Sigma}Q^2)\\ &= \gamma b\,\overline{\operatorname{tr}}(A\Sigma G^2)+o_p(1)\\ &= b\gamma\,\overline{\operatorname{tr}}(\Sigma GAG)+o_p(1)=b\tau_A+o_p(1),\end{aligned}$$

and this proves the remaining trace convergence. □

### C.4 Proof of Theorem 3.3

*Proof of Theorem 3.3.* By Theorem 3.1 of [15], the teacher risk has the deterministic equivalent

$$\mathcal{R}(\lambda_t)=\kappa_t^2 b_t q_{2,0}+\sigma^2 b_t t_{2,0}+\sigma^2. \tag{42}$$

The next two below lemmas give

$$\begin{aligned}R(\lambda_t)-C(\lambda_t,\lambda_s)\xrightarrow{\text{p}}\mathcal{R}(\lambda_t)-\mathcal{C}(\lambda_t,\lambda_s)&=\kappa_s\big(-\kappa_t q_{1,1}+\kappa_t^2 q_{2,1}+\kappa_t^2 b_t t_{2,1}q_{2,0}+\sigma^2 b_t t_{2,1}\big),\\ D(\lambda_t,\lambda_s)\xrightarrow{\text{p}}\mathcal{D}(\lambda_t,\lambda_s)&=\kappa_s^2\big(q_{0,2}+b_s\widetilde{q}_{0,2}-2\kappa_t(q_{1,2}+b_s\widetilde{q}_{1,2})+\kappa_t^2(q_{2,2}+b_s\widetilde{q}_{2,2})\\ &\quad+b_t(\kappa_t^2 q_{2,0}+\sigma^2)\big(t_{2,2}+b_s\gamma_t\,\overline{\operatorname{tr}}(\Sigma_t\Sigma_s G_t^2 G_s^2)\big)\big).\end{aligned}$$

Since $\mathcal{D}(\lambda_t,\lambda_s)=\mathcal{R}(\lambda_t)+\mathcal{R}_{\mathsf{pd}}(\lambda_s)-2\mathcal{C}(\lambda_t,\lambda_s)$, these identities yield closed-form expressions for $\mathcal{R}_{\mathsf{pd}}(\lambda_s)$ and $\mathcal{C}(\lambda_t,\lambda_s)$:

$$\mathcal{C}(\lambda_t,\lambda_s)=\sigma^2+\kappa_t^2 b_t q_{2,0}+\sigma^2 b_t t_{2,0}-\kappa_s(-\kappa_t q_{1,1}+\kappa_t^2 q_{2,1}+\kappa_t^2 b_t t_{2,1}q_{2,0}+\sigma^2 b_t t_{2,1}), \tag{43}$$

$$\begin{aligned}\mathcal{R}_{\mathsf{pd}}(\lambda_s)&=\kappa_s^2\big(q_{0,2}+b_s\widetilde{q}_{0,2}-2\kappa_t(q_{1,2}+b_s\widetilde{q}_{1,2})+\kappa_t^2(q_{2,2}+b_s\widetilde{q}_{2,2})\\ &\quad+b_t(\kappa_t^2 q_{2,0}+\sigma^2)\big(t_{2,2}+b_s\gamma_t\,\overline{\operatorname{tr}}(\Sigma_t\Sigma_s G_t^2 G_s^2)\big)\big)-\mathcal{R}(\lambda_t)+2\mathcal{C}(\lambda_t,\lambda_s).\end{aligned} \tag{44}$$

In the nontrivial case $r^2+\sigma^2>0$, we have the scalars $\kappa_t,\kappa_s,b_t,b_s,q_{2,0},t_{2,2}>0$, and

$$\begin{aligned}q_{0,2}-2\kappa_t q_{1,2}+\kappa_t^2 q_{2,2}&=\sum_{i=1}^p\frac{\sigma_i v_i}{(\widetilde{\sigma}_i+\kappa_s)^2}\left(1-\frac{2\kappa_t}{\sigma_i+\kappa_t}+\frac{\kappa_t^2}{(\sigma_i+\kappa_t)^2}\right)\\ &=\sum_{i=1}^p\frac{\sigma_i v_i}{(\widetilde{\sigma}_i+\kappa_s)^2}\left(\frac{\sigma_i}{\sigma_i+\kappa_t}\right)^2\\ &=\sum_{i=1}^p\frac{\sigma_i^3 v_i}{(\sigma_i+\kappa_t)^2(\widetilde{\sigma}_i+\kappa_s)^2}\geq 0,\\ \widetilde{q}_{0,2}-2\kappa_t\widetilde{q}_{1,2}+\kappa_t^2\widetilde{q}_{2,2}&=\sum_{i=1}^p\frac{\widetilde{\sigma}_i v_i}{(\widetilde{\sigma}_i+\kappa_s)^2}\left(1-\frac{2\kappa_t}{\sigma_i+\kappa_t}+\frac{\kappa_t^2}{(\sigma_i+\kappa_t)^2}\right)\\ &=\sum_{i=1}^p\frac{\widetilde{\sigma}_i v_i}{(\widetilde{\sigma}_i+\kappa_s)^2}\left(\frac{\sigma_i}{\sigma_i+\kappa_t}\right)^2\\ &=\sum_{i=1}^p\frac{\sigma_i^2\widetilde{\sigma}_i v_i}{(\sigma_i+\kappa_t)^2(\widetilde{\sigma}_i+\kappa_s)^2}\geq 0.\end{aligned}$$

Then, the limiting denominator $\mathcal{D}(\lambda_t,\lambda_s)=\mathcal{R}(\lambda_t)+\mathcal{R}_{\mathsf{pd}}(\lambda_s)-2\mathcal{C}(\lambda_t,\lambda_s)$ is strictly larger than 0 since

$$\begin{aligned}&\kappa_s^2\Big(q_{0,2}+b_s\widetilde{q}_{0,2}-2\kappa_t(q_{1,2}+b_s\widetilde{q}_{1,2})+\kappa_t^2(q_{2,2}+b_s\widetilde{q}_{2,2})+b_t(\kappa_t^2 q_{2,0}+\sigma^2)\big(t_{2,2}+b_s\gamma_t\,\overline{\operatorname{tr}}(\Sigma_t\Sigma_s G_t^2 G_s^2)\big)\Big)\\ &=\underbrace{\kappa_s^2}_{>0}\Big(\underbrace{q_{0,2}-2\kappa_t q_{1,2}+\kappa_t^2 q_{2,2}}_{\geq 0}+\underbrace{b_s(\widetilde{q}_{0,2}-2\kappa_t\widetilde{q}_{1,2}+\kappa_t^2\widetilde{q}_{2,2})}_{\geq 0}+\underbrace{b_t(\kappa_t^2 q_{2,0}+\sigma^2)\big(t_{2,2}+b_s\gamma_t\,\overline{\operatorname{tr}}(\Sigma_t\Sigma_s G_t^2 G_s^2)\big)}_{>0}\Big).\end{aligned}$$

Substituting these deterministic equivalents into the oracle identities in Proposition 3.1 gives

$$\xi^\star(\lambda_t,\lambda_s) \xrightarrow{\mathrm{p}} \frac{\mathcal{R}(\lambda_t)-\mathcal{C}(\lambda_t,\lambda_s)}{\mathcal{R}(\lambda_t)+\mathcal{R}_{\mathsf{pd}}(\lambda_s)-2\mathcal{C}(\lambda_t,\lambda_s)},$$

$$R^\star_{\mathsf{pm}}(\lambda_t,\lambda_s) \xrightarrow{\mathrm{p}} \mathcal{R}(\lambda_t) - \frac{\big(\mathcal{R}(\lambda_t)-\mathcal{C}(\lambda_t,\lambda_s)\big)^2}{\mathcal{R}(\lambda_t)+\mathcal{R}_{\mathsf{pd}}(\lambda_s)-2\mathcal{C}(\lambda_t,\lambda_s)}.$$

□

**Lemma C.2** (Deterministic equivalent for $R-C$)**.** *Under Assumption 3.2, as $n_t, n_s, p \to \infty$ with*

$$0 < \liminf \gamma_t \le \limsup \gamma_t < \infty, \qquad 0 < \liminf \gamma_s \le \limsup \gamma_s < \infty,$$

*for each fixed $\lambda_t > 0$ and $\lambda_s > 0$,*

$$R(\lambda_t) - C(\lambda_t,\lambda_s) \xrightarrow{\mathrm{p}} \kappa_s\big(-\kappa_t q_{1,1} + \kappa_t^2 q_{2,1} + \kappa_t^2 b_t t_{2,1} q_{2,0} + \sigma^2 b_t t_{2,1}\big). \tag{45}$$

*Proof of Lemma C.2.* Recall that

$$\beta_{\lambda_t} := Q_{\lambda_t} X^\top y/n_t, \qquad \beta_{\mathsf{pd},\lambda_s} := \widetilde{Q}_{\lambda_s}\widetilde{\Sigma}\,\beta_{\lambda_t} = (I_p - \lambda_s \widetilde{Q}_{\lambda_s})\beta_{\lambda_t},$$

Therefore,

$$R(\lambda_t) - C(\lambda_t,\lambda_s) = (\beta_{\lambda_t}-\beta)^\top \Sigma_t (\beta_{\lambda_t} - \beta_{\mathsf{pd},\lambda_s}) = \lambda_s\,(\beta_{\lambda_t}-\beta)^\top \Sigma_t \widetilde{Q}_{\lambda_s}\beta_{\lambda_t}. \tag{46}$$

By the fresh sample resolvent deterministic equivalent,

$$\lambda_s \widetilde{Q}_{\lambda_s} \asymp \kappa_s G_s,$$

hence

$$R(\lambda_t) - C(\lambda_t,\lambda_s) = \kappa_s\,(\beta_{\lambda_t}-\beta)^\top \Sigma_t G_s\,\beta_{\lambda_t} + o_p(1). \tag{47}$$

Now expand

$$\beta_{\lambda_t} - \beta = -\lambda_t Q_{\lambda_t}\beta + Q_{\lambda_t} X^\top \varepsilon/n_t, \qquad \beta_{\lambda_t} = (I_p - \lambda_t Q_{\lambda_t})\beta + Q_{\lambda_t} X^\top \varepsilon/n_t.$$

With $A_s := \Sigma_t G_s$, this gives

$$(\beta_{\lambda_t}-\beta)^\top A_s \beta_{\lambda_t} = -\lambda_t\,\beta^\top Q_{\lambda_t} A_s \beta + \lambda_t^2 \beta^\top Q_{\lambda_t} A_s Q_{\lambda_t}\beta + \frac{1}{n_t^2}\varepsilon^\top X Q_{\lambda_t} A_s Q_{\lambda_t} X^\top \varepsilon + \mathrm{Lin}(\lambda_t), \tag{48}$$

where $\mathrm{Lin}(\lambda_t)$ is linear in $\varepsilon$. By the same leave-one-out concentration argument used in Lemma B.3 of [15], we have $\mathrm{Lin}(\lambda_t) = o_p(1)$.

Apply Lemma C.1 with $A_s = \Sigma_t G_s$. Because $A_s$ commutes with $\Sigma_t$,

$$\tau_{A_s} = \gamma_t\ \overline{\mathrm{tr}}(\Sigma_t^2 G_t^2 G_s) = t_{2,1}.$$

Also,

$$\beta^\top G_t A_s \beta = \beta^\top G_t \Sigma_t G_s \beta = q_{1,1}, \qquad \beta^\top G_t A_s G_t \beta = \beta^\top G_t^2 \Sigma_t G_s \beta = q_{2,1}.$$

Therefore,

$$-\lambda_t\,\beta^\top Q_{\lambda_t} A_s \beta \xrightarrow{\mathrm{p}} -\kappa_t\, q_{1,1},$$

$$\lambda_t^2 \beta^\top Q_{\lambda_t} A_s Q_{\lambda_t}\beta \xrightarrow{\mathrm{p}} \kappa_t^2 q_{2,1} + \kappa_t^2 b_t t_{2,1}\, q_{2,0},$$

and

$$\frac{1}{n_t^2}\varepsilon^\top X Q_{\lambda_t} A_s Q_{\lambda_t} X^\top \varepsilon \xrightarrow{\mathrm{p}} \sigma^2 \cdot \frac{1}{n_t}\,\mathrm{tr}(A_s \widehat{\Sigma} Q_{\lambda_t}^2) \xrightarrow{\mathrm{p}} \sigma^2 b_t t_{2,1}.$$

Thus,

$$\mathcal{R}(\lambda_t) - \mathcal{C}(\lambda_t,\lambda_s) = \kappa_s\big(-\kappa_t q_{1,1} + \kappa_t^2 q_{2,1} + \kappa_t^2 b_t t_{2,1} q_{2,0} + \sigma^2 b_t t_{2,1}\big).$$

□

**Lemma C.3** (Deterministic equivalent for $D$)**.** *Under Assumption 3.2, as $n_t, n_s, p \to \infty$ with*

$$0 < \liminf \gamma_t \le \limsup \gamma_t < \infty, \qquad 0 < \liminf \gamma_s \le \limsup \gamma_s < \infty,$$

*for each fixed $\lambda_t > 0$ and $\lambda_s > 0$,*

$$D(\lambda_t, \lambda_s) \xrightarrow{\text{p}} \mathcal{D}(\lambda_t, \lambda_s) = \kappa_s^2\big(q_{0,2} + b_s\widetilde{q}_{0,2} - 2\kappa_t(q_{1,2} + b_s\widetilde{q}_{1,2}) + \kappa_t^2(q_{2,2} + b_s\widetilde{q}_{2,2}) + b_t(\kappa_t^2 q_{2,0} + \sigma^2)\big(t_{2,2} + b_s\gamma_t\,\overline{\mathrm{tr}}(\Sigma_t\Sigma_s G_t^2 G_s^2)\big)\big), \tag{49}$$

*where $b_s = \big(\gamma_s\,\overline{\mathrm{tr}}(\Sigma_s\Sigma_t G_s^2)\big)/\big(1 - \gamma_s\,\overline{\mathrm{tr}}(\Sigma_s^2 G_s^2)\big)$.*

*Proof of Lemma C.3.* We first note the exact identities

$$\beta_{\mathsf{pd},\lambda_s} = (I_p - \lambda_s\widetilde{Q}_{\lambda_s})\beta_{\lambda_t}, \qquad \beta_{\lambda_t} - \beta_{\mathsf{pd},\lambda_s} = \lambda_s\widetilde{Q}_{\lambda_s}\beta_{\lambda_t},$$

where $\widetilde{Q}_{\lambda_s} := (\widetilde{\Sigma} + \lambda_s I_p)^{-1}$ and $\widetilde{\Sigma} = \widetilde{X}^\top\widetilde{X}/n_s$. Writing $e_{\lambda_t} := \beta_{\lambda_t} - \beta$, we have

$$D(\lambda_t, \lambda_s) = (\beta_{\lambda_t} - \beta_{\mathsf{pd},\lambda_s})^\top\Sigma_t(\beta_{\lambda_t} - \beta_{\mathsf{pd},\lambda_s}) = \lambda_s^2\,\beta_{\lambda_t}^\top\widetilde{Q}_{\lambda_s}\Sigma_t\widetilde{Q}_{\lambda_s}\beta_{\lambda_t}.$$

Applying the fresh-sample analogue of Lemma C.1 with $A = \Sigma_t$, which commutes with $\Sigma_s$ by Assumption 3.2, gives

$$\lambda_s\widetilde{Q}_{\lambda_s} \asymp \kappa_s G_s, \qquad \lambda_s^2\widetilde{Q}_{\lambda_s}\Sigma_t\widetilde{Q}_{\lambda_s} \asymp \kappa_s^2 G_s^2\Sigma_t + \kappa_s^2\frac{\gamma_s\,\overline{\mathrm{tr}}(\Sigma_s\Sigma_t G_s^2)}{1 - \gamma_s\,\overline{\mathrm{tr}}(\Sigma_s^2 G_s^2)}G_s\Sigma_s G_s.$$

Hence,

$$D(\lambda_t, \lambda_s) = \beta_{\lambda_t}^\top A_2\beta_{\lambda_t} + o_p(1),$$
$$A_2 := \kappa_s^2 G_s^2\Sigma_t + \kappa_s^2 b_s G_s\Sigma_s G_s, \tag{50}$$

where we denote

$$b_s := \frac{\gamma_s\,\overline{\mathrm{tr}}(\Sigma_s\Sigma_t G_s^2)}{1 - \gamma_s\,\overline{\mathrm{tr}}(\Sigma_s^2 G_s^2)}.$$

Note that

$$e_{\lambda_t} = \beta_{\lambda_t} - \beta = -\lambda_t Q_{\lambda_t}\beta + Q_{\lambda_t}X^\top\varepsilon/n_t.$$

For any deterministic positive semidefinite insertion $A$ that commutes with $\Sigma_t$, Lemma C.1 gives

$$e_{\lambda_t}^\top A e_{\lambda_t} \xrightarrow{\text{p}} \kappa_t^2\beta^\top G_t A G_t\beta + b_t\,\gamma_t\;\overline{\mathrm{tr}}(A\Sigma_t G_t^2)\,\big(\kappa_t^2 q_{2,0} + \sigma^2\big),$$

and

$$e_{\lambda_t}^\top A\beta \xrightarrow{\text{p}} -\kappa_t\,\beta^\top G_t A\beta,$$

where linear terms in $\varepsilon$ are $o_p(1)$ by the same leave-one-out concentration argument used in Lemma B.3 of [15].

Using $\beta_{\lambda_t} = \beta + e_{\lambda_t}$, we therefore obtain

$$e_{\lambda_t}^\top A\beta_{\lambda_t} = e_{\lambda_t}^\top A\beta + e_{\lambda_t}^\top A e_{\lambda_t} \xrightarrow{\text{p}} -\kappa_t\,\beta^\top G_t A\beta + \kappa_t^2\beta^\top G_t A G_t\beta + b_t\,\gamma_t\;\overline{\mathrm{tr}}(A\Sigma_t G_t^2)\,\big(\kappa_t^2 q_{2,0} + \sigma^2\big),$$

and

$$\beta_{\lambda_t}^\top A\beta_{\lambda_t} = \beta^\top A\beta + 2e_{\lambda_t}^\top A\beta + e_{\lambda_t}^\top A e_{\lambda_t}$$
$$\xrightarrow{\text{p}} \beta^\top A\beta - 2\kappa_t\,\beta^\top G_t A\beta + \kappa_t^2\beta^\top G_t A G_t\beta + b_t\,\gamma_t\;\overline{\mathrm{tr}}(A\Sigma_t G_t^2)\,\big(\kappa_t^2 q_{2,0} + \sigma^2\big).$$

Substituting $A = A_2$ yields

$$\beta^\top A_2\beta = \kappa_s^2 q_{0,2} + \kappa_s^2 b_s\widetilde{q}_{0,2}, \qquad \beta^\top G_t A_2\beta = \kappa_s^2 q_{1,2} + \kappa_s^2 b_s\widetilde{q}_{1,2}, \qquad \beta^\top G_t A_2 G_t\beta = \kappa_s^2 q_{2,2} + \kappa_s^2 b_s\widetilde{q}_{2,2},$$

and

$$\gamma_t\,\overline{\mathrm{tr}}(A_2\Sigma_t G_t^2) = \kappa_s^2 t_{2,2} + \kappa_s^2 b_s\gamma_t\,\overline{\mathrm{tr}}(\Sigma_t\Sigma_s G_t^2 G_s^2).$$

Finally,

$$\mathcal{D}(\lambda_t, \lambda_s) = \kappa_s^2\big(q_{0,2} + b_s\widetilde{q}_{0,2} - 2\kappa_t(q_{1,2} + b_s\widetilde{q}_{1,2}) + \kappa_t^2(q_{2,2} + b_s\widetilde{q}_{2,2}) + b_t(\kappa_t^2 q_{2,0} + \sigma^2)\big(t_{2,2} + b_s\gamma_t\,\overline{\mathrm{tr}}(\Sigma_t\Sigma_s G_t^2 G_s^2)\big)\big).$$

This proves (49). □

### C.5 Background on deterministic equivalents

We use an anisotropic (bilinear-form) notion of deterministic equivalent.

**Definition C.4** (Deterministic equivalent)**.** Let $A = A_p$ be a (possibly random) $p \times p$ matrix and $\overline{A} = \overline{A}_p$ a deterministic $p \times p$ matrix. We write $A \asymp \overline{A}$ if, for every pair of deterministic vectors $u = u_p$ and $v = v_p$ with $\|u\|_2, \|v\|_2 = O(1)$,

$$u^\top (A - \overline{A}) v \xrightarrow{\mathrm{p}} 0.$$

If $A(\theta), \overline{A}(\theta)$ depend on a parameter $\theta$ in an open set $\Theta$, we write $A(\theta) \asymp \overline{A}(\theta)$ *uniformly on compact subsets of* $\Theta$ if the convergence above holds uniformly over $\theta$ in any compact $K \subset \Theta$.

This notion implies trace convergence whenever operator norms are uniformly bounded. In particular, if $A \asymp \overline{A}$ and $\|A\|_{\mathrm{op}}, \|\overline{A}\|_{\mathrm{op}} = O_{\mathbb{P}}(1)$, then $\overline{\mathrm{tr}}(A) - \overline{\mathrm{tr}}(\overline{A}) \xrightarrow{\mathrm{p}} 0$.

We will also use that uniform deterministic equivalents can be differentiated.

**Lemma C.5** (Differentiate a deterministic equivalent)**.** *Let $A(z)$ be random and $\overline{A}(z)$ deterministic, both analytic in a complex neighborhood $U$ of $z_0$. If $A(z) \asymp \overline{A}(z)$ locally uniformly on $U$, then $A'(z_0) \asymp \overline{A}'(z_0)$.*

*Proof.* For any deterministic $u$ and $v$, apply Cauchy's integral formula to $u^\top (A(z) - \overline{A}(z)) v$ on a sufficiently small circle around $z_0$ contained in $U$. Local uniform convergence makes the resulting contour integral $o_p(1)$. □

For the next two lemmas, let $\widehat{\Sigma} = X^\top X/n$ be a generic sample covariance matrix with population covariance $\Sigma$ and aspect ratio $\gamma = p/n$, and let $Q_\lambda = (\widehat{\Sigma} + \lambda I_p)^{-1}$. A standard anisotropic local law for sample-covariance resolvents gives the following deterministic equivalent; see, e.g., [59, 35, 50].

**Lemma C.6** (Scaled resolvent deterministic equivalent)**.** *Under Assumption 3.2, for each fixed $\lambda > 0$,*

$$\lambda Q_\lambda \;\asymp\; \kappa(\lambda)\,(\Sigma + \kappa(\lambda) I_p)^{-1} \;=\; \kappa G, \tag{51}$$

*where $\kappa = \kappa(\lambda) > 0$ is the unique solution to*

$$\kappa \;=\; \lambda + \gamma\,\kappa\;\overline{\mathrm{tr}}\big(\Sigma(\Sigma + \kappa I_p)^{-1}\big), \tag{52}$$

*and $G = (\Sigma + \kappa I_p)^{-1}$. The equivalence extends locally uniformly to a complex neighborhood of each $\lambda > 0$ and, in particular, holds uniformly for $\lambda$ in compact subsets of $(0, \infty)$.*

Define

$$t_2(\lambda) := \gamma\,\overline{\mathrm{tr}}(\Sigma^2 G^2), \qquad b(\lambda) := \frac{1}{1 - t_2(\lambda)}.$$

**Lemma C.7** (Derivative of the fixed-point solution)**.** *The map $\lambda \mapsto \kappa(\lambda)$ is differentiable for $\lambda > 0$ and*

$$\kappa'(\lambda) = b(\lambda) = \frac{1}{1 - t_2(\lambda)}.$$

*Proof.* Differentiate (52) and use $I_p - \kappa G = \Sigma G$ to obtain

$$\kappa'(\lambda)\big(1 - \gamma\,\overline{\mathrm{tr}}(\Sigma^2 G^2)\big) = 1.$$

□

## D Proofs in Section 4

### D.1 Proof of Theorem 4.1

*Proof of Theorem 4.1.* **Step 1:** Upper bound the size of the "teacher-degeneracy" set.

By Lemma C.2,

$$\mathcal{R}(\lambda_t) - \mathcal{C}(\lambda_t, \lambda_s) = \kappa_s\big( - \kappa_t q_{1,1} + \kappa_t^2 q_{2,1} + \kappa_t^2 b_t t_{2,1} q_{2,0} + \sigma^2 b_t t_{2,1}\big).$$

The equation $\mathcal{R}(\lambda_t) - \mathcal{C}(\lambda_t, \lambda_s) = 0$ is therefore equivalent to

$$-\kappa_t q_{1,1} + \kappa_t^2 q_{2,1} + \kappa_t^2 b_t t_{2,1} q_{2,0} + \sigma^2 b_t t_{2,1} = 0. \tag{53}$$

Write the eigendecompositions $\Sigma_t = U^\top \operatorname{diag}(\sigma_1, \ldots, \sigma_p) U$ and $\Sigma_s = U^\top \operatorname{diag}(\widetilde{\sigma}_1, \ldots, \widetilde{\sigma}_p) U$, and the squared signal-covariance alignment $v_i := ((U\beta)_i)^2$. Equation (53) becomes

$$\begin{aligned}
&\sum_{i=1}^{p} \left( \frac{-\kappa_t \sigma_i v_i}{(\sigma_i + \kappa_t)(\widetilde{\sigma}_i + \kappa_s)} + \frac{\kappa_t^2 \sigma_i v_i}{(\sigma_i + \kappa_t)^2 (\widetilde{\sigma}_i + \kappa_s)} + (\kappa_t^2 b_t q_{2,0} + \sigma^2 b_t) \frac{1}{n_t} \frac{\sigma_i^2}{(\sigma_i + \kappa_t)^2 (\widetilde{\sigma}_i + \kappa_s)} \right) = 0 \\
\Leftrightarrow\ &\sum_{i=1}^{p} \frac{\sigma_i^2 \big[ -\kappa_t v_i + (\kappa_t^2 b_t q_{2,0} + \sigma^2 b_t)/n_t \big]}{(\sigma_i + \kappa_t)^2} \frac{1}{\widetilde{\sigma}_i + \kappa_s} = 0.
\end{aligned} \tag{54}$$

For $i \in [p]$, define,

$$c_i := \frac{\kappa_t \sigma_i^2}{(\sigma_i + \kappa_t)^2} \big[ \frac{\kappa_t^2 b_t q_{2,0} + \sigma^2 b_t}{\kappa_t n_t} - v_i \big].$$

Then (54) becomes

$$H(\kappa_s) := \sum_{i=1}^{p} \frac{c_i}{\widetilde{\sigma}_i + \kappa_s} = \sum_{j=1}^{\widetilde{m}} \frac{d_j}{\widetilde{\sigma}_{(j)} + \kappa_s} = 0, \tag{55}$$

where $d_j := \sum_{k:\, \widetilde{\sigma}_k = \widetilde{\sigma}_{(j)}} c_k$.

As $\lambda_s$ ranges over $(0, \infty)$, the corresponding fixed-point solution $\kappa_s$ ranges over an interval. Thus, $\lambda_t \in \Lambda_t$ if and only if $H$ vanishes throughout this interval. Clearing the denominators gives

$$\sum_{j=1}^{\widetilde{m}} d_j \prod_{k \neq j} (\widetilde{\sigma}_{(k)} + \kappa_s) = 0.$$

The left-hand side is a polynomial that vanishes on an interval and is therefore identically zero. For any $j \in [\widetilde{m}]$, evaluating it at $\kappa_s = -\widetilde{\sigma}_{(j)}$ yields

$$d_j \prod_{k \neq j} (\widetilde{\sigma}_{(k)} - \widetilde{\sigma}_{(j)}) = 0.$$

The eigenvalues $\widetilde{\sigma}_{(1)}, \ldots, \widetilde{\sigma}_{(\widetilde{m})}$ are distinct, so $d_j = 0$. Conversely, if every $d_j = 0$, then $H$ is identically zero and $\lambda_t \in \Lambda_t$.

We provide a crude upper bound on the cardinality of $\Lambda_t$ using only the weaker necessary condition,

$$\sum_{j=1}^{\widetilde{m}} d_j = \sum_{i=1}^{p} c_i = 0.$$

This condition is equivalent to

$$\begin{aligned}
&\sum_{i=1}^{p} \frac{\kappa_t \sigma_i^2}{(\sigma_i + \kappa_t)^2} \left( \frac{\kappa_t^2 b_t q_{2,0} + \sigma^2 b_t}{\kappa_t n_t} - v_i \right) = 0, \\
\Leftrightarrow\quad &\frac{b_t}{n_t} (\kappa_t^2 q_{2,0} + \sigma^2) \sum_{i=1}^{p} \frac{\sigma_i^2}{(\sigma_i + \kappa_t)^2} - \kappa_t \sum_{i=1}^{p} \frac{\sigma_i^2 v_i}{(\sigma_i + \kappa_t)^2} = 0.
\end{aligned} \tag{56}$$

For convenience, write $\kappa_t$ as $\kappa$. Let $m$ be the number of distinct eigenvalues of $\Sigma_t$, denoted by $\{\sigma_{(1)}, \ldots, \sigma_{(m)}\}$. Define

$$P(\kappa) := \prod_{j=1}^{m} (\sigma_{(j)} + \kappa),$$

$$Q(\kappa) := \sum_{i=1}^{p} \sigma_i^2 \prod_{j:\sigma_{(j)} \neq \sigma_i} (\sigma_{(j)} + \kappa)^2,$$

$$R(\kappa) := \sum_{i=1}^{p} \sigma_i v_i \prod_{j:\sigma_{(j)} \neq \sigma_i} (\sigma_{(j)} + \kappa)^2,$$

$$S(\kappa) := \sum_{i=1}^{p} \sigma_i^2 v_i \prod_{j:\sigma_{(j)} \neq \sigma_i} (\sigma_{(j)} + \kappa)^2.$$

Then

$$\sum_{i=1}^{p} \frac{\sigma_i^2}{(\sigma_i + \kappa)^2} = \frac{Q(\kappa)}{P(\kappa)^2}, \quad \sum_{i=1}^{p} \frac{\sigma_i^2 v_i}{(\sigma_i + \kappa)^2} = \frac{S(\kappa)}{P(\kappa)^2}, \quad q_{2,0} = \sum_{i=1}^{p} \frac{\sigma_i v_i}{(\sigma_i + \kappa)^2} = \frac{R(\kappa)}{P(\kappa)^2},$$

$$b_t = \frac{1}{1 - t_{2,0}} = \frac{1}{1 - \frac{1}{n_t} \sum_{i=1}^{p} \frac{\sigma_i^2}{(\sigma_i + \kappa)^2}} = \frac{P(\kappa)^2}{P(\kappa)^2 - Q(\kappa)/n_t}.$$

Substituting these identities into the degeneracy condition and clearing the denominator gives

$$\frac{Q(\kappa)}{n_t} \big(\kappa^2 R(\kappa) + \sigma^2 P(\kappa)^2\big) - \kappa S(\kappa)(P(\kappa)^2 - Q(\kappa)/n_t) = 0. \tag{57}$$

By construction, $\deg P = m$ and $\deg Q, \deg R, \deg S \leq 2m - 2$, so this polynomial has degree at most $4m - 1$. When $\beta \neq 0$, its leading coefficient is $-\sum_{i=1}^{p} \sigma_i^2 v_i < 0$. If $\beta = 0$ and $\sigma^2 > 0$, the polynomial reduces to $\sigma^2 Q(\kappa) P(\kappa)^2 / n_t$ and is again nonzero. Thus, in the nontrivial case $r^2 + \sigma^2 > 0$, the degeneracy condition has at most $4m - 1$ solutions. Because $\lambda_t \mapsto \kappa_t$ is one-to-one, the same bound applies to the corresponding values of $\lambda_t$.

**Step 2:** Upper bound the size of the "student-tie" set.

We will prove a stronger result that when $\lambda_t \notin \Lambda_t$, the size of $\Lambda_s(\lambda_t)$ is upper bounded by the number of sign changes in the sequence $\{d_1, \ldots, d_{\widetilde{m}}\}$ after deleting zero entries, where the $d_j$ are defined after (55). Thus, it is clear that $\Lambda_s(\lambda_t) \leq \widetilde{m} - 1$ if this is true.

Let $j_1 < \cdots < j_K$ index the nonzero entries of $(d_1, \ldots, d_{\widetilde{m}})$, and let $M$ be the number of sign changes in $(d_{j_1}, \ldots, d_{j_K})$. Because $\lambda_t \notin \Lambda_t$, Theorem 4.1 ensures that $K \geq 1$. Suppose, for contradiction, that $H$ has $r > M$ distinct positive roots $0 < x_1 < \cdots < x_r$.

Let $\mathcal{I}$ contain the indices $a$ at which $d_{j_a}$ and $d_{j_{a+1}}$ have opposite signs. For each $a \in \mathcal{I}$, choose $\eta_a \in (\widetilde{\sigma}_{(j_a)}, \widetilde{\sigma}_{(j_{a+1})})$ and set

$$Q(t) := c_0 \prod_{a \in \mathcal{I}} (t - \eta_a),$$

where the sign of $c_0 \neq 0$ is chosen so that

$$\operatorname{sign} Q(\widetilde{\sigma}_{(j_a)}) = \operatorname{sign} d_{j_a}, \qquad a = 1, \ldots, K.$$

This construction gives $\deg(Q) = M \leq r - 1$. Define

$$P(t) := \frac{Q(t)}{\prod_{\ell=1}^{r} (t + x_\ell)}.$$

Since $\deg(Q) \leq r - 1$, $P$ has the partial fraction decomposition

$$P(t) = \sum_{\ell=1}^{r} \frac{y_\ell}{t + x_\ell}, \qquad y_\ell := \frac{Q(-x_\ell)}{\prod_{k \neq \ell} (x_k - x_\ell)}.$$

Indeed,

$$Q(t) = \sum_{\ell=1}^{r} y_\ell \prod_{k \neq \ell} (t + x_k),$$

because the two polynomials have degree at most $r-1$ and agree at $t=-x_1,\dots,-x_r$.

Since $x_\ell > 0$, the denominator of $P(\widetilde{\sigma}_{(j)})$ is positive. Thus, $d_j P(\widetilde{\sigma}_{(j)}) > 0$ for every nonzero $d_j$, and

$$\sum_{j=1}^{\widetilde{m}} d_j P(\widetilde{\sigma}_{(j)}) > 0.$$

On the other hand,

$$\sum_{j=1}^{\widetilde{m}} d_j P(\widetilde{\sigma}_{(j)}) = \sum_{\ell=1}^{r} y_\ell \sum_{j=1}^{\widetilde{m}} \frac{d_j}{\widetilde{\sigma}_{(j)} + x_\ell} = \sum_{\ell=1}^{r} y_\ell H(x_\ell) = 0, \tag{58}$$

which is a contradiction. Therefore, $H$ has at most $M$ positive roots. □

### D.2 Proof of Proposition 4.2

*Proof of Proposition 4.2.* When $\Sigma_s = I_p$, all $\widetilde{\sigma}_i$ equal 1, so (55) becomes

$$H(\kappa_s) = \frac{\sum_{i=1}^{p} c_i}{1+\kappa_s}.$$

If the numerator is nonzero, $H(\kappa_s) = 0$ has no solution; hence, strict improvement holds for every $\lambda_s > 0$. □

### D.3 Proof of Theorem 4.3

*Proof of Theorem 4.3.* Fix $0 < \gamma_{s,1} < \gamma_{s,2}$ and an arbitrary $\lambda_{s,2} > 0$. Let $\kappa_s$ be the corresponding solution to the student fixed-point equation at $\gamma_{s,2}$ (same as (6)),

$$\kappa_s = \lambda_{s,2} + \gamma_{s,2}\kappa_s \ \overline{\mathrm{tr}}\big(\Sigma_s(\Sigma_s + \kappa_s I_p)^{-1}\big).$$

Recall that $G_s = (\Sigma_s + \kappa_s I_p)^{-1}$. Equivalently, we have

$$\lambda_{s,2} = \kappa_s \left(1 - \gamma_{s,2} \operatorname{tr}(\Sigma_s G_s)\right). \tag{59}$$

Keep the $\kappa_s$ fixed. Now define $\lambda_{s,1} := \kappa_s \left(1 - \gamma_{s,1} \operatorname{tr}(\Sigma_s G_s)\right)$. Since $\gamma_{s,1} < \gamma_{s,2}$, $\lambda_{s,1} > \lambda_{s,2} > 0$. Moreover, by construction, the student fixed-point equation at $(\gamma_{s,1}, \lambda_{s,1})$ has the same solution $\kappa_s$. Thus, $\kappa_s$ and $G_s$ are the same under two aspect ratios with two regularization values $\lambda_{s,1}, \lambda_{s,2}$.

We recall the PMSD risk in Theorem 3.3,

$$\mathcal{R}^\star_{\mathsf{pm}}(\lambda_t, \lambda_s) = \mathcal{R}(\lambda_t) - \frac{\big(\mathcal{R}(\lambda_t) - \mathcal{C}(\lambda_t, \lambda_s)\big)^2}{\mathcal{R}(\lambda_t) + \mathcal{R}_{\mathsf{pd}}(\lambda_s) - 2\mathcal{C}(\lambda_t, \lambda_s)} \tag{60}$$

From Theorem 3.3,

$$\mathcal{R}(\lambda_t) - \mathcal{C}(\lambda_t, \lambda_s) = \kappa_s(-\kappa_t q_{1,1} + \kappa_t^2 q_{2,1} + b_t t_{2,1}(\kappa_t^2 q_{2,0} + \sigma^2)),$$

which does not depend on $\gamma_s$ when $\kappa_s$ is held fixed. Also, from Theorem 3.3,

$$\mathcal{R}(\lambda_t) + \mathcal{R}_{\mathrm{pd}}(\lambda_s) - 2\mathcal{C}(\lambda_t, \lambda_s) = \kappa_s^2 \left(B_0 + b_s(\gamma_s) B_1\right),$$

where

$$\begin{aligned}
B_0 &:= q_{0,2} - 2\kappa_t q_{1,2} + \kappa_t^2 q_{2,2} + b_t \left(\kappa_t^2 q_{2,0} + \sigma^2\right) t_{2,2} \geq 0, \\
B_1 &:= \widetilde{q}_{0,2} - 2\kappa_t \widetilde{q}_{1,2} + \kappa_t^2 \widetilde{q}_{2,2} + b_t \left(\kappa_t^2 q_{2,0} + \sigma^2\right) \gamma_t \operatorname{tr}\left(\Sigma_t \Sigma_s G_t^2 G_s^2\right) \geq 0, \\
b_s(\gamma_s) &= \frac{\gamma_s \operatorname{tr}(\Sigma_s \Sigma_t G_s^2)}{1 - \gamma_s \operatorname{tr}(\Sigma_s^2 G_s^2)}.
\end{aligned}$$

Note that $B_0$ and $B_1$ also do not depend on $\gamma_s$ when $\kappa_s$ is held fixed. For $\gamma_{s,1} < \gamma_{s,2}$, we have $b_s(\gamma_{s,1}) < b_s(\gamma_{s,2})$. Therefore,

$$\mathcal{R}^\star_{\mathsf{pm}}(\gamma_{s,1}, \lambda_{s,1}) \leq \mathcal{R}^\star_{\mathsf{pm}}(\gamma_{s,2}, \lambda_{s,2}).$$

We thus have,

$$\inf_{\lambda_s>0} \mathcal{R}^\star_{\mathsf{pm}}(\gamma_{s,1},\lambda_s) \le \mathcal{R}^\star_{\mathsf{pm}}(\gamma_{s,1},\lambda_{s,1}) \le \mathcal{R}^\star_{\mathsf{pm}}(\gamma_{s,2},\lambda_{s,2}).$$

Since $\lambda_{s,2} > 0$ is arbitrary, taking the infimum over $\lambda_{s,2} > 0$ gives

$$\inf_{\lambda_s>0} \mathcal{R}^\star_{\mathsf{pm}}(\gamma_{s,1},\lambda_s) \le \inf_{\lambda_s>0} \mathcal{R}^\star_{\mathsf{pm}}(\gamma_{s,2},\lambda_s).$$

□

# E Proofs in Section 5

## E.1 Proof of Theorem 5.1

We prove a slightly more general finite-candidate version of Theorem 5.1. This version is useful when the same calibration set is used to choose among finitely many student regularization parameters $\lambda_s$, unlabeled sample sizes $n_s$, or other finite hyperparameter choices.

The argument is straightforward: once the teacher and fresh-$X$ pure-distilled student are fixed, calibration estimates only the risks and residual correlation of two fixed predictors. Thus, no high-dimensional condition on $p/n_{\mathrm{cal}}$ is needed.

Let $\mathcal{G}$ denote the sigma-field generated by the deployed teacher, the fresh unlabeled covariates, and all fitted pure-distilled students under consideration. Conditionally on $\mathcal{G}$, the teacher and student predictors are fixed functions. For a single candidate $f_{\mathsf{pd}}$, let $(X_i^{\mathrm{cal}}, Y_i^{\mathrm{cal}})_{i=1}^{n_{\mathrm{cal}}}$ be an independent labeled calibration sample drawn from the target test distribution. Define

$$e_i := Y_i^{\mathrm{cal}} - f(X_i^{\mathrm{cal}}), \qquad e_i^{\mathsf{pd}} := Y_i^{\mathrm{cal}} - f_{\mathsf{pd}}(X_i^{\mathrm{cal}}).$$

The oracle quantities are

$$R := \mathbb{E}[(e)^2 \mid \mathcal{G}], \qquad R_{\mathsf{pd}} := \mathbb{E}[(e^{\mathsf{pd}})^2 \mid \mathcal{G}], \qquad C := \mathbb{E}[e e^{\mathsf{pd}} \mid \mathcal{G}],$$

where $(e, e^{\mathsf{pd}})$ denotes an independent test residual pair. The calibration estimates are

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

We first prove the following finite-candidate result.

**Theorem E.1** (Calibration consistency, finite-candidate version)**.** *Fix a finite collection of candidate pure-distilled students*

$$f_{\mathsf{pd}}^{(1)}, \ldots, f_{\mathsf{pd}}^{(J)}$$

*with $J < \infty$, and condition on $\mathcal{G}$. For each $j \in [J]$, define*

$$R_{\mathsf{pd},j} := \mathbb{E}[(e_j^{\mathsf{pd}})^2 \mid \mathcal{G}], \qquad C_j := \mathbb{E}[e e_j^{\mathsf{pd}} \mid \mathcal{G}], \qquad D_j := R + R_{\mathsf{pd},j} - 2C_j,$$

*where $e_j^{\mathsf{pd}} := Y - f_{\mathsf{pd}}^{(j)}(X)$ for an independent target-distribution pair $(X, Y)$. For the calibration observations, let*

$$e_{i,j}^{\mathsf{pd}} := Y_i^{\mathrm{cal}} - f_{\mathsf{pd}}^{(j)}(X_i^{\mathrm{cal}}),$$

*and define*

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

*Assume that there exists $d_0 > 0$ such that*

$$\min_{j\in[J]} D_j \ge d_0,$$

*and that the conditional variances of*

$$(e)^2, \qquad (e_j^{\mathsf{pd}})^2, \qquad e e_j^{\mathsf{pd}}, \qquad j \in [J],$$

*are bounded uniformly over $j \in [J]$. Then, as $n_{\mathrm{cal}} \to \infty$,*

$$\max_{j\in[J]} \left|\widehat{\xi}^{\star}_{\mathrm{cal},j} - \xi^{\star}_{j}\right| \xrightarrow{\mathrm{p}} 0, \qquad \max_{j\in[J]} \left|\widehat{R}^{\star}_{\mathsf{pm},\mathrm{cal},j} - R^{\star}_{\mathsf{pm},j}\right| \xrightarrow{\mathrm{p}} 0,$$

*where*

$$\xi^{\star}_{j} := \frac{R - C_j}{D_j}, \qquad R^{\star}_{\mathsf{pm},j} := R - \frac{(R - C_j)^2}{D_j},$$

*and $\widehat{\xi}^{\star}_{\mathrm{cal},j}$, $\widehat{R}^{\star}_{\mathsf{pm},\mathrm{cal},j}$ are the corresponding plug-in calibration estimates. In particular, Theorem 5.1 follows by taking $J = 1$.*

*Proof.* We first establish consistency of the three empirical quantities. For the teacher risk estimate, conditional on $\mathcal{G}$,

$$\mathbb{E}[\widehat{R}^{\mathrm{cal}} \mid \mathcal{G}] = R, \qquad \mathrm{Var}(\widehat{R}^{\mathrm{cal}} \mid \mathcal{G}) = \frac{1}{n_{\mathrm{cal}}} \mathrm{Var}((e)^2 \mid \mathcal{G}).$$

The assumed variance bound and Chebyshev's inequality give

$$\widehat{R}^{\mathrm{cal}} - R \xrightarrow{\mathrm{p}} 0.$$

The same argument, followed by a union bound over the fixed candidate set, gives

$$\max_{j\in[J]} \left|\widehat{R}^{\mathrm{cal}}_{\mathsf{pd},j} - R_{\mathsf{pd},j}\right| \xrightarrow{\mathrm{p}} 0, \qquad \max_{j\in[J]} \left|\widehat{C}^{\mathrm{cal}}_{j} - C_j\right| \xrightarrow{\mathrm{p}} 0.$$

Therefore,

$$\max_{j\in[J]} \left|\widehat{D}^{\mathrm{cal}}_{j} - D_j\right| \xrightarrow{\mathrm{p}} 0, \qquad \widehat{D}^{\mathrm{cal}}_{j} := \widehat{R}^{\mathrm{cal}} + \widehat{R}^{\mathrm{cal}}_{\mathsf{pd},j} - 2\widehat{C}^{\mathrm{cal}}_{j}.$$

Since $\min_j D_j \geq d_0$, we have

$$\min_{j\in[J]} \widehat{D}^{\mathrm{cal}}_{j} \geq d_0/2$$

with probability tending to one.

Thus, with probability tending to one, every plug-in denominator is bounded away from zero. Applying the continuous mapping theorem to each candidate and taking the maximum over the fixed finite set yields

$$\max_{j\in[J]} \left|\widehat{\xi}^{\star}_{\mathrm{cal},j} - \xi^{\star}_{j}\right| \xrightarrow{\mathrm{p}} 0, \qquad \max_{j\in[J]} \left|\widehat{R}^{\star}_{\mathsf{pm},\mathrm{cal},j} - R^{\star}_{\mathsf{pm},j}\right| \xrightarrow{\mathrm{p}} 0.$$

□

This finite-candidate consistency statement immediately implies that calibration can also be used for finite hyperparameter selection.

**Corollary E.2** (Selecting among finitely many candidates)**.** *Under the assumptions of Theorem E.1, let*

$$\widehat{j} \in \operatorname*{argmin}_{j\in[J]} \widehat{R}^{\star}_{\mathsf{pm},\mathrm{cal},j}.$$

*Then*

$$R^{\star}_{\mathsf{pm},\widehat{j}} - \min_{j\in[J]} R^{\star}_{\mathsf{pm},j} \xrightarrow{\mathrm{p}} 0.$$

*If the oracle minimizer $j^{\star} \in \operatorname{argmin}_j R^{\star}_{\mathsf{pm},j}$ is unique and separated by a positive gap, then*

$$\mathbb{P}(\widehat{j} = j^{\star}) \to 1.$$

*Proof.* Let $\Delta_{\mathrm{cal}} := \max_{j\in[J]} |\widehat{R}^{\star}_{\mathsf{pm},\mathrm{cal},j} - R^{\star}_{\mathsf{pm},j}|$. By Theorem E.1, $\Delta_{\mathrm{cal}} \xrightarrow{\mathrm{p}} 0$. By definition of $\widehat{j}$,

$$\widehat{R}^{\star}_{\mathsf{pm},\mathrm{cal},\widehat{j}} \leq \widehat{R}^{\star}_{\mathsf{pm},\mathrm{cal},j} \qquad \text{for every } j \in [J].$$

Let $j^{\star} \in \operatorname{argmin}_j R^{\star}_{\mathsf{pm},j}$. Then

$$R^{\star}_{\mathsf{pm},\widehat{j}} \leq \widehat{R}^{\star}_{\mathsf{pm},\mathrm{cal},\widehat{j}} + \Delta_{\mathrm{cal}} \leq \widehat{R}^{\star}_{\mathsf{pm},\mathrm{cal},j^{\star}} + \Delta_{\mathrm{cal}} \leq R^{\star}_{\mathsf{pm},j^{\star}} + 2\Delta_{\mathrm{cal}}.$$

Thus

$$0 \le R^\star_{\mathsf{pm},\widehat{j}} - \min_{j\in[J]} R^\star_{\mathsf{pm},j} \le 2\Delta_{\mathrm{cal}} \xrightarrow{\mathrm{p}} 0.$$

If the minimizer is unique, let

$$g := \min_{j\neq j^\star}\big(R^\star_{\mathsf{pm},j} - R^\star_{\mathsf{pm},j^\star}\big) > 0$$

denote its separation gap. The event $\Delta_{\mathrm{cal}} < g/2$ then forces $\widehat{j} = j^\star$. This event has probability tending to one. □

### E.2 Vector-valued predictors

The same calibration method extends directly to vector-valued prediction, which is used in our squared-loss linear-probing experiments. Suppose $Y \in \mathbb{R}^K$ and predictors $f : \mathcal{X} \to \mathbb{R}^K$ are evaluated by squared risk

$$R(f) := \mathbb{E}\big[\|Y - f(X)\|_2^2\big].$$

For two fixed predictors $f$ and $f_{\mathsf{pd}}$, define vector residuals

$$e := Y - f(X), \qquad e^{\mathsf{pd}} := Y - f_{\mathsf{pd}}(X),$$

and set

$$R := \mathbb{E}[\|e\|_2^2], \qquad R_{\mathsf{pd}} := \mathbb{E}[\|e^{\mathsf{pd}}\|_2^2], \qquad C := \mathbb{E}[\langle e, e^{\mathsf{pd}}\rangle].$$

Then the same affine-risk expansion gives

$$\xi^\star = \frac{R - C}{R + R_{\mathsf{pd}} - 2C}, \qquad R^\star_{\mathsf{pm}} = R - \frac{(R-C)^2}{R + R_{\mathsf{pd}} - 2C}.$$

The calibration estimates are obtained by replacing scalar squares and products with squared Euclidean norms and inner products:

$$\widehat{R}^{\mathrm{cal}} = \frac{1}{n_{\mathrm{cal}}}\sum_{i=1}^{n_{\mathrm{cal}}} \|e_i\|_2^2, \quad \widehat{R}^{\mathrm{cal}}_{\mathsf{pd}} = \frac{1}{n_{\mathrm{cal}}}\sum_{i=1}^{n_{\mathrm{cal}}} \|e_i^{\mathsf{pd}}\|_2^2, \quad \widehat{C}^{\mathrm{cal}} = \frac{1}{n_{\mathrm{cal}}}\sum_{i=1}^{n_{\mathrm{cal}}} \langle e_i, e_i^{\mathsf{pd}}\rangle.$$

The proof of Theorem E.1 applies directly under the corresponding finite-variance assumptions on

$$\|e\|_2^2, \qquad \|e^{\mathsf{pd}}\|_2^2, \qquad \langle e, e^{\mathsf{pd}}\rangle.$$

A note on the computational cost: The calibration estimator is "one-shot" in the following sense. For a fixed candidate student $f_{\mathsf{pd}}$, after training $f_{\mathsf{pd}}$ once on fresh pseudo-labels, tuning $\xi$ requires only

$$\big(f(x_i^{\mathrm{cal}}), f_{\mathsf{pd}}(x_i^{\mathrm{cal}}), y_i^{\mathrm{cal}}\big)_{i=1}^{n_{\mathrm{cal}}},$$

followed by three empirical averages and the closed-form plug-in formula. There is no retraining over candidate mixing weights $\xi$. If one considers finitely many candidate students, such as different student penalties $\lambda_s$ or different fresh unlabeled sample sizes $n_s$, each student must be trained once, but the same calibration set can evaluate all candidates by Corollary E.2.

## F Proofs in Section 6

### F.1 Preliminaries

For readability, we write $\rho$ in the finite sample systems below. When $n\rho$ is not an integer, this notation denotes $\rho_n := \lfloor n\rho\rfloor/n$; since $\rho_n \to \rho$, this does not affect the limiting conclusions.

Recall the training objective for the student classifier based on the teacher's hard labels:

$$\theta_{\mathsf{pd}} := \operatorname*{argmin}_\theta \underbrace{\frac{1}{2n}\sum_{i=1}^{2n} \mathrm{CE}(\widehat{y}_i, \sigma(\langle\theta, \phi(x_i)\rangle)) + \frac{\lambda}{2}\|\theta\|^2}_{\mathcal{L}_1(\theta)}. \tag{61}$$

The loss can be rewritten as

$$\mathcal{L}_1(\theta) = \frac{1}{2n}\sum_{i=1}^{2n}\Big(\log\left(1+e^{\langle\theta,\phi(x_i)\rangle}\right) - \widehat{y}_i\langle\theta,\phi(x_i)\rangle\Big) + \frac{\lambda}{2}\|\theta\|^2, \tag{62}$$

$$\nabla\mathcal{L}_1(\theta) = \frac{1}{2n}\sum_{i=1}^{2n}\Big(\sigma(\langle\theta,\phi(x_i)\rangle) - \widehat{y}_i\Big)\phi(x_i) + \lambda\theta. \tag{63}$$

Setting $\nabla\mathcal{L}_1(\theta) = 0$ gives

$$\theta_{\mathsf{pd}} = \sum_{i=1}^{2n}\underbrace{\frac{1}{2n\lambda}\Big(\widehat{y}_i - \sigma(\langle\theta_{\mathsf{pd}},\phi(x_i)\rangle)\Big)}_{:=\alpha(x_i,\widehat{y}_i)}\phi(x_i) = \sum_{i=1}^{2n}\alpha(x_i,\widehat{y}_i)\phi(x_i). \tag{64}$$

It follows that

$$y_{\mathsf{pd}}(x_i) = \sigma(\langle\theta_{\mathsf{pd}},\phi(x_i)\rangle) = \widehat{y}_i - 2n\lambda\alpha(x_i,\widehat{y}_i). \tag{65}$$

The following result characterizes these coefficients.

**Lemma F.1** (Adapted from Lemma 2 of [17])**.** *Under Assumption 6.1, let $\bar{\lambda}_n := 2n\lambda$. Then*

$$\alpha(x_i,\widehat{y}_i) = \begin{cases} -\widehat{\alpha}_n & \textit{for } i\in\mathcal{S}_{1,\mathrm{bad}},\\ \alpha_n & \textit{for } i\in\mathcal{S}_{1,\mathrm{good}},\\ \widehat{\alpha}_n & \textit{for } i\in\mathcal{S}_{0,\mathrm{bad}},\\ -\alpha_n & \textit{for } i\in\mathcal{S}_{0,\mathrm{good}}. \end{cases}.$$

*Here, $\alpha_n \geq 0$ and $\widehat{\alpha}_n \geq 0$ solve*

$$\begin{cases} \sigma\Big(cn(\alpha_n - (\alpha_n+\widehat{\alpha}_n)\rho) - (1-c)\widehat{\alpha}_n\Big) = \bar{\lambda}_n\widehat{\alpha}_n\\ \sigma\Big(cn(\alpha_n - (\alpha_n+\widehat{\alpha}_n)\rho) + (1-c)\alpha_n\Big) = 1-\bar{\lambda}_n\alpha_n \end{cases}. \tag{66}$$

By Lemma F.1, the student predictions on the training set are

$$y_{\mathsf{pd}}(x_i) = \begin{cases} \bar{\lambda}_n\widehat{\alpha}_n & \text{for } i\in\mathcal{S}_{1,\mathrm{bad}},\\ 1-\bar{\lambda}_n\alpha_n & \text{for } i\in\mathcal{S}_{1,\mathrm{good}},\\ 1-\bar{\lambda}_n\widehat{\alpha}_n & \text{for } i\in\mathcal{S}_{0,\mathrm{bad}},\\ \bar{\lambda}_n\alpha_n & \text{for } i\in\mathcal{S}_{0,\mathrm{good}}. \end{cases} \tag{67}$$

For $y_{\mathsf{pd}}$ to achieve $100\%$ training accuracy with respect to the ground-truth labels, we require $\bar{\lambda}_n\widehat{\alpha}_n > 0.5$ and $\bar{\lambda}_n\alpha_n < 0.5$.

In Section 6.1, we train a second student on the first student's soft labels $y_{\mathsf{pd}}(\cdot)$ over the same unlabeled set $\mathcal{D}_{\mathrm{unlab}}$. Its training objective is

$$\theta^{(2)}_{\mathsf{pd}} := \operatorname*{argmin}_{\theta}\underbrace{\frac{1}{2n}\sum_{i=1}^{2n}\mathrm{CE}(y_{\mathsf{pd}}(x_i),\sigma(\langle\theta,\phi(x_i)\rangle)) + \frac{\lambda}{2}\|\theta\|^2}_{\mathcal{L}_2(\theta)}. \tag{68}$$

Setting $\nabla\mathcal{L}_2(\theta) = 0$ gives

$$\theta^{(2)}_{\mathsf{pd}} = \sum_{i=1}^{2n}\underbrace{\frac{1}{2n\lambda}\Big(y_{\mathsf{pd}}(x_i) - \sigma(\langle\theta^{(2)}_{\mathsf{pd}},\phi(x_i)\rangle)\Big)}_{:=\beta_i}\phi(x_i) = \sum_{i=1}^{2n}\beta_i\phi(x_i). \tag{69}$$

The following lemma characterizes the coefficients $\beta_i$.

**Lemma F.2** (Adapted from Lemma 3 of [17]). *Under Assumption 6.1, let $\bar{\lambda}_n := 2n\lambda$. Then*

$$\beta_i = \begin{cases} -\widehat{\beta}_n & \text{for } i \in \mathcal{S}_{1,\mathrm{bad}}, \\ \beta_n & \text{for } i \in \mathcal{S}_{1,\mathrm{good}}, \\ \widehat{\beta}_n & \text{for } i \in \mathcal{S}_{0,\mathrm{bad}}, \\ -\beta_n & \text{for } i \in \mathcal{S}_{0,\mathrm{good}} \end{cases}. \tag{70}$$

*Here, $\beta_n$ and $\widehat{\beta}_n$ solve*

$$\begin{cases} \sigma\big(cn(\beta_n - (\beta_n + \widehat{\beta}_n)\rho) - (1-c)\widehat{\beta}_n\big) = \bar{\lambda}_n\widehat{\alpha}_n + \bar{\lambda}_n\widehat{\beta}_n \\ \sigma\big(cn(\beta_n - (\beta_n + \widehat{\beta}_n)\rho) + (1-c)\beta_n\big) = 1 - \bar{\lambda}_n\alpha_n - \bar{\lambda}_n\beta_n, \end{cases} \tag{71}$$

*where $(\alpha_n, \widehat{\alpha}_n)$ solve* (66)*. If $\bar{\lambda}_n\widehat{\alpha}_n < 0.5$ and $\bar{\lambda}_n\alpha_n < 0.5$, then $\beta_n, \widehat{\beta}_n \geq 0$.*

By Lemma F.2, the second student's predictions on the training set are

$$y^{(2)}_{\mathsf{pd}}(x_i) = \begin{cases} \bar{\lambda}_n\widehat{\alpha}_n + \bar{\lambda}_n\widehat{\beta}_n & \text{for } i \in \mathcal{S}_{1,\mathrm{bad}}, \\ 1 - \bar{\lambda}_n\alpha_n - \bar{\lambda}_n\beta_n & \text{for } i \in \mathcal{S}_{1,\mathrm{good}}, \\ 1 - \bar{\lambda}_n\widehat{\alpha}_n - \bar{\lambda}_n\widehat{\beta}_n & \text{for } i \in \mathcal{S}_{0,\mathrm{bad}}, \\ \bar{\lambda}_n\alpha_n + \bar{\lambda}_n\beta_n & \text{for } i \in \mathcal{S}_{0,\mathrm{good}}. \end{cases} \tag{72}$$

For $y^{(2)}_{\mathsf{pd}}$ to achieve $100\%$ training accuracy, we require $\bar{\lambda}_n(\widehat{\alpha}_n + \widehat{\beta}_n) > 0.5$ and $\bar{\lambda}_n(\alpha_n + \beta_n) < 0.5$.

To distinguish the affine scores from the thresholded classifiers in (15) and (17), define

$$s_{\mathsf{pm},\xi}(x) := (1-\xi)\widehat{y}(x) + \xi y_{\mathsf{pd}}(x), \qquad s^{(2)}_{\mathsf{pm},\xi}(x) := (1-\xi)y_{\mathsf{pd}}(x) + \xi y^{(2)}_{\mathsf{pd}}(x).$$

### F.2 Proof of Theorem 6.2

*Proof of Theorem 6.2.* We first show that when $\bar{\lambda}_n \to \bar{\lambda} \in (0, \infty)$, there exists $\rho_0 \in (0, 0.5)$ such that $y_{\mathsf{pd}}$ has $100(1-\rho)\%$ population accuracy for every $\rho \in (\rho_0, 0.5)$, matching the teacher.

By Lemma F.1,

$$\begin{cases} \sigma\Big(cn(\alpha_n - (\alpha_n + \widehat{\alpha}_n)\rho) - (1-c)\widehat{\alpha}_n\Big) = \bar{\lambda}_n\widehat{\alpha}_n \\ \sigma\Big(cn(\alpha_n - (\alpha_n + \widehat{\alpha}_n)\rho) + (1-c)\alpha_n\Big) = 1 - \bar{\lambda}_n\alpha_n \end{cases} \tag{73}$$

The regularized logistic objective is strictly convex, and the feature spans for the two classes are orthogonal, so this system has a unique solution $(\alpha_n, \widehat{\alpha}_n)$. Because the sigmoid takes values in $(0,1)$, both $\alpha_n$ and $\widehat{\alpha}_n$ are strictly positive.

Let $T_n := \alpha_n - (\alpha_n + \widehat{\alpha}_n)\rho$. Suppose, for contradiction, that $\widehat{\alpha}_n \leq \alpha_n$. Using $\sigma(z) = 1 - \sigma(-z)$ and the monotonicity of the sigmoid,

$$\begin{aligned} &\sigma(cnT_n - (1-c)\widehat{\alpha}_n) = \bar{\lambda}_n\widehat{\alpha}_n \leq \bar{\lambda}_n\alpha_n = \sigma(-cnT_n - (1-c)\alpha_n) \\ \Rightarrow\quad &cnT_n - (1-c)\widehat{\alpha}_n \leq -cnT_n - (1-c)\alpha_n \\ \Rightarrow\quad &2cnT_n \leq (1-c)(\widehat{\alpha}_n - \alpha_n) \leq 0. \end{aligned} \tag{74}$$

This is impossible because $\rho < 0.5$ and $\widehat{\alpha}_n \leq \alpha_n$ imply $T_n > 0$. Hence, $\widehat{\alpha}_n > \alpha_n$. Moreover,

$$\bar{\lambda}_n(\widehat{\alpha}_n + \alpha_n) = \sigma(cnT_n - (1-c)\widehat{\alpha}_n) + 1 - \sigma(cnT_n + (1-c)\alpha_n) \leq 1,$$

because $\alpha_n, \widehat{\alpha}_n \geq 0$.

Since $\bar{\lambda}_n \to \bar{\lambda} > 0$ and $\bar{\lambda}_n\alpha_n, \bar{\lambda}_n\widehat{\alpha}_n \in (0,1)$, the sequence $(\alpha_n, \widehat{\alpha}_n)$ is bounded. Fix $\rho \in (0, 0.5)$ and let

$$r := \frac{\rho}{1-\rho} \in (0,1).$$

Pass to an arbitrary convergent subsequence, still indexed by $n$, with limit $(\alpha, \widehat{\alpha})$. Then $0 \leq \alpha \leq \widehat{\alpha}$, $\bar{\lambda}(\alpha + \widehat{\alpha}) \leq 1$, and

$$T := \lim_n T_n = \alpha - (\alpha + \widehat{\alpha})\rho.$$

We claim that $T = 0$.

- If $T > 0$, both sigmoid arguments in (73) diverge to $+\infty$. Hence, $\bar{\lambda}\widehat{\alpha} = 1 - \bar{\lambda}\alpha = 1$, so $\widehat{\alpha} > 0$ and $\alpha = 0$, contradicting $T > 0$.
- If $T < 0$, both arguments diverge to $-\infty$. Hence, $\bar{\lambda}\widehat{\alpha} = 1 - \bar{\lambda}\alpha = 0$, so $\widehat{\alpha} = 0$ and $\alpha = 1/\bar{\lambda} > 0$, contradicting $0 \le \alpha \le \widehat{\alpha}$.

Therefore, $T = 0$, and every subsequential limit satisfies

$$\alpha = r\widehat{\alpha}. \tag{75}$$

The sequence $nT_n$ is also bounded. Otherwise, pass to a subsequence along which it diverges to $+\infty$ or $-\infty$ and then to a further subsequence along which $(\alpha_n, \widehat{\alpha}_n)$ converges. In the first case, (73) gives $\bar{\lambda}\widehat{\alpha} = 1$ and $\bar{\lambda}\alpha = 0$; in the second, it gives $\bar{\lambda}\widehat{\alpha} = 0$ and $\bar{\lambda}\alpha = 1$. Both conclusions contradict (75) because $r > 0$.

Consequently, every subsequence of $(\alpha_n, \widehat{\alpha}_n, nT_n)$ has a further convergent subsequence. Let $(\alpha, \widehat{\alpha}, N)$ be any such subsequential limit. Taking limits in (73) gives

$$\begin{cases} \sigma\Big(cN - (1-c)\widehat{\alpha}\Big) = \bar{\lambda}\widehat{\alpha} \\ \sigma\Big(cN + (1-c)\alpha\Big) = 1 - \bar{\lambda}\alpha \end{cases} \quad \Leftrightarrow \quad \begin{cases} \sigma\Big(cN - (1-c)\widehat{\alpha}\Big) = \bar{\lambda}\widehat{\alpha} \\ \sigma\Big(-cN - (1-c)\alpha\Big) = \bar{\lambda}\alpha. \end{cases} \tag{76}$$

We now show that this subsequential limit is unique. Let $\mathrm{logit}(u) := \log\{u/(1-u)\}$. The first equation in (76) implies $0 < \bar{\lambda}\widehat{\alpha} < 1$, so the logit below is well defined. The first equation in (76) gives

$$cN = \mathrm{logit}(\bar{\lambda}\widehat{\alpha}) + (1-c)\widehat{\alpha}.$$

Substituting this identity and $\alpha = r\widehat{\alpha}$ into the second equation gives

$$g_r(\widehat{\alpha}) := \sigma\big(-\mathrm{logit}(\bar{\lambda}\widehat{\alpha}) - (1-c)(1+r)\widehat{\alpha}\big) - \bar{\lambda}r\widehat{\alpha} = 0. \tag{77}$$

On $h \in (0, 1/\bar{\lambda})$, the function $g_r(h)$ is strictly decreasing, with

$$\lim_{h \downarrow 0} g_r(h) = 1, \qquad \lim_{h \uparrow 1/\bar{\lambda}} g_r(h) = -r.$$

Thus, (77) has a unique solution, denoted $h_r$. It follows that every subsequential limit is the same:

$$\widehat{\alpha} = h_r, \qquad \alpha = rh_r, \qquad N = N_r := \frac{\mathrm{logit}(\bar{\lambda}h_r) + (1-c)h_r}{c}.$$

Therefore, the full sequence $(\alpha_n, \widehat{\alpha}_n, nT_n)$ converges to $(rh_r, h_r, N_r)$.

It remains to study the limit as $\rho \uparrow 0.5$, or equivalently as $r \uparrow 1$. The unique zero $h_r$ of $g_r$ depends continuously on $r \in (0, 1]$, because $g_r(h)$ is continuous in $(r, h)$ and strictly decreasing in $h$. Indeed, if $r_k \to r$, the endpoint limits above ensure that every subsequential limit of $h_{r_k}$ lies in $(0, 1/\bar{\lambda})$ and solves $g_r(h) = 0$; uniqueness then forces that limit to equal $h_r$. At $r = 1$, the continuous extension of the limiting equations in (76) becomes

$$\sigma(cN_1 - (1-c)h_1) = \bar{\lambda}h_1, \qquad \sigma(-cN_1 - (1-c)h_1) = \bar{\lambda}h_1.$$

The strict monotonicity of the sigmoid implies $N_1 = 0$, and the first equation then gives

$$0 < \bar{\lambda}h_1 = \sigma(-(1-c)h_1) < 0.5.$$

Consequently,

$$N_r \to 0, \qquad h_r \to h_1 > 0 \qquad \text{as } \rho \uparrow 0.5.$$

Hence, there exists $\rho_0 \in (0, 0.5)$ such that, for every $\rho \in (\rho_0, 0.5)$,

$$cN_r - (1-c)h_r < 0, \qquad \bar{\lambda}h_r < 0.5.$$

Since $\alpha = rh_r < h_r$, we also have $\bar{\lambda}\alpha < 0.5$. Equation (67) therefore shows that $y_{\mathsf{pd}}$ misclassifies exactly the samples in $\mathcal{S}_{1,\mathrm{bad}}$ and $\mathcal{S}_{0,\mathrm{bad}}$, and hence has $100(1-\rho)\%$ population accuracy.

**Mixing achieves $100\%$ accuracy for all $\rho < 0.5$.** We now prove the second statement of the theorem. Under the assumed regime $\bar{\lambda}_n \to \bar{\lambda} \in (0, \infty)$, we can choose $\xi$ so that thresholding the affine score $(1-\xi)\widehat{y}(x) + \xi y_{\mathsf{pd}}(x)$ yields $100\%$ population accuracy as $n \to \infty$. The limiting affine scores are

$$s_{\mathsf{pm},\xi}(x_i) = (1-\xi)\widehat{y}_i + \xi y_{\mathsf{pd}}(x_i) = \begin{cases} \xi\bar{\lambda}\widehat{\alpha} & \text{for } i \in \mathcal{S}_{1,\mathrm{bad}}, \\ 1 - \xi\bar{\lambda}\alpha & \text{for } i \in \mathcal{S}_{1,\mathrm{good}}, \\ 1 - \xi\bar{\lambda}\widehat{\alpha} & \text{for } i \in \mathcal{S}_{0,\mathrm{bad}}, \\ \xi\bar{\lambda}\alpha & \text{for } i \in \mathcal{S}_{0,\mathrm{good}}. \end{cases}$$

Since $\widehat{\alpha} > \alpha$, there are three algebraic cases:

- If $\bar{\lambda}\widehat{\alpha} > 0.5 > \bar{\lambda}\alpha$, then (67) shows that the PD student already achieves $100\%$ accuracy as $n \to \infty$, so we choose $\xi = 1$.
- If $0.5 \geq \bar{\lambda}\widehat{\alpha} > \bar{\lambda}\alpha$, choose $\xi > 1$ such that $\xi\bar{\lambda}\widehat{\alpha} > 0.5$ and $\xi\bar{\lambda}\alpha < 0.5$. Such a choice exists because $\widehat{\alpha} > \alpha$.
- The case $\bar{\lambda}\widehat{\alpha} > \bar{\lambda}\alpha \geq 0.5$ is impossible because $\bar{\lambda}(\widehat{\alpha} + \alpha) \leq 1$.

□

### F.3 Proof of Theorem 6.3

*Proof of Theorem 6.3.* As established in the proof of Theorem 6.2,

$$\alpha = r\widehat{\alpha}, \qquad r := \frac{\rho}{1-\rho} \in (0,1), \qquad \bar{\lambda}(\widehat{\alpha} + \alpha) \leq 1.$$

In particular, $\widehat{\alpha} > \alpha > 0$. Write

$$P := \bar{\lambda}\widehat{\alpha}, \qquad Q := \bar{\lambda}\alpha = rP.$$

If $P > 0.5$, then $Q < 0.5$ because $P + Q \leq 1$. Thus, $y_{\mathsf{pd}}$ already achieves $100\%$ accuracy, and choosing $\xi = 0$ recovers the classifier induced by $y_{\mathsf{pd}}$. It remains to consider

$$0 < Q < P \leq 0.5. \tag{78}$$

This includes the boundary case $P = 0.5$.

The equations in (71) imply

$$0 < \bar{\lambda}_n(\widehat{\alpha}_n + \widehat{\beta}_n) < 1, \qquad 0 < \bar{\lambda}_n(\alpha_n + \beta_n) < 1.$$

Hence, $(\beta_n, \widehat{\beta}_n)$ is bounded. Let $(\beta, \widehat{\beta})$ be an arbitrary subsequential limit, and define

$$W_n := \beta_n - (\beta_n + \widehat{\beta}_n)\rho, \qquad W := \lim_n W_n = (1-\rho)\beta - \rho\widehat{\beta}.$$

We claim that $W = 0$. If $W > 0$, both sigmoid arguments in (71) diverge to $+\infty$, giving

$$P + \bar{\lambda}\widehat{\beta} = 1, \qquad Q + \bar{\lambda}\beta = 0.$$

Thus, $\widehat{\beta} > 0$ and $\beta < 0$, contradicting $W > 0$. If $W < 0$, both arguments diverge to $-\infty$, giving

$$P + \bar{\lambda}\widehat{\beta} = 0, \qquad Q + \bar{\lambda}\beta = 1.$$

Thus, $\widehat{\beta} < 0$ and $\beta > 0$, contradicting $W < 0$. Therefore, every subsequential limit satisfies

$$\beta = r\widehat{\beta}. \tag{79}$$

The sequence $nW_n$ is bounded as well. Otherwise, pass to a subsequence along which it diverges to $+\infty$ or $-\infty$ and then to a further subsequence along which $(\beta_n, \widehat{\beta}_n)$ converges. The limiting equations would again force $\beta$ and $\widehat{\beta}$ to have opposite signs, contradicting (79). Consequently, every subsequence of $(\beta_n, \widehat{\beta}_n, nW_n)$ has a further convergent subsequence.

Let $(\beta, \widehat{\beta}, M)$ be any such subsequential limit. Using (75) and (79), the limiting equations are

$$\begin{cases} \sigma(cM - (1-c)\widehat{\beta}) = \bar{\lambda}(\widehat{\alpha} + \widehat{\beta}), \\ \sigma(-cM - (1-c)r\widehat{\beta}) = \bar{\lambda}r(\widehat{\alpha} + \widehat{\beta}). \end{cases} \tag{80}$$

In particular, $0 < \bar{\lambda}(\widehat{\alpha} + \widehat{\beta}) < 1$, so the logit below is well defined. The first equation implies

$$cM = \operatorname{logit}\big(\bar{\lambda}(\widehat{\alpha} + \widehat{\beta})\big) + (1-c)\widehat{\beta}.$$

For $h \in (-\widehat{\alpha}, 1/\bar{\lambda} - \widehat{\alpha})$, define

$$k_r(h) := \sigma\big(-\operatorname{logit}\big(\bar{\lambda}(\widehat{\alpha} + h)\big) - (1-c)(1+r)h\big) - \bar{\lambda}r(\widehat{\alpha} + h). \tag{81}$$

Substituting the expression for $M$ into the second equation in (80) shows that $k_r(\widehat{\beta}) = 0$. The function $k_r$ in (81) is strictly decreasing, with

$$\lim_{h \downarrow -\widehat{\alpha}} k_r(h) = 1, \qquad \lim_{h \uparrow 1/\bar{\lambda} - \widehat{\alpha}} k_r(h) = -r.$$

Thus, $k_r(h) = 0$ has a unique solution, denoted $b_r$. Moreover,

$$k_r(0) = 1 - P - Q > 0,$$

where the strict inequality follows from (78) and $Q = rP < P$. Therefore,

$$b_r > 0, \qquad \widehat{\beta} = b_r, \qquad \beta = rb_r < b_r.$$

The value of $M$ is then uniquely determined by the first equation in (80). Hence, the full sequence $(\beta_n, \widehat{\beta}_n, nW_n)$ converges to this unique limit.

The limiting affine scores are

$$s^{(2)}_{\mathsf{pm},\xi}(x_i) = \begin{cases} P + \xi\bar{\lambda}b_r & \text{for } i \in \mathcal{S}_{1,\mathrm{bad}}, \\ 1 - Q - \xi\bar{\lambda}rb_r & \text{for } i \in \mathcal{S}_{1,\mathrm{good}}, \\ 1 - P - \xi\bar{\lambda}b_r & \text{for } i \in \mathcal{S}_{0,\mathrm{bad}}, \\ Q + \xi\bar{\lambda}rb_r & \text{for } i \in \mathcal{S}_{0,\mathrm{good}}. \end{cases}$$

Choose $\xi$ such that

$$\frac{0.5 - P}{\bar{\lambda}b_r} < \xi < \frac{0.5 - Q}{\bar{\lambda}rb_r}.$$

This interval is nonempty because $P > Q$ and $b_r > rb_r > 0$; its lower endpoint is zero when $P = 0.5$. This choice makes the first two scores exceed $0.5$ and the last two fall below $0.5$, proving the claim. □

# G Additional theoretical results

## G.1 Non-monotonicity in the fresh unlabeled sample size

In this subsection, we consider the isotropic and same-$\lambda$ setting, i.e., $\lambda_t = \lambda_s = \lambda$. We demonstrate that the optimal fresh-$X$ PMSD risk is *not* globally monotone in $\gamma_s$. Instead, it has a simple one-dimensional structure and is unimodal in $\gamma_s$.

Assume $\Sigma_t = \Sigma_s = I_p$ and $\beta \sim \mathcal{N}(0, (r^2/p)I_p)$, where $r^2 > 0$, and let

$$\lambda^\star := \gamma_t \frac{\sigma^2}{r^2}.$$

For the labeled sample, let $\kappa_t = \kappa_t(\lambda)$ be the unique positive solution to

$$\kappa_t = \lambda + \frac{\gamma_t \kappa_t}{1 + \kappa_t},$$

and set

$$b_t := \Big(1 - \frac{\gamma_t}{(1+\kappa_t)^2}\Big)^{-1}.$$

For the fresh unlabeled sample, let $\kappa_s = \kappa_s(\lambda)$ be the unique positive solution to

$$\kappa_s = \lambda + \frac{\gamma_s \kappa_s}{1+\kappa_s},$$

and define

$$c_s := \Big(1 - \frac{\gamma_s}{(1+\kappa_s)^2}\Big)^{-1}.$$

Let $\overline{\mathcal{R}}(\lambda) := \mathcal{R}(\lambda) - \sigma^2$ denote the excess teacher risk. Define the excess PD-student risk $\overline{\mathcal{R}}_{\mathsf{pd}}$ and the excess correlation $\bar{\mathcal{C}}$ similarly.

In the isotropic model, the limiting teacher risk in (42) simplifies to

$$\overline{\mathcal{R}}(\lambda) = \frac{b_t}{(1+\kappa_t)^2}\big(r^2\kappa_t^2 + \sigma^2\gamma_t\big). \tag{82}$$

Define also

$$\Theta_t(\lambda) := \overline{\mathcal{R}}(\lambda) - r^2\frac{\kappa_t}{1+\kappa_t}, \qquad T_t(\lambda) := \overline{\mathcal{R}}(\lambda) + r^2\frac{1-\kappa_t}{1+\kappa_t}. \tag{83}$$

We are now ready to state the key result in this subsection.

**Proposition G.1** (Dependence of optimal PMSD risk on $\gamma_s$)**.** *In the isotropic setting, for each fixed $\lambda > 0$,*

$$\overline{\mathcal{R}}^\star_{\mathsf{pm}}(\lambda;\gamma_s) := \mathcal{R}^\star_{\mathsf{pm}}(\lambda;\gamma_s) - \sigma^2 = \overline{\mathcal{R}}(\lambda) - \frac{\Theta_t(\lambda)^2}{T_t(\lambda)}\,\frac{1}{c_s(\lambda)}. \tag{84}$$

*Equivalently,*

$$\overline{\mathcal{R}}^\star_{\mathsf{pm}}(\lambda;\gamma_s) = \overline{\mathcal{R}}(\lambda) - K_t(\lambda)\,\frac{1}{c_s(\lambda)}, \qquad K_t(\lambda) := \frac{\Theta_t(\lambda)^2}{T_t(\lambda)} \geq 0. \tag{85}$$

*Moreover:*

*(a) if $\lambda = \lambda^\star$, then $K_t(\lambda) = 0$ and*

$$\mathcal{R}^\star_{\mathsf{pm}}(\lambda;\gamma_s) \equiv \mathcal{R}(\lambda) \qquad \textit{for all } \gamma_s > 0;$$

*(b) if $\lambda \neq \lambda^\star$, then*

$$\frac{\partial}{\partial\gamma_s}\overline{\mathcal{R}}^\star_{\mathsf{pm}}(\lambda;\gamma_s) = -K_t(\lambda)\,\frac{4(\gamma_s-\lambda-1)}{\big(\gamma_s+\lambda+1+\sqrt{(\gamma_s+\lambda-1)^2+4\lambda}\big)^2\sqrt{(\gamma_s+\lambda-1)^2+4\lambda}}. \tag{86}$$

*Hence $\gamma_s \mapsto \overline{\mathcal{R}}^\star_{\mathsf{pm}}(\lambda;\gamma_s)$ is strictly increasing on $(0, 1+\lambda)$, strictly decreasing on $(1+\lambda, \infty)$, and has a unique maximum at*

$$\gamma_s = 1 + \lambda.$$

*In particular, the optimal PMSD risk is* not *globally monotone in the amount of unlabeled data.*

Since smaller $\gamma_s = p/n_s$ corresponds to more unlabeled data, Proposition G.1 shows that adding more fresh unlabeled data is *not* uniformly beneficial. For fixed $\lambda \neq \lambda^\star$:

- if $\gamma_s < 1 + \lambda$, then decreasing $\gamma_s$ (adding more unlabeled data) decreases the optimal PMSD risk;
- if $\gamma_s > 1 + \lambda$, then decreasing $\gamma_s$ initially *increases* the optimal PMSD risk, until the turning point $\gamma_s = 1 + \lambda$ is reached.

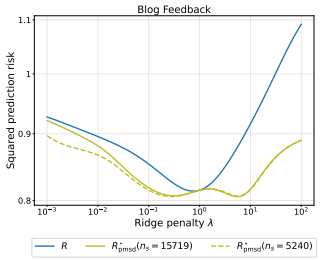
Thus, the fresh-$X$ PMSD risk is unimodal, not monotone, in the amount of unlabeled data.

*Proof of Proposition G.1.* Let

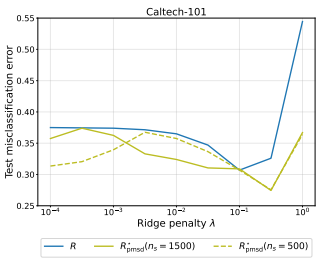
$$\Delta_\lambda := \beta_\lambda - \beta, \qquad \Delta_{\mathsf{pd},\lambda} := \beta_{\mathsf{pd},\lambda} - \beta.$$

All limits below are taken in the proportional asymptotic regime with $\gamma_t$ and $\gamma_s$ fixed. In the isotropic in-distribution setting,

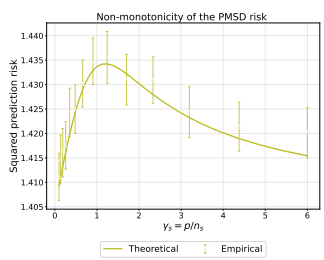
$$\overline{\mathcal{R}}(\lambda) = \lim \mathbb{E}\|\Delta_\lambda\|_2^2, \qquad \bar{\mathcal{C}}(\lambda) = \lim \mathbb{E}\langle\Delta_\lambda, \Delta_{\mathsf{pd},\lambda}\rangle, \qquad \overline{\mathcal{R}}_{\mathsf{pd}}(\lambda) = \lim \mathbb{E}\|\Delta_{\mathsf{pd},\lambda}\|_2^2.$$




(a) Ridge regression on the Blog Feedback dataset with $p = 280$, $n_t = 2619$.




(b) Linear probing on Caltech-101 using pretrained ResNet-34 features with $p = 512$, $n_t = 1500$, $n_{\text{cal}} = 500$, and corruption rate $\rho = 0.4$; see Section 6 for definitions.




(c) Same-$\lambda$ isotropic design and isotropic signal. Setting: $\lambda = 0.2$, $p = 400$, $n_t = 800$, and $r^2 = \sigma^2 = 1$.

Figure 4: Non-monotonicity of the PMSD risk with the amount of fresh unlabeled data in the same-$\lambda$, in-distribution setting.

Let $M_u := \widetilde{\Sigma}(\widetilde{\Sigma} + \lambda I_p)^{-1}$. Then

$$\beta_{\mathsf{pd},\lambda} = M_u \beta_\lambda, \qquad \Delta_{\mathsf{pd},\lambda} = M_u \Delta_\lambda + (M_u - I_p)\beta.$$

Because $\widetilde{\Sigma}$ is independent of the labeled sample, $M_u$ is independent of $(\Delta_\lambda, \beta)$. Under isotropy, $M_u$ is orthogonally invariant, and therefore

$$\Delta_\lambda^\top M_u \Delta_\lambda \to s_u \, \|\Delta_\lambda\|_2^2, \qquad \Delta_\lambda^\top M_u^2 \Delta_\lambda \to s_{2,u} \, \|\Delta_\lambda\|_2^2,$$
$$\Delta_\lambda^\top M_u \beta \to s_u \, \langle \Delta_\lambda, \beta \rangle, \qquad \beta^\top (M_u - I_p)^2 \beta \to (1 - 2s_u + s_{2,u}) \, \|\beta\|_2^2,$$

where

$$s_u = \lim \frac{1}{p} \operatorname{tr}(M_u) = \frac{1}{1+\kappa_s}, \quad s_{2,u} = \lim \frac{1}{p} \operatorname{tr}(M_u^2),$$

and

$$1 - 2s_u + s_{2,u} = \lim \frac{1}{p} \operatorname{tr}\big((I_p - M_u)^2\big) = \frac{\kappa_s^2 c_s}{(1+\kappa_s)^2}.$$

Similarly, on the labeled side,

$$s_t = \frac{1}{1+\kappa_t}, \qquad \langle \Delta_\lambda, \beta \rangle \to -r^2(1 - s_t) = -r^2 \frac{\kappa_t}{1+\kappa_t}.$$

Consequently,

$$T_t(\lambda) = \lim \mathbb{E}\|\beta_\lambda\|_2^2 > 0.$$

Expanding the fresh-$X$ correlation term gives

$$\bar{\mathcal{C}}(\lambda) = s_u \, \overline{\mathcal{R}}(\lambda) + (s_u - 1) \lim \langle \Delta_\lambda, \beta \rangle = s_u \, \overline{\mathcal{R}}(\lambda) + r^2(1 - s_u)(1 - s_t).$$

Therefore,

$$\begin{aligned} \overline{\mathcal{R}}(\lambda) - \bar{\mathcal{C}}(\lambda) &= (1 - s_u)(\overline{\mathcal{R}}(\lambda) - r^2(1 - s_t)) \\ &= (1 - s_u) \, \Theta_t(\lambda). \end{aligned} \tag{87}$$

Next, expanding the fresh-$X$ PD risk gives

$$\overline{\mathcal{R}}_{\mathsf{pd}}(\lambda) = s_{2,u} \overline{\mathcal{R}}(\lambda) + r^2\big(1 - s_{2,u} + 2s_t(s_{2,u} - s_u)\big).$$

Hence the discrepancy denominator is

$$\begin{aligned} \overline{\mathcal{D}}(\lambda) &:= \overline{\mathcal{R}}(\lambda) + \overline{\mathcal{R}}_{\mathsf{pd}}(\lambda) - 2\bar{\mathcal{C}}(\lambda) \\ &= (1 - 2s_u + s_{2,u})(\overline{\mathcal{R}}(\lambda) + r^2(2s_t - 1)) \\ &= (1 - 2s_u + s_{2,u}) \, T_t(\lambda). \end{aligned}$$

Using

$$1 - s_u = \frac{\kappa_s}{1+\kappa_s}, \qquad 1 - 2s_u + s_{2,u} = \frac{\kappa_s^2 c_s}{(1+\kappa_s)^2},$$

we obtain

$$\frac{(\overline{\mathcal{R}}(\lambda) - \bar{\mathcal{C}}(\lambda))^2}{\overline{\mathcal{D}}(\lambda)} = \frac{\Theta_t(\lambda)^2}{T_t(\lambda)} \cdot \frac{1}{c_s(\lambda)}.$$

Substituting this into the oracle decomposition

$$\overline{\mathcal{R}}^{\star}_{\mathsf{pm}}(\lambda) = \overline{\mathcal{R}}(\lambda) - \frac{(\overline{\mathcal{R}}(\lambda) - \bar{\mathcal{C}}(\lambda))^2}{\overline{\mathcal{D}}(\lambda)}$$

proves (84).

To identify the flat case, note that

$$\Theta_t(\lambda) = \overline{\mathcal{R}}(\lambda) - r^2(1 - s_t) = r^2\lambda^2 m_{2,t}(\lambda) + \sigma^2\gamma_t\big(m_{1,t}(\lambda) - \lambda m_{2,t}(\lambda)\big) - r^2\lambda m_{1,t}(\lambda),$$

where

$$m_{k,t}(\lambda) := \int \frac{1}{(u+\lambda)^k}\, dF_{\mathrm{MP},\gamma_t}(u), \qquad k \in \{1,2\},$$

are the labeled resolvent moments. Rearranging,

$$\Theta_t(\lambda) = (\sigma^2\gamma_t - r^2\lambda)\big(m_{1,t}(\lambda) - \lambda m_{2,t}(\lambda)\big).$$

Since

$$m_{1,t}(\lambda) - \lambda m_{2,t}(\lambda) = \int \frac{u}{(u+\lambda)^2}\, dF_{\mathrm{MP},\gamma_t}(u) > 0,$$

we have $\Theta_t(\lambda) = 0$ if and only if

$$\lambda = \lambda^\star := \gamma_t \frac{\sigma^2}{r^2}.$$

This proves part (a).

For part (b), when $\lambda \neq \lambda^\star$ we have $K_t(\lambda) > 0$, so it remains to differentiate $1/c_s$. Using the explicit isotropic root

$$\kappa_s = \frac{\lambda + \gamma_s - 1 + \sqrt{(\lambda+\gamma_s-1)^2 + 4\lambda}}{2},$$

we obtain

$$\frac{1}{c_s} = 1 - \frac{\gamma_s}{(1+\kappa_s)^2} = 1 - \frac{4\gamma_s}{\big(\gamma_s + \lambda + 1 + \sqrt{(\gamma_s+\lambda-1)^2+4\lambda}\big)^2}.$$

Direct differentiation yields

$$\frac{\partial}{\partial\gamma_s}\Big(\frac{1}{c_s}\Big) = \frac{4(\gamma_s - \lambda - 1)}{\big(\gamma_s + \lambda + 1 + \sqrt{(\gamma_s+\lambda-1)^2+4\lambda}\big)^2 \sqrt{(\gamma_s+\lambda-1)^2+4\lambda}}.$$

Substituting into

$$\frac{\partial}{\partial\gamma_s}\overline{\mathcal{R}}^{\star}_{\mathsf{pm}}(\lambda;\gamma_s) = -K_t(\lambda)\frac{\partial}{\partial\gamma_s}\Big(\frac{1}{c_s}\Big)$$

proves (86) and the sign characterization. □

### G.2 Non-identifiability of tuning without labeled data

Recall that, for squared prediction risk with test covariance $\Sigma \succ 0$, the oracle weight is

$$\xi^\star = \frac{R - C}{D} = \frac{\langle \beta_{\lambda_t} - \beta,\ \beta_{\lambda_t} - \beta_{\mathsf{pd},\lambda_s}\rangle_\Sigma}{\|\beta_{\lambda_t} - \beta_{\mathsf{pd},\lambda_s}\|_\Sigma^2},$$

provided $D(\lambda_t, \lambda_s) > 0$. Thus, for known $\Sigma$, the denominator depends only on the observable pair $(\beta_{\lambda_t}, \beta_{\mathsf{pd},\lambda_s})$, whereas the numerator still depends on the unknown target parameter $\beta$.

Consider the setting where the teacher's training covariates $X$ is drawn from an isotropic distribution. The teacher ridge estimator with penalty $\lambda_t$ is,

$$\widehat{\beta}_{\lambda_t} = \left(\frac{X^\top X}{n_t} + \lambda_t I_p\right)^{-1} \left(\frac{X^\top X}{n_t}\beta + \frac{X^\top \varepsilon}{n_t}\right)$$

Under classical asymptotic regime with fixed $p$ and $n_t \to \infty$, we have

$$\widehat{\beta}_{\lambda_t} \to \frac{\beta}{1+\lambda_t}.$$

Now consider two response models,

$$y_1 = X\beta_1 + \varepsilon_1 \qquad \text{and} \qquad y_2 = X\beta_2 + \varepsilon_2,$$

in which the teachers are trained using two different penalties $\lambda_1$ and $\lambda_2$, respectively. Suppose that

$$\frac{\beta_1}{1+\lambda_1} = \frac{\beta_2}{1+\lambda_2}$$

and the training penalty is unknown to the user.

The two teachers then have the same asymptotic limit. Consequently, given the same fresh unlabeled covariate distribution and a fixed student penalty $\lambda_s$, the corresponding pure-distilled students also have the same asymptotic limit.

However, the population-optimal mixing weight under model $j \in \{1, 2\}$ is

$$\xi_j^\star = \frac{\left\langle \widehat{\beta}_{\lambda_j} - \beta_j,\, \widehat{\beta}_{\lambda_j} - \widehat{\beta}_{\mathrm{pd},\lambda_s} \right\rangle_\Sigma}{\left\| \widehat{\beta}_{\lambda_j} - \widehat{\beta}_{\mathrm{pd},\lambda_s} \right\|_\Sigma^2}.$$

Although the teacher and student are asymptotically identical under the two models, the target signals $\beta_1$ and $\beta_2$ are different. Therefore, provided that

$$\left\langle \beta_1 - \beta_2,\, \widehat{\beta}_{\lambda_j} - \widehat{\beta}_{\mathrm{pd},\lambda_s} \right\rangle_\Sigma \neq 0,$$

the two models have different optimal mixing weights. Hence, the same observable teacher and student can correspond to different values of $\xi^\star$, showing that the optimal mixing weight is not identifiable from the teacher and fresh unlabeled data alone.

For an explicit concrete example, fix a nonzero vector $\theta$ and distinct penalties $\lambda_1, \lambda_2$, and set

$$\beta_j = (1+\lambda_j)\theta.$$

Under response model $y_j = X\beta_j + \varepsilon_j$, the ridge teacher trained with penalty $\lambda_j$ converges to $\theta$ in both worlds. It is easy to show that the observable teacher and student predictions have the same asymptotic law under the two models, but $\xi_1^\star \neq \xi_2^\star$.

### G.3 Logistic regression under heavy label noise

When $\rho > 0.5$, incorrectly labeled neighbors outnumber correctly labeled ones. The resulting label averaging prevents the student from correcting misclassified samples and also causes it to misclassify samples that the teacher labeled correctly. In fact, both $y_{\mathsf{pd}}$ and $y_{\mathsf{pd}}^{(2)}$ achieve $0\%$ population accuracy as $n \to \infty$. Nevertheless, the PMSD student $y_{\mathsf{pm},\xi}^{(2)}$ can achieve asymptotic $100\%$ population accuracy. See Figures 15 and 17 for synthetic illustrations with $\rho > 0.5$.

**Theorem G.2.** *Under Assumption 6.1, let $\rho > 0.5$ and take the regularization parameter to be a sequence $\lambda_n = \Theta(1)$. Then, as $n \to \infty$, both $y_{\mathsf{pd}}$ and $y_{\mathsf{pd}}^{(2)}$ achieve $0\%$ population accuracy. On the other hand, there exists $\xi > 1$ such that the PMSD student $y_{\mathsf{pm},\xi}^{(2)}$ achieves $100\%$ population accuracy.*

In this regime, the first PD student $y_{\mathsf{pd}}$ misclassifies both the samples that the original teacher labeled correctly and those it labeled incorrectly. Although the soft predictions of the second student $y_{\mathsf{pd}}^{(2)}$ move in the correct direction relative to those of $y_{\mathsf{pd}}$, the shift is insufficient to alter the induced hard predictions. With an appropriate mixing weight $\xi > 1$, the PMSD student amplifies this corrective signal and extrapolates far enough to change the final predictions.

*Proof of Theorem G.2.* For this proof, write $\lambda_n$ for the regularization parameter at sample size $n$, so that $\bar{\lambda}_n = 2n\lambda_n$ and $\lambda_n = \Theta(1)$. Let $T_n := \alpha_n - (\alpha_n + \widehat{\alpha}_n)\rho$. The equations in (66) force $\alpha_n, \widehat{\alpha}_n > 0$. We first show that $\alpha_n > \widehat{\alpha}_n$ when $\rho > 0.5$. Suppose instead that $\alpha_n \le \widehat{\alpha}_n$. Using $\sigma(z) = 1 - \sigma(-z)$ and the monotonicity of the sigmoid, we have

$$\begin{aligned}
&\sigma(-cnT_n - (1-c)\alpha_n) = \bar{\lambda}_n\alpha_n \le \bar{\lambda}_n\widehat{\alpha}_n = \sigma(cnT_n - (1-c)\widehat{\alpha}_n) \\
\Rightarrow\quad & -cnT_n - (1-c)\alpha_n \le cnT_n - (1-c)\widehat{\alpha}_n \\
\Rightarrow\quad & 2cnT_n \ge (1-c)(\widehat{\alpha}_n - \alpha_n) \ge 0,
\end{aligned}$$

which is a contradiction because $T_n = (1-\rho)\alpha_n - \rho\widehat{\alpha}_n < 0$ (recall that $\rho > 0.5$). Thus, we must have $\alpha_n > \widehat{\alpha}_n$.

Because $\lambda_n = \Theta(1)$, we have $\bar{\lambda}_n \to \infty$. Define the scaled coefficients

$$A_n := \bar{\lambda}_n\alpha_n, \qquad \widehat{A}_n := \bar{\lambda}_n\widehat{\alpha}_n, \qquad N_n := nT_n.$$

The equations in (66) imply $A_n, \widehat{A}_n \in (0,1)$, so $\alpha_n, \widehat{\alpha}_n \to 0$. Moreover,

$$2\lambda_n N_n = (1-\rho)A_n - \rho\widehat{A}_n. \tag{88}$$

Hence, $(A_n, \widehat{A}_n, N_n)$ is bounded.

Starting from an arbitrary subsequence, pass to a further subsequence along which

$$(A_n, \widehat{A}_n, N_n, \lambda_n) \to (A, \widehat{A}, N, \lambda_\star), \qquad \lambda_\star \in (0, \infty).$$

Taking limits in (66) gives

$$\widehat{A} = \sigma(cN), \qquad A = \sigma(-cN), \qquad A + \widehat{A} = 1. \tag{89}$$

Since $\alpha_n > \widehat{\alpha}_n$, we have $A \ge \widehat{A}$. If $N > 0$, then (89) gives $\widehat{A} > A$, a contradiction. If $N = 0$, then $A = \widehat{A} = 0.5$, whereas (88) gives

$$0 = \frac{1-2\rho}{2},$$

again a contradiction because $\rho > 0.5$. Therefore,

$$N < 0, \qquad \widehat{A} < 0.5 < A. \tag{90}$$

The original subsequence was arbitrary, so every subsequential limit satisfies (90). Consequently, for all sufficiently large $n$, $\widehat{A}_n < 0.5 < A_n$. By (67), the first PD student therefore has asymptotic $0\%$ population accuracy.

For the second student, define

$$B_n := \bar{\lambda}_n\beta_n, \qquad \widehat{B}_n := \bar{\lambda}_n\widehat{\beta}_n, \qquad M_n := n\big((1-\rho)\beta_n - \rho\widehat{\beta}_n\big).$$

Equation (71) implies

$$0 < \widehat{A}_n + \widehat{B}_n < 1, \qquad 0 < A_n + B_n < 1,$$

so $(B_n, \widehat{B}_n)$ is bounded and $\beta_n, \widehat{\beta}_n \to 0$. We also have the exact identity

$$2\lambda_n M_n = (1-\rho)B_n - \rho\widehat{B}_n, \tag{91}$$

which shows that $M_n$ is bounded.

Passing to a further subsequence if necessary, suppose that

$$(B_n, \widehat{B}_n, M_n) \to (B, \widehat{B}, M).$$

Taking limits in (71) gives

$$\widehat{A} + \widehat{B} = \sigma(cM), \qquad A + B = \sigma(-cM). \tag{92}$$

Combining these equations with $A + \widehat{A} = 1$ yields

$$B + \widehat{B} = 0. \tag{93}$$

Taking limits in (91) and using (93) gives

$$2\lambda_\star M = B. \tag{94}$$

We claim that $B < 0 < \widehat{B}$. If $B > 0$, then (94) gives $M > 0$. On the other hand, $A + B > A > 0.5$, so (92) gives $M < 0$, a contradiction. If $B = 0$, then $\widehat{B} = M = 0$, and (92) gives $A = \widehat{A} = 0.5$, contradicting (90). Therefore, $B < 0 < \widehat{B}$, and (94) gives $M < 0$. It follows from (92) that

$$\widehat{A} + \widehat{B} < 0.5 < A + B. \tag{95}$$

Because the original subsequence was arbitrary, for all sufficiently large $n$, $\widehat{A}_n + \widehat{B}_n < 0.5 < A_n + B_n$. Equation (72) therefore shows that the second PD student also has asymptotic $0\%$ population accuracy.

At any subsequential limit, the affine scores are

$$s^{(2)}_{\mathsf{pm},\xi}(x_i) = \begin{cases} \widehat{A} + \xi\widehat{B} & \text{for } i \in \mathcal{S}_{1,\mathrm{bad}}, \\ 1 - A - \xi B & \text{for } i \in \mathcal{S}_{1,\mathrm{good}}, \\ 1 - \widehat{A} - \xi\widehat{B} & \text{for } i \in \mathcal{S}_{0,\mathrm{bad}}, \\ A + \xi B & \text{for } i \in \mathcal{S}_{0,\mathrm{good}}. \end{cases}$$

Thus, it suffices to choose

$$\xi > \max\left\{ \frac{0.5 - \widehat{A}}{\widehat{B}}, \frac{A - 0.5}{-B} \right\}. \tag{96}$$

Both ratios in (96) exceed 1 because

$$\widehat{A} + \widehat{B} < 0.5 < A + B.$$

Finally, let $\mathcal{K}$ denote the compact set of all subsequential limits of $(A_n, \widehat{A}_n, B_n, \widehat{B}_n)$. On $\mathcal{K}$, we have $\widehat{B} > 0$ and $-B > 0$, so the two ratios in (96) are continuous and have a finite maximum. Choosing $\xi$ larger than this maximum gives $\xi > 1$ and makes all four groups classify correctly for all sufficiently large $n$. Hence, the PMSD student has asymptotic $100\%$ population accuracy. □

### G.4 Optimal fresh-$X$ covariance structure

Our framework also raises a natural design question: given the teacher predictor, from which distribution should the fresh unlabeled covariates be sampled to minimize the PMSD risk? Within the class of covariances $\Sigma_s$ that are simultaneously diagonalizable with $\Sigma_t$, this amounts to maximizing the teacher-to-PMSD risk reduction in Theorem 3.3,

$$\mathcal{R}(\lambda_t) - \mathcal{R}^\star_{\mathsf{pm}}(\lambda_t, \lambda_s) = \frac{\big(\mathcal{R}(\lambda_t) - \mathcal{C}(\lambda_t, \lambda_s)\big)^2}{\mathcal{R}(\lambda_t) + \mathcal{R}_{\mathsf{pd}}(\lambda_s) - 2\mathcal{C}(\lambda_t, \lambda_s)}. \tag{97}$$

Because this quantity also depends on $\lambda_t$ and $\lambda_s$, the optimal choice $\Sigma_s^\star$ generally varies with the regularization parameters. Jointly optimizing $\Sigma_s$ and $\lambda_s$ for a given teacher is considerably more involved, and we leave this problem for future work.

We instead consider the underparameterized regime $\gamma_t, \gamma_s < 1$ and the post-asymptotic ridgeless limit $\lambda_t, \lambda_s \to 0$. Recall that $\Sigma_t = U^\top \operatorname{diag}(\sigma_1, \ldots, \sigma_p) U$, $\Sigma_s = U^\top \operatorname{diag}(\widetilde{\sigma}_1, \ldots, \widetilde{\sigma}_p) U$, and $v_i = ((U\beta)_i)^2$. For each $i \in [p]$, define $w_i = \widetilde{\sigma}_i^{-1}$ and let $w = (w_1, \ldots, w_p)^\top$. The limiting risk reduction in (97) is invariant under the rescaling $\Sigma_s \mapsto c\Sigma_s$ for $c > 0$, so we impose the normalization $p^{-1} \sum_{i=1}^p w_i = 1$. Define

$$s_i := \sigma_i v_i + \frac{\sigma^2 \gamma_t}{p(1 - \gamma_t)}, \quad t_i := \frac{\gamma_s}{2p(1 - \gamma_s)} \left( v_i + \frac{\sigma^2 \gamma_t}{p\sigma_i (1 - \gamma_t)} \right), \quad i = 1, \ldots, p,$$

and

$$S = \operatorname{diag}(s_1, \ldots, s_p), \quad t = (t_1, \ldots, t_p)^\top, \quad \theta = (\sigma_1, \ldots, \sigma_p)^\top.$$

**Proposition G.3.** *Under Assumption 3.2, further assume that $\gamma_t, \gamma_s < 1$ and $\sigma^2 > 0$, and consider the ridgeless limit $\lambda_t, \lambda_s \to 0$. Then maximizing the risk reduction in* (97) *over the spectrum $\{\widetilde{\sigma}_1, \dots, \widetilde{\sigma}_p\}$ of the positive definite matrix $\Sigma_s$ is equivalent, up to positive rescaling, to the optimization problem*

$$\begin{aligned} \underset{w \in \mathbb{R}^p_+}{\text{minimize}} \quad & w^\top \big(S + t\theta^\top + \theta t^\top\big) w, \\ \text{subject to} \quad & 1_p^\top w = p, \end{aligned} \tag{98}$$

*where $w_i = \widetilde{\sigma}_i^{-1}$ for $i \in [p]$.*

This is a linearly constrained quadratic program in $w$ and can be solved numerically using standard quadratic programming methods. In general, the spectrum of the optimal $\Sigma_s^\star$ depends on the teacher covariance, the signal-covariance alignment, the noise level, and the data aspect ratios. In particular, matching the teacher covariance need not be optimal: $\Sigma_s^\star$ need not be proportional to $\Sigma_t$. This differs from a recent result on optimal *labeled*-data selection for ridgeless estimators [58], reflecting the different role of the fresh sample in our prediction-only setting.

*Proof of Proposition G.3.* Let

$$W := \Sigma_s^{-1} = U^\top \operatorname{diag}(w_1, \dots, w_p) U.$$

For $a \in \{t, s\}$, the fixed-point equation can be rewritten as

$$\lambda_a = \kappa_a \big(1 - \gamma_a \, \overline{\operatorname{tr}}\big(\Sigma_a (\Sigma_a + \kappa_a I_p)^{-1}\big)\big).$$

Because

$$0 \le \overline{\operatorname{tr}}\big(\Sigma_a (\Sigma_a + \kappa_a I_p)^{-1}\big) \le 1$$

and $\gamma_a < 1$, the factor in parentheses is bounded below by $1 - \gamma_a > 0$. Hence,

$$0 < \kappa_a \le \frac{\lambda_a}{1 - \gamma_a},$$

so

$$\kappa_t, \kappa_s \to 0 \qquad \text{as } \lambda_t, \lambda_s \to 0.$$

It follows that

$$G_t \to \Sigma_t^{-1}, \qquad G_s \to \Sigma_s^{-1} = W, \qquad b_t \to \frac{1}{1 - \gamma_t}.$$

Moreover,

$$b_s = \frac{\gamma_s \, \overline{\operatorname{tr}}(\Sigma_s \Sigma_t G_s^2)}{1 - \gamma_s \, \overline{\operatorname{tr}}(\Sigma_s^2 G_s^2)} \to \frac{\gamma_s}{1 - \gamma_s} \, \overline{\operatorname{tr}}(\Sigma_t W). \tag{99}$$

Recall that the risk reduction has the form $(\mathcal{R} - \mathcal{C})^2 / \mathcal{D}$. Substituting these limits into the expressions for $\mathcal{R} - \mathcal{C}$ and $\mathcal{D}$ in Theorem 3.3, we obtain

$$\lim_{\lambda_t, \lambda_s \to 0} \frac{\mathcal{R} - \mathcal{C}}{\kappa_s} = \frac{\sigma^2 \gamma_t}{1 - \gamma_t} \, \overline{\operatorname{tr}}(W), \tag{100}$$

and

$$\begin{aligned} \lim_{\lambda_t, \lambda_s \to 0} \frac{\mathcal{D}}{\kappa_s^2} = {} & \beta^\top W \Sigma_t W \beta + \frac{\gamma_s}{1 - \gamma_s} \, \overline{\operatorname{tr}}(\Sigma_t W) \beta^\top W \beta \\ & + \frac{\sigma^2 \gamma_t}{1 - \gamma_t} \Big\{ \overline{\operatorname{tr}}(W^2) + \frac{\gamma_s}{1 - \gamma_s} \, \overline{\operatorname{tr}}(\Sigma_t W) \, \overline{\operatorname{tr}}(\Sigma_t^{-1} W) \Big\}. \end{aligned} \tag{101}$$

After canceling the common factor $\kappa_s^2$, the limiting risk reduction is the square of the right-hand side of (100) divided by the right-hand side of (101).

We next express the right-hand side of (101) in terms of $w$. In the common eigenbasis of $\Sigma_t$ and $\Sigma_s$, we have

$$\beta^\top W \Sigma_t W \beta = \sum_{i=1}^p \sigma_i v_i w_i^2, \qquad \beta^\top W \beta = \sum_{i=1}^p v_i w_i,$$

and

$$\overline{\mathrm{tr}}(\Sigma_t W) = \frac{1}{p}\sum_{i=1}^{p}\sigma_i w_i, \qquad \overline{\mathrm{tr}}(W^2) = \frac{1}{p}\sum_{i=1}^{p} w_i^2, \qquad \overline{\mathrm{tr}}(\Sigma_t^{-1} W) = \frac{1}{p}\sum_{i=1}^{p}\frac{w_i}{\sigma_i}.$$

Substituting these identities into (101) gives

$$\sum_{i=1}^{p}\left(\sigma_i v_i + \frac{\sigma^2\gamma_t}{p(1-\gamma_t)}\right)w_i^2 + \frac{\gamma_s}{p(1-\gamma_s)}\left(\sum_{i=1}^{p}\sigma_i w_i\right)\left(\sum_{i=1}^{p} v_i w_i\right) + \frac{\sigma^2\gamma_t}{p(1-\gamma_t)}\cdot\frac{\gamma_s}{p(1-\gamma_s)}\left(\sum_{i=1}^{p}\sigma_i w_i\right)\left(\sum_{i=1}^{p}\frac{w_i}{\sigma_i}\right).$$

By the definitions of $S$, $t$, and $\theta$, this expression equals

$$w^\top\left(S + t\theta^\top + \theta t^\top\right)w.$$

The squared numerator determined by (100) and the denominator in (101) are both homogeneous of degree two in $w$. We may therefore impose the normalization $1_p^\top w = p$, under which $\overline{\mathrm{tr}}(W) = 1$ and the squared numerator equals the positive constant $\tau_t^2$. Hence, maximizing the limiting risk reduction is equivalent to

$$\begin{aligned} \underset{w\in\mathbb{R}^p_+}{\text{minimize}} \quad & w^\top\left(S + t\theta^\top + \theta t^\top\right)w, \\ \text{subject to} \quad & 1_p^\top w = p. \end{aligned}$$

□

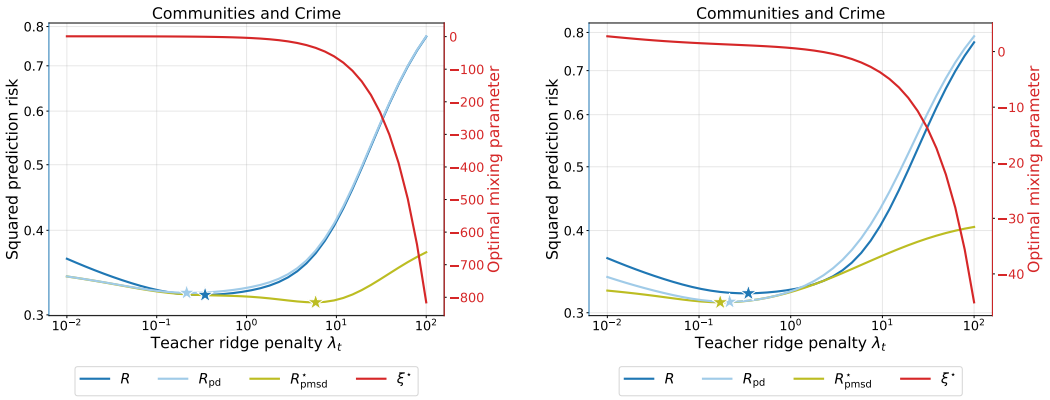
# H Additional experiments

Code for reproducing the results is available at `https://github.com/hhd357/prediction_only_mixing_distillation`.

## H.1 Real-world regression tasks with different student regularization levels

We report additional ridge regression results on Communities and Crime and Blog Feedback. For each dataset, we compare fresh covariates drawn from the teacher's training distribution with fresh covariates drawn from an isotropic Gaussian distribution, fixing or tuning $\lambda_s$ as in the captions.

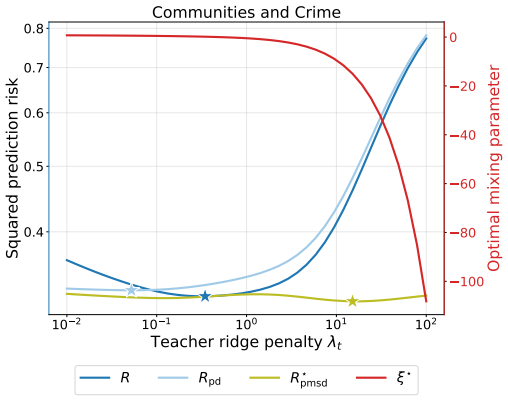
### H.1.1 Communities and Crime



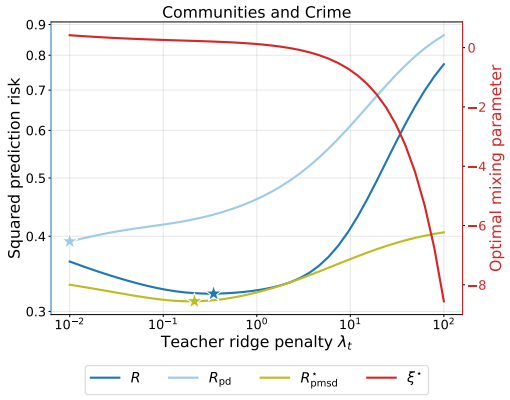

(a) Fresh unlabeled covariates are drawn from the same distribution as the teacher's training covariates.

(b) Fresh unlabeled covariates are drawn i.i.d. from $\mathcal{N}(0, I_p)$.

Figure 5: Communities and Crime with $n_t = 400$, $n_s = 800$, and $p = 99$. The student regularization is fixed at $\lambda_s = 0.1$ for all values of $\lambda_t$.

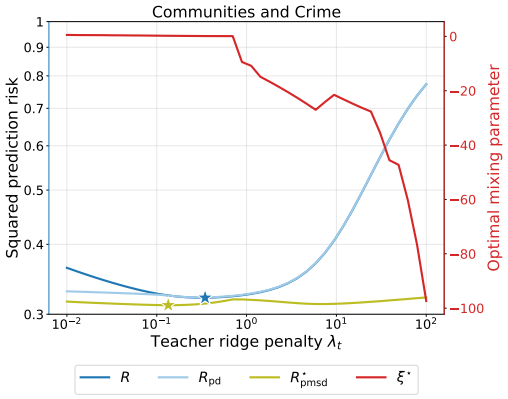



(a) Fresh unlabeled covariates are drawn from the same distribution as the teacher's training covariates.

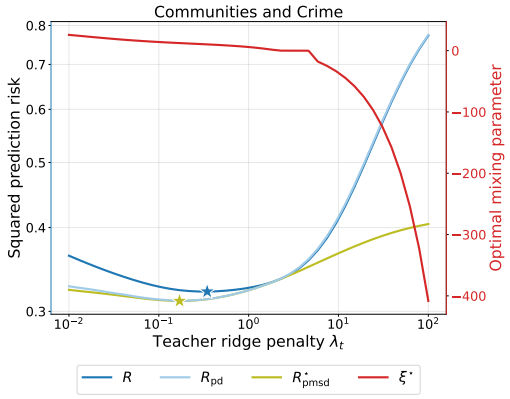



(b) Fresh unlabeled covariates are drawn i.i.d. from $\mathcal{N}(0, I_p)$.

Figure 6: Communities and Crime with $n_t = 400$, $n_s = 800$, and $p = 99$. The student regularization is fixed at $\lambda_s = 1$ for all values of $\lambda_t$.



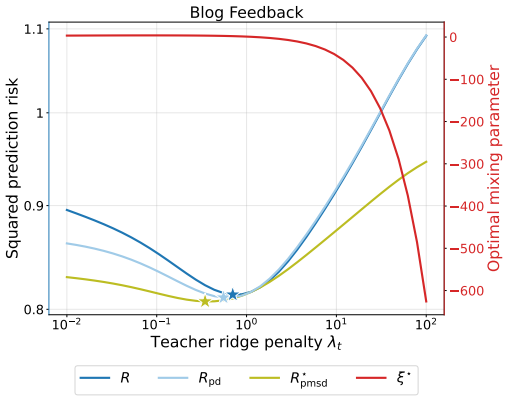

(a) Fresh unlabeled covariates are drawn from the same distribution as the teacher's training covariates.



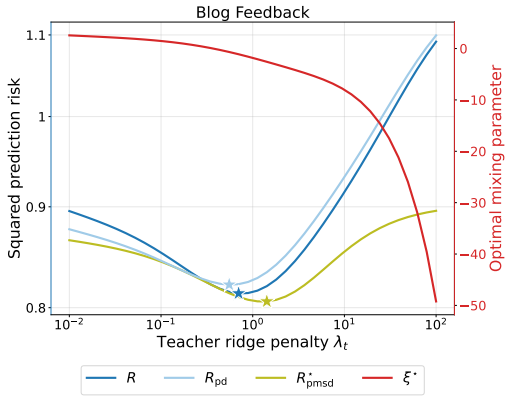

(b) Fresh unlabeled covariates are drawn i.i.d. from $\mathcal{N}(0, I_p)$.

Figure 7: Communities and Crime with $n_t = 400$, $n_s = 800$, and $p = 99$. At each value of $\lambda_t$, the student regularization $\lambda_s$ is tuned separately over a grid for the PD and PMSD students.

### H.1.2 Blog Feedback

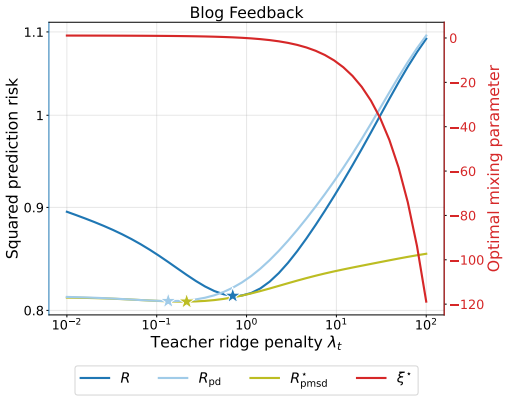



(a) Fresh unlabeled covariates are drawn from the same distribution as the teacher's training covariates.

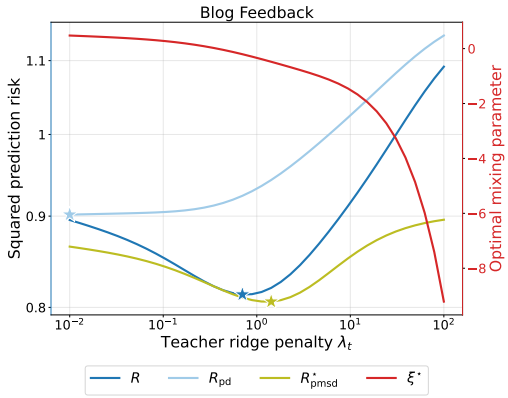



(b) Fresh unlabeled covariates are drawn i.i.d. from $\mathcal{N}(0, I_p)$.

Figure 8: Blog Feedback with $n_t = 2619$, $n_s = 5240$, and $p = 280$. The student regularization is fixed at $\lambda_s = 0.1$ for all values of $\lambda_t$.

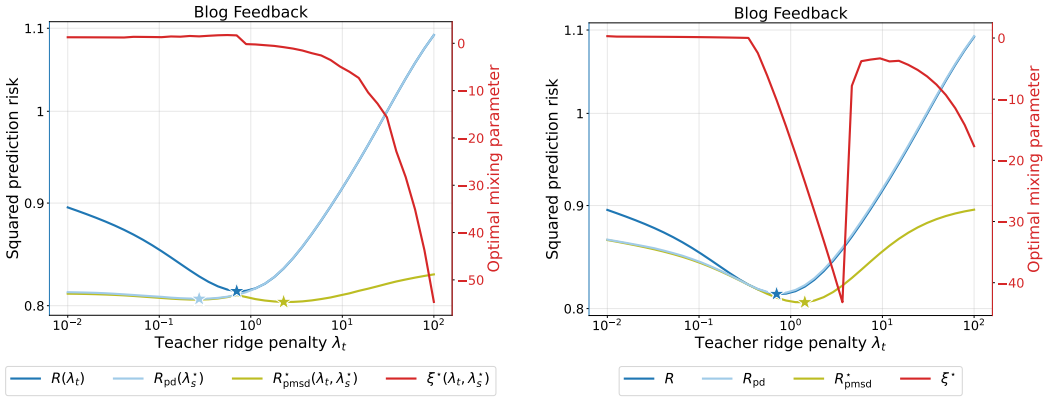



(a) Fresh unlabeled covariates are drawn from the same distribution as the teacher's training covariates.




(b) Fresh unlabeled covariates are drawn i.i.d. from $\mathcal{N}(0, I_p)$.

Figure 9: Blog Feedback with $n_t = 2619$, $n_s = 5240$, and $p = 280$. The student regularization is fixed at $\lambda_s = 1$ for all values of $\lambda_t$.




(a) Fresh unlabeled covariates are drawn from the same distribution as the teacher's training covariates.

(b) Fresh unlabeled covariates are drawn i.i.d. from $\mathcal{N}(0, I_p)$.

Figure 10: Blog Feedback with $n_t = 2619$, $n_s = 5240$, and $p = 280$. At each value of $\lambda_t$, the student regularization $\lambda_s$ is tuned separately over a grid for the PD and PMSD students.

## H.2 Linear probing experiments on real datasets

The following tables report test classification accuracy and the estimated optimal mixing weight across regularization levels. In every experiment, the mixing weight is selected using a calibration set over the grid $[-20, 20]$; see Section I for the dataset and training details.

### H.2.1 Caltech-101

Table 2: Test classification accuracy on Caltech-101 with random corruption rate $\rho = 0.2$.

| Regularization level $\lambda$ | $10^{-4}$ | $10^{-3.5}$ | $10^{-3}$ | $10^{-2.5}$ | $10^{-2}$ | $10^{-1.5}$ | $10^{-1}$ | $10^{-0.5}$ | $10^{0}$ |
|---|---|---|---|---|---|---|---|---|---|
| Teacher (%) | 68.7 | 68.7 | 68.8 | 68.9 | 69.5 | 70.9 | 72.8 | 68.4 | 46.0 |
| PD student (%) | 68.9 | 69.7 | 71.4 | 72.4 | 73.6 | 71.9 | 63.5 | 42.7 | 27.0 |
| Optimal PMSD student (%) | 68.9 | 70.8 | 71.5 | 72.4 | 73.4 | 72.5 | 72.8 | 76.1 | 63.0 |
| Estimated optimal mixing weight $\widehat{\xi}^{\star}$ | 5.5 | 2.5 | 1.1 | 0.9 | 0.9 | 0.7 | $-0.1$ | $-20.0$ | $-17.8$ |

Table 3: Test classification accuracy on Caltech-101 with random corruption rate $\rho = 0.4$.

| Regularization level $\lambda$ | $10^{-4}$ | $10^{-3.5}$ | $10^{-3}$ | $10^{-2.5}$ | $10^{-2}$ | $10^{-1.5}$ | $10^{-1}$ | $10^{-0.5}$ | $10^{0}$ |
|---|---|---|---|---|---|---|---|---|---|
| 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 optimal mixing weight $\widehat{\xi}^{\star}$ | 4.9 | 5.5 | 2.2 | 1.4 | 1.1 | 0.9 | $-0.1$ | $-4.2$ | $-18.6$ |

Table 4: Test classification accuracy on Caltech-101 with random corruption rate $\rho = 0.6$.

| Regularization level $\lambda$ | $10^{-4}$ | $10^{-3.5}$ | $10^{-3}$ | $10^{-2.5}$ | $10^{-2}$ | $10^{-1.5}$ | $10^{-1}$ | $10^{-0.5}$ | $10^{0}$ |
|---|---|---|---|---|---|---|---|---|---|
| Teacher (%) | 49.3 | 49.3 | 49.3 | 49.6 | 50.1 | 52.2 | 57.1 | 59.8 | 36.2 |
| PD student (%) | 49.6 | 50.0 | 51.6 | 54.1 | 57.4 | 58.3 | 52.5 | 33.4 | 23.5 |
| Optimal PMSD student (%) | 49.5 | 52.4 | 54.3 | 55.9 | 57.0 | 58.2 | 57.2 | 65.5 | 60.4 |
| Estimated optimal mixing weight $\widehat{\xi}^{\star}$ | 0.8 | 8.4 | 3.2 | 1.6 | 1.7 | 1.0 | 0.2 | $-10.6$ | $-20.0$ |

### H.2.2 Caltech-256

Table 5: Test classification accuracy on Caltech-256 with random corruption rate $\rho = 0.2$.

| Regularization level $\lambda$ | $10^{-4}$ | $10^{-3.5}$ | $10^{-3}$ | $10^{-2.5}$ | $10^{-2}$ | $10^{-1.5}$ | $10^{-1}$ | $10^{-0.5}$ | $10^{0}$ |
|---|---|---|---|---|---|---|---|---|---|
| Teacher (%) | 65.0 | 65.1 | 65.2 | 65.5 | 66.6 | 67.5 | 61.5 | 35.8 | 14.1 |
| PD student (%) | 65.7 | 66.5 | 68.0 | 68.6 | 67.8 | 58.4 | 28.4 | 11.4 | 6.0 |
| Optimal PMSD student (%) | 66.9 | 67.7 | 68.2 | 68.5 | 68.1 | 68.0 | 67.6 | 50.8 | 25.6 |
| Estimated optimal mixing weight $\widehat{\xi}^{\star}$ | 5.6 | 2.7 | 1.2 | 0.9 | 0.5 | $-1.9$ | $-20.0$ | $-20.0$ | $-20.0$ |

Table 6: Test classification accuracy on Caltech-256 with random corruption rate $\rho = 0.4$.

| Regularization level $\lambda$ | $10^{-4}$ | $10^{-3.5}$ | $10^{-3}$ | $10^{-2.5}$ | $10^{-2}$ | $10^{-1.5}$ | $10^{-1}$ | $10^{-0.5}$ | $10^{0}$ |
|---|---|---|---|---|---|---|---|---|---|
| Teacher (%) | 56.8 | 56.8 | 57.0 | 57.5 | 58.4 | 60.3 | 55.0 | 33.4 | 13.0 |
| PD student (%) | 57.4 | 58.2 | 59.7 | 60.9 | 60.3 | 49.3 | 25.5 | 10.4 | 5.4 |
| Optimal PMSD student (%) | 59.0 | 59.5 | 60.3 | 61.2 | 60.4 | 60.6 | 61.5 | 45.7 | 23.5 |
| Estimated optimal mixing weight $\widehat{\xi}^{\star}$ | 7.4 | 2.8 | 1.6 | 1.2 | 0.6 | $-0.8$ | $-20.0$ | $-20.0$ | $-20.0$ |

Table 7: Test classification accuracy on Caltech-256 with random corruption rate $\rho = 0.6$.

| Regularization level $\lambda$ | $10^{-4}$ | $10^{-3.5}$ | $10^{-3}$ | $10^{-2.5}$ | $10^{-2}$ | $10^{-1.5}$ | $10^{-1}$ | $10^{-0.5}$ | $10^{0}$ |
|---|---|---|---|---|---|---|---|---|---|
| Teacher (%) | 43.5 | 43.6 | 43.9 | 44.4 | 45.6 | 48.0 | 45.1 | 27.5 | 11.6 |
| PD student (%) | 44.1 | 44.9 | 46.7 | 48.3 | 48.0 | 39.4 | 19.6 | 9.5 | 3.8 |
| Optimal PMSD student (%) | 45.8 | 46.5 | 47.5 | 48.6 | 47.6 | 48.3 | 51.5 | 39.6 | 21.5 |
| Estimated optimal mixing weight $\widehat{\xi}^{\star}$ | 6.0 | 2.5 | 1.9 | 1.1 | 0.6 | $-2.1$ | $-20.0$ | $-20.0$ | $-19.4$ |

### H.2.3 CIFAR-100 (random corruption)

Table 8: Test classification accuracy on CIFAR-100 with random corruption rate $\rho = 0.2$.

| Regularization level $\lambda$ | $10^{-4}$ | $10^{-3.5}$ | $10^{-3}$ | $10^{-2.5}$ | $10^{-2}$ | $10^{-1.5}$ | $10^{-1}$ | $10^{-0.5}$ | $10^{0}$ |
|---|---|---|---|---|---|---|---|---|---|
| Teacher (%) | 53.8 | 53.9 | 54.0 | 54.4 | 55.2 | 57.1 | 58.2 | 50.6 | 35.9 |
| PD student (%) | 54.0 | 54.6 | 55.4 | 56.8 | 57.9 | 56.9 | 46.4 | 18.4 | 2.4 |
| Optimal PMSD student (%) | 55.0 | 55.3 | 55.9 | 56.9 | 57.9 | 57.6 | 59.7 | 54.8 | 44.2 |
| Estimated optimal mixing weight $\widehat{\xi}^\star$ | 6.6 | 2.4 | 1.5 | 1.2 | 0.9 | 0.7 | $-20.0$ | $-20.0$ | $-20.0$ |

Table 9: Test classification accuracy on CIFAR-100 with random corruption rate $\rho = 0.4$.

| Regularization level $\lambda$ | $10^{-4}$ | $10^{-3.5}$ | $10^{-3}$ | $10^{-2.5}$ | $10^{-2}$ | $10^{-1.5}$ | $10^{-1}$ | $10^{-0.5}$ | $10^{0}$ |
|---|---|---|---|---|---|---|---|---|---|
| Teacher (%) | 47.8 | 47.9 | 48.0 | 48.5 | 49.8 | 52.7 | 54.2 | 45.4 | 31.7 |
| PD student (%) | 48.3 | 49.3 | 50.6 | 52.3 | 54.0 | 52.6 | 39.9 | 16.0 | 2.3 |
| Optimal PMSD student (%) | 50.5 | 50.6 | 51.5 | 53.3 | 54.0 | 53.5 | 56.1 | 49.5 | 40.3 |
| Estimated optimal mixing weight $\widehat{\xi}^\star$ | 6.9 | 4.4 | 2.1 | 1.5 | 1.0 | 0.8 | $-20.0$ | $-14.4$ | $-20.0$ |

Table 10: Test classification accuracy on CIFAR-100 with random corruption rate $\rho = 0.6$.

| Regularization level $\lambda$ | $10^{-4}$ | $10^{-3.5}$ | $10^{-3}$ | $10^{-2.5}$ | $10^{-2}$ | $10^{-1.5}$ | $10^{-1}$ | $10^{-0.5}$ | $10^{0}$ |
|---|---|---|---|---|---|---|---|---|---|
| Teacher (%) | 38.0 | 38.0 | 38.1 | 38.6 | 40.1 | 43.9 | 46.3 | 37.3 | 24.3 |
| PD student (%) | 38.3 | 39.1 | 41.0 | 43.6 | 46.4 | 44.7 | 30.5 | 13.6 | 3.3 |
| Optimal PMSD student (%) | 40.9 | 41.6 | 43.0 | 45.4 | 46.9 | 44.9 | 48.2 | 43.1 | 34.0 |
| Estimated optimal mixing weight $\widehat{\xi}^\star$ | 10.2 | 4.8 | 2.5 | 1.7 | 1.3 | 0.5 | $-1.7$ | $-20.0$ | $-20.0$ |

### H.2.4 CIFAR-100 (hierarchical corruption)

Table 11: Test classification accuracy on CIFAR-100 with hierarchical corruption rate $\rho = 0.4$.

| Regularization level $\lambda$ | $10^{-4}$ | $10^{-3.5}$ | $10^{-3}$ | $10^{-2.5}$ | $10^{-2}$ | $10^{-1.5}$ | $10^{-1}$ | $10^{-0.5}$ | $10^{0}$ |
|---|---|---|---|---|---|---|---|---|---|
| Teacher (%) | 44.9 | 45.0 | 45.0 | 45.3 | 46.3 | 49.5 | 51.1 | 42.7 | 28.4 |
| PD student (%) | 45.2 | 46.1 | 47.6 | 50.0 | 51.6 | 50.9 | 39.4 | 17.5 | 4.4 |
| Optimal PMSD student (%) | 47.0 | 47.6 | 48.9 | 49.9 | 51.6 | 50.9 | 53.5 | 47.0 | 35.9 |
| Estimated optimal mixing weight $\widehat{\xi}^\star$ | 13.0 | 5.7 | 2.5 | 2.1 | 1.3 | 0.9 | $-15.6$ | $-20.0$ | $-20.0$ |

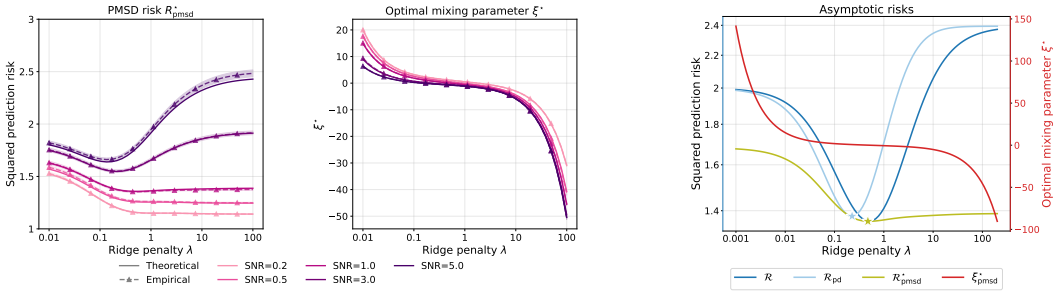
Table 12: Test classification accuracy on CIFAR-100 with hierarchical corruption rate $\rho = 0.6$.

| Regularization level $\lambda$ | $10^{-4}$ | $10^{-3.5}$ | $10^{-3}$ | $10^{-2.5}$ | $10^{-2}$ | $10^{-1.5}$ | $10^{-1}$ | $10^{-0.5}$ | $10^{0}$ |
|---|---|---|---|---|---|---|---|---|---|
| Teacher (%) | 31.3 | 31.3 | 31.5 | 31.8 | 32.5 | 35.8 | 39.0 | 32.5 | 22.4 |
| PD student (%) | 31.4 | 32.2 | 33.4 | 35.4 | 38.4 | 39.1 | 29.5 | 15.3 | 3.8 |
| Optimal PMSD student (%) | 33.6 | 34.0 | 35.3 | 37.0 | 38.7 | 39.2 | 40.3 | 36.9 | 27.8 |
| Estimated optimal mixing weight $\widehat{\xi}^\star$ | 13.0 | 4.8 | 2.4 | 1.8 | 1.5 | 1.0 | $-4.3$ | $-15.6$ | $-20.0$ |

### H.3 Synthetic experiments for ridge regression

We next compare the empirical performance of PMSD with its theoretical predictions under several covariance and regularization settings.

#### H.3.1 Empirical versus theoretical predictions




(a) Theoretical versus empirical risks and mixing weights.

(b) Same-$\lambda$ balanced regime with $\gamma_t = \gamma_s$.

Figure 11: Setting: $\Sigma_t = \Sigma_s$ and $\lambda_t = \lambda_s = \lambda$. Both panels use an AR1 covariance design, and the true signal $\beta$ aligns with the top eigenvectors of $\Sigma_t$. We set $p = 200$, $n_t = n_s = 400$, and $\sigma^2 = 1$ in both panels; the right panel has $r^2 = 1$.

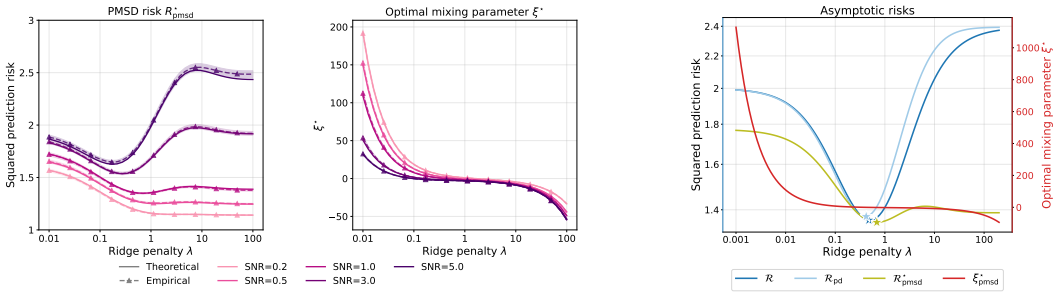



Figure 12: Empirical and theoretical results when $\Sigma_t$ follows an AR1 covariance model, $\Sigma_s = 10I_p$, and $\lambda_t = \lambda_s$. The signal $\beta$ aligns with the top eigenvectors of $\Sigma_t$. Both panels use $p = 200$, $n_t = n_s = 400$, and $\sigma^2 = 1$; the right panel has $r^2 = 1$.

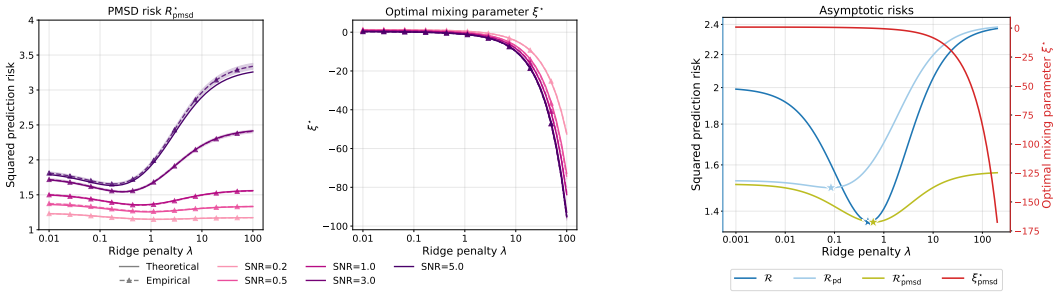



Figure 13: Empirical and theoretical results when $\Sigma_t = \Sigma_s$ follow the same AR1 covariance model and $\lambda_s = 1$. The signal $\beta$ aligns with the top eigenvectors of $\Sigma_t$. Both panels use $p = 200$, $n_t = n_s = 400$, and $\sigma^2 = 1$; the right panel has $r^2 = 1$.

## H.4 Synthetic experiments for logistic regression

### H.4.1 Constant feature correlation

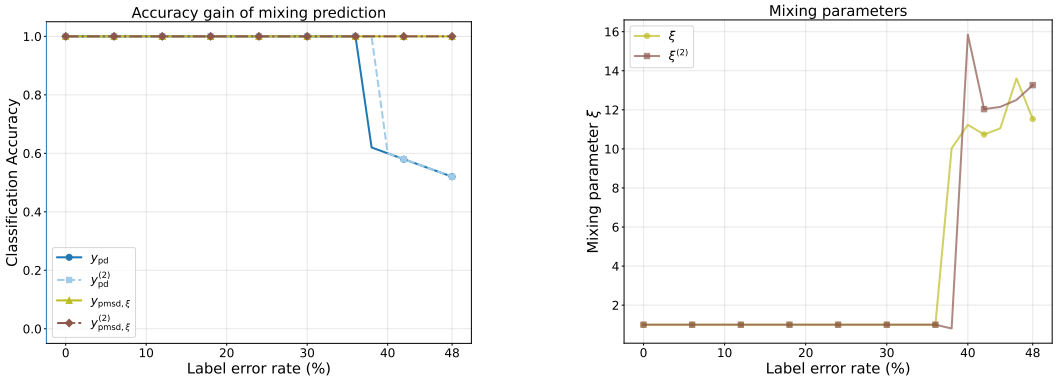



Figure 14: Synthetic logistic regression with constant feature correlation $c = 0.25$, $\lambda = 0.01$, and $n = 500$ samples per class; here, $\rho < 0.5$.

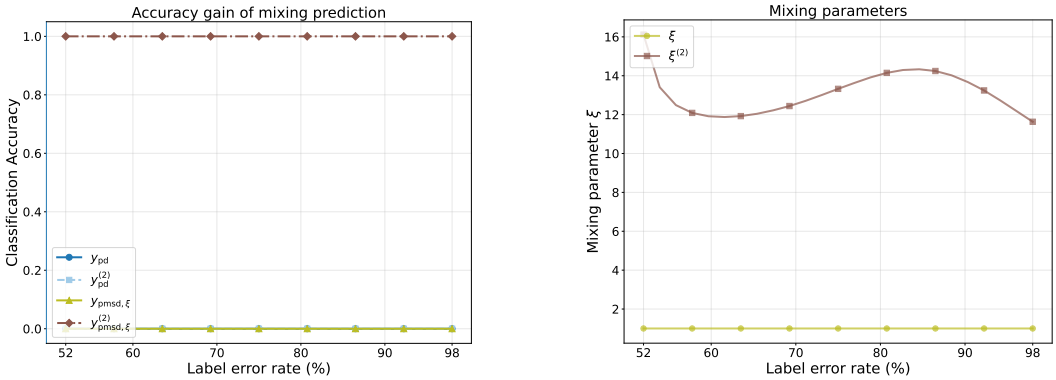



Figure 15: Synthetic logistic regression with constant feature correlation $c = 0.25$, $\lambda = 10$, and $n = 1500$ samples per class; here, $\rho > 0.5$.

### H.4.2 Uniform feature correlation

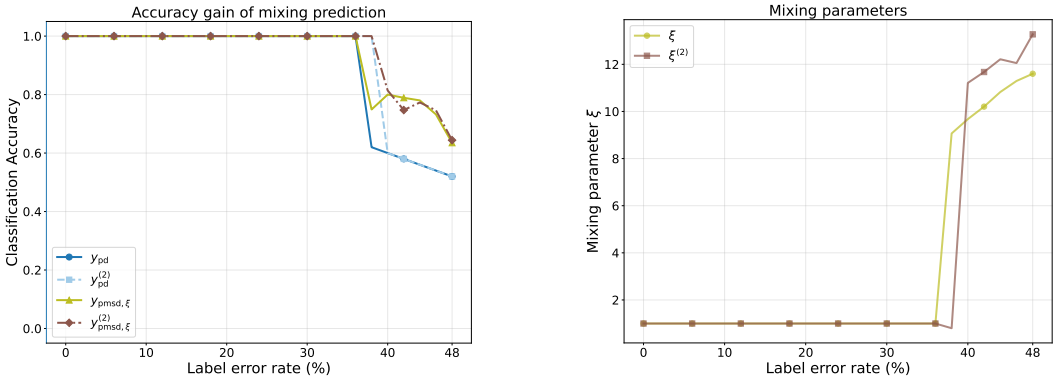



Figure 16: Synthetic logistic regression with uniform feature correlation (defined in Section I.3), $\lambda = 0.01$, and $n = 500$ samples per class; here, $\rho < 0.5$.

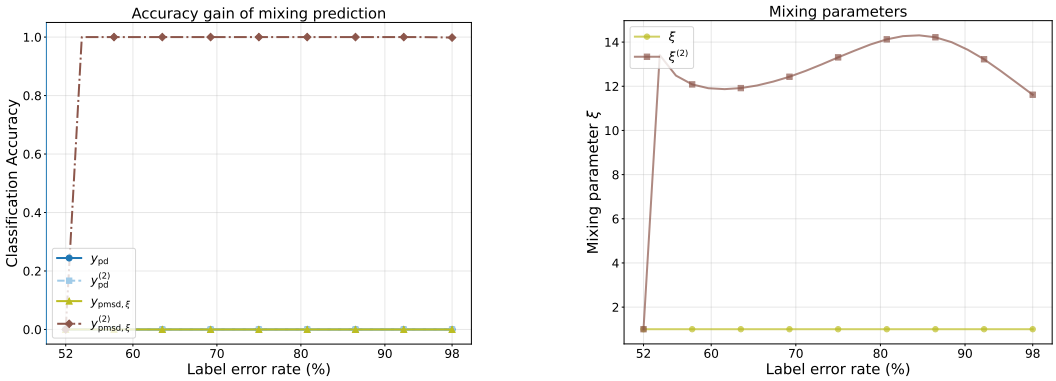
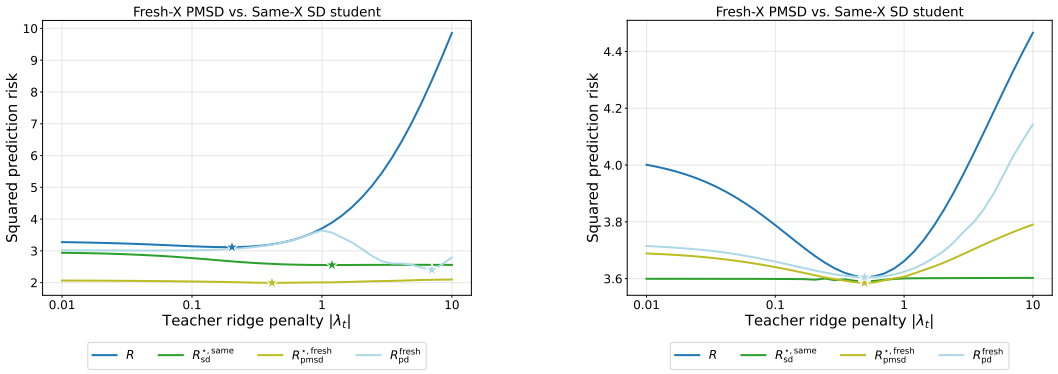



Figure 17: Synthetic logistic regression with uniform feature correlation (defined in Section I.3), $\lambda = 10$, and $n = 1500$ samples per class; here, $\rho > 0.5$.

## H.5 Fresh-X prediction-mixed versus same-X loss-mixed SD




(a) One-spike covariance model.

(b) Isotropic covariance model.

Figure 18: **Fresh-$X$ PMSD versus same-$X$ loss-mixed SD.** Setting: $p = 400$, $n_t = 200$, $n_s = 4000$, and $\Sigma_t = \Sigma_s$. The $n_s$ fresh unlabeled samples are used only by the fresh-$X$ PD and PMSD students. For each value of $|\lambda_t|$ on the x-axis, we evaluate both signs of the teacher regularization, tune $\lambda_s$ over a grid on $[-100, 100]$ in each case, and report the smaller risk. The signal is $\beta = au_1 + \sqrt{r^2 - a^2}v$, where $a = 1.7$, $r = 2$, $u_1$ is the principal direction of $\Sigma_t$, and $v$ is orthogonal to $u_1$.

Fresh-$X$ PMSD and the same-$X$ loss-mixed SD methods studied by [17, 49, 15, 39] operate with different information and therefore have different performance limits. Consider a teacher whose regularization parameter $\lambda_t$ is mistuned for the target risk. A same-$X$ student can reuse the original labeled sample to correct the teacher, whereas prediction-only PMSD can use only the teacher's predictions on fresh unlabeled covariates. Consequently, fresh unlabeled data need not recover information lost in the teacher, regardless of the fresh sample size. Figure 18 illustrates both the benefits and the limitations of PMSD relative to same-$X$ loss mixing.

In the one-spike covariance model, [39] show that same-$X$ loss-mixed SD, with jointly optimized $\lambda_t$, $\lambda_s$, and $\xi \in \mathbb{R}$ and with negative regularization allowed, is optimal within the spectral shrinkage class given the labeled sample $\mathcal{D}_{\text{lab}} = (X, y)$. Once additional fresh unlabeled data $\mathcal{D}_{\text{unlab}}$ are available, the fresh-$X$ PMSD student, which uses only the teacher's ridge predictions, can nevertheless outperform this optimal same-$X$ student; see Figure 18a. The gains are especially pronounced in the overparameterized regime $p > n_t$.

The comparison reverses in the isotropic design setting, where the optimally tuned teacher is optimal within the spectral shrinkage class. Because the same-$X$ loss-mixed student is itself a spectral shrinkage estimator [39], it cannot outperform the optimally tuned teacher. If the given teacher is mistuned, however, the same-$X$ student can attain the optimal ridge risk by setting $\lambda_s = \lambda_t^\star$ and effectively ignoring the teacher's predictions while using the original training data. In contrast, PMSD relies only on the mistuned teacher's predictions on fresh unlabeled covariates. It can substantially improve upon this teacher but generally cannot attain the optimal ridge risk or outperform the optimally tuned teacher; see Figure 18b.

# I Experiment details

## I.1 Ridge regression experiments

For the real-world ridge regression experiments in the left panel of Figure 2 and the middle and right panels of Figure 3, we use UCI Communities and Crime [57], UCI Blog Feedback [10], and UCI Airfoil [7]. We divide each dataset into a teacher training set, a fresh unlabeled set, a calibration set, and a test set, with the sizes specified below or in the corresponding figure captions. After handling missing values as described below, we center and standardize all variables using the means and standard deviations computed on the teacher training set.

**UCI Communities and Crime.** The UCI Communities and Crime dataset [57] combines socioeconomic data from multiple sources. It consists of 1,994 samples with 127 covariates. The response, "ViolentCrimesPerPop," is the rate of violent crimes per 100,000 population. We remove 5 nonpredictive identifiers and 23 covariates from the LEMAS survey that contain 1,675 missing values across the 1,994 samples. The remaining 99 covariates have no missing values and are used in our experiments.

**UCI Blog Feedback.** The UCI Blog Feedback dataset [10] was constructed from crawled and processed HTML documents of blog posts. We use its training split, which contains 52,397 samples, and reserve 30% of these samples for testing. The sizes of the teacher training, fresh unlabeled, and calibration sets are given in the corresponding figure captions.

**UCI Airfoil.** The UCI Airfoil Self-Noise dataset [7] contains aerodynamic and acoustic measurements from NASA tests of two- and three-dimensional airfoil blade sections. The continuous response is the sound pressure level, measured in decibels. The dataset contains $n = 1503$ samples and $p = 5$ covariates: `frequency`, `attack-angle`, `free-stream-velocity`, `chord-length`, and `suction-side-displacement-thickness`. We reserve 30% of the samples for testing. The sizes of the teacher training, fresh unlabeled, and calibration sets are given in the corresponding figure captions.

**CIFAR-10.** The CIFAR-10 dataset [38] consists of $32 \times 32$ color images from 10 classes, with 50,000 training images and 10,000 test images. We use subsets of the training and test splits, with sizes specified in Figure 3.

For the CIFAR-10 ridge regression experiment in the left panel of Figure 3, we extract 512-dimensional features using a ResNet-18 [29] pretrained on ImageNet and available in PyTorch. We perform no additional data preprocessing and fit ridge regression to these frozen representations. The resulting predictor is a vector-valued function $f : \mathbb{R}^{512} \to \mathbb{R}^{10}$. Using the one-hot label vector $y \in \mathbb{R}^{10}$ as the response, we define the aggregate squared risk by summing over the 10 output dimensions:

$$R(f) = \mathbb{E}_{(x,y)}\left[\sum_{k=1}^{10}(y_k - f_k(x))^2\right], \tag{102}$$

where $f_k : \mathbb{R}^{512} \to \mathbb{R}$ is the ridge predictor for the $k$th component of the one-hot response.

### I.2 Linear probing experiments

For the linear probing experiments on Caltech-101, Caltech-256, and CIFAR-100, we extract features using a ResNet-34 [29] pretrained on ImageNet and available in PyTorch. We then train a linear classifier with a softmax output on top of these frozen features. Results appear in the right panel of Figure 2, Table 1, and Section H.2.

**Training details.** The teacher is trained by minimizing cross-entropy against the corrupted one-hot labels. The PD student is trained by minimizing cross-entropy against the teacher's softmax predictions. Each linear classifier is trained for 200 epochs using stochastic gradient descent with momentum 0.9 and a learning-rate decay factor of 0.98. The initial learning rates are $0.001$ for the teacher and $0.05$ for the student. We perform no data preprocessing, and all experiments are run on an NVIDIA T4 GPU.

**Caltech-101.** Caltech-101 [26] is an object recognition dataset with 101 object categories plus a background category. It contains approximately 9,000 images, with 40 to 800 images per category, and is therefore highly imbalanced. We draw disjoint subsets containing $n_t = 1500$ teacher training samples, $n_s = 4500$ fresh unlabeled samples, $n_{\mathrm{cal}} = 500$ calibration samples, and 2,000 test samples.

**Caltech-256.** Caltech-256 [26] contains 30,607 images spanning 256 object categories. We draw disjoint subsets containing $n_t = 5000$ teacher training samples, $n_s = 15000$ fresh unlabeled samples, $n_{\mathrm{cal}} = 1000$ calibration samples, and 6,000 test samples.

**CIFAR-100.** CIFAR-100 [38] contains 60,000 images from 100 classes. These classes are grouped into 20 superclasses, each containing five related fine-grained classes; for example, beaver, dolphin, otter, seal, and whale belong to the same superclass. We draw disjoint subsets containing $n_t = 5000$ teacher training samples, $n_s = 15000$ fresh unlabeled samples, $n_{\mathrm{cal}} = 1000$ calibration samples, and 9,000 test samples.

**Random corruption.** Under random label corruption with rate $\rho$, we select a fraction $\rho$ of the teacher's training samples uniformly at random and replace each label with one drawn uniformly from the remaining classes.

**Hierarchical corruption.** This scheme models errors between semantically similar classes. Because only CIFAR-100 provides the required superclass structure, we use hierarchical corruption only for this dataset. At corruption rate $\rho$, we select a fraction $\rho$ of the teacher's training samples uniformly at random and replace each label with one drawn uniformly from the other classes in the same superclass.

### I.3 Synthetic experiments

#### I.3.1 Ridge regression

For the synthetic ridge experiments in Figure 11 and the additional experiments in Section H.3, we use the following covariance and signal models.

- Isotropic covariance: $\Sigma = I_p$.
- One-spike covariance: $\Sigma = I_p + 5vv^\top$, where $v = g/\|g\|_2$ and $g \sim \mathcal{N}(0, I_p)$.
- AR1 covariance: $\Sigma_{ij} = \rho^{|i-j|}$ for all $i, j$, with correlation parameter $\rho = 0.25$.
- Isotropic signal: $\beta \sim \mathcal{N}(0, (r^2/p)I_p)$.
- Top-aligned signal: let $m \in (0, 1)$ be the alignment proportion, $a \in (0, 1)$ the alignment factor, and $k = \lfloor mp \rfloor$, where $1 \le k \le p - 1$. If $V = [v_1, \dots, v_p]$ contains the eigenvectors of $\Sigma$ in decreasing order of their eigenvalues, define

$$\Sigma_\beta := pV \operatorname{diag}\Big(\underbrace{\frac{a}{k}, \dots, \frac{a}{k}}_{k}, \underbrace{\frac{1-a}{p-k}, \dots, \frac{1-a}{p-k}}_{p-k}\Big) V^\top, \qquad \beta \sim \mathcal{N}\Big(0, \frac{r^2}{p}\Sigma_\beta\Big).$$

  Thus, a proportion $a$ of the expected signal energy lies in the top $k$ eigendirections. In Figure 11, we set $m = 0.1$ and $a = 0.9$.

The empirical curves in Figure 11 are averaged over 30 independent simulations.

#### I.3.2 Logistic regression

For the synthetic logistic regression experiments, we follow [17] and construct a block-diagonal kernel matrix

$$K := \operatorname{diag}(K_1, K_0) \in \mathbb{R}^{2n \times 2n},$$

where the two blocks correspond to the ground-truth classes and have zero cross-class correlation. Each block has unit diagonal. In the constant feature-correlation setting, every off-diagonal entry equals $c = 0.25$. In the uniform feature-correlation setting, the off-diagonal entries are taken from $(1/n)ZZ^\top$, where the entries of $Z \in \mathbb{R}^{n \times n}$ are drawn independently from the uniform distribution on $[0.3, 0.7]$. The second block is constructed analogously. Both constructions yield a positive semidefinite kernel matrix.

We train kernel logistic regression in the dual parameterization. For pseudo-labels $q = (q_1, \dots, q_{2n}) \in [0, 1]^{2n}$ and dual variables $\alpha \in \mathbb{R}^{2n}$, let $f = K\alpha$ and apply the sigmoid componentwise to obtain $\widehat{p} = \sigma(f)$. We minimize the regularized binary cross-entropy objective

$$\mathcal{L}_q(\alpha) \;=\; -\frac{1}{2n} \sum_{i=1}^{2n} \Big[ q_i \log \widehat{p}_i + (1 - q_i) \log(1 - \widehat{p}_i) \Big] + \frac{\lambda}{2}\, \alpha^\top K \alpha. \tag{103}$$

For the first PD student, $q_i = \widehat{y}_i$ is the teacher's hard pseudo-label; for the second PD student, $q_i = y_{\mathsf{pd}}(x_i)$ is the first student's soft prediction. This is the dual form of the training objectives in Equations (14) and (16). We minimize it using L-BFGS-B [11] with analytic gradients and initialize at $\alpha = 0$.

## NeurIPS Paper Checklist

1. **Claims**
   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?
   Answer: [Yes]
   Justification: All the claims made in the abstract and introduction are justified both theoretically and empirically in Section 3, Section 5 and Section 6.
   Guidelines:
   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.
2. **Limitations**
   Question: Does the paper discuss the limitations of the work performed by the authors?
   Answer: [Yes]
   Justification: The settings and assumptions are discussed and explained when we mention them.
   Guidelines:
   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.
3. **Theory assumptions and proofs**
   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?
   Answer: [Yes]
   Justification: All the assumptions for each of the results are included in the main text, and the complete proofs for each of the results are included in the appendix.
   Guidelines:
   - The answer [N/A] means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.

- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**
   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?
   Answer: [Yes]
   Justification: The supplementary materials include the code and instructions needed to reproduce the experimental results in this paper.
   Guidelines:
   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**
   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?
   Answer: [Yes]
   Justification: We use publicly available data. The code and instructions for reproducing experimental results in this paper are included in the supplementary materials.
   Guidelines:
   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.

- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**
   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?
   Answer: [Yes]
   Justification: The experimental details are included in the appendix.
   Guidelines:
   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.
7. **Experiment statistical significance**
   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?
   Answer: [No]
   Justification: We run the experiments over many datasets and corruption rates using only one single seed: seed 2026 for all experiments, not cherry-picked.
   Guidelines:
   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.
   - It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
   - For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
   - If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.
8. **Experiments compute resources**
   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?
   Answer: [Yes]
   Justification: The compute resources are described in the README file of the submitted code in the anonymous github link.
   Guidelines:
   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.

- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**
   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?
   Answer: [Yes]
   Justification: We conform with the NeurIPS Code of Ethics.
   Guidelines:
   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**
    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?
    Answer: [N/A]
    Justification: The paper provides a theoretical study and does not have immediate societal impacts.
    Guidelines:
    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
    - If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**
    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?
    Answer: [No]
    Justification: The paper does not pose any risks that require safeguards.
    Guidelines:
    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?
Answer: [Yes]
Justification: The paper correctly cites papers of related assets.
Guidelines:

- The answer [N/A] means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**
Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?
Answer: [No]
Justification: The paper does not release new assets.
Guidelines:

- The answer [N/A] means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**
Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?
Answer: [No]
Justification: The paper does not involve crowdsourcing or research with human subjects
Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**
Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?
Answer: [No]
Justification: The paper does not involve crowdsourcing or research with human subjects
Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.

- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [N/A]

    Justification: LLM is not used for core methodology, scientific rigor, or originality of the research.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.