WHETHER TO DISTILL ACROSS MODALITIES: AN EXACT GAIN IDENTITY AND A DECISION RULE

Anonymous authors
Paper under double-blind review

ABSTRACT

Cross-modal knowledge distillation trains a student in one modality using a teacher in another, but teacher accuracy alone does not determine whether distillation helps. Although mixing teacher outputs with labels can reduce the effect of noise in the training labels, teacher outputs may also introduce variation unexplained by the student's input and alter the student's bias. To characterize this trade-off, we derive an exact gain identity for students with a fixed linear solution map under squared loss. The gain over supervised training on labels alone is quadratic in the mixing weight, and we identify conditions under which six scalars determine the gain curve and optimal weight. We further prove that criteria which ignore the student's labels do not determine whether distillation is beneficial. We therefore develop a gain estimator that reuses the student's existing labeled data together with unlabeled pairs to decide whether and how much to distill, without additional labels or a validation sweep. Experiments on six real-world multi-modal datasets show that existing distillation methods can degrade student performance, whereas our gain-based rule reduces that degradation and, in our main policy comparison, improves on supervised training on average.

1 INTRODUCTION

Cross-modal knowledge distillation (CMKD) trains a student in one modality using a teacher in another, while requiring only the student modality at inference (Gupta et al., 2016). We consider mixing labels with teacher outputs (Hinton et al., 2015), where $\lambda \in [0, 1]$ controls the teacher's contribution and $\lambda = 0$ gives supervised training on labels alone. Although teacher accuracy provides a natural guide to teacher selection, a more accurate teacher need not produce a better student (Cho & Hariharan, 2019; Mirzadeh et al., 2020; Xue et al., 2022). Moreover, an unsuitable mixing weight λ can make the student worse than supervised training. Prior studies have examined when cross-modal transfer is beneficial (Xie et al., 2025; Kwon & Ham, 2026). We characterize the gain as a function of λ and identify the information needed to decide whether and how much to distill.

A key difficulty is that the teacher and student can rely on different information. For example, a video teacher may use visual cues that improve prediction but cannot be recovered by an audio student. Although mixing teacher outputs with labels can reduce the effect of noise in the labels (Menon et al., 2021), teacher outputs may also introduce variation unexplained by the student's input. Moreover, the systematic discrepancy between labels and teacher outputs within the student's own view can either offset or amplify the bias already present under supervised training. Consequently, teacher accuracy alone is insufficient to assess the benefit of CMKD.

To characterize these effects, we derive an exact gain identity for students with a fixed linear solution map under squared loss, expressing the reduction in population risk relative to supervised training as a quadratic in λ. We identify conditions under which six scalars determine the gain curve and optimal weight λ^* (Propositions 1 and 2). We further show that the teacher's risk and the label noise do not determine whether CMKD is beneficial (Theorem 1). Even a *label-blind* criterion, which may use everything except the labels, cannot determine it (Theorem 2).

We therefore develop a gain estimator that reuses the student's existing labeled data together with unlabeled pairs in which both modalities are observed. The resulting rule selects λ while accounting for estimation uncertainty and returns $\lambda = 0$ when no improvement is predicted. The procedure requires no additional labels and no validation sweep.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

We evaluate the analysis and decision rule on six real-world multi-modal datasets. The experiments support the gain analysis and show that fixed λ values and existing distillation methods can degrade student performance. In the main policy comparison, our rule reduces degradation from distillation with a fixed λ and yields positive average gains over supervised training. The gain-based decision also reduces performance degradation when applied to the evaluated distillation objectives.

Our contributions are summarized as follows.

1. We derive an exact gain identity and a six-scalar characterization of the gain curve and optimal weight under explicit conditions. We extend the analysis to smooth losses through a local quadratic surrogate.
2. We prove two impossibility results: the teacher's risk and the label noise do not determine whether CMKD is beneficial, and a label-blind criterion cannot determine it.
3. We develop a gain-based decision rule requiring no additional labels or validation sweep and evaluate the rule on six multi-modal datasets.

2 Related Work

Statistical perspectives. Teacher accuracy and fidelity do not fully explain student generalization (Cho & Hariharan, 2019; Mirzadeh et al., 2020; Stanton et al., 2021). Menon et al. (2021) study class-probability estimation and bias–variance trade-offs, while Dao et al. (2021) develop cross-fitting and loss correction through semiparametric inference. Their fixed teachers have no output variation given the shared input, whereas our gain identity includes teacher-output variation unexplained by the student's input, its covariance with label noise, and changes in student bias. Safaryan et al. (2023) derive a closed-form weight minimizing a term controlling the SGD convergence neighborhood, while PPI++ tunes a correction weight to minimize the asymptotic variance of statistical estimates (Angelopoulos et al., 2023). Our exact gain identity instead targets the fitted student's population risk under squared loss with a fixed linear solution map.

Distillation objectives. Methods match intermediate features (Romero et al., 2015; Tian et al., 2019) or modify output matching and weighting (Zhao et al., 2022; Huang et al., 2022; Li et al., 2023; Sun et al., 2024; Zhou et al., 2021). WTTM uses sample-adaptive matching of transformed teacher outputs (Zheng & Yang, 2024), while PLD incorporates labels into a ranking objective (Bassam et al., 2026). Selective CMKD also limits unreliable transfer using uncertainty or reliability estimates (Ding et al., 2024; Hou et al., 2026). Our rule sets the mixing weight from the estimated gain rather than introducing another matching objective.

Teacher selection and transfer criteria. The modality focusing hypothesis emphasizes task-relevant teacher information shared across modalities (Xue et al., 2022), while NTDH examines teacher–student disagreement on non-target classes (Chen et al., 2024). CCH compares teacher–student and student–label mutual information to analyze when CMKD helps in a Gaussian asymptotic setting (Xie et al., 2025). Teacher-ranking methods use logit ratios (Gosal et al., 2026) or gradient-based scores (Panigrahi et al., 2026). Teacher quality is also studied in adversarial distillation (Lee & Chung, 2026), language-model pretraining (Lu & Liu, 2026), and reasoning distillation (Yang et al., 2026). We adapt ranking scores to distill-or-not decisions and score a gradient-covariance proxy in GRACE's spirit rather than GRACE itself (Section 6.2 and Appendix F).

Privileged information and cross-modal transfer. Learning with privileged information uses training-only inputs (Vapnik & Vashist, 2009; Lopez-Paz et al., 2015), while CMKD transfers supervision through paired modalities (Gupta et al., 2016). Recent methods construct surrogate targets (Xia & Wainwright, 2024) or jointly train teachers and students (Shi et al., 2026). Closer to our setting, Kwon & Ham (2026) analyze when conditional-mean transfer from unlabeled pairs and residual correction with target labels improve prediction.

We do not claim conditional-mean projection, cross-fitting, or label-based correction as new (Dao et al., 2021; Angelopoulos et al., 2023; Kwon & Ham, 2026). The weight-selection results above choose a weight to minimize a stated quantity, and the nearest analyses of cross-modal transfer give conditions under which it helps, asymptotically (Xie et al., 2025) or for the conditional-mean transfer above. Our focus is the information needed to make that choice. We characterize the gain for a fitted student, establish which information cannot determine whether CMKD is beneficial, and develop an estimator using the student's existing labels.

3 THE DISTILLATION GAIN IDENTITY

Training data and student. Over n training samples observed in both modalities, let $X^b \in \mathbb{R}^{n\times d}$ collect the student's inputs and let X^a collect the teacher's inputs. Write $y, t \in \mathbb{R}^n$ for the labels and the teacher's outputs, each teacher output computed from its row of X^a. The student has a linear solution map

$$\hat{\theta}_\lambda = L\big((1-\lambda)y + \lambda t\big), \qquad L \in \mathbb{R}^{d\times n}, \qquad \lambda \in [0,1],$$

with L depending on X^b but not on y, t or λ, as in ridge regression with a fixed penalty. At $\lambda = 0$ the student is trained on labels alone and at $\lambda = 1$ on teacher outputs alone.

Conditional means and residuals. Let $h = y - t$ be the label–teacher difference, so that $\hat{\theta}_\lambda = L(y - \lambda h)$. We hold X^b and the teacher fixed, which also fixes L. Expectations are then over y and t. Here t is random despite the fixed teacher, because it depends on X^a, which X^b does not determine. Split y and h into their means and residuals, $y = m_y + u_y$ and $h = m_h + u_h$ with $m_y = \mathbb{E}[y]$ and $m_h = \mathbb{E}[h]$. The second moments are $\Omega_{yy} = \mathbb{E}[u_y u_y^\top]$, $\Omega_{hh} = \mathbb{E}[u_h u_h^\top]$ and $\Omega_{yh} = \mathbb{E}[u_y u_h^\top]$. We assume only that these moments are finite, and allow u_y and u_h to be correlated.

Population risk and gain. For a test pair (x_0^b, y_0), let $\mathcal{R}(\theta) = \mathbb{E}[(y_0 - (x_0^b)^\top\theta)^2]$ be the risk of the linear predictor θ, averaged over the test pair. Let θ^* be the best linear predictor minimizing $\mathcal{R}$. The expected risk of the fitted student $\hat{\theta}_\lambda$ and its gain over supervised training ($\lambda = 0$) are

$$R(\lambda) = \mathbb{E}\big[\mathcal{R}(\hat{\theta}_\lambda)\big], \qquad G(\lambda) = R(0) - R(\lambda),$$

where the expectation runs over y and t, on which $\hat{\theta}_\lambda$ depends. A positive gain means that distillation improves on the supervised baseline. Let $\Sigma = \mathbb{E}[x_0^b (x_0^b)^\top]$ be the test-input second moment and write $\|\cdot\|_\Sigma$ for the norm it induces. The supervised student has mean Lm_y, so let $b = Lm_y - \theta^*$ be its bias, and write $K = L^\top \Sigma L$. Excess risk is a squared distance, $\mathcal{R}(\theta) - \mathcal{R}(\theta^*) = \|\theta - \theta^*\|_\Sigma^2$, so $R(\lambda) = \mathcal{R}(\theta^*) + \mathbb{E}\|\hat{\theta}_\lambda - \theta^*\|_\Sigma^2$ and $\mathcal{R}(\theta^*)$ cancels in G (Appendix B).

Proposition 1 (The gain in matrix form)**.** *The supervised excess risk R_0 and the gain $G(\lambda)$ are*

$$R_0 = R(0) - \mathcal{R}(\theta^*) = b^\top \Sigma b + \operatorname{tr}(K\Omega_{yy}), \tag{1}$$

$$G(\lambda) = 2\lambda A - \lambda^2 D, \tag{2}$$

$$A = b^\top \Sigma L m_h + \operatorname{tr}(K\Omega_{yh}), \qquad D = m_h^\top K m_h + \operatorname{tr}(K\Omega_{hh}) \geq 0. \tag{3}$$

Cauchy–Schwarz gives $A^2 \leq R_0 D$, so $D = 0$ forces $A = 0$ and every weight has zero gain, in which case we select $\lambda^* = 0$. For $D > 0$ the optimal weight on $[0,1]$ is $\lambda^* = [A/D]_{[0,1]}$.

Appendix B gives the proof. The identity is exact rather than a local approximation, so λ^* is exactly the clipped ratio of A to D. It also holds coordinatewise for multiple outputs, with a single shared weight determined by the summed A and D.

The mean discrepancy is $m_h = m_y - q$, where $q = \mathbb{E}[t \mid X^b]$ is the teacher's output projected onto the student's view. Write $u_t = t - q$ for the part of the teacher's output that X^b does not explain, so that $u_h = u_y - u_t$. Suppose the residuals are uncorrelated across samples, with the same moments in each sample i, $\sigma_y^2 = \operatorname{Var}(u_{y,i})$, $\sigma_t^2 = \operatorname{Var}(u_{t,i})$ and $\sigma_{yt} = \operatorname{Cov}(u_{y,i}, u_{t,i})$. Then $\Omega_{yy} = \sigma_y^2 I$ and, since $u_h = u_y - u_t$, $\Omega_{yh} = (\sigma_y^2 - \sigma_{yt})I$ and $\Omega_{hh} = (\sigma_y^2 - 2\sigma_{yt} + \sigma_t^2)I$. Let $T = \operatorname{tr}(K)$ and let $\psi = Lm_h$ be the mean discrepancy filtered through the learner L, giving

$$\begin{aligned} A &= T(\sigma_y^2 - \sigma_{yt}) + \beta, \qquad & D &= T(\sigma_y^2 - 2\sigma_{yt} + \sigma_t^2) + \delta, \\ \beta &= b^\top \Sigma \psi, & \delta &= \psi^\top \Sigma \psi. \end{aligned} \tag{4}$$

Corollary 1 (Conditions for a beneficial weight)**.** *Under the conditions of Proposition 1, let $D > 0$. By Eq. 2, a beneficial weight λ exists exactly when $A > 0$, while a given $\lambda > 0$ is beneficial exactly when $2A > \lambda D$.*

Remark 1 (The normalized gain)**.** *Let $D > 0$, $\lambda^* = A/D \in (0,1]$ and let $z = \lambda/\lambda^*$ be the normalized weight. Rearranging Eq. 2 gives the normalized gain $G(\lambda)/G(\lambda^*) = 2z - z^2$.*

Proposition 2 (A six-scalar characterization of the gain)**.** *Under the conditions of Proposition 1 and the per-sample moment conditions for Eq. 4, the six scalars* $(T, \sigma_y^2, \sigma_{yt}, \sigma_t^2, \beta, \delta)$ *determine the gain curve* G *and the optimal weight* λ^**, taking* $\lambda^* = 0$ *at* $D = 0$*. That is, problems that share all six have the same* G *and* λ^**.*

Corollary 2 (The same-view case)**.** *Under the conditions of Proposition 2, suppose the teacher is a function of the student's view, with* $T\sigma_y^2 > 0$*,* $\delta > 0$*. Then* $q = \mathbb{E}[t \mid X^b] = t$*, the teacher's residual is* $u_t = t - q = 0$*,* $\sigma_t^2 = \sigma_{yt} = 0$ *and the mean discrepancy is* $m_h = m_y - t$*. Eq. 4 then gives*

$$\lambda^* = \left[\frac{T\sigma_y^2 + \beta}{T\sigma_y^2 + \delta}\right]_{[0,1]}, \qquad \psi = Lm_h = L(m_y - t). \tag{5}$$

For a fixed student, λ^* *depends on the teacher only through* ψ*.*

Appendix B proves this, and shows there that the weight does not rise when the discrepancy is scaled up. In the same-view setting the teacher's squared risk on the training design is $\sigma_y^2 + n^{-1}\|m_h\|^2$, which involves m_h only through its magnitude, while λ^* depends on $\beta = b^\top \Sigma \psi$, which involves its direction. Reversing m_h keeps the risk unchanged and flips the sign of β. Therefore two teachers of identical squared risk and accuracy can have optimal weights 0 and 1, already in the setting of Menon et al. (2021) and Safaryan et al. (2023). In the CMKD setting even the same ψ is not enough, since two teachers can still differ in σ_{yt} and σ_t^2, on which A and D also depend (Appendix E).

Regret and the decision to distill. A and D must themselves be estimated, since the six scalars of Proposition 2 are population quantities. Call $G(\lambda^*) - G(\lambda)$ the *regret* of a weight λ. For $\lambda^* = A/D \in (0, 1]$, completing the square in Eq. 2 gives

$$G(\lambda^*) - G(\lambda) = D(\lambda - \lambda^*)^2, \tag{6}$$

an exactly quadratic penalty on the weight error. Not distilling is the weight $\lambda = 0$, with regret $G(\lambda^*) = D\lambda^{*2}$, whereas $\lambda = (1+\varepsilon)\lambda^*$ has regret $\varepsilon^2 D\lambda^{*2}$ (Proposition 3). We show that an estimate $\hat{\lambda}$ is worse in expectation than not distilling exactly when $\mathbb{E}[(\hat{\lambda} - \lambda^*)^2] > \lambda^{*2}$ (Proposition 4). Appendix C proves both propositions.

Fixed and estimated weights. Propositions 3 and 4 quantify regret on the gain curve. Selecting the weight from the student's training data introduces covariance terms those results do not cover. Section 5 therefore gives a decision rule for this setting, which Section 6 evaluates on real tasks.

4 The Information a Criterion Needs

Section 3 identifies the quantities determining the optimal weight and reduces the decision to distill to the sign of A (Corollary 1). From Eq. 4, A splits into two terms, $T(\sigma_y^2 - \sigma_{yt})$ for how much noise distilling removes and $\beta = b^\top \Sigma \psi$ for which way distilling shifts the student's bias. Distilling shifts the fitted student by $-\lambda\psi$ in expectation, moving its bias to $b - \lambda\psi$. β is positive exactly when that shift points back toward θ^*, a matter of the direction of b rather than its size. We now ask whether a selection criterion can reach β, and answer with two impossibility results.

To make that question precise, we restrict attention to criteria that use no additional labels, since those would allow a validation sweep. Within that restriction, a *label-blind* criterion is a function of (X^a, X^b, t) alone, using no labels and nothing fitted to them.

Impossibility when sharing the risk and the discrepancy. On a balanced design $x \in \{-1, +1\}$ with a ridge slope, we construct two problems that share the teacher's risk, the label noise σ_y^2, the mean teacher–label mismatch $\delta_{\mathrm{raw}} = n^{-1}\|m_h\|_2^2$, the supervised excess risk R_0 and the filtered discrepancy ψ itself. They differ in the sign of the labels' dependence on x and in the teacher, which flips the direction of b while leaving ψ unchanged. Their optimal weights λ^* are then 0 and 1, and the shared quantities therefore do not determine whether to distill. A selector reading only them must answer the same way in both.

Theorem 1 (Indistinguishable summaries)**.** *For every even* $n > 3$ *there exist two problems that share the filtered discrepancy* ψ*, the teacher risk, the supervised excess risk* R_0 *and* $(T, \sigma_y^2, \sigma_{yt}, \sigma_t^2, \delta_{\mathrm{raw}})$*.*

They differ in β, and their optimal weights are 0 and 1. Let $\hat{\lambda}$ be any selector reading only the shared quantities, randomized or not, and write G_+ for $G(\lambda^*)$ in the problem with $\lambda^* = 1$. In the worse of the two problems its expected regret satisfies

$$\frac{\mathbb{E}\big[G(\lambda^*) - G(\hat{\lambda})\big]}{G_+} \;\geq\; \frac{9(n-1)^2}{16n^2}.$$

Appendix B gives the proof. The bound is 43% at $n = 8$ and rises with n toward $9/16 \approx 56\%$. Roughly half of the gain that was available is lost, whatever the criterion answers.

Impossibility when sharing the teacher and the inputs. The teachers of Theorem 1 have different outputs, so a criterion reading the teacher directly would separate its problems. On the three-point design $x \in \{-1, 0, 1\}$ with a ridge slope, we construct two problems in which one teacher reads the same inputs as the student, $X^a = X^b = X$, so the joint distribution of (X^a, X^b, t) is identical in both and a label-blind criterion sees nothing that separates them. They differ only in the means of their Bernoulli labels, which moves β and δ. Their optimal weights λ^* are again 0 and 1, and no function of (X^a, X^b, t) therefore determines whether to distill.

Theorem 2 (Indistinguishable teachers)**.** *There exist two problems that share the entire joint distribution of (x, t), together with the supervised excess risk R_0, the teacher risk, the noise term $\mathrm{tr}(K\Omega_{yy})$ and $(T, \bar{\sigma}_y^2, \sigma_{yt}, \sigma_t^2, \delta_{\mathrm{raw}})$, where $\bar{\sigma}_y^2$ is the conditional variance averaged over samples rather than a common one. They differ in β and in δ, and their optimal weights are 0 and 1. Let $\hat{\lambda}$ be any selector reading only these shared quantities and write G_+ for $G(\lambda^*)$ in the problem with $\lambda^* = 1$. In the worse of the two its expected regret satisfies*

$$\frac{\mathbb{E}\big[G(\lambda^*) - G(\hat{\lambda})\big]}{G_+} \;\geq\; \frac{68131 - 1533\sqrt{1302}}{20628} \approx 0.621.$$

Appendix B gives the proof, the shared values and the constant in closed form. Here β and δ both differ, and the shared δ_{raw} does not determine either. A criterion that may read everything except the labels, however many unlabeled pairs it draws, still loses 62% of the available gain.

What a criterion needs. The two results show what is missing from the information they fix, and our rule uses the student's existing labels to estimate β and the other scalars of Eq. 4. Reaching β needs the discrepancy m_h, which the labels supply. Cross-modality adds σ_{yt} and σ_t^2, which the same-view case sets to zero (Corollary 2) and teacher accuracy does not disentangle. Of the three, $\sigma_t^2 = \mathbb{E}[\mathrm{Var}(t \mid X^b)]$ is free of the labels and the other two are not. Section 5 proposes such a criterion, reading the labels the student already trains on.

5 A Gain-Based Rule for Whether to Distill

Section 4 motivates using the student's labels to estimate the gain, and the bias alignment β in particular. Since β is a population quantity, we estimate it, together with the other five scalars of Eq. 4, using the student's n labeled samples and separate unlabeled pairs. Our rule therefore needs no additional labels, and returns a weight, or 0 when the estimated A is not positive. Algorithm 1 gives the rule in pseudocode, Appendix A states it in full, and the rest of this section explains it.

Estimation of the scalars. Cross-fitted ridges on the fixed student view X^b give $\hat{m}_y$ for $m_y = \mathbb{E}[y]$ and $\hat{m}_h$ for the mean discrepancy $m_h = \mathbb{E}[y - t] = \mathbb{E}[y - t \mid X^b]$. $\hat{\beta}$ follows from $\hat{m}_h$ and the student's bias $\hat{b}$ by Eq. 4. The unlabeled pairs provide the bridge $\hat{q}$ for $q = \mathbb{E}[t \mid X^b]$, which gives $\hat{\sigma}_t^2$ and $\hat{\sigma}_{yt}$ through the residual $t - \hat{q}$ and $\hat{\delta}_{\mathrm{raw}}$ through $\hat{m}_y - \hat{q}$. Therefore every quantity comes from the n labeled samples, or from unlabeled pairs and teacher outputs.

Compute and constants. Our rule trains the student once, or five times where T is measured, while a validation sweep trains one student per point of its weight grid, seven points in our runs. On a ridge student that difference narrows, since a fixed linear map makes the weight path affine and two fits generate the whole grid, but the sweep still takes its validation split out of the same n the student trains on, which our rule leaves to training. Against those savings, our rule has *one* constant,

Algorithm 1: Pseudocode of the proposed rule, stated in full in Appendix A

Input: labeled samples (X_ℓ, y) with the teacher's outputs t_ℓ on them, and unlabeled pairs X_u with the teacher's outputs t_u

Output: a weight in the swept grid $\Lambda \subset [0, 1]$

*// **estimate** the scalars that Eq. 4 needs*

$\hat{q} \leftarrow$ ridge of t_u on X_u; $\hat{\sigma}_t^2 \leftarrow$ residual of that fit *// the bridge, no labels*

$\hat{\sigma}_y^2, \hat{\sigma}_{yt}, \hat{\delta}_{\mathrm{raw}} \leftarrow$ cross-fitted on (X_ℓ, y, t_ℓ) with $\hat{q}$

$\hat{m}_h \leftarrow$ cross-fitted regression of $y - t_\ell$ on X_ℓ; $\hat{\beta} \leftarrow \hat{b}^\top \Sigma L \hat{m}_h$ *// $\hat{b}$ is the student's bias*

$T \leftarrow \mathrm{tr}(L^\top \Sigma L)$ in closed form, by a ridge proxy, or by half-fits *// the student's variance factor*

*// **decide** whether any weight is beneficial*

$\hat{A} \leftarrow T(\hat{\sigma}_y^2 - \hat{\sigma}_{yt}) + \hat{\beta}$; $\hat{D} \leftarrow T(\hat{\sigma}_y^2 - 2\hat{\sigma}_{yt} + \hat{\sigma}_t^2) + \hat{\delta}_{\mathrm{raw}}$ *// Eq. 4*

if $\hat{D} \le 0$ *or* $\hat{A} \le 0$ **then return** 0 *// do not distill*

*// **choose** the weight, then discount it for its own noise*

$\hat{\lambda} \leftarrow \left[\hat{A}/\hat{D}\right]_{[0,1]}$

if *the teacher's targets go on the unlabeled pairs* **then**

  $\hat{w} \leftarrow \left[\hat{A}/(\hat{D} + \gamma\hat{A})\right]_{[0,9]}$; $\hat{\lambda} \leftarrow \hat{w}/(1 + \hat{w})$

$\hat{s} \leftarrow \hat{A}^2/(\hat{A}^2 + \hat{\sigma}_{\mathrm{j}}^2 + \hat{b}_{\mathrm{j}}^2)$ *// the shrinkage, by jackknife*

return the nearest point of the swept grid Λ to $\hat{s}\hat{\lambda}$ *// the shrunk weight*

the ridge penalty ρ_b of the bridge $\hat{q}$, set to 100 throughout, with the fold counts fixed rather than tuned. Appendix H.7 sweeps it over four decades and the mean gain stays positive at every value, and the constant therefore does not carry the result, although the tail does move. The student's own penalty ρ_s is not counted, since it defines the learner whose risk the identity describes.

The rule and the alarm. We call *the rule* the shrunk weight $\hat{s}\hat{\lambda}$ that Algorithm 1 returns after projecting it onto the grid, and Section 6 reports it as our policy. An associated *alarm* is the sign test $\hat{G}(\lambda) = 2\lambda\hat{A} - \lambda^2\hat{D} < 0$ (Corollary 1), which indicates whether to distill at a chosen weight. Section 6 measures it on its own.

What the rule substitutes. The shrinkage needs the mean squared error of $\hat{\lambda}$ about λ^*, which the rule cannot evaluate, so it uses a jackknife contrast as a proxy for that error (Proposition 4). The denominator uses $\delta_{\mathrm{raw}} = n^{-1}\|m_h\|_2^2$ in place of δ, which changes the mean by 0.012 points against the unbiased alternative of Appendix D. The identity conditions on q while the procedure estimates it, which the disjoint folds make legitimate, and that buys separation and not accuracy, since every nuisance error remains. Whether the estimates still decide well is therefore a matter of measurement, and Section 6 measures it on six datasets.

6 Experiments

Experimental settings. We evaluate on UCI-HAR (Anguita et al., 2013), RAVDESS (Livingstone & Russo, 2018), NYU-Depth v2 (Silberman et al., 2012), VGGSound (Chen et al., 2020), MIMIC-CXR (Johnson et al., 2019) and Parkinsons telemonitoring (Tsanas et al., 2009), spanning audio, video, image, radiology text, depth and inertial sensing, and contributing sixteen teacher-to-student directions with both directions of every dataset included. Unless stated otherwise a gain is the reduction in test squared error relative to supervised training, as a percentage of that baseline. The cross-entropy and published-objective comparisons are in accuracy percentage points. A *cell*, or *setting*, is one configuration of direction, teacher, student, label budget and seed. Every cell is swept over the whole weight grid, with a separately trained student at each weight, and n counts *student* labels throughout. Appendix F gives the full setup and the populations, Appendix A lists our rule's variants by run, and Appendix G collects the tables this section cites but does not present.

Experimental questions. We ask first how the gain distillation offers compares with the loss a fixed weight risks, whether Section 3's curve describes that trade in the data, and whether teacher accuracy locates the settings where the loss falls (Section 6.1). We then ask how well the sign of the estimated gain decides, and whether criteria without labels or published objectives match it (Section 6.2). Finally we ask whether acting on that sign beats the customary weight and a validation sweep (Section 6.3), and how far that holds outside the pooled design (Section 6.4).

Regime	Cells	Best λ	λ=0.5	Ratio
linear student (ridge)	340	$+2.26\%$	-17.35%	$7.7\times$
neural student (512–256 MLP)	120	$+5.40\%$	-10.99%	$2.0\times$
teacher family varied (five classes)	420	$+1.73\%$	-36.94%	$21.3\times$
student size varied (23K–2.8M)	160	$+5.02\%$	-12.75%	$2.5\times$
teacher label budget varied (40–1280)	360	$+2.09\%$	-19.68%	$9.4\times$
end-to-end student (1D CNN, raw signals)	104	$+6.00\%$	-18.95%	$3.2\times$

Table 1: **The gain at the best weight and at the customary $\lambda = 0.5$.** One row per regime, each entry averaging $100(R_{\text{sup}} - R)/R_{\text{sup}}$ over that regime's cells, with R the unclipped squared error. *Ratio* is the size of the loss against the gain, from 2.0 to 22.7 times here and in the two pair settings of Table 15, with a median of 8.6. Appendix F gives the regimes in full and Appendix H.14 repeats the comparison under clipped predictions.

6.1 The Asymmetry and the Limits of Teacher Accuracy

We first measure the gain at the best weight, chosen per setting by a full grid search, and at the customary $\lambda = 0.5$. Table 1 reports both in six regimes, which vary the student, its size, the teacher family and the teacher's label budget. The best weight gains little, from $+1.7\%$ to $+6.0\%$ of the student's risk across the six regimes, with a median of $+1.12\%$ over their $1{,}504$ cells. The fixed weight loses several times as much, -11% to -37%, and degrades the student on average in every one of the six regimes. In every regime the loss at $\lambda = 0.5$ is at least $2.0\times$ the gain at the best weight. We call this gap the *asymmetry*, and it is not particular to one student, teacher or budget.

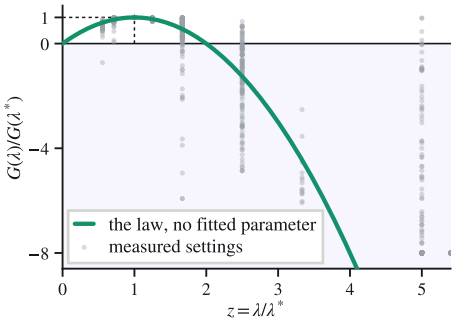


Figure 1: **The predicted curve against the measurements.** The line is Remark 1, with no fitted parameter, and the measurements are normalized by each setting's grid optimum. Each of the $1{,}048$ settings sits at the customary $\lambda = 0.5$, and the 170 settings past the frame are drawn on its edge.

Section 3 accounts for that asymmetry, and we test it directly. Remark 1 writes the gain as $2z - z^2$ of its own maximum, where $z = \lambda/\lambda^*$. The gain is therefore capped at that maximum, while beyond $2\lambda^*$ the loss can greatly exceed it. The curve has no fitted parameter, and over the $1{,}048$ cells of these six regimes in which some weight helps, the predicted and the measured gains agree in rank at Spearman 0.953 (Figure 1 and Table 31). On the 144 paired cells that meet its premise exactly, a ridge student taking the teacher's targets on its own samples (Appendix F), the measured and the predicted curves agree to $R^2 = 1.000$ (Table 15).

A more accurate teacher does not make distilling safe. The customary weight degrades the student in 37% of the 605 cells whose teacher is more accurate than the student (Table 9), and teacher accuracy does not identify which cells those are. Over the 113 cells whose teacher is 10 to 40% more accurate, the gain at $\lambda = 0.5$ runs from -8.7% to $+35.2\%$ (Table 6), and nine matched pairs of teachers have outcomes of opposite sign (Figure 3). By Section 3 the bias alignment β can change the decision even at equal teacher risk, and computing it needs the labels. Section 6.2 measures how far criteria without labels can stand in.

6.2 Empirical Limits of Label-Blind Criteria

Decision criteria. We next ask which signals decide whether to distill. Table 2 compares our alarm with twelve baselines by AUC on the same 144 cells, each asked whether distilling at $\lambda = 0.5$ will make the student worse. Our alarm, the sign of the estimated gain, scores best at 0.834 among all thirteen, including the five published criteria it is scored against. The second best is the plain "is the teacher more accurate than the student" baseline, at 0.775. The alarm also reaches 0.863 over all 802 cells, beyond the 144 on which every criterion can be computed.

Section 4 proved that teacher accuracy cannot determine the decision, and here it is the second best signal, which is not a contradiction. Theorem 1 is a worst-case construction, while Remark 2 shows, in the same-view case, how a small discrepancy bounds $|\beta|$, the term that can flip the sign of A. Teacher accuracy is therefore a usable signal where the discrepancy is small, and this paper

Generic signal	AUC	Published criterion	AUC
teacher / student risk	0.775	secondary soft probs. (SSP) (Tan & Liu, 2022)	0.742
view similarity (CKA)	0.742	gradient covariance, GRACE-like (Panigrahi et al., 2026)	0.718
teacher confidence	0.711	NTDH (Chen et al., 2024)	0.724
prediction correlation	0.699	CCH (Xie et al., 2025)	0.693
teacher–student agreement	0.638	Ratio 1-2 (Gosal et al., 2026)	0.526
teacher margin	0.636		
teacher output spread	0.600	**ours:** $\hat{G}(0.5)$	0.834

Table 2: **Evaluation of decision criteria.** The same 144 cells and one question, whether distilling at $\lambda = 0.5$ will make the student worse. AUC needs a continuous score, so ours is $-\hat{G}(0.5)$ rather than its sign, and each baseline is scored in the direction, high or low, that predicts the loss better.

		544 live settings			all 802		
Policy	Fits	Mean	Worst dir.	Hurt	Mean	Worst dir.	Hurt
always, $\lambda = 0.5$	1	-3.72%	-89.78%	5/16	-26.47%	-172.50%	11/16
never distill	1	$+0.00\%$	$+0.00\%$	0/16	$+0.00\%$	$+0.00\%$	0/16
skip if $\hat{G}(0.5) < 0$	1	$+1.55\%$	-2.20%	1/16	$+0.61\%$	-1.89%	6/16
ours, shrunk	**1**	$+1.55\%$	$+0.00\%$	0/16	$+0.89\%$	-0.47%	4/16
best per cell, grid search	8	$+3.64\%$	$+0.75\%$	0/16	$+2.47\%$	$+0.32\%$	0/16

Table 3: **Policy comparison.** *Mean* and *worst dir.* are the change in the student's risk, averaged over the cells and the worst of the sixteen directions, *hurt* counts the harmed directions, and *fits* the student trainings each policy needs. **Left:** the 544 of 802 settings in which some weight beats the supervised student, a restriction that conditions on the oracle. **Right:** all 802. Appendix F states how that population is composed.

adds the reason it fails elsewhere. The accuracy threshold has to be measured somewhere, and its quality tracks the labels it is given, 0.775 on a labeled test set, 0.602 on the only fold a cross-modal practitioner has, and 0.738 when bought at 20% of the student's labels (Table 10). Our alarm reads the labels the student already trains on, and its 0.834 spends none of them. This ordering is consistent with Theorem 2, since every label-blind criterion in the table scores below our alarm. On the 340 main-ridge settings, removing $\hat{\beta}$ reduces the alarm's AUC from 0.831 to 0.804 and the unshrunk policy's mean gain by 0.70 points (Table 11).

Published objectives. A better distillation objective does not supply the decision either. We implement seven published objectives at a common temperature, with one exception, and unit scale (Appendix F), and every one is worse than not distilling on all six directions, by 3.69 to 22.77 accuracy points (Table 20). Picking the best weight for each cell after the fact does not repair the average. PLD (Bassam et al., 2026) carries no distillation weight at all, and it too makes the student worse in all six directions. What these methods lack is therefore a decision rather than a better weight. Applied on top of them, our alarm reduces the loss in every case, though never into a gain (Appendix G).

6.3 Policy Evaluation of the Decision Rule

We now run our rule as a policy, declining when $\hat{A}$ is not positive and otherwise distilling at the weight it returns, as Algorithm 1 states. Table 3 compares it with distilling always at $\lambda = 0.5$, with never distilling, and with the best weight per cell from a full grid search. Over the 802 cells distilling at $\lambda = 0.5$ raises the student's risk by 26.47% on average. Distilling at $\lambda = 0.5$ only when $\hat{G}(0.5) > 0$ lifts that average to $+0.61\%$, and our rule, which also chooses the weight and shrinks it, reaches $+0.89\%$. Paired comparisons across these cells also favour our rule, with Wilcoxon signed-rank $p = 6.9 \times 10^{-52}$ against always distilling and $p = 6.5 \times 10^{-36}$ against never distilling. Cells within a dataset are not independent draws, so we read these p-values as consistency checks, and a bootstrap over the six source datasets puts the mean at $[+0.29, +2.15]$ (Appendix H.12).

The average is not a guarantee, and our rule still loses on 147 of the 802 cells and on four of the sixteen directions. Nevertheless the positive average is not the work of one dataset or one weighting, since weighting the six datasets equally rather than the cells still gives $+1.39\%$ and deleting a dataset or a teacher family at a time leaves the sign unchanged (Appendix H.12). Restoring every cell the pooled design excludes leaves the pooled mean at $+0.89\%$ (Table 5).

Against a validation sweep. The natural alternative to a fixed $\lambda = 0.5$ is a sweep over the same grid on a validation split. At a matched label budget that split comes out of the student's n labels, and over three budgets the sweep loses 1.67 points to our rule ($p = 1.7 \times 10^{-13}$, Table 12). Given those labels back and a refit, the sweep draws level on the mean, -0.14 points at $p = 0.554$, which we read as a small observed difference rather than as equivalence. Even then its worst cell is more than twice as bad, -20.77% against our -9.69%. The fit counts in the table are what our implementation runs, not what a sweep requires, since the weight path is affine on this student (Section 5). The split size is a further choice, worth up to 1.69 points and made with no labels left to make it on (Table 13). What a sweep costs and our rule does not is therefore the split and its size, with extra fits where the solution map is not affine.

Decision against weight. The value comes from the decision more than from the weight. Replacing our weight with the best one lifts the mean to 87% of what an oracle decision and an oracle weight reach together, while replacing our decision instead reaches 64%, against 43% for the rule as deployed (Table 28). Each swap repairs every error in the half it replaces, so the split between the two depends on which half is repaired first (Appendix H.10).

6.4 Scope and Boundary Conditions

Table 8 asks how far the policy claim of Table 3 reaches beyond its pooled design. Several regimes outside that design retain positive gains, while two incur losses. In those two our rule distills and loses, though it limits the damage, -10.32% against -35.76% for $\lambda = 0.5$ in the worse of the two, where the teacher is $3.26\times$ worse than the student. In the adversarial-teacher regime the deployed weight sits well above the grid optimum, and the loss there is one of magnitude rather than of the decision to distill. Appendix H.1 reports both in full.

Cross-modal work usually labels the *unlabeled pairs* rather than the student's own samples. Doing so roughly doubles the asymmetry, raising the gain available and the damage together, and our alarm transfers unchanged. Weighting the pairs by $w = \lambda/(1-\lambda)$ makes the design move with the weight, and Proposition 5(i) gives the matrix law that replaces the quadratic (Appendix H.2).

No direction was dropped after seeing an outcome, and restoring the one regime scoped out after its result was known (Appendix H.13) gives $+0.79\%$ (Table 5).

7 Conclusion

We study the decision of whether to distill in the cross-modal setting and at what weight. First, under a fixed linear solution map and squared loss, we derive the exact gain identity $G(\lambda) = 2\lambda A - \lambda^2 D$ and show that six scalars, $(T, \sigma_y^2, \sigma_{yt}, \sigma_t^2, \beta, \delta)$, determine that curve (Propositions 1 and 2). The decision is therefore determined by the sign of A, and the optimal weight is the clipped ratio A/D. We identify β as the term of A that carries the direction of the student's bias rather than its size, and that can flip that sign. Then, two impossibility results show that the sign is underdetermined by the teacher's risk and the label noise (Theorem 1), and by any criterion that reads no labels (Theorem 2), since both leave β unidentified. We therefore propose a decision rule (Algorithm 1) that estimates β and the remaining scalars from the labels the student already trains on together with unlabeled pairs, and returns a weight, or zero when no weight is predicted to help. On six multi-modal datasets the measurements support the analysis, and our estimated gain decides whether to distill better than the published criteria and the label-blind baselines. The practical consequence is to estimate the gain before training and distill only at the weight the rule returns, or not at all.

Limitations and future work. The gain identity of Proposition 1 is exact under a fixed linear solution map and squared loss, and smooth losses only through the local quadratic surrogate of Proposition 6. In practice the estimator built on it still decides well on the neural students we evaluate, though deriving the identity under weaker assumptions is a natural next step. The proposed rule returns a weight as well as a decision, and while that weight improves on the customary one, estimating it more tightly remains open. It is also estimated on the same data the student then trains on, which our regret analysis does not cover. A separate-fold comparison yields similar gains (Appendix H.11).

AI USE STATEMENT

An LLM-based assistant supported coding, running and aggregating experiments, formatting LaTeX tables, checking proofs for correctness, and editing the manuscript. All experimental results, mathematical arguments and proofs were verified independently by the authors, who take full responsibility for the final content.

ETHICS STATEMENT

MIMIC-CXR is credentialed clinical data released under a data use agreement. It is read from a read-only mount, and no artifact derived from it is redistributed. The other five datasets are public benchmarks, and the paper raises no further ethical concern.

REPRODUCIBILITY STATEMENT

Appendix B proves the identity, the normalized gain, the corollary and both theorems. Appendices C and D carry the arguments for the extensions where they are stated. Appendix F gives the experimental details, including the datasets, the models and the budgets. Appendix A states Algorithm 1 in full as Algorithm 2. Reproducing the MIMIC-CXR experiments requires the reader's own PhysioNet credentials.

REFERENCES

Anastasios N Angelopoulos, John C Duchi, and Tijana Zrnic. Ppi++: Efficient prediction-powered inference. *arXiv preprint arXiv:2311.01453*, 2023.

Davide Anguita, Alessandro Ghio, Luca Oneto, Xavier Parra, Jorge Luis Reyes-Ortiz, et al. A public domain dataset for human activity recognition using smartphones. In *Esann*, volume 3, pp. 3–4, 2013.

Ejafa Bassam, Dawei Zhu, and Kaigui Bian. Pld: A choice-theoretic list-wise knowledge distillation. *Advances in Neural Information Processing Systems*, 38:136090–136112, 2026.

Honglie Chen, Weidi Xie, Andrea Vedaldi, and Andrew Zisserman. Vggsound: A large-scale audio-visual dataset. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 721–725. IEEE, 2020.

Yilong Chen, Zongyi Xu, Xiaoshui Huang, Shanshan Zhao, Xinqi Jiang, Xinyu Gao, and Xinbo Gao. Non-target divergence hypothesis: Toward understanding domain gaps in cross-modal knowledge distillation. *arXiv preprint arXiv:2409.02438*, 2024.

Jang Hyun Cho and Bharath Hariharan. On the efficacy of knowledge distillation. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 4793–4801. IEEE, 2019.

Tri Dao, Govinda M Kamath, Vasilis Syrgkanis, and Lester Mackey. Knowledge distillation as semiparametric inference. *arXiv preprint arXiv:2104.09732*, 2021.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pp. 248–255. Ieee, 2009.

Rui Ding, Meng Yang, and Nanning Zheng. Selective transfer learning of cross-modality distillation for monocular 3d object detection. *IEEE Transactions on Circuits and Systems for Video Technology*, 34(10):9925–9938, 2024.

Oswin Gosal, Edwin Arkel Rios, Augusto Christian Surya, Fernando Mikael, Bo-Cheng Lai, and Min-Chun Hu. How to choose your teacher for fine grained image recognition. *arXiv preprint arXiv:2605.15689*, 2026.

Saurabh Gupta, Judy Hoffman, and Jitendra Malik. Cross modal distillation for supervision transfer. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2827–2836, 2016.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

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

Bojing Hou, Ruohao Li, Yitong Zhu, Hongjun Liu, Luwen Yu, and Yuyang Wang. Biokd: Selective physiology-to-video knowledge distillation via reliability gate for emotion recognition. *arXiv preprint arXiv:2608.06023*, 2026.

Tao Huang, Shan You, Fei Wang, Chen Qian, and Chang Xu. Knowledge distillation from a stronger teacher. *Advances in neural information processing systems*, 35:33716–33727, 2022.

Alistair EW Johnson, Tom J Pollard, Seth J Berkowitz, Nathaniel R Greenbaum, Matthew P Lungren, Chih-ying Deng, Roger G Mark, and Steven Horng. Mimic-cxr, a de-identified publicly available database of chest radiographs with free-text reports. *Scientific data*, 6(1):317, 2019.

Oh-Ran Kwon and Daeyoung Ham. Black-box knowledge transfer across distinct feature sets. *arXiv preprint arXiv:2608.12403*, 2026.

Hongsin Lee and Hye Won Chung. Toward understanding adversarial distillation: Why robust teachers fail. *arXiv preprint arXiv:2605.21999*, 2026.

Zheng Li, Xiang Li, Lingfeng Yang, Borui Zhao, Renjie Song, Lei Luo, Jun Li, and Jian Yang. Curriculum temperature for knowledge distillation. In *Proceedings of the AAAI conference on artificial intelligence*, volume 37, pp. 1504–1512, 2023.

Steven R Livingstone and Frank A Russo. The ryerson audio-visual database of emotional speech and song (ravdess): A dynamic, multimodal set of facial and vocal expressions in north american english. *PloS one*, 13(5):e0196391, 2018.

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

Taiming Lu and Zhuang Liu. Strong teacher not needed? on distillation in llm pretraining. *arXiv preprint arXiv:2605.23857*, 2026.

Aditya K Menon, Ankit Singh Rawat, Sashank Reddi, Seungyeon Kim, and Sanjiv Kumar. A statistical perspective on distillation. In *International Conference on Machine Learning*, pp. 7632–7642. PMLR, 2021.

Seyed Iman Mirzadeh, Mehrdad Farajtabar, Ang Li, Nir Levine, Akihiro Matsukawa, and Hassan Ghasemzadeh. Improved knowledge distillation via teacher assistant. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pp. 5191–5198, 2020.

Abhishek Panigrahi, Bingbin Liu, Sadhika Malladi, Sham Kakade, and Surbhi Goel. In good graces: Principled teacher selection for knowledge distillation. In *International Conference on Learning Representations*, volume 2026, pp. 54779–54822, 2026.

Adriana Romero, Nicolas Ballas, Samira Ebrahimi Kahou, Antoine Chassang, Carlo Gatta, and Yoshua Bengio. FitNets: Hints for thin deep nets. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2015.

Mher Safaryan, Alexandra Peste, and Dan Alistarh. Knowledge distillation performs partial variance reduction. *Advances in Neural Information Processing Systems*, 36:75229–75258, 2023.

Jiahao Shi, Omar Hagrass, and Jason M Klusowski. Coupled training with privileged information and unlabeled data. *arXiv preprint arXiv:2605.23268*, 2026.

Nathan Silberman, Derek Hoiem, Pushmeet Kohli, and Rob Fergus. Indoor segmentation and support inference from rgbd images. In *European conference on computer vision*, pp. 746–760. Springer, 2012.

Samuel Stanton, Pavel Izmailov, Polina Kirichenko, Alexander A Alemi, and Andrew G Wilson. Does knowledge distillation really work? *Advances in neural information processing systems*, 34:6906–6919, 2021.

Shangquan Sun, Wenqi Ren, Jingzhi Li, Rui Wang, and Xiaochun Cao. Logit standardization in knowledge distillation. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 15731–15740. IEEE, 2024.

Chao Tan and Jie Liu. Improving knowledge distillation with a customized teacher. *IEEE Transactions on Neural Networks and Learning Systems*, 35(2):2290–2299, 2022.

Yonglong Tian, Dilip Krishnan, and Phillip Isola. Contrastive representation distillation. *arXiv preprint arXiv:1910.10699*, 2019.

Athanasios Tsanas, Max Little, Patrick McSharry, and Lorraine Ramig. Accurate telemonitoring of parkinson's disease progression by non-invasive speech tests. *Nature Precedings*, pp. 1–1, 2009.

Vladimir Vapnik and Akshay Vashist. A new learning paradigm: Learning using privileged information. *Neural networks*, 22(5-6):544–557, 2009.

Eric Xia and Martin J Wainwright. Prediction aided by surrogate training. *arXiv preprint arXiv:2412.09364*, 2024.

Rongrong Xie, Yizhou Xu, and Guido Sanguinetti. Information-theoretic criteria for knowledge distillation in multimodal learning. *arXiv preprint arXiv:2510.13182*, 2025.

Zihui Xue, Zhengqi Gao, Sucheng Ren, and Hang Zhao. The modality focusing hypothesis: Towards understanding crossmodal knowledge distillation. *arXiv preprint arXiv:2206.06487*, 2022.

Yuming Yang, Mingyoung Lai, Wanxu Zhao, Xiaoran Fan, Zhiheng Xi, Mingqi Wu, Chiyue Huang, Jun Zhao, Haijun Lv, Jian Tong, et al. Which reasoning trajectories teach students to reason better? a simple metric of informative alignment. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 42123–42150, 2026.

Borui Zhao, Quan Cui, Renjie Song, Yiyu Qiu, and Jiajun Liang. Decoupled knowledge distillation. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 11943–11952. IEEE, 2022.

Kaixiang Zheng and En-Hui Yang. Knowledge distillation based on transformed teacher matching. In *International Conference on Learning Representations*, volume 2024, pp. 44963–44979, 2024.

Helong Zhou, Liangchen Song, Jiajie Chen, Ye Zhou, Guoli Wang, Junsong Yuan, and Qian Zhang. Rethinking soft labels for knowledge distillation: A bias-variance tradeoff perspective. *arXiv preprint arXiv:2102.00650*, 2021.

A THE RULE IN FULL

Algorithm 1 in the main text names each step, and the version below gives the estimator used for it, including the two penalties, the branch that measures T on the deployed student, and the design correction for the pair setting.

The fold counts are fixed rather than tuned: three for the cross-fitted scalars, five two-fold splits averaged for β, and five for the jackknife. They set the precision of an estimate rather than what is estimated.

Algorithm 2: The rule in full, the estimator behind each line of Algorithm 1

Input: labeled fold (X_ℓ, y) of n samples with the teacher's outputs t_ℓ on them; unlabeled pairs X_u with the teacher's outputs t_u; bridge penalty ρ_b; the student's own ridge penalty ρ_s

Output: a weight in the swept grid $\Lambda \subset [0, 1]$, where 0 means do not distill

*// **estimate** the scalars that Eq. 4 needs*

$\hat{q} \leftarrow$ ridge of t_u on X_u at ρ_b; $\hat{\sigma}_t^2 \leftarrow$ cross-fitted residual of that fit *// the bridge, no labels*

$\hat{\sigma}_y^2, \hat{\sigma}_{yt}, \hat{\delta}_{\text{raw}} \leftarrow$ cross-fitted on (X_ℓ, y, t_ℓ) at ρ_s, with $\hat{q}$

$\hat{m}_h \leftarrow$ cross-fitted ridge of $y - t_\ell$ on X_ℓ at ρ_s *// m_h directly, never as a difference of two fits*

$\hat{\beta} \leftarrow \hat{b}^\top \Sigma \hat{\psi}$ with $M = (X_\ell^\top X_\ell + \rho_s P)^{-1}$, $\hat{b} = -\rho_s M P \hat{\theta}$ and $\hat{\psi} = M X_\ell^\top \hat{m}_h$ *// P zeroes the intercept*

if *the run takes T in closed form or from the ridge proxy* (Table 4) **then**

  $T \leftarrow \text{tr}(L^\top \Sigma L)$, on the student's own features or on the frozen ones

else

  $f_1, f_2 \leftarrow$ the deployed student fit on two disjoint halves, over two independent splits *// four fits*

  $T \leftarrow \mathbb{E}\|f_1 - f_2\|^2/(4\hat{\sigma}_y^2)$ *// Remark 3*

*// **decide** whether any weight is beneficial*

$\hat{A} \leftarrow T(\hat{\sigma}_y^2 - \hat{\sigma}_{yt}) + \hat{\beta}$; $\hat{D} \leftarrow T(\hat{\sigma}_y^2 - 2\hat{\sigma}_{yt} + \hat{\sigma}_t^2) + \hat{\delta}_{\text{raw}}$

if $\hat{D} \le 0$ *or* $\hat{A} \le 0$ **then return** 0 *// do not distill*

*// **choose** the weight, then discount it for its own noise*

$\tilde{\lambda} \leftarrow \hat{A}/\hat{D}$; $\hat{\lambda} \leftarrow \left[\tilde{\lambda}\right]_{[0,1]}$ *// the jackknife below uses the raw ratio $\tilde{\lambda}$*

if *the teacher's targets go on the unlabeled pairs* **then**

  $\gamma \leftarrow$ teacher-weighted mean of the spectrum of S_u against $S_\ell + \rho_s P$ *// no labels; weighted by t_u*

  $\hat{w} \leftarrow \left[\hat{A}/(\hat{D} + \gamma\hat{A})\right]_{[0,9]}$; $\hat{\lambda} \leftarrow \hat{w}/(1 + \hat{w})$ *// Proposition 5(iii), R_0 dropped*

$\hat{\sigma}_{\text{j}}, \hat{b}_{\text{j}} \leftarrow$ jackknife of $\hat{A} - \tilde{\lambda}\hat{D}$ over five folds of the same labeled samples, at the full-fit $\tilde{\lambda}$ *// a per-fold $\tilde{\lambda}$ would vanish*

$\hat{s} \leftarrow \hat{A}^2/(\hat{A}^2 + \hat{\sigma}_{\text{j}}^2 + \hat{b}_{\text{j}}^2)$; $\hat{\lambda}_{\text{cont}} \leftarrow \hat{s}\hat{\lambda}$ *// shrunk; the gated variant returns $\hat{\lambda}$*

return the nearest point of the swept grid Λ to $\hat{\lambda}_{\text{cont}}$ *// this is the policy evaluated*

The estimators. Algorithm 2 names each step, and this paragraph fixes the formulas so that a reader reimplements the same quantities. Write $\langle a, b\rangle$ for $n^{-1}\sum_i a_i^\top b_i$, which sums over outputs and averages over samples, so a shared weight sees the summed A and D of Proposition 1. S_ℓ, S_u are sums of outer products and g_ℓ, g_u sums of cross-products, so w weights the extra rows against the labeled ones one for one.

The bridge and $\hat{\sigma}_t^2$. $\hat{q}$ is a ridge of t_u on X_u at ρ_b. For $\hat{\sigma}_t^2$ the pair pool is split three times. A fifth is held out and two disjoint halves of the rest are fitted, giving $\hat{q}_a$ and $\hat{q}_b$. Then $\hat{\sigma}_t^2 = \langle t - \hat{q}_a, t - \hat{q}_b\rangle$ is scored on the held-out fifth, so neither residual is in-sample. The three repetitions are averaged.

The labeled scalars. The labeled fold is split into $K = 3$ folds. For each evaluation fold, two *disjoint* other folds give ridge fits $\hat{m}_{y,1}, \hat{m}_{y,2}$ at ρ_s. Writing $\hat{u}_{y,j} = y - \hat{m}_{y,j}$ and $\hat{u}_t = t - \hat{q}$ on that fold, the three scalars are $\hat{\sigma}_y^2 = \langle \hat{u}_{y,1}, \hat{u}_{y,2}\rangle$, $\hat{\sigma}_{yt} = \frac{1}{2}(\langle \hat{u}_{y,1}, \hat{u}_t\rangle + \langle \hat{u}_{y,2}, \hat{u}_t\rangle)$ and $\hat{\delta}_{\text{raw}} = \langle \hat{m}_{y,1} - \hat{q}, \hat{m}_{y,2} - \hat{q}\rangle$. Two disjoint copies remove the variance a single fit squared would carry, and they do not make the product unbiased. Writing $a_j = \hat{m}_{y,j} - m_y$ on the evaluation fold, and for samples independent across folds, $\mathbb{E}[(y - \hat{m}_{y,1})(y - \hat{m}_{y,2}) \mid x, \hat{m}_{y,1}, \hat{m}_{y,2}] = \text{Var}(y \mid x) + a_1 a_2$, so each fit's bias survives as a product. $\hat{\delta}_{\text{raw}}$ additionally puts one $\hat{q}$ into both of its factors, so the bridge's error enters it squared. $\hat{\sigma}_t^2$ uses the two disjoint copies above instead, so its bridge error enters as a product of two different ones. These are cross-fitted product estimators, not unbiased ones, and exact enumeration on a small ridge confirms the bias is nonzero. The three evaluation folds are averaged.

$\hat{\beta}$. $\hat{m}_h$ is a cross-fitted ridge of $y - t_\ell$ on X_ℓ at ρ_s, five two-fold splits averaged. Σ is the second moment of the student's view over the whole training pool, which reads X^b only and so costs no labels. $\hat{\theta}$ is the supervised fit on the labeled fold, and $\hat{b} = -\rho_s MP\hat{\theta}$.

γ. With $H_\ell = S_\ell + \rho_s P$ and $S_u v_j = \gamma_j H_\ell v_j$, the eigenvectors normalized as in Proposition 5 so that $v_j^\top H_\ell v_j = 1$, the teacher-weighted mean is $\sum_j \gamma_j e_j / \sum_j e_j$ with $e_j = \|v_j^\top g_u\|^2$ the teacher's pull in coordinate j. Without that normalization the weights are not determined, since rescaling v_j rescales e_j. When $\sum_j e_j$ underflows the rule falls back to $\mathrm{tr}(S_u)/\mathrm{tr}(H_\ell)$. It uses the teacher's outputs on the pairs and no labels.

The jackknife. Five folds of the labeled samples. Each deleted fold gives $\hat{A}^{(k)}, \hat{D}^{(k)}$ by refitting the labeled scalars and $\hat{\beta}$ on the retained samples, with the bridge and Σ held fixed. With $c_k = \hat{A}^{(k)} - \tilde{\lambda}\hat{D}^{(k)}$ at the full-fit raw ratio $\tilde{\lambda} = \hat{A}/\hat{D}$, $\hat{\sigma}_{\mathrm{j}}^2 = \frac{K-1}{K}\sum_k (c_k - \bar{c})^2$ and $\hat{b}_{\mathrm{j}} = (K-1)(\bar{c} - c_{\mathrm{full}})$.

One definition. The weight the rule returns is computed by a single function that the experiments and the table generator both call: abstain when $\hat{D} \le 0$ or $\hat{A} \le 0$, otherwise clip $\hat{A}/\hat{D}$ to $[0, 1]$, scale it by $\hat{s}$, and project onto the swept grid.

Variants by run. Table 4 is the specification per run, because the exceptions are easier to check than to read. The pooled comparison uses the corresponding labeled-sample variants.

run	estimates	T	weight	shrink applied to	reports
main ridge, label budgets	$\hat{A}, \hat{D}$	closed form	$[\hat{A}/\hat{D}]_{[0,1]}$	$\hat{\lambda}$	policy
teacher families	$\hat{A}, \hat{D}$	closed form	$[\hat{A}/\hat{D}]_{[0,1]}$	$\hat{\lambda}$	policy
student capacity (MLP)	$\hat{A}, \hat{D}$	ridge proxy	$[\hat{A}/\hat{D}]_{[0,1]}$	$\hat{\lambda}$	policy
main neural (MLP)	$\hat{A}, \hat{D}$	ridge proxy	$[\hat{A}/\hat{D}]_{[0,1]}$	$\hat{\lambda}$	policy
VGGSound (own samples)	$\hat{A}, \hat{D}$	measured	$[\hat{A}/\hat{D}]_{[0,1]}$	$\hat{\lambda}$	policy
VGGSound (pairs)	$\hat{A}, \hat{D}, \gamma$	ridge proxy	design-corrected $\hat{w}$	$\hat{w}/(1+\hat{w})$	policy
MIMIC (pairs)	$\hat{A}, \hat{D}, \gamma$	measured, else proxy	design-corrected $\hat{w}$	$\hat{w}/(1+\hat{w})$	policy
cross-entropy (Table 18)	$\hat{A}, \hat{D}$	closed form	clipped to the grid	·	policy
unlabeled pairs, scorecard	$\hat{A}, \hat{D}$	closed form	fixed $\lambda = 0.5$	·	alarm only
pair law (Table 16)	$\hat{A}, \hat{D}, \gamma$	closed form	design-corrected $\hat{w}$	·	policy
withheld pretrained pairs	$\hat{A}, \hat{D}, \gamma$	closed form	design-corrected $\hat{w}$	$\hat{w}$, then $\hat{w}/(1+\hat{w})$	policy
end-to-end CNNs	·	·	·	·	gain and loss only

Table 4: **The rule's variants, by run.** Where shrinkage is applied matters: shrinking $\hat{w}$ and then converting gives $\hat{s}\hat{w}/(1+\hat{s}\hat{w})$, while shrinking the converted weight gives $\hat{s}\hat{w}/(1+\hat{w})$, and at $\hat{s} = 1/2$, $\hat{w} = 1$ those are $1/3$ and $1/4$. The end-to-end rows estimate no scalars and carry no policy arm.

Edge cases and evaluation. $\hat{D} \le 0$ returns 0, as does $\hat{A} \le 0$. The measured-T branch is undefined when $\hat{\sigma}_y^2$ is at or below zero, and is not attempted below 24 samples; in either case the rule falls back to the closed-form T. The projection Π_Λ is part of the policy rather than a reporting convenience. Every arm in this paper is scored at a grid point, so no student is trained off the grid, and a small positive $\hat{\lambda}_{\mathrm{cont}}$ can project to 0 and abstain. With $\hat{A} = 0.1$, $\hat{D} = 1$ and $\hat{\sigma}_{\mathrm{j}}^2 + \hat{b}_{\mathrm{j}}^2 = 0.04$ the continuous weight is 0.02 and the deployed one is 0, while at $\hat{D} = 0.1$ it is 0.2 and survives. A continuous weight equidistant from two grid points takes the smaller, so the projection is a function and ties fall toward abstaining.

Which weight each number reports. Every risk number in the paper is the projected weight's. Among the weight statistics, the compression analysis of Appendix H.10 uses $\hat{\lambda}_{\mathrm{cont}}$, since projection would confound the spread it measures with the grid's, while the decomposition of Table 28 uses the projected weight, which is the one deployed. The over- and under-distilling counts are reported at each stage separately, because the answer depends on the stage: the clipped proposal $[\hat{A}/\hat{D}]_{[0,1]}$, the shrunk continuous weight $\hat{s}\hat{\lambda}$, and the projected weight are three different statistics and Appendix H.10 gives all three.

B PROOFS

Proposition 1. Since θ^* solves the normal equations $\mathbb{E}[x_0^b(y_0 - (x_0^b)^\top\theta^*)] = 0$, expanding $\mathcal{R}(\theta)$ about θ^* leaves no cross term, so $\mathcal{R}(\theta) - \mathcal{R}(\theta^*) = \|\theta - \theta^*\|_\Sigma^2$. Write the mixed estimator's error as $b + Lu_y - \lambda(Lm_h + Lu_h)$ and expand $\|\cdot\|_\Sigma^2$. Taking expectations, the risk is $R(\lambda) - \mathcal{R}(\theta^*) =$

$R_0 - 2\lambda A + \lambda^2 D$ with $R_0 = b^\top \Sigma b + \mathrm{tr}(K\Omega_{yy})$, $A = b^\top \Sigma L m_h + \mathrm{tr}(K\Omega_{yh})$ and $D = m_h^\top K m_h + \mathrm{tr}(K\Omega_{hh})$. Subtracting from the value at $\lambda = 0$ gives Eq. 2. Imposing homoscedastic moments gives Eq. 4. The identity holds coordinatewise for multiple outputs because Σ acts on the output index trivially.

Proposition 2. Eq. 4 writes A and D in the six scalars, so $G(\lambda) = 2\lambda A - \lambda^2 D$ and $\lambda^* = [A/D]_{[0,1]}$ are functions of them alone.

Remark 1. Completing the square in Eq. 2 gives Eq. 6, which is $G(\lambda^*) - G(\lambda) = D(\lambda - \lambda^*)^2$, and dividing Eq. 2 by $G(\lambda^*) = A^2/D$ gives $2z - z^2$.

Monotonicity in the scale of the discrepancy. Under the conditions of Corollary 2, replacing ψ by $c\psi$ with $c > 1$ does not increase λ^*, since β scales with c and δ with c^2. The clip is not cosmetic here: the unclipped ratio rises before it falls whenever $\beta > 0$, which is why the statement is about λ^* and not about A/D. Fix the student, so $a = T\sigma_y^2 > 0$, b and L are fixed, and scale the discrepancy $m_y - t$ along a fixed direction by a size $c > 0$. Because L is linear the filtered discrepancy scales with it, $\psi = c\,\psi_0$ for a fixed ψ_0. Then $\beta = c\beta_0$ with $\beta_0 = b^\top \Sigma \psi_0$, and $\delta = c^2\delta_0$ with $\delta_0 = \psi_0^\top \Sigma \psi_0 > 0$, so $\lambda^*(c) = [f(c)]_{[0,1]}$ with $f(c) = (a + c\beta_0)/(a + c^2\delta_0)$ and $f(0) = 1$.

The sign of $f'(c)$ is the sign of $N(c) = a\beta_0 - 2ac\delta_0 - c^2\beta_0\delta_0$. Suppose $0 < f(c) < 1$, which is where the clip is inactive. From $f(c) < 1$ and $c > 0$ we get $\beta_0 < c\delta_0$, hence $\beta_0 - 2c\delta_0 < -c\delta_0 < 0$ and $a(\beta_0 - 2c\delta_0) < -ac\delta_0$. Therefore $N(c) < -ac\delta_0 - c^2\beta_0\delta_0 = -c\delta_0\,(a + c\beta_0)$, and $f(c) > 0$ makes $a + c\beta_0 > 0$, so $N(c) < 0$. Thus f is strictly decreasing wherever it lies in $(0, 1)$.

It remains to show that the clipped weight cannot return. If $\beta_0 \geq 0$ then $a + c\beta_0 > 0$ for all c, so $f > 0$ throughout and $f \to 0$. If $\beta_0 < 0$ then f reaches 0 at $c = a/|\beta_0|$ and $a + c\beta_0 < 0$ beyond it, so $f < 0$ and the clip holds it at 0. With $f(0) = 1$, λ^* is therefore 1 until f first falls below 1, strictly decreasing until it reaches 0, and 0 after. The unclipped f is *not* monotone: $N(0) = a\beta_0 > 0$ whenever $\beta_0 > 0$, so f rises above 1 before falling, which is why the statement is about λ^* and not about A/D.

Accuracy is not sufficient even in the same-view case. Take a balanced design $x_i \in \{-1, +1\}$ with $n = 8$, $Y \mid x \sim \mathrm{Bernoulli}(\frac{1}{2} + \frac{x}{4})$, a known intercept $\frac{1}{2}$ and a ridge penalty 8 on the slope, and compare the two same-view teachers $t_1(x) = \frac{1}{2} + \frac{x}{8}$ and $t_2(x) = \frac{1}{2} + \frac{3x}{8}$. Both have squared risk $\frac{13}{64}$ and the same accuracy at threshold $\frac{1}{2}$. Their filtered discrepancies are $\psi_1 = \frac{1}{16}$ and $\psi_2 = -\frac{1}{16}$, so $\beta_1 = -\frac{1}{128}$ and $\beta_2 = +\frac{1}{128}$ against a common $T\sigma_y^2 = \frac{3}{512}$ and $\delta = \frac{1}{256}$, giving $A_1 = -\frac{1}{512}$ and $A_2 = \frac{7}{512}$ with $D = \frac{5}{512}$. The optimal weights are 0 and 1. The two discrepancies point in opposite directions, so this is consistent with the monotonicity above, which fixes the direction.

Theorem 1. The construction in full. Take an even balanced design $x_i \in \{-1, +1\}$, a known intercept $1/2$ and a ridge penalty n on the slope, so that the fitted slope is $(2n)^{-1} \sum_i x_i(y_i - 1/2)$, with the test covariate balanced as well. In problem $-$, $Y \mid x \sim \mathrm{Bernoulli}(1/2 + x/4)$ and $t = 1/2$. In problem $+$, $Y \mid x \sim \mathrm{Bernoulli}(1/2 - x/4)$ and $t = 1/2 - x/2$. The teacher, the supervised student and every mixed student predict in $[0, 1]$.

Both problems have filtered discrepancy $\psi = 1/8$ and baseline slope biases $-1/8$ and $+1/8$. Substituting into Eq. 4 gives $A_\mp = \kappa \mp \nu$ and $D = \kappa + \nu$ with $\nu = 1/64$ and $\kappa = 3/(64n)$, so $A_- < 0$ for every even $n > 3$ and that problem's optimum is 0 while the other's is 1. The respective regrets at a common decision λ are $2(\nu - \kappa)\lambda + (\nu + \kappa)\lambda^2$ and $(\nu + \kappa)(1 - \lambda)^2$. The former increases and the latter decreases on $[0, 1]$, and they intersect at $\lambda_\dagger = (1 + 3/n)/4$, and their common value, $c_n = 2(\nu - \kappa)\lambda_\dagger + (\nu + \kappa)\lambda_\dagger^2 = 9(n - 1)^2(n + 3)/(1024n^3)$, is the minimax regret, which tends to $9/1024$. Every selector reading the agreed quantities incurs at least c_n in the worse of the two problems, and the selector that plays $\lambda_\dagger$ attains it. Everything obtainable in the problem whose optimum is 1 is $G(\lambda^*) = \nu + \kappa \to 1/64$, which is the scale the bound should be read against, and as a share of that gain the bound is $9(n - 1)^2/(16n^2)$, tending to $9/16$. For randomized weights, both expected regrets are at least their values at $\mathbb{E}\lambda$ by convexity. Teacher risk is $\sigma_y^2 + \delta_{\mathrm{raw}} = 1/4$ in both problems, and the supervised squared bias and variance also coincide. For $n = 8$ the bound is $4851/524288$, which is 43% of the $\nu + \kappa = 11/512$ obtainable there. That share rises with n and

stays below its limit $9/16$ at every finite even $n > 3$, because $\nu + \kappa$ approaches $1/64$ from above while the bound approaches $9/1024$.

Theorem 2. The design is $X = (-1, 0, 1)^\top$ with one observation per point and x_0 uniform on the same three values, so $\Sigma = \frac{2}{3}$. The student minimizes $\frac{1}{3}\|y - X\theta\|^2 + \theta^2$, giving $L = (-\frac{1}{5}, 0, \frac{1}{5})$ and $T = \mathrm{tr}(L^\top \Sigma L) = \frac{4}{75}$. The single teacher is $t(x) = (\frac{1}{5}, \frac{1}{5}, \frac{4}{5})$, and $y \mid x$ is Bernoulli with mean $m_- = (\frac{1}{2}, \frac{1}{20}, \frac{1}{20})$ in problem $-$ and $m_+ = (\frac{1}{20}, \frac{19}{20}, \frac{1}{2})$ in problem $+$.

Those means give conditional variances $(\frac{1}{4}, \frac{19}{400}, \frac{19}{400})$ and $(\frac{19}{400}, \frac{19}{400}, \frac{1}{4})$, which are not constant across samples. The row-moment condition behind Eq. 4 therefore fails here, and the noise term is computed from Ω_{yy} directly, as $\mathrm{tr}(K\Omega_{yy}) = \frac{119}{15000}$ in both problems, rather than as $T\sigma_y^2 = \frac{23}{3750}$. Everything else follows by direct computation. Both problems have the row-average conditional variance $\bar{\sigma}_y^2 = \frac{23}{200}$, written with a bar because it is not the common per-row variance σ_y^2 of Section 3, $\sigma_{yt} = \sigma_t^2 = 0$ because the teacher is a function of x, $\delta_{\mathrm{raw}} = \frac{9}{40}$, supervised excess risk $\frac{241}{12000}$ and teacher risk $\frac{17}{50}$, all identical. They differ in β, equal to $-\frac{189}{10000}$ and $+\frac{27}{10000}$, and in δ, equal to $\frac{147}{5000}$ and $\frac{3}{5000}$, so $A = -\frac{329}{30000}$ and $+\frac{319}{30000}$ against $D = \frac{14}{375}$ and $\frac{16}{1875}$. Everything obtainable in problem $+$ is $G(1) = \frac{191}{15000}$. The two problems share the distribution of (x, t) by construction, so any functional of that distribution agrees between them.

The minimax step differs from Theorem 1's, because here $A/D = 319/256 > 1$ in problem $+$, so that problem's optimum is the clipped weight 1 and everything obtainable in it is $G(1) = 2A - D$ rather than A^2/D. The two regrets at a common decision λ are $\frac{329}{15000}\lambda + \frac{14}{375}\lambda^2$ and $\frac{191}{15000} - \frac{319}{15000}\lambda + \frac{16}{1875}\lambda^2$, both convex, so a randomized weight cannot beat its own mean. They cross at $\lambda = \sqrt{1302}/36 - \frac{3}{4}$, where the common value is $68131/1620000 - 511\sqrt{1302}/540000 \approx 0.007911$, which is 62.1% of $\frac{191}{15000}$. The students fitted in the two problems have slope means $-\frac{9}{100}$ and $+\frac{9}{100}$, the realized slope taking one of $-\frac{1}{5}, 0, \frac{1}{5}$, which is why criteria reading the fitted student are not covered.

Two limits of the theorems. The asymptotic statement retains a fixed population regularization, so it is not a lower bound for consistent ridge with a vanishing penalty. And an algorithm that observes labels, such as a validation sweep, is not restricted to the information shared in these constructions, which is why Section 6.3 compares against a sweep by measurement rather than by theorem. In the second construction the shared δ_{raw} pins neither of the two quantities that separate the problems, and Theorem 1 holds δ equal as well, so there β is the only difference.

C Extensions

Proposition 3 (Regret at a mis-specified weight)**.** *Deciding not to distill when one should costs $D\lambda^{*2} = A^2/D$. For any ε with $(1+\varepsilon)\lambda^* \in [0, 1]$, distilling at $\lambda = (1+\varepsilon)\lambda^*$ costs $\varepsilon^2 A^2/D$. Their ratio is exactly ε^2, with no approximation and no condition beyond $D > 0$ and $\lambda^* = A/D \in (0, 1]$.*

Remark 2 (An accurate teacher bounds the term that can flip the decision)**.** *Under the conditions of Proposition 1, Cauchy–Schwarz on $\beta = b^\top \Sigma \psi$ gives*

$$|\beta| \;\le\; \|b\|_\Sigma \, \|\psi\|_\Sigma \;=\; \sqrt{\big(R_0 - \mathrm{tr}(K\Omega_{yy})\big)\,\delta}, \qquad \delta = m_h^\top K m_h \;\le\; n\|K\|_2 \, \delta_{\mathrm{raw}}.$$

In the same-view case of Corollary 2 the teacher's squared risk on the training design is $\sigma_y^2 + \delta_{\mathrm{raw}}$, so a teacher whose risk is close to the label noise has a small δ_{raw} and therefore a small $|\beta|$. Then A lies within that bound of $T(\sigma_y^2 - \sigma_{yt})$, and its sign agrees with that term whenever the term's magnitude exceeds the bound.

Teacher accuracy is therefore informative about the decision without determining it. Theorem 1 lives where the alignment can change the sign of A, its two problems sharing the teacher's risk while β changes sign.

Proposition 4 (Declining to distill, and shrinkage as a restricted family)**.** *Let $\lambda^* = A/D \in (0, 1]$ and let $\mu_\lambda = \mathbb{E}\hat{\lambda} - \lambda^*$ and $\tau_\lambda^2 = \mathbb{E}[(\hat{\lambda} - \lambda^*)^2]$ be the bias and the mean squared error of the plug-in weight $\hat{\lambda} = \hat{A}/\hat{D}$, assumed finite. Then (i) distilling at $\hat{\lambda}$ has expected curve regret $D\tau_\lambda^2$ against an*

abstain regret of A^2/D, so declining dominates exactly when $D\tau_\lambda > |A|$; and (ii) over the family $\lambda = s\hat{\lambda}$ with s fixed,

$$\mathbb{E}(s\hat{\lambda} - \lambda^*)^2 = (1-s)^2\lambda^{*2} + s^2\tau_\lambda^2 - 2s(1-s)\lambda^*\mu_\lambda,$$

so the first two terms upper-bound the error on $s \in [0,1]$ whenever λ^* and μ_λ share a sign, and that bound is minimized at $s^\star = \lambda^{*2}/(\lambda^{*2} + \tau_\lambda^2)$, which satisfies $s^\star \le 1/2$ exactly when declining weakly dominates.

Proofs of Propositions 3 and 4. Both follow from Eq. 6. Setting $\lambda = 0$ gives the abstain regret A^2/D, and setting $\lambda = (1+\varepsilon)\lambda^*$ gives $\varepsilon^2 A^2/D$, so the ratio is ε^2. Part (i) is $D\tau_\lambda^2$ against $A^2/D = D\lambda^{*2}$. For Proposition 4(ii), decompose $\mathbb{E}(s\hat{\lambda} - \lambda^*)^2 = s^2 \operatorname{Var}(\hat{\lambda}) + \big((s-1)\lambda^* + s\mu_\lambda\big)^2$ and expand the square, using $\tau_\lambda^2 = \operatorname{Var}(\hat{\lambda}) + \mu_\lambda^2$. The subtracted term $2s(1-s)\lambda^*\mu_\lambda$ is non-negative on $s \in [0,1]$ when $\lambda^*\mu_\lambda \ge 0$, since $s(1-s) \ge 0$ there, so dropping it leaves an upper bound. Differentiating the bound in s gives $s^\star = \lambda^{*2}/(\lambda^{*2} + \tau_\lambda^2)$, and $s^\star \le 1/2$ rearranges to $\lambda^{*2} \le \tau_\lambda^2$, which is (i). This is a statement about a fixed s chosen with λ^* and τ_λ known. It motivates the shrinkage form the rule ships, and does not carry over to it, because the shipped factor is random, is computed from the same data, and does not come with $\lambda^*\mu_\lambda \ge 0$.

The rule cannot evaluate τ_λ, and substitutes the jackknife variance $\hat{\sigma}_{\mathrm{j}}^2$ and bias $\hat{b}_{\mathrm{j}}$ of $\hat{A} - \tilde{\lambda}\hat{D}$ for the raw ratio $\tilde{\lambda} = \hat{A}/\hat{D}$, recomputing $\hat{A}$ and $\hat{D}$ on each fold while holding $\tilde{\lambda}$ at its full-fit value, since a per-fold $\tilde{\lambda}$ would make the contrast vanish identically. That substitution gives a hard-gate threshold $|\hat{A}| < \sqrt{\hat{\sigma}_{\mathrm{j}}^2 + \hat{b}_{\mathrm{j}}^2}$ and a shrinkage factor $\hat{s} = \hat{A}^2/(\hat{A}^2 + \hat{\sigma}_{\mathrm{j}}^2 + \hat{b}_{\mathrm{j}}^2)$, and the deployed rule uses the second, with Table 14 measuring it against the first.

Even with the denominator fixed at D, so that $D^2\tau_\lambda^2 = \operatorname{Var}(\hat{A}) + (\mathbb{E}\hat{A} - A)^2$ exactly, a resampling estimate does not recover the second term: a bias shared by every resample, as a persistent nuisance or regularization bias is, is invisible to the jackknife however large its variance estimate. The jackknife approximates estimation variability and the part of the bias that moves with the sample, and Section 6 measures whether the rule using it dominates.

The gap between the curve and the policy. Propositions 3 and 4 are about regret on the gain curve. Writing $a_{\mathcal{D}} = (Ly - \theta^*)^\top \Sigma L h$ and $d_{\mathcal{D}} = \|Lh\|_\Sigma^2$ for the sample quantities whose means are A and D, a policy that selects $\hat{\lambda}$ on the same draw has expected gain $2\mathbb{E}[\hat{\lambda}a_{\mathcal{D}}] - \mathbb{E}[\hat{\lambda}^2 d_{\mathcal{D}}]$, which differs from $\mathbb{E}[G(\hat{\lambda})]$ by

$$2\operatorname{Cov}(\hat{\lambda}, a_{\mathcal{D}}) - \operatorname{Cov}(\hat{\lambda}^2, d_{\mathcal{D}}).$$

Cross-fitting the nuisances keeps each out of the samples it is applied to, which gives independence when those samples are independent across folds, and does not make this covariance vanish, and nothing here bounds it: the sign of the realized gain can differ from the frozen curve's, which is why Section 6 establishes the policy claim by measurement and Appendix H.11 compares the two protocols directly.

Remark 3 (T is measurable under a variance-scaling hypothesis)**.** *Under the common row-moment conditions of Eq. 4, Proposition 1 makes the student's variance contribution exactly $T\sigma_y^2$. Fit the student on two disjoint halves of the labeled fold, giving f_1 and f_2, for which*

$$\mathbb{E}\|f_1 - f_2\|^2 = \|\mathbb{E}f_1 - \mathbb{E}f_2\|^2 + \operatorname{Var} f_1 + \operatorname{Var} f_2 - 2\operatorname{tr}\operatorname{Cov}(f_1, f_2)$$

in general. Assume (i) $\mathbb{E}f_1 = \mathbb{E}f_2$, (ii) $\operatorname{Var} f_1 = \operatorname{Var} f_2 = \operatorname{Var}_{n/2}$ with $\operatorname{Cov}(f_1, f_2) = 0$, and (iii) $\operatorname{Var}_{n/2} = 2\operatorname{Var}_n$. Then $\mathbb{E}\|f_1 - f_2\|^2 = 2\operatorname{Var}_{n/2}$ and $T_n = \mathbb{E}\|f_1 - f_2\|^2/(4\sigma_y^2)$.

T can therefore be measured on the deployed student rather than assumed from a proxy, under three hypotheses that are not automatic. Equal means alone are not enough, and (ii) holds for two disjoint halves only when the fits share no randomness, no nuisance estimate and no group. A ridge whose penalty is held fixed instead of scaling with n violates (iii), so the measurement is a calibration under a stated scaling rather than an exact reading of T for an arbitrary student.

The measurement costs two extra fits of the student per repetition, four at the default, and no labels beyond the ones it is already training on. Table 17 reports what it decides at $n = 20{,}000$.

Proposition 5 (The pair setting, where the design moves with the weight)**.** *Put the teacher's targets on n_u unlabeled pairs weighted by $w \geq 0$, so that $\hat{\theta}(w) = (S_\ell + wS_u + \rho_s P)^{-1}(g_\ell + wg_u)$ with S_ℓ, g_ℓ the labeled rows' second moment and cross-moment, S_u the design mass the extra rows add and g_u the pull their targets exert. Assume $S_\ell + \rho_s P \succ 0$ and let $\gamma_1 \leq \cdots \leq \gamma_d$ solve the generalized eigenproblem $S_u v = \gamma(S_\ell + \rho_s P)v$ with ρ_s the student's own penalty, let V collect the eigenvectors, and write $N = V^\top \Sigma V$ and $\sigma_j = N_{jj}$. Normalize V so that $V^\top(S_\ell + \rho_s P)V = I$, let $\mathcal{E}(w) = \mathbb{E}\|\hat{\theta}(w) - \theta^*\|_\Sigma^2$ be the expected excess risk, and let $Q(w) = R_0 - 2A_{\mathrm{p}}w + D_{\mathrm{p}}w^2$ with the coefficients defined in (i). Then*

(i) (the estimator decouples, with no condition on the spectrum) *write $u = V^\top(g_\ell - (S_\ell + \rho_s P)\theta^*)$ for the supervised error and $r = V^\top(g_u - S_u\theta^*)$ for the teacher's pull, both in the generalized coordinates. Then $\hat{\theta}(w) - \theta^* = V\zeta$ with $\zeta_j = (u_j + wr_j)/(1 + w\gamma_j)$, and $\mathcal{E}(w) = \mathbb{E}[\zeta^\top N \zeta]$ exactly, for every N. The coefficients of Q are the pair-specific moments $R_0 = \mathbb{E}[u^\top N u]$, which is the supervised excess risk, $A_{\mathrm{p}} = -\mathbb{E}[u^\top N r]$ and $D_{\mathrm{p}} = \mathbb{E}[r^\top N r]$. The coordinatewise form $\mathcal{E}(w) = \mathbb{E}\sum_j \sigma_j \zeta_j^2$ holds when N is diagonal, which is the case when the test metric is proportional to $S_\ell + \rho_s P$ and in particular for an isotropic design, and otherwise with a pathwise error of*

$$\left|\zeta^\top N \zeta - \sum_j \sigma_j \zeta_j^2\right| \;\leq\; \left\|N - \operatorname{diag} N\right\|_2 \|\zeta\|_2^2;$$

(ii) (a two-sided bracket, under the diagonal metric of (i)) $Q(w)/(1 + \gamma_d w)^2 \leq \mathcal{E}(w) \leq Q(w)/(1 + \gamma_1 w)^2$ *for all $w \geq 0$, since each term of (i) is decreasing in γ_j; and*

(iii) (exact collapse) *if the γ_j share a common value γ then $(1 + w\gamma_j)^{-1}$ factors out of ζ, so $\mathcal{E}(w) = Q(w)/(1 + \gamma w)^2$ for every N, and the w^2 terms cancel in the stationarity condition, whose closed-form root is $w^\star = (A_{\mathrm{p}} + \gamma R_0)/(D_{\mathrm{p}} + \gamma A_{\mathrm{p}})$. That root is the minimizer when $D_{\mathrm{p}} + \gamma A_{\mathrm{p}} > 0$ and it lands in the admissible range. Otherwise the infimum is at an endpoint of that range, and on the unbounded domain $[0, \infty)$ it may be approached only as $w \to \infty$.*

Reading the three parts. The generalized coordinates of (i) aggregate rather than coincide with the u_y and m_h of Section 3. Its error bound splits the two sources cleanly, since $\|N - \operatorname{diag} N\|_2$ is built from the two designs and the test moment alone while $\|\zeta\|_2^2$ carries the labels, the teacher and θ^*. The bracket of (ii) can fail without its diagonal metric, and a two-coordinate metric with off-diagonal $-9/10$ breaks the upper bound already at $w = 1$. The endpoint case of (iii) is why the rule searches a bounded range.

The pair setting is not the labeled-sample one. This is a different problem from the one Section 3 analyzes, and the two should not be run together. Labeled-row mixing holds the design fixed and moves the target, which is why the quadratic is exact there. Pair training adds teacher-supervised rows, so the design moves with the weight. Taking $\gamma = 0$ here does not recover the first, because all the generalized eigenvalues vanish only when $S_u = 0$, and then $X_u = 0$ and $g_u = 0$ as well, so there is no teacher term left. The two weights are also different parameters: the experiments put weight $w = \lambda/(1 - \lambda)$ on the extra rows, so w ranges over $[0, \infty)$ where λ ranges over $[0, 1)$.

What the measurements add. The condition in (i) is not a technicality, and the measurements show its effect directly. On an isotropic design, where the condition holds by construction, the law fits at $R^2 = 1.000000$, and on real features, where $V^\top \Sigma V$ is not diagonal, the best collapse reaches 0.9428. Two things separate those numbers, off-diagonal mass and spectral spread, and the second alone is enough, since a spread spectrum is not represented by any single γ even when N is diagonal. The spectrum itself is a property of the two *designs* and costs no labels and no student fit, though the summary the rule picks from it is weighted by the teacher's outputs, so the summary is label-blind rather than design-only. And part (ii) is what makes a single γ usable *under the diagonal metric it assumes*, because there it places the truth between two curves of the same family. On the real features, where N is not diagonal, that bracket does not apply and the collapse is an empirical approximation, which is what Table 16 measures.

The two substitutions the rule makes. The shipped rule makes two substitutions on top of that collapse, and neither is a consequence of the proposition. It estimates A_{p} and D_{p} by the labeled-sample scalars $\hat{A}$ and $\hat{D}$ of Section 5, which are computed on the labeled samples and are not the pair coefficients of (i), and it drops the R_0 term in (iii) and uses $\hat{A}/(\hat{D} + \gamma\hat{A})$. Dropping R_0 is a conservative approximation and not a limit of Eq. 2, since $A_{\mathrm{p}}^2 \le R_0 D_{\mathrm{p}}$ forbids sending R_0 to zero at fixed A_{p}, and R_0 carries the baseline's bias as well as its variance, so what the substitution gives up is not only the variance the extra rows remove. It is also the side the measurements support, because keeping the term selects a mean weight of 1.063 against an optimum of 0.160, so it overshoots by nearly seven times and saturates at the top of the grid. Table 16 reports what the two substitutions together are worth.

Proposition 6 (A local quadratic surrogate for smooth losses)**.** *Let the student minimize a twice-differentiable loss ℓ with a linear predictor and population risk $\mathcal{R}_\ell$.*

(i) (the population statement) *at a local minimum θ_ℓ, write $H = \frac{1}{2}\nabla^2\mathcal{R}_\ell(\theta_\ell)$, which for a linear predictor is half the Gauss–Newton matrix and returns Σ for squared loss. A local minimum gives $H \succeq 0$, which a stationary point does not, and the excess risk of θ is $\|\theta - \theta_\ell\|_H^2 + o(\|\theta - \theta_\ell\|^2)$.*

(ii) (the fixed quadratic surrogate) *the rule does not have θ_ℓ, and freezes the curvature at a pilot fit instead, giving a quadratic surrogate Q whose half-Hessian J we require to satisfy $J \succeq 0$. The optimum is read as in Proposition 1 where $D > 0$. Along the path $\hat{\theta}_\lambda = \hat{\theta}_0 - \lambda Lh$, which is affine in λ,*

$$Q(\hat{\theta}_0) - Q(\hat{\theta}_\lambda) = \lambda\,\nabla Q(\hat{\theta}_0)^\top Lh - \lambda^2\|Lh\|_J^2$$

exactly*, with no remainder. Taking expectations recovers Eq. 2 with $2A = \mathbb{E}[\nabla Q(\hat{\theta}_0)^\top Lh]$ and $D = \mathbb{E}\|Lh\|_J^2$, both under the pilot's own conditioning. When the pilot sits at the surrogate's minimizer θ_Q the numerator collapses to $b_Q^\top JLm_h + \mathrm{tr}(L^\top JL\,\Omega_{yh})$ with $b_Q = \mathbb{E}[\hat{\theta}_0] - \theta_Q$, which is Proposition 1's form centred at θ_Q rather than at the squared-loss θ^*.*

The gain in (ii) is the surrogate's, not the loss's, and J need not equal the H of (i) unless the pilot is at θ_ℓ.

What $J \succeq 0$ buys. The condition makes $\|\cdot\|_J$ a seminorm and $D \ge 0$. A population local minimum supplies it and an arbitrary pilot does not, and without it the displayed identity still holds while the norm reading of $\|\cdot\|_J$ and the clipped optimum do not. The case $D = 0$ is then decided by the sign of A at the endpoints, since the squared-risk implication $D = 0 \Rightarrow A = 0$ does not carry over: $Q(\theta) = \theta$ with $J = 0$, $\hat{\theta}_0 = 0$ and $Lh = 1$ gives $D = 0$ with $A = 1/2$.

Two conditions on the pilot. Two conditions are worth separating. A pilot that is not stationary leaves the linear term in (ii), which is why the numerator is written as a gradient contraction rather than as a bias inner product. And the pilot in our cross-entropy runs is fitted on the student's own labeled fold, so the moments defining A and D there are conditional on that fit rather than on a fit made independently of that fold, as Section 3 assumes for the teacher. Fitting the pilot on a disjoint fold removes the second condition and we do not do it, because the runs reuse the fold on purpose.

What the shipped cross-entropy rule is. The rule we ship for cross-entropy is not an estimator of A_Q. It maps the problem into the working coordinates of a pilot fit and then applies the squared-loss scalars of Section 5 unchanged, which is a metric-calibrated approximation of (ii) rather than an instance of it, and Section 6.4 evaluates it as one. The gap has three parts and not one: the multiclass problem is scalarized, the fitted path is not affine, and the coordinates are taken at $T = 1$ while the objective that is optimized carries the teacher at $T_{\mathrm{KD}} = 4$, whose gradient $(1-\lambda)(p_1 - y) + \lambda T(p_T - q_T)$ and curvature $(1-\lambda)W(p_1) + \lambda W(p_T)$ both move with λ. For a binary logistic head the surrogate is one reweighting. Writing η_i for the linear predictor and p_i for its fitted probabilities, the working weights are $\omega_i = p_i(1-p_i)$, the working response is $\tilde{y}_i = \eta_i + (y_i - p_i)/\omega_i$ and the working teacher is $\tilde{t}_i = \eta_i + (t_i - p_i)/\omega_i$, so the design scales by $\sqrt{\omega}$ and $\tilde{y} - \tilde{t} = (y - t)/\omega$. Those weights come from a pilot fit and are therefore functions of the labels. Freezing them makes the map affine in λ, which is what the surrogate needs, and it does not make the map independent of the training labels, which the identity's conditioning already allows for.

Two ways the surrogate differs from the loss. The surrogate is not the loss, and the experiments measure both differences. The fitted parameters of a smooth loss are not an affine function of its targets, so the surrogate's gain is not the loss's gain, and an intercept-only logistic model with label mean $\frac{1}{4}$ and teacher probability $\frac{3}{4}$ already has $R'''(0) = -\frac{32}{9}$. And for a softmax head with more than two classes the curvature is the block $\mathrm{diag}(p_i) - p_i p_i^\top$, which the class-averaged scalar $\overline{p_i(1-p_i)}$ the experiments use approximates rather than reproduces. The scalar is exact only for the one-logit parameterization of the binary case, where the contrast $\eta = z_1 - z_2$ is the single coordinate. Carrying a scalar weight on each of two logits is already not exact, since at $p = \frac{1}{2}$ the block gives $d^\top W d = 1$ along $d = (1, -1)^\top$ against $\frac{1}{2}$ for $\frac{1}{4}I$, and 0 along $(1, 1)^\top$ against $\frac{1}{2}$. Both are why Section 6.4 reports the cross-entropy regime as a boundary condition rather than as a case the identity covers.

D THE DENOMINATOR

Proposition 7 (The denominator is a squared norm)**.** *Let $\hat{\psi} = MX_\ell^\top(y - t)$ be the raw residual put through the learner, with no smoothing. Then $\mathbb{E}\hat{\psi} = \psi$ and $\mathrm{Cov}(\hat{\psi}) = MX_\ell^\top \Omega_{hh} X_\ell M$ with $\Omega_{hh} = (\sigma_y^2 - 2\sigma_{yt} + \sigma_t^2)I$, so $\mathbb{E}\|\hat{\psi}\|_\Sigma^2 = \delta + (\sigma_y^2 - 2\sigma_{yt} + \sigma_t^2)T = D$, so the plug-in is unbiased for D whenever Σ is fixed with respect to the labeled fold.*

None of δ, δ_{raw}, σ_t^2 or σ_{yt} has to be estimated to obtain D this way. This matters because δ is the one quantity the rule does not estimate directly, which is why Section 5 substitutes δ_{raw} for it. Proposition 7 removes the need to estimate it, and measured, it does not help.

The numerator. The numerator does not simplify the same way. $A = b^\top \Sigma \psi + T(\sigma_y^2 - \sigma_{yt})$ contains $b = Lm_y - \theta^*$, and θ^* is the unknown optimum. With ridge, $b = L(m_y - X_\ell \theta^*) - \rho_s MP\theta^*$, which reduces to $-\rho_s MP\theta^*$ when the linear model is well specified and is generally nonzero either way. The asymmetry is the one Theorem 1 describes, since the part that reads straight off the data is not the part that carries the sign. Two further points follow. The statistic is A rather than the predicted gain A^2/D. To first order with the denominator held fixed, the delta method puts the z-score on the gain at half the one on A, so only one of them comes from the regret comparison. The uncertainty σ is also estimable at no extra labeled cost, by jackknifing $\hat{A} - \tilde{\lambda}\hat{D}$ over five folds of the same samples. The raw ratio $\tilde{\lambda} = \hat{A}/\hat{D}$ is held at its full-fit value there, so that the contrast does not vanish identically. The clipped and pair-transformed weights the rule deploys are not $\tilde{\lambda}$, and the vanishing argument is about the raw ratio alone. Those folds are not the ones the scalars are cross-fitted on, and refitting on them buys no labels beyond the student's own.

E WHAT TEACHER ACCURACY IS BLIND TO

Before any theorem, the identity already determines what kind of object the decision is. Distillation replaces part of the student's labels with the teacher's predictions, so it alters three quantities at once, namely the label noise the teacher *removes*, the noise the teacher *adds*, and the systematic error the teacher *is*. The terms of Eq. 4 name all three. Here $\sigma_y^2 - \sigma_{yt}$ carries what the teacher removes, σ_t^2 what it adds, and δ_{raw} with β what it systematically is, although $\sigma_y^2 - \sigma_{yt}$ is the slope contribution of those variances rather than a literal amount removed and may be negative. The optimal weight is where their trade balances.

The two sweeps quoted in Section 3. Both are calculations on a synthetic student, reported to show what the terms do rather than to measure anything. Draw a $d = 6$ standard Gaussian design X over $n = 400$ samples and a standard Gaussian θ^* from `numpy` seed 4, take a well-specified linear mean $m_y = X\theta^*$, set $\Sigma = I$ and $\sigma_y^2 = 1$, and fit the student by ridge with penalty 8, so that $L = (X^\top X + 8I)^{-1} X^\top$ and the student is biased, $b = Lm_y - \theta^* \neq 0$. For the same-view sweep, take the teacher $t = X\theta^*/2$, scale its discrepancy $m_y - t$ by $s \in \{0.05, 0.1, 0.25, 0.5, 1, 2\}$ and read $\lambda^* = [(T\sigma_y^2 + \beta)/(T\sigma_y^2 + \delta)]_{[0,1]}$ at each s, which runs from 0.807 down to 0. For the cross-modal sweep, hold $s = 0.10$, so that β and δ never move, and evaluate λ^* from Eq. 4 at the four pairs $(\sigma_t^2, \sigma_{yt}) \in \{(0, 0), (3, 0), (0.3, 0.30), (1.0, 0.45)\}$, which gives 0.547, 0.180, 0.430 and

0.225. What is held fixed across those four is the projected teacher mean q and the discrepancy m_h that the weight is supposed to read. The teacher's own squared risk is not fixed, since $\mathbb{E}[(y-t)^2] = \delta_{\mathrm{raw}} + \sigma_y^2 - 2\sigma_{yt} + \sigma_t^2$ moves with σ_{yt} and σ_t^2, and the point is that the weight moves by more than a third of its range while the quantity the corollary says determines it does not.

All three appear in the teacher's own risk on the training design, since $n^{-1}\mathbb{E}[\|y-t\|^2 \mid X^b] = \delta_{\mathrm{raw}} + \sigma_y^2 - 2\sigma_{yt} + \sigma_t^2$ with $\delta_{\mathrm{raw}} = n^{-1}\|m_h\|_2^2$. That quantity is a fixed-design average rather than the population teacher risk over the test distribution, and one scalar cannot separate the three terms or read the alignment the weight needs. Teacher accuracy is therefore not blind but *selectively* blind, and measurably so (Figure 4). Across 420 settings the teacher's risk relative to the student's tracks the systematic term at rank correlation 0.892 and the added-noise term at only 0.128, and the added-noise term is itself associated with the optimal weight, at rank correlation 0.327.

The asymmetry replicates on an independent design. Sweeping only the teacher's training budget across twelve directions, which is 360 settings in which only the teacher's budget varies, the same three correlations read 0.867, 0.192 and 0.399. Raw explanatory power, however, does *not* improve. The rank of the optimal weight is explained at $R^2 = 0.175$ by accuracy and $R^2 = 0.264$ by the three terms in the first design, and at 0.307 against 0.309 in the second, where the teacher's budget drives everything so that accuracy proxies the whole state. We therefore claim the asymmetry and not the improvement. Accuracy is informative and insufficient, which is precisely what Theorem 1 asserts. We do not claim it is uninformative.

F Experimental setup

Scale of the evaluation. The evaluation covers 3,906 distinct settings across twenty runs, reaching all 48,755 clips and 141 classes of the VGGSound feature release. It varies the student from ridge regression to a 2.8M-parameter network, the teacher across five function classes including an ImageNet-pretrained ResNet-50, the label budget from 25 to 20,000, the representation from frozen to end-to-end, and the loss from squared error to cross-entropy.

The criteria each theorem covers. Theorem 2 holds the joint distribution of (x, t) fixed, so it closes every criterion computed from the student's features and the teacher's outputs alone. Those are teacher confidence, margin and output spread, the Ratio 1-2 criterion (Gosal et al., 2026) and the secondary-soft-probability metric it improves on (Tan & Liu, 2022), linear CKA between the views, and GRACE (Panigrahi et al., 2026) in the form we score. Linear CKA is the one whose input needs saying: the implementation scores $\mathrm{CKA}(X^a, X^b)$ between the two raw views, which is not in general a function of (x, t). In Theorem 2's construction the teacher is a function of the student's own view and both problems take $X^a = X^b = X$, so the joint distribution of the two raw views is shared and any view-similarity score takes the same value in both. Equating t alone would not do it: a teacher can ignore a coordinate, and two teacher views agreeing on t while differing on an unused coordinate give linear CKA of 1 and nearly 0.

The teacher-to-student risk ratio is the eighth baseline. It reads labels, so it is not label-blind. Theorem 1 equates the *population* risk ratio across its two problems, which closes any selector given that summary. The empirical ratio a practitioner computes from finite labels is a different statistic: in that construction the two problems put different distributions on it, since the teacher of one predicts $\frac{1}{2}$ everywhere and can never be exactly right while the other's is exactly right with probability $(3/4)^2$ on two evaluation labels. That baseline is therefore answered by measurement in Table 2 and not by the theorem. Of the remaining four in Table 2, teacher–student agreement, prediction correlation and NTDH (Chen et al., 2024) read the fitted student, and CCH (Xie et al., 2025) reads the labels.

The gradient-covariance proxy. GRACE (Panigrahi et al., 2026) scores a teacher by the spectrum of the student's gradients on its outputs, taken at a *pretrained* student. What we score is a proxy in that spirit rather than their statistic, and Table 2 labels it as such. With $g_i = -x_i\tau_i^\top$ the gradient of a zero-initialized half-squared-loss head, we form $\hat{\Sigma}_g$ as the *centred* covariance of $\{g_i\}$, stabilize it as $\hat{\Sigma}_g + \nu d\, I$, and score with the same centred second moment on both sides. The Kronecker product $-x_i\tau_i^\top$ is projected on both factors, so the score depends on the teacher, and it separates the three teachers in all 48 direction-and-seed groups. Their definition normalizes each gradient to unit norm on the inverse side, leaves the evaluation side uncentred, and stabilizes with $\nu/d\, I$, which are three

separate differences. No reparameterization of ν absorbs the first two, so this is a different statistic rather than a tuning of theirs. Evaluating at initialization is a fourth difference, and the one that matters for the theorem: there the score is available in closed form and is a function of the student's features and the teacher's outputs alone, which places it inside Theorem 2. We therefore read its result as a statement about this proxy, not as a verdict on GRACE.

Implementations of the published objectives. Plain KD is the convex mixture $(1-\alpha)\mathrm{CE} + \alpha T_{\mathrm{KD}}^2 \mathrm{KL}$, so $\alpha > 1$ leaves that parameterization and makes the cross-entropy coefficient negative. Whether the objective is then unbounded below depends on the teacher and the temperature as well, the asymptotic slope in the true class's logit being $1 - \alpha + \alpha T_{\mathrm{KD}} q_y$. We sweep each objective only over the range its own form admits, which leaves every number in Table 20 unchanged, since the tuned column never selected the region we removed. Five of the others take α as a positive global scale on a term added to the cross-entropy, where $\alpha > 1$ is well posed, and PLD has no such parameter at all. We implement each objective's loss as its paper writes it and run them all at a common $T_{\mathrm{KD}} = 4$ and unit scale, which is not the same as reproducing each paper's tuned hyperparameters: WTTM, for one, reports different γ and β per teacher-student pair. CTKD is the exception to the common temperature, since a learned temperature is what it contributes: its Global-T module is initialized at 4 and trained from there. DKD, DIST and LSKD are run from the authors' released loss code. CTKD, WTTM and PLD are implemented from their equations, the contribution in each being a procedure rather than a copyable loss. Each keeps the internal constants its own paper fixes, while the temperature and the global scale are replaced by the common protocol above, and each is swept over $\alpha \in \{0.1, 0.25, 0.5, 1, 2, 4\}$, restricted to $\alpha \leq 1$ for plain KD. The grid excludes $\alpha = 0$, so the tuned column of Table 20 reports the best *non-zero* weight rather than the option of not distilling, which none of these objectives contains.

Datasets and views. UCI-HAR (Anguita et al., 2013) supplies three raw inertial streams, namely total acceleration, body acceleration and gyroscope, kept strictly apart, so the 561-dimensional engineered vector shipped with the dataset is *not* used, since it mixes the sensors and would leak the teacher's view into the student. The three streams give six ordered directions, split by subject over 30 subjects in every run but one: the end-to-end 1D-CNN runs split the concatenated windows at random instead, so all 30 subjects appear on both sides and those runs measure generalization to new windows of the same subjects rather than to new subjects. They also standardize each channel over all windows rather than over the training samples alone. Both are properties of that one producer. The frozen-feature and MLP runs take their splits from the subject identifiers. RAVDESS (Livingstone & Russo, 2018) pairs audio with video over $1{,}440$ clips and 8 emotions, split by actor. NYU-Depth v2 (Silberman et al., 2012) pairs RGB with depth over $1{,}449$ frames, split by scene over 464 scenes. The target is the presence of each of the 40 semantic classes, a 40-dimensional multi-label vector, and RGB and depth are the two views of it rather than one being predicted from the other. VGGSound (Chen et al., 2020) pairs audio with video, and we use its released feature archive in full, $48{,}755$ clips over 141 classes, which is a subset of the full corpus. It has no grouping variable, so its splits are random. MIMIC-CXR (Johnson et al., 2019) pairs a frontal chest radiograph with its radiology report over $64{,}775$ studies, split by subject. Parkinsons telemonitoring (Tsanas et al., 2009) pairs the clinical motor-UPDRS score with the voice-feature vector, split by subject. Every dataset contributes both directions. In every one of them the label is held fixed, and what reverses is which view the teacher reads and which the student reads. The reverse direction is therefore a different transfer problem rather than a relabelling. What a depth teacher knows about a room's contents that an RGB student cannot recover is not what an RGB teacher knows that a depth student cannot.

Two datasets whose labels depend on one view. MIMIC-CXR's 14 CheXpert findings are extracted *from the reports*, so in the frozen-feature direction the teacher's view is the source of the labels, and in the ImageNet-teacher experiment of Table 19 the *student's* view is. Parkinsons' target is total UPDRS and motor UPDRS is one of its components, so in the forward direction the teacher reads a part of the label. Neither is an accident of our pipeline, and both are properties of the datasets. For MIMIC-CXR the label-derived teacher plausibly flatters the customary weight, since it is the direction where distillation helps most, and dropping MIMIC-CXR in Table 24 does make that weight look worse. A single deletion is consistent with that reading rather than an identification of it. Parkinsons at a relative teacher risk of 0.734 is the harmed end of Table 6, where a teacher

holding part of the label still loses 6.9% at $\lambda = 0.5$. RAVDESS is the complement: its emotion label is the one the clip was acted to, assigned when the corpus was recorded rather than read off either view.

Features. The frozen-feature student is ridge on features extracted once, and what "features" means differs by dataset, as stated below. UCI-HAR and Parkinsons use engineered summaries, namely per-channel time statistics and 32 FFT bands for HAR, and the dataset's own voice descriptors for Parkinsons. RAVDESS uses the MST-Distill 156-d audio feature and a CLIP ViT-B/16 video embedding pooled over 15 frames. VGGSound pools the released audio and visual features to 128 and 512 dimensions. MIMIC-CXR uses a frozen MedSigLIP-448 image embedding (1152-d) and a frozen Bio-ClinicalBERT report embedding (768-d). MedSigLIP's pretraining may overlap MIMIC-CXR, which we could not resolve exactly, and Table 24 reports the dataset-deletion sensitivity. NYU-Depth uses ResNet-34 `layer3` maps for both views, average-pooled over space to 256 dimensions each.

Students and teachers. The ridge on those features is the student the identity is written for. The neural students are a 512–256 MLP trained by AdamW and four capacities from 23K to 2.8M parameters. The end-to-end students are 1D CNNs trained from raw signals, on UCI-HAR's inertial channels and on RAVDESS's audio sequence. Teachers are ridge, gradient boosting, k-NN, random forest, an MLP, and an ImageNet-pretrained ResNet-50 (Deng et al., 2009; He et al., 2016). Teacher quality is swept in two designs: 50 to 1600 labels over twelve directions, and 40 to 1280 over the same twelve with five seeds. Four of the sixteen directions are absent from both, and for an arithmetic reason rather than a chosen one: NYU-Depth and RAVDESS have about 870 samples in the training pool, and a sweep that must fit 250 student labels, up to 1280 teacher labels and 200 unlabeled pairs into that pool does not fit. The condition is evaluated before any outcome is seen.

The scope of the label budget n. For the ridge students n is every label used. The MLP students of Tables 3 and 23 are trained with early stopping, and their checkpoint is selected on a group-disjoint validation split holding 15% of the dataset's groups, drawn from outside the pool n comes from. Those labels are outside n. The remaining neural students do not use a validation split at all: the end-to-end CNNs and the pretrained-pair students run a fixed 60 epochs, and the VGGSound and MIMIC heads a fixed 40, so for them n is every label used. Where the split is used it is part of the training pipeline that every arm in those runs shares, including the supervised baseline and the sweep, so it is not an advantage our rule holds over its comparisons. No label of that split reaches the weight decision, which is what Section 5 means by costing none. A comparison held strictly to n would have to fix the epoch count in advance or take the validation split out of n for every arm. We did neither, and report the budget as n plus that shared early-stopping split.

The populations this paper reports on. Ten populations appear, and three of them have 144 cells without being the same cells, so each is named by where it is reported.

Cells	Where	What it is
$1{,}504$	Table 1	the six regimes of the asymmetry, every cell swept over the grid
$1{,}048$	Figure 1, Table 31	those of the $1{,}504$ in which some weight helps, which the normalized gain needs
802	Table 3	the pooled policy population, composed below
544	Table 3, left block	those of the 802 in which some weight beats the supervised student
340	Tables 11, 22 and 24 (upper)	the main ridge run at one validation fraction
420	Tables 23, 24 (lower)	the five teacher families
222	Tables 12, 13, 14 and 28	the label-budget run at one validation fraction, $n = 25$ to 150, which is not a subset of the 340
144	Tables 2 and 10	the criterion comparison, on which every criterion can be computed
144	Tables 15 and 16	the pair experiment, which shares no cell with the criterion comparison
144	Appendix H.7	the bridge-penalty sweep

The 802 cells of Table 3 are what remains of 1040 candidates after 7 for $\hat{D} \leq 0$, 71 for a repeat of a setting another run already contributes, and 160 for no stored scalars.

Population	Cells	**Ours**	λ=0.5
pooled, as reported	802	$+\mathbf{0.89}\%$	-26.47%
+ the 7 with $\hat{D} \leq 0$, at $\lambda = 0$	809	$+\mathbf{0.88}\%$	-26.22%
+ the 160 whose run stores no scalars	969	$+\mathbf{1.01}\%$	-24.00%
+ the 360 of the withheld regime	1,329	$+\mathbf{0.89}\%$	-17.62%

Table 5: **What the pooled design's exclusions are worth.** The three exclusions above, and the regime Appendix H.13 scopes out, restored one group at a time and cumulatively. The repeats are not restored, since a setting two runs both evaluate is one setting. Each row adds the previous row's cells back. Our arm stays between $+0.88\%$ and $+1.01\%$, and restoring everything returns it to the $+0.89\%$ the main text reports, while restoring the withheld regime alone gives $+0.79\%$. The customary weight looks less bad as the harder populations come back, so the margin narrows while the sign does not move. On the seven degenerate cells our rule returns $\lambda = 0$ and gains nothing while the customary weight gains $+2.27\%$, so restoring them is the direction that costs us, and it costs 0.26 points of a 27.4-point margin.

Grids, budgets and seeds. Most runs sweep the distillation weight over the seven-point grid $\{0, 0.1, 0.2, 0.3, 0.5, 0.7, 0.9\}$. Four kinds of run depart from it, each for a stated reason. The two teacher-quality sweeps use ten points down to 0.05, because their optima are small. One run uses seventeen points down to 0.025 to measure what the coarse grid censors, since a grid cannot report an optimum below its smallest non-zero point. The MIMIC pair experiment uses ten points down to 0.002, because its optimum is 0.0056 and the seven-point grid cannot represent it. The unlabeled-pair run uses ten evenly spaced points to 0.9, because it reads the grid through $w = \lambda/(1-\lambda)$ and an even spacing in λ is what spreads w over its range. Appendix H reports what that censoring costs. Student label budgets run from $n = 25$ to $n = 20{,}000$. Seeds are three to five per cell in most runs, two in the student-capacity sweep, eight where the student learns its representation, and twenty-four in the two runs with a pretrained teacher, which are small enough to run deep. The bridge penalty is $\rho_b = 100$ throughout, while the student's own ridge penalty is $\rho_s = \max(1, d)$ and is not ours to count, since the practitioner sets it either way.

G Additional tables and figures

Reading the tables. Unless a caption states otherwise, every percentage in this appendix is a change in the student's risk against its own supervised risk, with positive meaning the student improved. *Worst cell* and *worst dir.* are the minimum of that quantity over cells and over the sixteen teacher-to-student directions, *best* λ is the gain the best weight on the grid attains, and *cells* counts evaluated configurations. Tables 18 and 20 report accuracy percentage points instead, and their captions state it.

Setting	Teacher / student risk	Best λ	λ^*	KD at $\lambda = 0.5$
RAVDESS ← **ResNet-50**	0.882	$+8.23\%$	0.70	-8.69%
Parkinsons	0.734	$+0.00\%$	0.00	-6.86%
HAR tacc→bacc	0.851	$+0.00\%$	0.00	-6.82%
107 further settings in the same band				
RAVDESS ← **ResNet-50**	0.672	$+30.23\%$	0.50	$+30.23\%$
RAVDESS ← **ResNet-50**	0.703	$+31.82\%$	0.50	$+31.82\%$
RAVDESS ← **ResNet-50**	0.608	$+35.23\%$	0.50	$+35.23\%$

Table 6: **Same teacher accuracy, opposite outcomes.** *Best* λ and λ^* are the gain the best weight on the grid attains and its location. Percentages are of the student's own risk. 113 settings whose teacher is 10–40% more accurate than the student: outcomes span -8.7% to $+35.2\%$, harmful in 34% of them.

H Additional results

This appendix holds the measurements the main text cites but does not carry: the two regimes with the least available gain, the weight's estimation error, and the checks a reader would run against the rule.

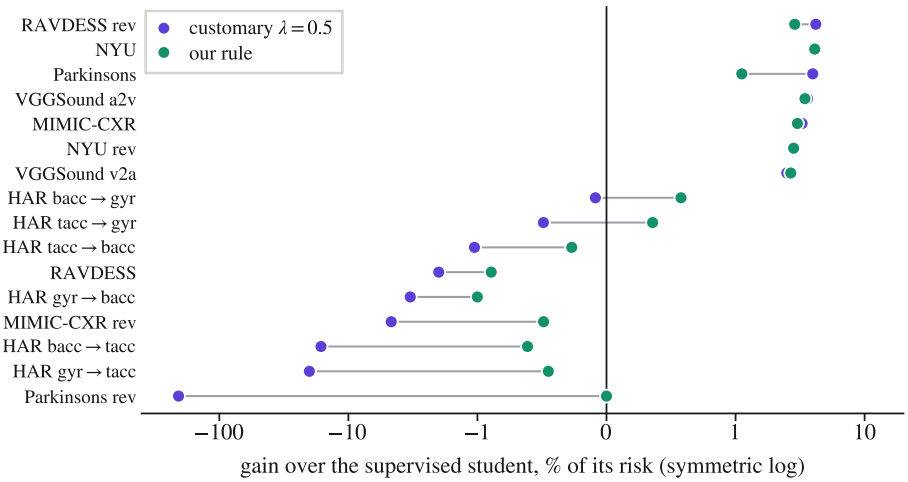


Figure 2: **Per-direction outcomes.** One row per direction, the two dots being its mean under the customary weight and under the rule, as a percentage of the supervised student's risk with positive meaning better. Rows are ordered by what the customary weight costs, and the axis is symmetric-log because the two policies differ by more than two orders of magnitude in what they can cost. The rule never falls below -1.00%, while the customary weight reaches -206.95%. Main-ridge run, 340 cells, which is not the pooled population of Table 3.

		544 live settings			all 802		
Policy	Fits	Mean	Worst dir.	Hurt	Mean	Worst dir.	Hurt
always, $\lambda = 0.5$	1	-3.72%	-89.78%	5/16	-26.47%	-172.50%	11/16
never distill	1	$+0.00\%$	$+0.00\%$	0/16	$+0.00\%$	$+0.00\%$	0/16
skip if $\hat{G}(0.5) < 0$	1	$+1.55\%$	-2.20%	1/16	$+0.61\%$	-1.89%	6/16
skip if $\hat{G}(\hat{\lambda}) \leq 0$	**1**	$+1.77\%$	-0.30%	2/16	$+0.73\%$	-1.32%	5/16
ours, shrunk	**1**	$+1.55\%$	$+0.00\%$	0/16	$+0.89\%$	-0.47%	4/16
best per cell, grid search	8	$+3.64\%$	$+0.75\%$	0/16	$+2.47\%$	$+0.32\%$	0/16

Table 7: **Policies, with the two intermediate variants of our own rule.** The main text reports the first two rows, the shrunk rule and the grid-search oracle. Declining on the sign at the customary weight and declining on the sign at $\hat{\lambda}$ are the two steps between them, and on the live subset the shrinkage exchanges 0.2 points of mean for a worst direction of zero.

H.1 The two regimes with the least available gain

The two are the adversarial-teacher regime of Table 19 and the cross-entropy regime of Table 18, and in both the rule distills and loses. In the adversarial-teacher regime the rule costs -10.32% against -35.76% for the customary weight, having deployed a weight well above the optimum. $\hat{A} > 0$ predicts only that some small weight would help. Improving at the weight actually chosen needs $2A > \lambda D$, and that is what fails here. Abstaining is not the right answer there either, since the best weight on the grid is 0.0056 and earns $+1.14\%$ (Table 19), so what the rule gets wrong is the size of the weight and not the decision to use one. Under cross-entropy the picture is the same at a smaller scale. On those cells the best temperature and weight on the grid are worth $+0.46$ accuracy points while the customary setting costs -7.40, so our -0.92 is a loss where little was available rather than none. Table 18 separates that loss from the metric: read under Σ the rule loses 5.80 accuracy points, and under the Gauss–Newton metric of Proposition 6 it loses 0.92. Under Σ the weights also sit above the grid optimum, in 114 of the 132 cells. Under the Gauss–Newton metric their mean falls below that optimum. We read the two regimes as the edge of the rule's reach: it is put into settings where little is available and it spends more than that little.

Two corrections to the gain at the best weight. Two corrections pull that gain in opposite directions and the larger one favours the reader. The best weight is chosen and scored on the same test set, so its gain is optimistic. Cross-fitting the choice against the score lowers it by 0.011 points, from 2.315% to 2.304%. Against that, the seven-point grid cannot represent an optimum below its smallest non-zero weight, and a seventeen-point grid raises the mean from 2.735% to 2.905%, which is fifteen times the first correction. The two are measured on different populations and cannot be combined into a net, and each bounds a different sensitivity of that gain.

Regime	Cells	$\lambda = 0.5$	Ours	
frozen features, $n = 25$–150	222	-14.99	$+1.49$	wins
five teacher families	420	-36.94	$+0.74$	wins
four student sizes	160	-12.75	$+1.69$	wins
teacher on the unlabeled pairs	144	-63.97	$+0.23$	wins
cross-entropy, 141 and 8 classes	132	-7.40	-0.92	limits the loss
VGGSound, 141 classes, 20k labels	18	$+8.12$	$+8.93$	wins
teacher $3.26\times$ worse, 2500 pairs	72	-35.76	-10.32	**loses**

Table 8: **Every regime, and what the rule does in it.** Our arm against the customary weight, with five regimes won outright, one where the loss is limited rather than avoided, and one that is lost. Of these rows, only the teacher-family one belongs to the pooled design of Table 3. 5 of 7 regimes won outright; units are points of risk except the cross-entropy row, which is accuracy. The two rows the rule does not win are the two with the least available to gain, consistent with the association Table 9 measures.

What an oracle weight could win	Cells	Share of all	Harmed
under 1%	247	41%	87%
1–3%	134	22%	5%
3–10%	163	27%	1%
over 10%	61	10%	3%

Table 9: **The harm rate against the gain available.** Rows split every setting by what the best weight on the grid could win in it. At $\lambda = 0.5$ the customary weight harms 87% of the band offering under 1% and 5% or less of every band above it. Across the grid that band's harm rate runs from 60% at $\lambda = 0.1$ to 98% at $\lambda = 0.9$, so no fixed weight escapes it. These cells measure that association rather than Remark 1 implying it, and the low band is 41% of all settings in which distillation would be attempted. 605 settings whose teacher is the better of the two; Spearman between the gain at the best weight and being harmed is -0.812. In the lowest band the oracle could win 0.122%, the customary weight loses a median -1.82%, and its harm rate is 86%, 96%, 100% in the 3 datasets that contribute.

H.2 The unlabeled-pair design

Labelling the unlabeled pairs rather than the student's own samples raises the gain at the best weight from $+1.83\%$ to $+2.82\%$ and the damage from -21.52% to -63.97%, so the asymmetry roughly doubles while our decision transfers unchanged (Table 15). Weighting those rows by $w = \lambda/(1-\lambda)$ makes the design depend on w, and Proposition 5(i) gives the matrix law that replaces the quadratic. Table 16 measures the single-γ collapse of that law, which is the approximation the rule uses, against the curve itself.

H.3 Per-direction detail

Table 22 gives every direction, Table 23 every teacher family and student size, and Table 24 the leave-one-dataset-out check. Figure 2 plots the two policies side by side per direction of the main-ridge run, where the rule never falls below -1.00% while the customary weight reaches -206.95%.

H.4 Teacher quality and matched pairs

Table 25 varies only the teacher's quality, and Table 26 shows that the matched pairs do not depend on the two tolerances that define them.

H.5 The value of the decision

Three measurements locate it. At the customary weight the alarm reaches AUC 0.863 over the 802 pooled cells, and deleting any one dataset leaves it between 0.830 and 0.890. The alarm holds at every weight on the grid and is sharpest around $\lambda = 0.7$ (Table 27), the regime in which harm is common and being wrong is expensive. Per dataset it is uneven, RAVDESS sitting at chance and NYU-Depth's 1.000 resting on six harmed cells, which is why the pooled figure is read through the deletions of Table 24 rather than through the six values. The value is concentrated where the customary weight is catastrophic and there is nothing to gain, which is the association Table 9 measures. And Table 28 separates the two contributions, which Appendix H.10 reads in full. The

Teacher accuracy measured on	Which costs	AUC	Transferred	Worst
a labeled test set	test labels the student does not have	0.775	-0.13%	-2.02%
the student's own fold	nothing, but the fit is in-sample	0.602	$+0.37\%$	-0.33%
a held-out fifth of n	20% of the student's labels	0.738	-0.31%	-2.88%
ours: $\hat{G}(0.5)$	**no additional labels**	$\mathbf{0.834}$	$\mathbf{+0.51}\%$	-2.62%

Table 10: **What the strongest baseline costs.** Teacher accuracy is not one criterion but three, depending on what one is allowed to measure it on. *AUC* is for the distill-or-not decision, *transferred* is the mean the resulting policy realises and *worst* its worst direction. The criterion's strength is proportional to the labels it is given. The conservative free version does have the better tail, but it obtains that tail by almost never distilling, which costs it the mean. Each threshold is calibrated on the other 11 directions and applied to the held-out one; ours is 0 by construction.

Numerator of $\hat{A}$	AUC at $\lambda = 0.5$	Mean	Worst cell
$T(\hat{\sigma}_y^2 - \hat{\sigma}_{yt})$, without $\hat{\beta}$	0.804	-0.40%	-31.67%
$T(\hat{\sigma}_y^2 - \hat{\sigma}_{yt}) + \hat{\beta}$	0.831	$+0.30\%$	-31.67%
the deployed rule (Algorithm 1)	—	$\mathbf{+0.88}\%$	$\mathbf{-6.91}\%$
best per cell, grid search	—	$+2.26\%$	$+0.00\%$

Table 11: **What the bias alignment is worth.** The numerator of $\hat{A}$ built up one term at a time, on the same cells and against the same question. *AUC* is for the decision at $\lambda = 0.5$, and the last two columns are the risk of acting on the weight each numerator implies. The stored scalars satisfy $\hat{A} = T(\hat{\sigma}_y^2 - \hat{\sigma}_{yt}) + \hat{\beta}$ to machine precision, so the ablation retrains no student. 340 main-ridge cells under the unclipped squared error the rest of the paper reports. Both unshrunk arms use the same denominator, grid and predicted-gain gate, and differ only in the bias-alignment term: adding it is worth $+0.70$ points, $p = 2.0 \times 10^{-16}$, better on 120 cells and worse on 27, and it changes the sign of $\hat{A}$ on 43.

shrinkage that produces the weight is the plug-in form Proposition 4(ii) motivates for a fixed oracle s, measured against the two hard gates in Table 14.

H.6 The published criteria on the decisions they were built for

Scored on the decision each was built for rather than ours, the published criteria do work: Ratio 1-2 picks the best of three teachers well above chance while sitting at chance on whether to distill at all (Table 29). The per-sample reweighting schemes are the same story about a different object, two of them rescued by about fifteen points once the global scale comes from our rule instead of from the fixed 0.5 (Table 30). Table 31 is the asymmetry law measured against its own prediction, with no fitted parameter.

H.7 Sensitivity to the bridge penalty

The bridge penalty ρ_b is the only constant the rule chooses, and $\hat{q}$ is the nuisance whose error this sweep can vary directly. The conditioning argument removes none of the nuisance errors, as Section 5 and Appendix H.10 both state. Both are measured on the same sweep, over 144 cells and five penalties from 1 to 10,000, at the plug-in weight: the gate and the projection, without the jackknife shrinkage, so that what varies over the sweep is the bridge and not the rest of the rule.

ρ_b	1	10	100	1,000	10,000
$\hat{A} \leq 0$, the gate fires	16.0%	16.0%	16.7%	17.4%	18.1%
weight 0, the cell is not distilled	26.4%	26.4%	26.4%	27.1%	27.8%
mean	$+0.911\%$	$+0.891\%$	$+0.861\%$	$+0.708\%$	$+0.819\%$
worst cell	-3.98%	-3.98%	-6.44%	-6.78%	-2.83%

Both rates vary by at most 2.1 points, three of the 144 cells change whether they are distilled at all, and the sweep moves the mean by 0.203 points and leaves it positive at every penalty. The two rates differ because the gate and the grid abstain separately: $\hat{A} \leq 0$ stops a sixth of the cells, and the projection sends a further tenth to weight zero on its own. The tail does move, from -2.83% to -6.78%, and we report that rather than claiming the worst case is stable. The sweep spans bridges

Policy	Fits	n=25	n=75	n=150	Worst cell	Worst dir.
distill at $\lambda = 0.5$	1	-11.03%	-15.09%	-19.97%	-434.87%	-215.52%
sweep, matched labels	7	$+0.39\%$	-0.14%	-0.96%	-24.84%	-8.16%
sweep, then refit on all n	8	$+2.37\%$	$+1.42\%$	$+0.94\%$	-20.77%	-2.13%
ours	1	$+2.14\%$	$+1.15\%$	$+1.09\%$	-9.69%	-1.36%

Table 12: **Against a validation sweep, at matched labels and given its labels back.** A sweep must hold out a validation split, which in the cross-modal regime comes out of the student's own budget. Held to the same labels it is beaten at every budget and never clears not distilling at $n \geq 75$. Given those labels back and a final refit, the two means are close and it is twice as bad in the tail. The *fits* column is what each implementation runs rather than what a sweep requires. On this student the weight path is affine (Appendix H.8). Ours against the matched-label sweep: $+1.67$ points, $p = 1.7 \times 10^{-13}$; against the refit sweep: -0.14 points, $p = 0.554$, at one fit against eight.

n	Validation split	Sweep, matched	Sweep, then refit	**Ours**
25	20%	$+0.69\%$	$+0.68\%$	$\mathbf{+2.14\%}$
	33%	$+0.39\%$	$+2.37\%$	$\mathbf{+2.14\%}$
75	20%	-0.03%	$+0.19\%$	$\mathbf{+1.15\%}$
	33%	-0.14%	$+1.42\%$	$\mathbf{+1.15\%}$
150	20%	-0.10%	$+0.73\%$	$\mathbf{+1.09\%}$
	33%	-0.96%	$+0.94\%$	$\mathbf{+1.09\%}$

Table 13: **What the sweep's split fraction is worth.** A sweep must decide how much of n to hold out, and the two fractions the run covers put its arms more than a point apart at the same budget. Deciding it would need labels of its own. Our rule holds nothing out and has no such choice. The same cells as Table 12, which reports the 33% rows. The fraction is worth up to 1.69 points to the sweep, and nothing is left over to choose it on. Our rule holds no split out, so its column repeats, and it is ahead of both sweep arms in 4 of the 6 settings.

from nearly unregularized to nearly constant, so it varies $\hat{q}$ over a wider range than a single fit would, and the rates stay stable across that range while the worst cell does not.

H.8 What a validation sweep costs on a fixed linear map

For the ridge student the weight path is affine, $\hat{\theta}_\lambda = (1-\lambda)\hat{\theta}_0 + \lambda\hat{\theta}_1$, so one factorization with two right-hand sides generates every weight on the grid. The squared-loss validation optimum is then available in closed form as $[\Delta p^\top (y_{\text{val}} - p_0)/\|\Delta p\|^2]_{[0,1]}$ with $p_0 = X_{\text{val}}\hat{\theta}_0$ and $\Delta p = X_{\text{val}}(\hat{\theta}_1 - \hat{\theta}_0)$. It takes $\lambda = 0$ when $\Delta p = 0$, where every weight gives the same validation loss. The seven and eight fits in Table 12 are therefore what our implementation runs, not a lower bound, and the rule's advantage on that student is the validation split rather than the compute. The shortcut does not reach the neural or end-to-end students, whose solution maps are not affine in the targets, nor the pair setting, where the design moves with the weight.

H.9 The two steps of the policy apart

Table 7 separates the gate from the shrinkage on the policy comparison of Table 3, for readers who want the two steps apart.

H.10 The weight's estimation error

The estimated weights have a narrower spread than the optimal ones, which we call *compression*, and a compressed $\hat{\lambda}$ costs less on the decision than on the weight, since the decision needs only the side of the threshold $A/D > \lambda/2$, and a compressed ratio crosses it only where the true ratio is near the threshold or the compression is severe. Figure 5 shows the two populations separating across that threshold at error rates of 38% and 14%. The weight's error is cell-specific, and the one substitution that does reduce it buys almost no risk. The direct denominator lowers the weight's mean squared error by 11% and moves the mean gain by 0.012 points.

Where the error comes from. Three separable measurements establish that the error is cell-specific. Averaging the weight over a setting's five seeds removes 7% of its squared error. Cross-

Policy	Mean	Gain $< 1\%$	Gain $> 3\%$	Worst cell
plug-in $\hat{\lambda}$, no abstain	$+0.74\%$	-3.85%	$+5.06\%$	-35.28%
+ hard gate $\lvert\hat{A}\rvert < \sqrt{\hat{\sigma}^2 + \hat{b}^2}$	$+0.85\%$	-1.39%	$+3.01\%$	-19.29%
+ hard gate $\hat{s} \le 1/2$	$+0.85\%$	-1.39%	$+3.01\%$	-19.29%
shrunk $\hat{s}\hat{\lambda}$, no hard gate	$+1.49\%$	-0.93%	$+3.79\%$	-9.69%

Table 14: **Proposition 4(ii), measured.** Four abstain policies on the same cells with the same plug-in weight. *Gain* $< 1\%$ and *gain* $> 3\%$ split the cells by what the best weight on the grid could win. The two hard gates are the same policy, verified rather than assumed, and the continuous shrinkage dominates both. The redundant test is the one on the sign of $\hat{A}$, which the clip already applies. The confidence gate is not redundant and is not the same decision as the shrinkage: at $\hat{A} = 0.1$, $\hat{D} = 0.1$ and $\hat{\sigma}_{\mathrm{j}}^2 + \hat{b}_{\mathrm{j}}^2 = 0.04$ the gate abstains while the shrinkage returns 0.2, which survives projection onto the grid. The two hard gates agree on all 222 evaluated cells, the equivalence of Proposition 4(ii) holding here up to the boundary $\lvert\hat{A}\rvert = \sqrt{\hat{\sigma}^2 + \hat{b}^2}$, which no cell reaches; the shrinkage improves the mean over all of them by $+0.64$ points and the worst cell from -19.29% to -9.69%, and it differs from them on 92 cells, where it is worth $+1.55$ points ($p = 3.3 \times 10^{-8}$ over all cells).

Where the teacher's output goes	Best λ	KD λ=0.5	Ratio	Our gate	AUC
on the student's labeled samples	$+1.83\%$	-21.52%	$11.7\times$	-0.25%	0.816
on the unlabeled pairs	$+2.82\%$	-63.97%	$22.7\times$	$+0.23\%$	0.821

Table 15: **The same comparison with the teacher on the unlabeled pairs.** Same directions, teachers and budgets, and the only change is whether the teacher's output lands on the student's n labeled samples or on 1000 unlabeled pairs. *Ratio* is the loss at $\lambda = 0.5$ against the gain at the best weight, *our gate* the mean our decision realises, and *AUC* its accuracy on the distill-or-not decision. Using the pairs is better, as the practice assumes, and the damage is larger too. The asymmetry roughly doubles, while our decision transfers unchanged with nothing re-derived. The $R^2 = 1.000$ reported here is a consistency check on the implementation rather than a proof of Proposition 1's premise, since a quadratic risk can come from a predictor that is not affine. The premise itself comes from the ridge closed form, and the accuracy of the six scalars is measured separately. The pairs raise the gain at the best weight in 85 of 144 cells, $+0.99$ points, $p = 2.3 \times 10^{-9}$; the gain curve is exactly quadratic on the labeled samples ($R^2 = 1.000$) and only approximately so on the pairs ($R^2 = 0.899$).

fitting keeps each nuisance out of the samples it is applied to and removes none of their estimation error, in $\hat{q}$, $\hat{m}_h$, $\hat{b}$ or the scalars alike. Rescaling every weight by the best linear map fitted on the oracle's own answers removes 18%, which bounds what an affine recalibration can buy. A richer one might buy more, but it would have to be fitted on answers a practitioner does not have. The two-copy estimator gives $\mathbb{E}\hat{\sigma}_y^2 = \sigma_y^2 + \mathrm{bias}(n_{\mathrm{f}})^2$ when the two copies carry the same mean bias, and the product of their biases otherwise, which the lean below reads.

The direct denominator. Proposition 7 estimates D without that decomposition, and the weight it implies keeps 0.92 of the oracle's spread against the shipped rule's 0.74 and ranks with the oracle at 0.52 against 0.38. On the same cells it returns $+0.890\%$ against the shipped rule's $+0.878\%$. The compression is real, it is removable, and removing it does not pay.

Which way the estimate leans. Where the two copies carry the same mean bias the contamination in $\hat{\sigma}_y^2$ is a squared bias and therefore one-signed. With unequal biases it is their product and can go either way. In the one-signed case, held in isolation it enters $\hat{A}$ and $\hat{D}$ with the same coefficient, and $\frac{d}{dc}\frac{A+c}{D+c} = (D - A)/(D + c)^2$, so it pulls the ratio toward one: upward where $A < D$, which is where the optimum is interior and where these cells sit. The observed lean is in that direction, which is consistent with an isolated $\hat{\sigma}_y^2$ perturbation rather than evidence that it causes the lean: the other nuisances move $\hat{\sigma}_{yt}$, $\hat{\delta}_{\mathrm{raw}}$ and $\hat{\beta}$ as well, and their net direction is not fixed by this argument. Which weight is counted decides what is seen. On the runs that store the deployed weight, one validation fraction per setting, the clipped proposal $[\hat{A}/\hat{D}]_{[0,1]}$ over-distills in 64% of the 642 cells where it differs from the oracle's ($p = 4 \times 10^{-12}$). After the shrinkage that falls to 52%, and the deployed weight is at 48% of 461, neither separated from an even split. The lean is a property of the estimate before shrinkage, and the shrinkage is what removes it, which is the same reading as the failure in Table 19. Of the components, $\hat{\sigma}_y^2$ is $\sqrt{n}$-consistent in simulation against a closed-form truth, and the ratio is not.

Law for the pair setting	γ from	Median R^2	Beats quadratic	p
$Q(w)$, with no design correction	$\gamma = 0$	0.7504	0/144	—
$Q(w)/(1+\gamma w)^2$	the trace ratio	0.8058	65/144	1.1×10^{-4}
$Q(w)/(1+\gamma w)^2$	teacher energy	-9.2240	40/144	8.3×10^{-17}
$Q(w)/(1+\gamma w)^2$	**the spectrum's median**	0.9428	126/144	7.7×10^{-15}
weight the law implies		mean	worst cell	share of best
$\hat{w} = \hat{A}/\hat{D}$, no design correction		-0.05%	-40.72%	-2%
$\hat{w} = \hat{A}/(\hat{D} + \gamma\hat{A})$		$+1.05\%$	-5.72%	37%
oracle weight w, from a grid search		$+2.82\%$	—	100%

Table 16: **The pair setting's law, and the summary it needs.** The upper block fits four candidate laws to the same measured weight curves, differing only in which summary of the spectrum γ they use, and the first sets $\gamma = 0$, which drops the design correction and leaves a quadratic in w. The lower block scores the weight each law implies on realized risk. The spectrum is a property of the two designs, so it costs no labels and no student fit, while the summaries are weighted by the teacher's outputs. Which summary to use is settled by these measurements rather than by the law. Where the metric is diagonal in these coordinates the exact stationarity condition weights coordinate j by $\sigma_j \mathbb{E}[(u_j + wr_j)(r_j - \gamma_j u_j)]/(1 + w\gamma_j)^3$, and otherwise it carries N's off-diagonal terms as well. That weight depends on the supervised error and on w itself, so no fixed summary locates the optimum in general. The measurements show that the median tracks the curve's shape and the teacher-weighted mean tracks the optimum's location on these designs. Weighting the *curve* by teacher energy is worst, because the risk divides coordinate j by $(1 + w\gamma_j)^2$ and suppresses what that weighting promotes.

Policy	n=1000	n=5000	n=20000	Worst cell
best λ, from a grid search	$+10.01\%$	$+10.36\%$	$+9.51\%$	$+7.37\%$
distill at $\lambda = 0.5$	$+10.01\%$	$+8.57\%$	$+5.79\%$	$+1.39\%$
ours, T from the ridge proxy	$+8.67\%$	$+7.27\%$	$+1.59\%$	$+0.00\%$
ours, T measured on the student	$+9.40\%$	$+8.57\%$	$+8.84\%$	$+5.86\%$
and the weight each T implies, against an oracle that reads the answer				
T from the ridge proxy	0.52	0.22	0.07	—
T measured on the student	0.63	0.50	0.46	—
oracle weight	0.50	0.30	0.30	—

Table 17: **All of the VGGSound release: 48,755 clips, 141 classes, up to 20,000 labels, nothing re-tuned.** Distilling helps at every weight here, so the decision has little to do and the question is the weight. A ridge proxy holds its penalty fixed and so breaks hypothesis (iii) of Remark 3: its T collapses with n, dragging $\hat{A} \to \hat{\beta}$ and $\hat{D} \to \hat{\delta}_{\text{raw}}$. These eighteen cells show that on the one dataset we can run at this scale, and measuring T removes it. The proxy's advantage falls away with the label budget and the measured version's does not. Two directions and three seeds per cell is a confirmation, not a scaling law.

The decision against the weight. Table 28 separates the decision from the weight by swapping one half at a time for an oracle. Our decision with an oracle weight reaches 87% of the gain at the best weight, an oracle decision with our weight reaches 64%, and the two together reach 43%. Holding our deployed decision fixed, replacing the weight by the oracle's recovers 78% of the gap that remains. Replacing the decision first and then the weight attributes 63% to it, so the two are not separable without choosing an order, and neither isolates compression from the weight's other errors. In absolute terms we capture 38.9% of that gain, measured as a mean of relative gains. Over the 221 cells where some weight beats the supervised student, the median recovers 81% of what was available. On the other 119 there is nothing to recover. The rule abstains on 55 of them and takes a positive weight on the remaining 64, where it costs 1.69% on average and 6.9% at worst.

H.11 Reusing the student's labels against a disjoint fold

Section 3 names the gap between the gain curve at a weight fixed in advance and the risk of a student trained at a weight chosen on its own samples, and nothing in the theory bounds it. The main ridge run bears on it, because it carries two weights against one grid of trained students, though the two arms differ in more than the reuse. One weight is cross-fitted on the samples the student trains on, and a second is estimated on a disjoint fold of 400 labels. Over the 221 live cells of that run, the disjoint weight returns $+2.14\%$ against the own-sample weight's $+1.95\%$, with a worst cell of -5.01% against -7.55% and a Wilcoxon $p = 0.027$. Both arms take the weight from the ratio

Policy	Test accuracy	Against no distillation	p
supervised, no distillation	44.16%	—	—
best (τ, λ) from a grid search	44.62%	$+0.46$	—
distill at $\lambda = 0.5$, $\tau = 4$	36.77%	$\mathbf{-7.40}$	2.2×10^{-22}
ours, scalars under Σ (the failure)	38.37%	$\mathbf{-5.80}$	2.6×10^{-19}
ours, scalars under the Gauss–Newton metric	43.25%	$\mathbf{-0.92}$	1.5×10^{-4}

Table 18: **Cross-entropy, under two metrics.** Under Σ the rule loses 5.80 accuracy points. In the Gauss–Newton working coordinates of Proposition 6 it loses 0.92, against 7.40 for the customary setting. The scored quantity is accuracy, for which the surrogate supplies no gain identity, so this locates most of the loss with the metric without establishing it as the only cause. Mean $|\hat{\lambda} - \lambda^*|$: 0.259 under Σ, 0.030 under the Gauss–Newton metric, over 132 cells.

Policy	Mean	Worst cell	Cells made worse
best λ, from a grid search	$+1.14\%$	$+0.00\%$	0/72
distill at $\lambda = 0.5$	-35.76%	-53.57%	72/72
ours	-10.32%	-26.90%	69/72

Table 19: **An adversarial teacher.** An ImageNet teacher on chest radiographs, a neural student on the radiology reports, and the teacher's outputs on $2{,}500$ unlabeled pairs, with the teacher's risk 3.26 times the student's. The teacher is far worse than the student, which is the case the decision exists to catch. The rule holds the loss to -10.32% against -35.76% for the customary weight and does not reach zero, so it limits the damage here rather than avoiding it. 72 cells; the teacher's risk is 3.26 times the student's, and the loss at $\lambda = 0.5$ is $31\times$ the gain at the best weight, the widest gap in this paper.

without the shrinkage, so neither is the deployed rule, which returns $+1.84\%$ on the same cells. The two protocols therefore differ by about a fifth of a point, a twentieth of the $+3.47\%$ the best weight attains on that population.

Three differences ride along with the reuse: the disjoint arm omits the β correction, its scalars are computed against a teacher refitted on that fold, and the extra fold is not paid for out of the student's label budget. The figure is therefore a comparison of two protocols, not a measurement of the reuse alone, and it does not bound the reuse in either direction.

H.12 The unit of the comparisons

What a test is over. The policy comparisons pair two policies on the same cell and run a Wilcoxon signed-rank test over cells, so the pairing is exact and the alternative is a shift in the paired difference. The one-sided weight counts use a binomial test instead. The cell is not an independent draw. Cells within a direction share the dataset's samples and its train and test split, the two directions of a dataset share its samples, and cells at different label budgets or seeds within one direction share the teacher. The unit of independence is therefore at most the *dataset*, of which there are six, which is few enough that any interval at that level is wide. Resampling those six with replacement over $20{,}000$ draws puts the pooled mean at $[+0.29, +2.15]$ for our rule, unchanged to two decimals across

Objective	$\alpha = 1$	α tuned on the outcome	With our alarm	Dirs. hurt
plain KD (Hinton et al., 2015)	-22.77	-4.33	-11.07	6/6
DKD (Zhao et al., 2022)	-16.62	-8.78	-8.20	6/6
DIST (Huang et al., 2022)	-16.63	-6.97	-7.80	6/6
CTKD (Li et al., 2023)	-11.02	-4.42	-5.90	6/6
LSKD (Sun et al., 2024)	-3.69	-0.28	-2.04	6/6
WTTM (Zheng & Yang, 2024)	-9.89	-5.41	-6.52	6/6
PLD (Bassam et al., 2026)	-12.19	-12.19	-7.33	6/6
not distilling	—	—	—	**0/6**

Table 20: **Seven published distillation objectives, as their papers define them.** Six cross-modal directions. Here α is the mixing parameter for plain KD and a global scale for the five objectives that add a term to the cross-entropy, and PLD has none. "Dirs. hurt" counts directions harmed at $\alpha = 1$, and the third column adds our alarm. Entries are accuracy percentage points, not the relative risk changes the other tables report. These objectives were built for same-modality distillation from strong end-to-end teachers, so the comparison is about what they contain rather than about their quality. Implementations in Appendix F. 30 cells, against the same student trained without a teacher. PLD has no distillation weight, so its first two columns coincide.

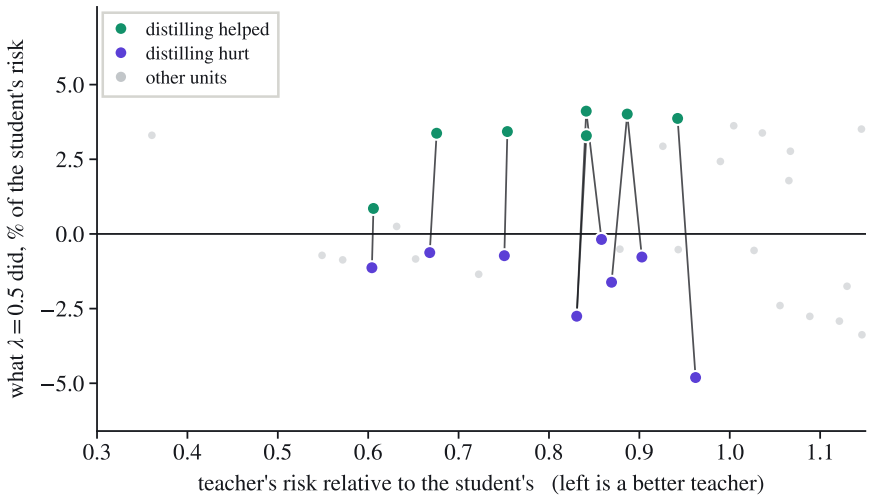


Figure 3: **The matched pairs after seed averaging.** Each point is one of 72 seed-averaged units, a direction at a teacher budget, placed by its teacher's risk relative to the student's and by what the customary weight did to it. The nine joined pairs are those whose teachers' risks relative to the student differ by at most 0.02 while their optimal weights differ by at least 0.4. *Every pair straddles zero*, so the horizontal axis, which is all the teacher's risk provides, does not distinguish the two ends of a pair.

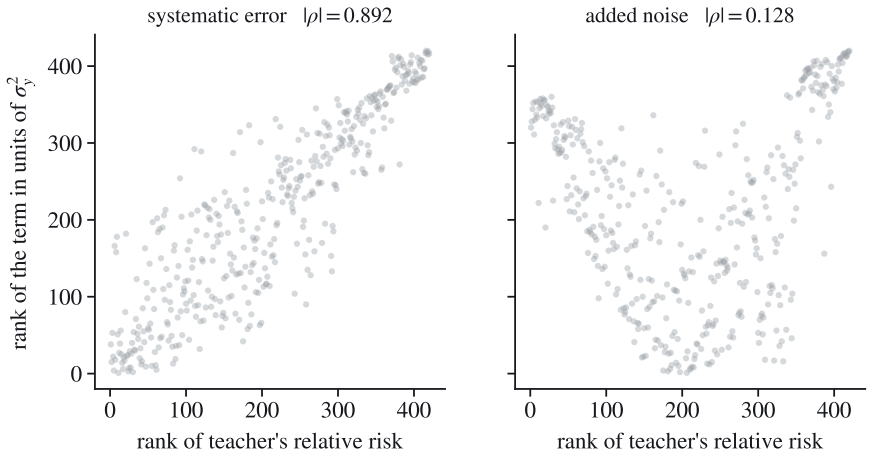


Figure 4: **Selective blindness.** Both panels rank the same 420 settings by the teacher's relative risk on the horizontal axis. Both terms are taken in units of the label noise. Against the teacher's risk, the systematic error the teacher *is* tracks at rank correlation 0.892 (left) while the noise it *adds* tracks at 0.128 (right). The added-noise term is also associated with the optimal weight, at rank correlation 0.327. Accuracy is therefore not blind but blind selectively, which is why two equally accurate teachers can need opposite weights.

seeds, and at $[-65, -4]$ for the customary weight. Six clusters cannot produce a negative resample once all six are positive, so what the interval adds is its width rather than a test, and the sign is the same information the deletions below carry.

The signed-rank reference distribution assumes the paired differences are independent, and ours are not, so the cell-level p-values are diagnostics of how consistent a difference is across cells and not calibrated tests. We do not rest a claim on one. Section 6.2's two comparisons against the accuracy threshold read this way, at $p = 5.9 \times 10^{-5}$ against the version bought at 20% of the student's labels and $p = 0.075$ against the version tuned on a labeled test set. Two comparisons carry a leave-one-dataset-out deletion, and the rest have no such check and are read as diagnostics only. Against the matched-label sweep the margin survives every deletion but varies widely, from $+0.09$ to $+2.93$ points over the 222 label-budget cells, while against the refit sweep it changes sign, from -0.54 to $+0.26$, which is why no claim rests on that comparison. Deleting the 84 cells whose teacher is an MLP leaves the rule at $+0.99\%$ against -14.09% for $\lambda = 0.5$. Against the customary weight, on the 544 live pooled cells, the rule stays between $+1.27\%$ and $+2.04\%$ and the customary

Hypothesis	Population	Raw p	kp	Holm
ours vs always distilling	the 802 pooled cells	6.9×10^{-52}	7.6×10^{-51}	yes
ours vs never distilling	the 802 pooled cells	6.5×10^{-36}	7.1×10^{-35}	yes
adding the bias alignment to the numerator	340 main-ridge cells	2.0×10^{-16}	2.2×10^{-15}	yes
ours vs the matched-label sweep	222 label-budget cells	1.7×10^{-13}	1.9×10^{-12}	yes
the unshrunk weight over-distills	642 cells	3.7×10^{-12}	4.0×10^{-11}	yes
pairs raise the best-weight gain over own-sample targets	144 paired settings	2.3×10^{-9}	2.6×10^{-8}	yes
shrinkage vs the hard gate	222 cells, 92 of them nonzero	3.3×10^{-8}	3.6×10^{-7}	yes
our AUC lead over the affordable accuracy threshold	144 criterion-comparison settings	5.9×10^{-5}	6.5×10^{-4}	yes
the two reuse protocols	221 live main-ridge cells	0.0270	0.297	no
our AUC lead over the outcome-tuned threshold	144 criterion-comparison settings	0.0750	0.825	no
ours vs the refit sweep	222 label-budget cells	0.554	1	no

Table 21: **Eleven comparisons, with a Holm correction.** Each with the population it was tested on, its raw p, the Bonferroni product kp at $k = 11$, and whether it survives Holm at 0.05. Tables 16 and 18 print further within-table comparisons that no claim rests on, and they are not in the family. The cells within a population are not independent draws, so these are the same diagnostics Appendix H.12 describes rather than calibrated tests. The three that do not survive are the two the paper reports as not separating and the reuse comparison, and no claim rests on them.

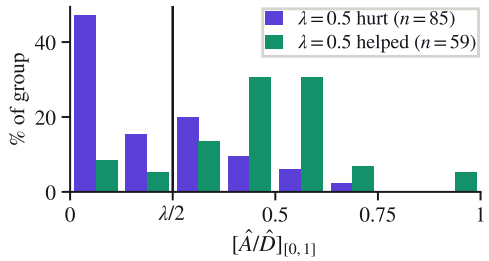


Figure 5: **The estimated ratio, split by outcome.** $\left[\hat{A}/\hat{D}\right]_{[0,1]}$, the clipped ratio the rule works with, over the 144 settings of Table 2, split by whether $\lambda = 0.5$ helped. The end bins therefore hold the cells the clip brings in. Right of $\lambda/2$ the alarm calls a gain at that fixed weight, which is not the rule's own decision.

weight between -0.02% and -5.25%, so each sign survives every deletion. Deletion is a robustness check and not a calibration either. Six all-positive clusters give a two-sided sign-flip probability of $2/2^6 = 0.031$, and replicating each cluster ten times would leave every deletion positive while adding no information. Where the coarser clustering does not support a claim we drop the claim rather than the clustering, which is why Section 6.2 states the alarm's own accuracy instead of its margin over the accuracy ratio: resampling whole directions puts that margin's interval across zero.

H.13 A REGIME WITHHELD AFTER ITS RESULT WAS KNOWN

One regime is not in the main tables and was removed after its result was known, which the sentence in Section 6 about dropping nothing does not cover. We report it here rather than leave the record in the pre-registration ledger alone.

The regime pairs an ImageNet-pretrained ResNet-50 teacher on video with a 1D CNN student on raw audio, with the teacher's targets on the unlabeled pairs. At 24 seeds and 72 cells the rule returned $+2.58\%$ against the customary weight's $+0.51\%$, separating no pair of policies. We pre-registered a five-fold increase in seeds, predicting the intervals would tighten around that estimate, and both predictions failed. At 120 seeds and 360 cells the rule returns $+0.56\%$ against the customary weight's -0.45%, the seed-clustered interval against never distilling is $[-2.01, +2.95]$ and against the customary weight $[-1.33, +3.15]$, and the rule's worst cell moves from -54.66% to -166.24%.

Having separated no pair of policies at either seed count, we scoped the submission to the settings in which the rule is evaluated, after the outcome was known. What the regime still shows is the stakes rather than the rule. The best weight attains $+15.42\%$ there and the customary weight's worst cell is -100.98%. The artifacts are kept.

H.14 CLIPPED RISK

Unclipped squared error lets a handful of catastrophic cells dominate any mean, which is why the main text reports worst-case columns beside every average. A practitioner would project predictions

Direction	λ^*	Best λ	KD λ=0.5	Ours
HAR bacc→tacc	0.02	$+0.06\%$	-16.30%	-0.61%
HAR gyr→tacc	0.02	$+0.09\%$	-20.05%	-0.45%
MIMIC-CXR rev	0.08	$+0.38\%$	-4.66%	-0.49%
HAR gyr→bacc	0.10	$+0.59\%$	-3.32%	-1.00%
HAR tacc→bacc	0.18	$+0.74\%$	-1.05%	-0.27%
HAR tacc→gyr	0.18	$+0.88\%$	-0.49%	$+0.36\%$
HAR bacc→gyr	0.22	$+1.10\%$	-0.09%	$+0.58\%$
RAVDESS	0.16	$+1.18\%$	-2.00%	-0.89%
NYU-Depth rev	0.44	$+2.88\%$	$+2.82\%$	$+2.82\%$
VGGSound v→a	0.37	$+2.99\%$	$+2.49\%$	$+2.68\%$
MIMIC-CXR	0.50	$+3.28\%$	$+3.28\%$	$+3.02\%$
VGGSound a→v	0.51	$+3.62\%$	$+3.62\%$	$+3.45\%$
NYU-Depth	0.52	$+4.14\%$	$+4.11\%$	$+4.11\%$
RAVDESS rev	0.58	$+4.37\%$	$+4.19\%$	$+2.87\%$
Parkinsons rev	0.05	$+4.59\%$	-206.95%	$+0.00\%$
Parkinsons	0.72	$+7.32\%$	$+3.96\%$	$+1.12\%$
mean	0.29	$+2.39\%$	-14.40%	$+1.08\%$

Table 22: **All sixteen directions, ordered by the gain at the best weight.** Six are worth less than 1% and eight less than 2%. The worst direction under the customary weight is 207 times worse than the worst under the rule, but that ratio is carried by one direction, and at the second-worst it is 22 times. Main-ridge run, 340 cells. The final row averages the sixteen directions equally, which on these cells gives -14.40% and $+1.08\%$ where weighting the cells gives -17.35% and $+0.88\%$.

Axis	Setting	Cells	Best λ	KD λ=0.5	Ours
teacher	ridge	84	$+2.25\%$	-17.64%	$+1.06\%$
	grad. boosting	84	$+1.82\%$	-8.71%	$+0.91\%$
	k-NN	84	$+1.89\%$	-13.30%	$+0.92\%$
	random forest	84	$+1.97\%$	-12.79%	$+0.81\%$
	MLP *(failed fit)*	84	$+0.73\%$	-132.25%	$+0.03\%$
student	tiny (23, 148 params)	40	$+4.95\%$	-15.48%	$+2.26\%$
	small (306, 739 params)	40	$+4.68\%$	-11.79%	$+1.53\%$
	wide (2, 799, 820 params)	40	$+5.15\%$	-17.12%	$+1.62\%$
	deep (1, 025, 177 params)	40	$+5.29\%$	-6.61%	$+1.33\%$
end to end	UCI-HAR, 1D CNN	96	$+5.03\%$	-21.73%	—
	RAVDESS, 1D CNN	8	$+17.62\%$	$+14.40\%$	—

Table 23: **The same rule across teacher families and student sizes.** Nothing is re-tuned. The end-to-end rows report the attainable gain and loss only, which the dash marks, and their HAR splits are not subject-disjoint (Appendix F), so they bound what is available under a changed representation rather than isolating the representation. The failed MLP teacher is an ordinary configuration whose fit diverged: on its worst cell the teacher's squared error is 153.6 against the student's 0.95, where the customary weight costs -1530% and the rule at worst -0.78%.

onto $[0, 1]$ instead, a step the identity does not model, so Table 32 repeats the headline comparison under it. Clipping compresses what there is to gain. The oracle falls from $+2.26\%$ to $+1.58\%$ on the ridge student and from $+5.40\%$ to $+3.94\%$ on the MLP, while the customary weight costs more rather than less, -17.35% to -18.95% and -10.99% to -14.32%. Our arm stays positive on both students. We report the less favourable metric beside the one the theory matches, because selecting between them after observing which is more favourable would be an instance of the decision this paper is about.

Cells scored	KD λ=0.5	Ours	Best λ	Share of best
ridge student, 340 cells, 6 datasets				
all of them	-17.35%	$+\mathbf{0.88}\%$	$+2.26\%$	38.9%
drop MIMIC-CXR (the pretraining overlap)	-20.22%	$+\mathbf{0.81}\%$	$+2.33\%$	34.8%
drop Parkinsons (worst at $\lambda = 0.5$)	-2.84%	$+\mathbf{0.93}\%$	$+1.62\%$	57.7%
worst of all 6 single drops	-2.84%	$+\mathbf{0.50}\%$	$+1.62\%$	24.1%
five teacher families, 420 cells, 6 datasets				
all of them	-36.94%	$+\mathbf{0.74}\%$	$+1.73\%$	43.0%
drop MIMIC-CXR (the pretraining overlap)	-41.38%	$+\mathbf{0.76}\%$	$+1.84\%$	41.2%
drop Parkinsons (worst at $\lambda = 0.5$)	-28.73%	$+\mathbf{0.87}\%$	$+1.29\%$	66.9%
worst of all 6 single drops	-18.77%	$+\mathbf{0.54}\%$	$+1.29\%$	33.3%

Table 24: **Leave one source dataset out.** A dataset and not a direction: UCI-HAR's six directions and VGGSound's two leave together with their source. Our policy stays positive in all twelve fits and $\lambda = 0.5$ stays negative in all twelve. Removing MIMIC-CXR, where a pretraining overlap could not be resolved exactly, makes the customary weight look *worse*.

Teacher risk / student risk	Settings	Median λ^*	Range	IQR
$[0, 0.5)$	30	0.90	$[0.00, 0.90]$	0.00
$[0.5, 0.8)$	56	0.20	$[0.00, 0.90]$	0.50
$[0.8, 1.0)$	66	0.40	$[0.00, 0.70]$	0.39
$[1.0, 1.5)$	111	0.30	$[0.00, 0.50]$	0.38
$[1.5, \infty)$	97	0.00	$[0.00, 0.15]$	0.05

Table 25: **Vary only the teacher's quality.** Rows band the settings by the teacher's relative risk. *Range* and *IQR* are of the optimal weight within each band. Teacher quality does carry information, at Spearman -0.554 with the optimal weight, and it does not come close to determining it: knowing the teacher is about a third better than the student leaves the optimal weight anywhere from 0.00 to 0.90.

Teacher risk ratios differ by at most	Optimal weights differ by	Pairs	Flip the outcome
0.01	≥ 0.3	6	6/6 (100%)
0.01	≥ 0.4	3	3/3 (100%)
0.02	≥ 0.3	16	16/16 (100%)
0.02	≥ 0.4	**9**	**9/9 (100%)**
0.02	≥ 0.5	1	1/1 (100%)
0.03	≥ 0.3	25	25/25 (100%)
0.03	≥ 0.4	12	12/12 (100%)
0.03	≥ 0.5	3	3/3 (100%)
0.05	≥ 0.3	37	36/37 (97%)
0.05	≥ 0.4	16	16/16 (100%)
0.05	≥ 0.5	4	4/4 (100%)
0.10	≥ 0.3	72	71/72 (99%)
0.10	≥ 0.4	30	30/30 (100%)
0.10	≥ 0.5	8	8/8 (100%)

Table 26: **The matched pairs do not depend on the thresholds we chose.** Two tolerances define a pair. Across all fourteen settings of the pair of thresholds, the share that flip the outcome is 97 to 100%, including the loosest, which admits 72 pairs rather than nine. These are pairs drawn from the 72 seed-averaged units of Figure 3, not those units themselves.

λ	Settings harmed	AUC of $-\hat{G}(\lambda)$	Precision	Recall
0.1	35%	0.757	63%	49%
0.2	42%	0.798	74%	57%
0.3	49%	0.823	78%	59%
0.5	58%	0.856	84%	68%
0.7	65%	0.870	85%	79%
0.9	76%	0.805	84%	83%

Table 27: **The alarm at every weight.** One row per candidate weight. AUC ranks the continuous score $-\hat{G}(\lambda)$, while precision and recall score the thresholded alarm $\hat{G}(\lambda) < 0$. The two are different objects, and a binary score could not reach these AUCs at these recalls. The alarm holds everywhere and is sharpest at $\lambda = 0.7$, in the regime where being wrong is expensive. By $\lambda = 0.9$ three quarters of settings are harmed, so there is less left to separate even as recall reaches its highest. Across 1,229 run records, including the teacher-quality sweep and without the deduplication Table 3 applies.

Decision	Weight	Mean	Share of best
oracle	oracle	$+3.48\%$	100%
ours	oracle	$+3.04\%$	87%
oracle	**ours**	$+2.22\%$	64%
ours	**ours**	$+1.49\%$	43%

Table 28: **Which half of the rule costs.** Swapping one half at a time for an oracle. The decision is worth 87% of the gain at the best weight and the weight is what costs. 222 cells, all read under the same unclipped metric. *Decision* is the rule's deployed abstention, not the sign of $\hat{A}$ alone, and *our weight* is the magnitude it proposes, taken on the positive grid points so that it is defined where the rule abstains. Replacing our weight by the oracle's recovers most of the gap; replacing our decision by the oracle's recovers much less.

Criterion	ρ within direction	Picks best teacher	Usable
NTDH (Chen et al., 2024)	$+0.113$	40%	40/48
CCH (Xie et al., 2025)	constant within a direction		0/48
Ratio 1-2 (Gosal et al., 2026)	$+0.475$	65%	40/48
teacher risk (uses test labels)	$+0.510$	69%	48/48
ours: $\hat{G}(0.5) = \hat{A} - \frac{1}{4}\hat{D}$	$+0.656$	67%	48/48

Table 29: **Each criterion on the decision it was built for.** Three teachers within one direction, trained on 40, 160 and 640 labeled pairs. ρ ranks the three within a direction, *picks best teacher* is how often the top-ranked one is the best, and *usable* records whether the criterion can express the choice at all. Ratio 1-2 works here and is at chance on the distill-or-not decision. CCH cannot express the teacher choice, because it reads the two modalities and the labels and never the teacher. Our signal does both jobs from one computation. Chance is $\rho = 0$ and 33% for three candidate teachers per direction.

Per-sample scheme	At a fixed scale 0.5	Best scale on the outcome	**With our scale**
teacher confidence	$+0.66\%$	$+1.14\%$	$\mathbf{+0.47\%}$
teacher–student agreement	$+0.30\%$	$+1.55\%$	$\mathbf{+0.36\%}$
teacher entropy	-13.89%	$+1.60\%$	$\mathbf{+0.51\%}$
most confident half	-14.10%	$+1.57\%$	$\mathbf{+0.54\%}$
no per-sample weighting, our scale	—	—	$\mathbf{+0.97\%}$

Table 30: **Four per-sample reweighting schemes, given our global scale.** Two are rescued outright, by about fifteen points, which is this paper's claim stated about somebody else's method, since those schemes were wrong about *how much* to trust rather than about *where*. No scheme reaches our uniform rule once the scale is right, though they do buy a better tail. 84 cells; the tuned column selects a constant on the outcome, which our column never sees.

Predicted $G(\lambda)/G(\lambda^*)$	Settings	Predicted (median)	Measured (median)
$(0, 1]$: helps	662	$+0.93$	$+0.97$
$(-3, 0]$	139	-1.25	-1.22
$(-15, -3]$	19	-4.44	-5.66
$(-\infty, -15]$	228	-15.00	-19.76

Table 31: **The ordering the law predicts.** Settings are binned by the gain Remark 1 predicts at the default weight $\lambda = 0.5$, read as a fraction of the gain at λ^*_{grid}, and the last column is the fraction measured. Normalizing by the grid optimum rather than the continuous one makes this an ordering check rather than an exact evaluation of the law. No fitted parameter. All 1048 settings, Spearman 0.953, no free parameter.

Policy	Student fits	Gain over not distilling	Share of oracle	Directions hurt
ridge student, 340 cells, clipped squared error				
always, $\lambda = 0.5$	1	-18.95%	-67.6%	9/16
sweep, matched labels	7	-2.01%	-77.0%	10/16
sweep, then refit on all n	8	$+0.44\%$	$+67.0\%$	9/16
ours, shrunk	**1**	$+0.10\%$	$+59.0\%$	8/16
ours, then a 3-point sweep	**4**	$+0.12\%$	$+57.6\%$	9/16
best per cell, grid search	—	$+1.58\%$	$+100.0\%$	0/16
MLP student, 120 cells, clipped squared error				
always, $\lambda = 0.5$	1	-14.32%	-61.1%	8/16
sweep, matched labels	7	-0.77%	-58.7%	10/16
sweep, then refit on all n	8	$+0.94\%$	$+60.7\%$	6/16
ours, shrunk	**1**	$+0.34\%$	$+46.3\%$	4/16
ours, then a 3-point sweep	**4**	$+0.90\%$	$+60.3\%$	6/16
best per cell, grid search	—	$+3.94\%$	$+100.0\%$	0/16

Table 32: **The same comparison under clipped squared error.** Predictions projected onto $[0, 1]$, the metric a practitioner deploys, which the identity does not model. *Ours, then a 3-point sweep* is an arm reported only here: it takes the three grid points nearest the unshrunk $\hat{\lambda}$, picks among them on a validation split, and refits on all n, at four student fits. There is less to gain under clipping and the customary weight costs more, while our arm stays positive on both students. Reported beside the theory-matched metric rather than in place of it. *Share of oracle* pools absolute reductions, $\sum(\text{sup} - \text{arm})/\sum(\text{sup} - \text{oracle})$, so it weights the high-risk directions and is not the ratio of the preceding column to the oracle's.