Label-Free Consistency Correction for Weak-to-Strong Generalization

Qi He
Department of Computer Science
University of Maryland, College Park
qhe123@umd.edu

Heng Huang
Department of Computer Science
University of Maryland, College Park
heng@umd.edu

Abstract

Most explanations of weak-to-strong (W2S) generalization focus on why a strong student may fail to exactly imitate weak-label errors. We study a complementary mechanism: before student training, weak-teacher predictions may be inconsistent along semantic constraints specified independently of the weak model. When such constraints are valid for the target, these inconsistencies give label-free lower bounds on the risk reduction achievable by correcting weak labels, rather than merely indicating unstructured prediction error. We formalize this mechanism using Markov and Laplacian operators. Under exact target validity, Markov trajectory inconsistency equals the squared-risk gain from smoothing; under approximate validity, it remains a conservative lower bound after an explicit validity penalty. We then solve the associated minimax correction problem, obtaining a resolvent correction that continuously attenuates constraint-violating directions and reduces to the identity when validity uncertainty is too large. These results yield validity-adjusted correction (VAC), a split-sample pipeline that selects corrections from weak-teacher queries and predeclared validity budgets before training the strong student. Controlled graph-signal experiments, CIFAR-10 classification, and ARC language-model experiments support this mechanism: valid constraints enable weak-label correction and student gain, while invalid controls show why raw inconsistency alone is insufficient.

1 Introduction

Weak-to-strong (W2S) generalization asks whether a stronger student trained only on weak-model outputs can improve on its weak supervisor [3]. Much existing theory explains this phenomenon through the student: its capacity, representation, optimization dynamics, data overlap, or training objective may prevent exact imitation of the weak supervisor [4, 7, 10, 14, 16, 22, 26, 27]. This paper studies a complementary set of sufficient conditions. Rather than attempting to explain all W2S gains, we ask when weak supervision already contains an independently specified signal of its own errors before the strong student is trained.

The mechanism studied here is the following. Weak-teacher predictions can be inconsistent along constraints that are specified independently of the weak model. Examples include image transformations that should preserve class labels, paraphrases that should preserve meaning, and preference comparisons that should be consistent with an underlying utility. Such constraints are closely related to consistency regularization and augmentation invariance in semi-supervised learning [2, 5, 13, 17, 23, 24], as well as utility-consistency structure in preference learning [11, 19, 21]. The key requirement is independent target validity: the constraint must be specified as a property of the target, not inferred from the weak teacher's behavior. If the weak teacher varies along a constraint where the target is known to vary little, that variation gives label-free evidence of weak-label error. If the target may vary

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

substantially along the same constraint, the same weak-teacher variation may reflect target variation, so the lower bound must be reduced by an explicit validity budget.

The paper develops this mechanism in three steps. First, we fix a single semantic constraint and ask what can be inferred before student training. For reversible Markov constraints, the expected change in the weak teacher along a constraint trajectory is exactly the squared-risk gain from smoothing under exact target validity. Under approximate validity, the same quantity remains a conservative lower bound after subtracting a geometry-dependent penalty. This step turns trajectory inconsistency into a label-free lower bound on a correctable component of weak-label error.

Second, we ask how much correction is uniformly safe once a constraint-violating component has been identified. Hard projection is appropriate only under exact validity, because it can remove legitimate target variation when validity is approximate. A minimax analysis under a Laplacian validity budget instead yields a correction that attenuates constraint-violating directions while preserving directions that may contain target variation, and it reduces to the identity correction when no positive worst-case gain is possible.

Third, we turn the fixed-constraint lower bound into an operational W2S procedure, in the same broad training paradigm as distillation and pseudo-labeling but with a different source of label modification [9, 25]. In applications, many candidate transformations, graphs, paraphrase generators, or filters may be available, and selecting the largest raw inconsistency can choose invalid constraints. We therefore use a lower-bound selection rule that chooses a candidate whose validity-adjusted gain is close to the best candidate's gain using only weak-teacher queries and validity budgets. We refer to the resulting selected-correction pipeline as validity-adjusted correction (VAC). A split-sample transfer bound then states when VAC yields an end-to-end W2S risk-improvement bound after student training, and a PAC-Bayes variant extends the same principle to large candidate families.

This paper makes three contributions. First, it identifies a label-free lower-bound formulation for weak-label error: weak-teacher trajectory inconsistency is an observable squared-risk gain under exact target validity and a conservative lower bound under approximate validity. Second, it solves the minimax correction problem under a Laplacian validity budget, yielding a resolvent correction when positive worst-case gain is possible and the identity map otherwise, with a stochastic-teacher extension that accounts for sampling variance. Third, it connects the correction lower bound to an end-to-end W2S workflow through ERM transfer, split-sample selection, PAC-Bayes candidate control, controlled graph-signal experiments, and CIFAR-10 and ARC neural experiments with invalid controls.

2 Related work

Weak-to-strong generalization. The scalable-oversight formulation of W2S shows that stronger models trained on weak labels can outperform their weak supervisors [3]. Existing theory explains W2S gain through misfit on weak labels and extensions beyond squared loss [4, 18], pseudolabel correction and coverage expansion [14], and overlap density, a data property indicating points that contain both weak-accessible and strong-accessible patterns [22]. Representation and intrinsic-dimension analyses explain W2S using discrepancies between weak and strong feature spaces [7, 26]. High-dimensional and random-feature analyses characterize how model size, feature mismatch, and training dynamics affect W2S [10, 16]. Loss-level work studies reverse KL or reverse cross-entropy as alternatives to forward imitation [27]. These theories primarily analyze conditions under which the student class, representation, data distribution, or objective prevents exact imitation of the weak supervisor. The present analysis is complementary to these explanations: it does not claim that all W2S gains arise from semantic inconsistency, but identifies conditions under which the weak teacher's own outputs contain an independently specified, label-free signal for correction before student training.

Distillation, self-training, and pseudo-labeling. The W2S setting is related to knowledge distillation, where a student is trained to match a teacher's soft predictions [9]. It also resembles self-training and pseudo-labeling, where unlabeled examples receive model-generated targets that are later used for training; Noisy Student training is a prominent large-scale instance of this pattern [25]. In these settings, performance depends on whether the student, augmentation procedure, or confidence filtering mitigates teacher noise. The present theory differs in the source of the label-modification

signal. It does not require ground-truth labels, confidence calibration, or an already strong student; instead, independently valid semantic constraints make certain weak-teacher inconsistencies observable evidence for correcting weak labels.

Additional related work on consistency regularization, graph smoothing, inverse problems, and preference consistency is discussed in Appendix J.

3 Setup and assumptions

Let $(\mathcal{X}, \mathcal{B}, P)$ be the input probability space, with sigma-field $\mathcal{B}$. On $L^2(P)$, write $\langle f, g\rangle = \mathbb{E}_{X\sim P}[f(X)g(X)]$, $\|f\|^2 = \langle f, f\rangle$, and let I denote the identity operator. The target is $f^\star \in L^2(P)$, the weak teacher's mean prediction is $t \in L^2(P)$, and the population squared risk is

$$\mathcal{R}(h) = \|h - f^\star\|^2_{L^2(P)}. \tag{1}$$

A label-free semantic constraint is represented by a Markov kernel $K(x, dz)$ and its Markov operator

$$Tf(x) = \int f(z)K(x, dz). \tag{2}$$

For a candidate family, denote by K_a, T_a, Λ_a, and β_a the corresponding kernel, Markov operator, semantic Laplacian, and validity budget; when a single candidate is fixed, the subscript is omitted. The kernel may be a random image transformation, a verified paraphrase transition, a retrieval-neighbor random walk, or an edge walk in a preference graph. The assumptions below separate three roles: how constraints are specified, how the induced averaging operator behaves, and how much target variation is allowed along the constraint. They do not assert that every proposed transformation is valid; rather, they specify when label-free inconsistency can be interpreted as evidence about weak-label error.

Assumption 1 (Label-free constraint specification)**.** *There is an auxiliary sigma-field $\mathcal{I}$ that contains the information used to specify the constraints but does not contain target labels. The candidate index set $\mathcal{A}$, kernels $\{K_a : a \in \mathcal{A}\}$, nonnegative validity budgets $\{\beta_a : a \in \mathcal{A}\}$, and any prior $\pi \in \Delta(\mathcal{A})$ are $\mathcal{I}$-measurable, where $\Delta(\mathcal{A})$ denotes the set of probability distributions over $\mathcal{A}$. The weak-teacher statistics used later for estimation or selection are computed after this specification, and the validity inequalities in Assumption 3 are not functions of the realized weak-teacher inconsistency statistics.*

Assumption 2 (Reversible positive Markov operator)**.** *For each $a \in \mathcal{A}$, K_a has invariant measure P and satisfies detailed balance,*

$$P(dx)K_a(x, dz) = P(dz)K_a(z, dx). \tag{3}$$

Thus the associated operator T_a is self-adjoint on $L^2(P)$. Assume $0 \preceq T_a \preceq I$ in the Loewner order on self-adjoint operators; equivalently, if a reversible kernel has negative spectrum, all results are applied to its lazy version $(I + T_a)/2$.

Assumption 3 (Target validity class)**.** *Let $\Lambda_a = I - T_a$. For each candidate a, the target satisfies*

$$f^\star \in \mathcal{F}_{\beta_a}(\Lambda_a) \quad \textit{where} \quad \mathcal{F}_{\beta_a}(\Lambda_a) = \{f \in L^2(P) : \left\|\Lambda_a^{1/2} f\right\| \le \beta_a\}. \tag{4}$$

Exact validity corresponds to $\beta_a = 0$. If a data-dependent upper bound $\bar{\beta}_a$ is used, the bounds are conditioned on the event

$$\mathcal{E}_\beta = \left\{\forall a \in \mathcal{A} : \left\|\Lambda_a^{1/2} f^\star\right\| \le \bar{\beta}_a\right\}, \tag{5}$$

and the lower-bound and selection rules use $\bar{\beta}_a$ in place of β_a.

Together, these assumptions separate specification, analysis, and validity. Assumption 1 prevents post hoc constraint choice based on large weak-teacher variation; Assumption 2 makes T_a a positive self-adjoint averaging operator on $L^2(P)$, which enables the spectral decomposition; and Assumption 3 bounds target variation along candidate a through β_a. Thus $\beta_a = 0$ gives exact invariance, larger β_a yields more conservative bounds, and a candidate has a positive label-free lower bound only when this budget is independently specified or validated.

Sources of validity budgets. Budgets may come from task structure, independent validity audits, or predeclared semantic priors. In the ARC experiment, answer-preserving option permutations give $\beta = 0$ on a corpus checked for position-dependent references. For the audited extension in Equation (5), the deployed transformation is frozen before a separate, teacher-blind calibration sample is used to bound target variation. For one-hot classification in the local probability-space geometry used in our implementation, a target-changing probability p_a gives $\beta^2_{a,\mathrm{local}} = 2p_a$, so a simultaneous upper bound $\bar{p}_a$ yields $\bar{\beta}_{a,\mathrm{local}} = \sqrt{2\bar{p}_a}$. When this audit uses gold labels, correction is label-free conditional on that independent calibration set. The CIFAR budgets are predeclared semantic priors, with guarantees conditional on their validity.

Definition 1 (Validity-adjusted correction)**.** Given a predeclared candidate family $\mathcal{A}$ and validity budgets $\{\beta_a : a \in \mathcal{A}\}$, the validity-adjusted correction (VAC) procedure selects and applies a weak-label correction using weak-teacher queries and the predeclared validity information. On an independent scoring sample, it evaluates each candidate correction h_a and computes a conservative score $\mathcal{L}_a$ for its validity-adjusted gain, then selects $\widehat{a} \in \arg\max_{a\in\mathcal{A}} \mathcal{L}_a$. For the end-to-end transfer guarantee, the selected correction $h_{\widehat{a}}$ is applied to a disjoint weak-labeled training sample, and the strong student is trained on these corrected weak labels.

4 Theory: from lower bounds to minimax correction

Following the sequence introduced above, this section first studies a fixed predeclared constraint, then optimizes the correction under approximate validity, and finally controls data-dependent selection and student training. Complete proofs are in the appendix.

4.1 Markov lower bounds and student transfer

The first stage fixes a semantic constraint and asks what weak-teacher queries imply about risk reduction. Fix an integer $k \geq 1$. Let $X_0 \sim P$ and, conditional on X_r, draw $X_{r+1} \sim K(X_r, \cdot)$ for $r = 0, \ldots, 2k-1$. Under Assumption 2, this gives a stationary reversible Markov chain $(X_0, \ldots, X_{2k})$ with transition operator T. Define the trajectory inconsistency

$$C_k(t) = \langle t, (I - T^{2k})t \rangle = \frac{1}{2}\mathbb{E}\big(t(X_0) - t(X_{2k})\big)^2. \tag{6}$$

This quantity is the expected squared change in the weak teacher after moving along the semantic chain. It is observable because it uses only weak-teacher evaluations on constraint-related inputs. If the target is invariant along the chain, meaning that $Tf^\star = f^\star$, then any variation of the weak teacher along this chain cannot be attributed to target variation and therefore identifies a component of weak-label error. To quantify the penalty for approximate validity, also define

$$A_k(t) = \left\| (\Lambda^\dagger)^{1/2}(I - T^k)t \right\|, \tag{7}$$

where $\Lambda^\dagger$ is the Moore–Penrose inverse of Λ. The factor $A_k(t)$ measures the sensitivity of this correction to target variation allowed by the validity budget; it is the term that converts exact invariance into approximate invariance.

Theorem 1 (Markov smoothing risk identity)**.** *Suppose Assumption 2 holds, $t \in L^2(P)$, and $u_k = T^k t$. Then the population gain from Markov smoothing is*

$$\mathcal{R}(t) - \mathcal{R}(u_k) = C_k(t) - 2\left\langle (I - T^k)t, f^\star \right\rangle. \tag{8}$$

Consequently, exact validity $\Lambda f^\star = 0$ gives

$$\mathcal{R}(t) - \mathcal{R}(u_k) = C_k(t), \tag{9}$$

while approximate validity $\left\|\Lambda^{1/2} f^\star\right\| \leq \beta$ gives

$$\mathcal{R}(t) - \mathcal{R}(u_k) \geq C_k(t) - 2\beta A_k(t) \geq C_k(t) - 2\beta\sqrt{kC_k(t)}. \tag{10}$$

The proof of Theorem 1 is given in Appendix A. The theorem should be read as a decomposition of the correction gain into an observable term and a target-validity term. Exact validity removes the

second term, so trajectory inconsistency equals squared-risk improvement. Approximate validity does not give equality, but it bounds the unobserved target-validity term by $2\beta A_k(t)$.

The trajectory inconsistency can be estimated without ground-truth labels. If $|t| \le B$ and $\widehat{C}_k = \frac{1}{2N}\sum_{i=1}^{N}(t(X_0^{(i)}) - t(X_{2k}^{(i)}))^2$ is computed from N independent stationary trajectories, then for $\alpha \in (0,1)$, with probability at least $1-\alpha$,

$$C_k(t) \ge \widehat{C}_k - 2B^2\sqrt{\frac{\log(1/\alpha)}{2N}}. \tag{11}$$

The next proposition makes explicit when such a population correction lower bound transfers to a trained student. It applies to Markov smoothing with $\bar{t} = u_k$, and also to any later selected correction whose conditional mean is fixed before the student-training sample is drawn.

Proposition 1 (Transfer radius)**.** *Let $\bar{t}$ be a corrected weak-label mean that is fixed before the student-training sample is drawn, or is selected on an independent split. Define its population correction gain*

$$G_{\bar{t}} := \mathcal{R}(t) - \mathcal{R}(\bar{t}) = \|t - f^\star\|^2 - \|\bar{t} - f^\star\|^2.$$

For any student predictor $\hat{h}$, the exact transfer identity is

$$\mathcal{R}(t) - \mathcal{R}(\hat{h}) = G_{\bar{t}} - 2\langle \bar{t} - f^\star, \hat{h} - \bar{t}\rangle - \|\hat{h} - \bar{t}\|^2.$$

Consequently, if the label-free lower-bound step gives $G_{\bar{t}} \ge \Gamma$ and $\|\hat{h} - \bar{t}\| \le \Delta$, then

$$\mathcal{R}(t) - \mathcal{R}(\hat{h}) \ge \Gamma - 2M\Delta - \Delta^2, \qquad M := \|\bar{t} - f^\star\|.$$

If $\Gamma > 0$, the trained student has smaller risk than the weak teacher whenever

$$\Delta < \tau(\Gamma, M) := \sqrt{M^2 + \Gamma} - M.$$

Appendix B proves Proposition 1 and derives the radius Δ from approximation, estimation, and optimization errors for approximate empirical risk minimization (ERM). The proposition shows that a population lower bound is not automatically inherited by the trained student. The population correction supplies a lower-bound margin Γ, while student training subtracts a sharp radius penalty. The critical radius $\tau(\Gamma, M)$ is the largest $L^2(P)$ distance from the corrected label that can be tolerated under worst-case alignment of the student error. Thus a small lower-bound margin requires accurate imitation of the corrected labels, whereas a larger margin can tolerate larger approximation, estimation, or optimization error.

Theorem 1 and Proposition 1 separate the two parts of the W2S argument. The first result gives a population lower bound for a corrected weak label, while the second result states when a student trained on that correction inherits the gain. This separation also explains why a positive label-free lower bound is not by itself sufficient: it must be larger than the approximation, estimation, and optimization errors incurred when learning the corrected labels.

4.2 The minimax correction problem

The lower-bound stage identifies when correction can improve the weak teacher, but it does not determine the correction magnitude. Hard projection removes every component outside the constraint subspace and is optimal when $\beta = 0$, but under approximate validity those directions may contain legitimate target variation. The next stage therefore chooses the corrected label by maximizing its worst-case gain over all targets satisfying the validity budget.

We now work in a finite-dimensional Hilbert space $\mathcal{H}$, with its inner product and norm. Fix $t \in \mathcal{H}$ and a self-adjoint $\Lambda \succeq 0$. For $\beta \ge 0$, define the validity class

$$\mathcal{F}_\beta(\Lambda) = \{f \in \mathcal{H} : \left\|\Lambda^{1/2} f\right\| \le \beta\}. \tag{12}$$

For any corrected label vector $h \in \mathcal{H}$, define its worst-case gain over the teacher t by

$$\Psi_\Lambda(h; t, \beta) = \inf_{f \in \mathcal{F}_\beta(\Lambda)} \{\|t - f\|^2 - \|h - f\|^2\}. \tag{13}$$

A positive value is equivalent to h having smaller squared risk than the teacher uniformly over all targets satisfying the validity budget.

Let Π_0 be the orthogonal projection onto $\ker(\Lambda)$ and $\Pi_+ = I - \Pi_0$. Let $\Lambda^\dagger$ denote the Moore–Penrose inverse of Λ. Choose an orthonormal eigenbasis $\{u_j\}$ of $\Pi_+\mathcal{H}$, with $\Lambda u_j = \lambda_j u_j$, $\lambda_j > 0$, and write $v_j = \langle v, u_j \rangle$ for spectral coordinates. Then $(\Lambda^\dagger)^{1/2} u_j = \lambda_j^{-1/2} u_j$, and $(\Lambda^\dagger)^{1/2}$ vanishes on $\ker(\Lambda)$.

Theorem 2 (Sharp minimax correction under Laplacian validity)**.** *Assume $0 < \beta < \left\|\Lambda^{1/2} t\right\|$. Then the unique maximizer of $\sup_{h \in \mathcal{H}} \Psi_\Lambda(h; t, \beta)$ is the resolvent smoother*

$$h^\star = \Pi_0 t + \theta(\Lambda + \theta I)^{-1} \Pi_+ t, \tag{14}$$

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

$$\beta^2 = \theta^2 \sum_{\lambda_j > 0} \frac{\lambda_j t_j^2}{(\lambda_j + \theta)^2}. \tag{15}$$

The optimal worst-case gain is

$$\Psi_\Lambda(h^\star; t, \beta) = \sum_{\lambda_j > 0} \left(\frac{\lambda_j}{\lambda_j + \theta} \right)^2 t_j^2, \tag{16}$$

and this value is attained in the minimax problem: for every correction h with finite worst-case gain, some $f_h \in \mathcal{F}_\beta(\Lambda)$ attains the infimum in (13); hence no rule depending only on (t, Λ, β) can have a uniformly larger worst-case gain.

The proof of Theorem 2 is given in Appendix C. The theorem formalizes why approximate validity leads to continuous attenuation rather than hard projection. Directions with large semantic variation are attenuated, but directions in which the target may vary are not forced to zero. When the validity budget is large enough to explain the teacher's inconsistency, the identity correction is minimax.

Two boundary cases and two auxiliary consequences clarify the theorem. If $\beta \geq \left\|\Lambda^{1/2} t\right\|$, the optimal value is zero and is achieved by $h = t$, so no nontrivial correction has positive worst-case gain. If $\beta = 0$, the limiting solution is hard projection $h^\star = \Pi_0 t$ with gain $\|\Pi_+ t\|^2$. Moreover, any correction with finite worst-case gain must preserve the unconstrained component $\Pi_0 t$; changing it makes $\Psi_\Lambda(h; t, \beta) = -\infty$, because $\mathcal{F}_\beta(\Lambda)$ imposes no restriction on $\ker(\Lambda)$. Finally, the proof yields the fixed-correction formula

$$\Psi_\Lambda(h; t, \beta) = 2\langle t, c \rangle - \|c\|^2 - 2\beta \left\|(\Lambda^\dagger)^{1/2} c\right\|, \qquad c = t - h, \quad c \perp \ker(\Lambda). \tag{17}$$

For hard projection $h_{\mathrm{HP}} = \Pi_0 t$, this becomes

$$\Psi_\Lambda(h_{\mathrm{HP}}; t, \beta) = \|\Pi_+ t\|^2 - 2\beta \left\|(\Lambda^\dagger)^{1/2} \Pi_+ t\right\|. \tag{18}$$

Thus hard projection has negative worst-case gain whenever the right-hand side is negative.

The deterministic minimax theorem treats the weak teacher as its mean prediction. When the weak supervisor is stochastic, as in LLM sampling or weak-judge evaluations, correction must also account for teacher sampling variance. The following spectral extension keeps the same validity-budget logic while adding the variance contribution of each semantic eigendirection.

Proposition 2 (Stochastic-teacher minimax spectral correction)**.** *Work on the positive semantic subspace where $\Lambda u_j = \lambda_j u_j$ with $\lambda_j > 0$. Let the weak teacher output $Y = t + \zeta$ with $\mathbb{E}\zeta = 0$ and arbitrary covariance Σ. For a spectral smoother $S_s u_j = s_j u_j$, $s_j \in [0, 1]$, define $\nu_j = \langle u_j, \Sigma u_j \rangle$, $a_j = t_j^2 + \nu_j$, and $J = \{j : t_j \neq 0\}$. The worst-case expected gain is*

$$\Gamma_\Sigma(s) = \sum_{\lambda_j > 0} (1 - s_j^2)(t_j^2 + \nu_j) - 2\beta \left(\sum_{\lambda_j > 0} \frac{(1 - s_j)^2 t_j^2}{\lambda_j} \right)^{1/2}. \tag{19}$$

For pure-noise modes with $t_j = 0$ and $\nu_j > 0$, every maximizer sets $s_j^\star = 0$. On the mean-carrying modes, define

$$\beta_c^2 = \sum_{j \in J} \frac{\lambda_j a_j^2}{t_j^2}.$$

If $\beta = 0$, then $s_j^\star = 0$ for all $j \in J$. If $0 < \beta < \beta_c$, the unique maximizer is interior and has

$$s_j^\star = \frac{t_j^2}{t_j^2 + \eta\lambda_j(t_j^2 + \nu_j)}, \qquad \beta^2 = \sum_{j \in J} \frac{\lambda_j t_j^2 (t_j^2 + \nu_j)^2}{[t_j^2 + \eta\lambda_j(t_j^2 + \nu_j)]^2}, \tag{20}$$

for the unique $\eta > 0$ solving the displayed equation. If $\beta \geq \beta_c$, then $s_j^\star = 1$ for all $j \in J$, so no mean-carrying component is corrected. Coordinates with $t_j = \nu_j = 0$ do not affect the objective, and setting $\nu_j = 0$ recovers the deterministic resolvent filter after reparametrizing $\eta = 1/\theta$.

The proof of Proposition 2 is given in Appendix D. This result identifies the exact decision-theoretic optimum for weak-label correction under approximate semantic validity. The deterministic solution is the resolvent filter $S_\theta = \theta(\Lambda + \theta I)^{-1}$: directions with larger Laplacian eigenvalues receive stronger attenuation, whereas low-eigenvalue directions that may contain target variation are preserved. The stochastic extension implies that high teacher sampling variance also favors attenuation, while a sufficiently large validity budget makes the identity map optimal on mean-carrying modes. For spectral smoothers, the covariance Σ need not commute with Λ; only the diagonal spectral variances $\nu_j = \langle u_j, \Sigma u_j \rangle$ enter the trace term. The hard-projection formula identifies the corresponding failure condition: if target-validity uncertainty is large relative to the teacher's constraint-violating energy, exact projection can increase worst-case risk.

4.3 Finite-query selection of constraints and filters

The final stage removes the assumption that the correction is fixed in advance. In applications, the constraint is often selected from a finite or parameterized family of transformation sets, paraphrase generators, graph builders, or smoothing parameters. Maximizing unadjusted inconsistency does not yield a valid lower bound, because a class-changing transformation can produce high inconsistency without implying weak-label error. The selection rule must therefore use a lower confidence bound on validity-adjusted gain.

Let $\mathcal{A} = \{1, \ldots, M\}$ be a finite set of candidates. Candidate a specifies $(\Lambda_a, h_a, \beta_a)$, where h_a is the corrected weak label and β_a is a validity budget. Require $c_a = t - h_a \in \mathrm{range}(\Lambda_a)$, so the correction preserves the teacher's nullspace component, and define

$$H_a = 2\langle t, c_a \rangle - \|c_a\|^2 = \|t\|^2 - \|h_a\|^2, \qquad D_a = \left\| (\Lambda_a^\dagger)^{1/2} c_a \right\|^2. \tag{21}$$

The term H_a is the observable quadratic gain associated with replacing t by h_a, while D_a measures the exposure of the removed component to target variation allowed by the semantic constraint. Their combination gives the validity-adjusted lower bound $\Gamma_a = H_a - 2\beta_a\sqrt{D_a}$.

Theorem 3 (Conservative finite-query selection)**.** *Let $S_a = \sqrt{D_a}$, so that $\Gamma_a = H_a - 2\beta_a S_a$, and fix $\delta \in (0, 1)$. Suppose that, with probability at least $1 - \delta$, the scoring sample gives numbers $\underline{H}_a, \overline{S}_a$ satisfying, simultaneously for all $a \in \mathcal{A}$,*

$$H_a - \varepsilon_H \leq \underline{H}_a \leq H_a, \qquad S_a \leq \overline{S}_a \leq S_a + \varepsilon_S. \tag{22}$$

Let $\bar{\beta}_a$ be the budget used by the selection rule, with $\beta_a \leq \bar{\beta}_a \leq \beta_a + \varepsilon_\beta$ on the same event, and choose

$$\widehat{a} \in \arg\max_{a \in \mathcal{A}} \mathcal{L}_a, \qquad \mathcal{L}_a = \underline{H}_a - 2\bar{\beta}_a \overline{S}_a. \tag{23}$$

Then every score is conservative, $\mathcal{L}_a \leq \Gamma_a$, and if $S_a \leq S_{\max}$ for all a,

$$\Gamma_{\widehat{a}} \geq \max_{a \in \mathcal{A}} \Gamma_a - \left(\varepsilon_H + 2\bar{\beta}_{\max}\varepsilon_S + 2\varepsilon_\beta S_{\max}\right), \qquad \bar{\beta}_{\max} = \max_a \bar{\beta}_a. \tag{24}$$

For exact budgets, $\varepsilon_\beta = 0$.

The proof of Theorem 3 is given in Appendix E. The theorem is stated at the level of conservative scores: any finite-candidate construction that lower-bounds H_a and upper-bounds S_a can be used for selection. A bounded empirical instantiation based on Hoeffding inequalities, together with the Markov and CIFAR score constructions used in the experiments, is given in Appendices E and I.

In the terminology of Definition 1, Theorem 3 justifies the scoring stage of VAC: the selected score remains conservative after data-dependent candidate selection, and its population lower-bound value

is close to the best validity-adjusted value in the candidate family. In the CIFAR experiment, VAC denotes the local probability-space correction selected by this rule; graph-resolvent variants are named explicitly.

The finite-query result controls the correction selected from weak-teacher queries, but student training introduces a second source of randomness. The next proposition states the independence conditions under which the selected population lower bound can be combined with the ERM transfer result.

Proposition 3 (Split-sample selected-correction transfer)**.** *Let $D_{\mathrm{build}}, D_{\mathrm{score}}, D_{\mathrm{train}}$ be mutually independent samples. Let D_{build} fix the candidate family and correction rules, and let D_{score} select $\widehat{a}$ using a valid lower-bound rule. Candidate a supplies corrected training labels with conditional mean $\bar{t}_a = h_a$; write $\widehat{h}_{\widehat{a}}$ for the student trained on these labels using D_{train}. For $\delta_{\mathrm{score}}, \delta_{\mathrm{train}} \in (0,1)$, suppose selection gives $\Gamma_{\widehat{a}} \geq \mathcal{L}_{\widehat{a}}$ with probability at least $1-\delta_{\mathrm{score}}$, and the radius condition $\|\widehat{h}_{\widehat{a}} - \bar{t}_{\widehat{a}}\| \leq \Delta_{\widehat{a}}$ holds with probability at least $1-\delta_{\mathrm{train}}$ conditional on $D_{\mathrm{build}}, D_{\mathrm{score}}$. Then, with probability at least $1-\delta_{\mathrm{score}}-\delta_{\mathrm{train}}$,*

$$\mathcal{R}(t) - \mathcal{R}(\widehat{h}_{\widehat{a}}) \geq \mathcal{L}_{\widehat{a}} - 2M_{\widehat{a}}\Delta_{\widehat{a}} - \Delta_{\widehat{a}}^2, \qquad M_{\widehat{a}} = \|\bar{t}_{\widehat{a}} - f^\star\|. \tag{25}$$

If $|\bar{t}_{\widehat{a}}|, |f^\star| \leq B$, then $M_{\widehat{a}} \leq 2B$, so a positive end-to-end lower bound is obtained when

$$\mathcal{L}_{\widehat{a}} > 4B\Delta_{\widehat{a}} + \Delta_{\widehat{a}}^2.$$

The proof of Proposition 3 is given in Appendix E. This proposition is the step from selected correction to a trained student: sample splitting makes the selected correction fixed relative to the student-training sample, so the population lower bound can be combined with the ERM transfer bound. If selection and student training share data, an additional joint complexity term is required; this is why the full-pool CIFAR run is reported only as an appendix ablation. A PAC-Bayes extension for large or continuous candidate families is stated and proved in Appendix F.

The same Hilbert-space mechanism also gives a preference-learning instance: the cyclic component of a weak preference flow gives a label-free lower bound only when target cyclic energy is independently bounded. The full statement, including the bounded-logit Bradley–Terry/logistic transfer inequality, is given in Appendix G.

5 Experiments

The empirical evaluation tests the proposed correction procedure rather than a general explanation of W2S. The controlled graph-signal experiment instantiates the Hilbert-space model, where target variation and teacher error components are known. The CIFAR-10 experiment then studies correction and student transfer in a split-sample neural W2S pipeline, including matched multiview-averaging baselines. The ARC experiment tests an exactly valid option-permutation constraint with pretrained language models. Additional CIFAR runs, alternative correction variants, validation-budget ablations, and non-split full-pool runs are reported in Appendix H.

5.1 Controlled graph-signal regression

The controlled experiment uses a finite semantic graph with paired nuisance copies and local circular edges. Target and teacher signals are generated in the graph-Laplacian eigenbasis, so the true validity budget β and the teacher's constraint-violating component are known. We compare Markov smoothing, hard projection, the minimax resolvent, weak-label cross-validated graph smoothing, and an invalid pair-averaging control.

Table 1: Controlled graph-signal experiment. Panel A reports population risk reduction relative to the raw teacher for all validity budgets, with mean ± standard deviation over 20 seeds; negative values indicate increased risk. Panel B reports finite-sample selection frequencies for all label-free sample sizes, averaged over validity budgets and seeds. Higher best-selection frequency and lower invalid-selection frequency are better.

Panel A: population risk reduction					
β	Raw	Markov $k = 8$	Hard proj.	Resolvent	Invalid pair
0.0	0.000 ± 0.000	0.250 ± 0.018	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	0.711 ± 0.016
0.05	0.000 ± 0.000	0.249 ± 0.021	0.201 ± 0.153	$\mathbf{0.843 \pm 0.025}$	0.488 ± 0.063
0.1	0.000 ± 0.000	0.245 ± 0.024	-2.194 ± 0.613	$\mathbf{0.684 \pm 0.046}$	-0.184 ± 0.207
0.2	0.000 ± 0.000	0.235 ± 0.032	-11.777 ± 2.451	$\mathbf{0.438 \pm 0.073}$	-2.879 ± 0.764
0.4	0.000 ± 0.000	$\mathbf{0.195 \pm 0.049}$	-50.107 ± 9.803	0.193 ± 0.083	-13.663 ± 2.942
0.8	0.000 ± 0.000	0.043 ± 0.084	-203.430 ± 39.211	$\mathbf{0.064 \pm 0.062}$	-56.816 ± 11.555
Panel B: finite-sample selection frequencies					
N	Unadj. best	VAC best	Unadj. invalid	VAC invalid	Invalid reduction
16	0.148	0.110	0.343	0.019	0.323
32	0.150	0.125	0.346	0.020	0.326
64	0.147	0.154	0.338	0.021	0.316
128	0.139	0.191	0.335	0.024	0.311
256	0.130	0.255	0.318	0.029	0.289
512	0.118	0.337	0.291	0.032	0.259

Table 1 is consistent with the three main predictions of the analysis. Exact-validity inconsistency is an exact risk-reduction statistic; hard projection becomes unsafe under approximate validity because it removes genuine target variation; and one-sided conservative scoring sharply reduces selection of the invalid control. Additional misspecification ablations in Appendix H show that treating approximate validity as exact causes over-smoothing and negative gain.

5.2 CIFAR-10 split-sample neural W2S experiment

The neural experiment uses the cats-versus-dogs subset of CIFAR-10. In each of ten seeds, a four-layer ConvNet weak teacher is trained on 500 labeled examples per class. The remaining weak-labeled pool is split into D_{score} with 2,200 examples and a disjoint D_{train} with 6,600 examples. VAC uses only weak-teacher queries on D_{score} and predeclared probability-space validity budgets to select a local correction; the ResNet-18 student is then trained for 200 epochs on weak or corrected weak labels from D_{train}. Across seeds, VAC selects the combined valid transformation in 9 seeds and random crop in 1 seed.

To separate averaging, budget-dependent shrinkage, and candidate selection, we compare Raw W2S, fixed test-time augmentation (TTA), fixed shrinkage, and full VAC using ten paired seeds and identical student-side augmentation. Fixed TTA averages teacher probabilities over the same eight transformed views as the combined VAC candidate, excluding the original view. Fixed shrinkage uses that candidate with $\beta = 0.06$, while full VAC selects the candidate on D_{score}.

Table 2: Matched split-sample CIFAR-10 comparison over 10 paired seeds. All methods use the same student-side augmentation. Values are mean ± standard deviation. Fixed TTA uses eight transformed views without budget adjustment or candidate selection.

Method	Weak-label Brier ↓	Student acc.	Student Brier ↓	Student NLL ↓
Raw W2S	0.4181 ± 0.0192	0.7505 ± 0.0100	0.3965 ± 0.0178	0.8535 ± 0.0582
Fixed TTA	0.3685 ± 0.0138	0.7513 ± 0.0124	0.3651 ± 0.0168	0.6350 ± 0.0340
Fixed shrink, $\beta = 0.06$	0.3732 ± 0.0149	0.7559 ± 0.0153	0.3655 ± 0.0231	0.6704 ± 0.0560
Full VAC	0.3741 ± 0.0148	0.7563 ± 0.0155	0.3658 ± 0.0229	0.6723 ± 0.0547

Table 2 shows that multiview weak-label correction improves student probability metrics: Fixed TTA reduces Brier by 0.03134 relative to Raw W2S (paired 95% CI $[0.02698, 0.03571]$). Full VAC improves accuracy over Raw W2S by 0.575 percentage points (95% CI $[0.040, 1.075]$) and reduces Brier by 0.03062 (95% CI $[0.02501, 0.03581]$). Relative to Fixed TTA, fixed shrinkage changes Brier by $+0.00039$ (95% CI $[-0.00541, 0.00583]$), and candidate selection adds little because VAC selects

the combined candidate in nine seeds. Thus multiview averaging accounts for the probability-risk improvement in this setting, with no additional Brier benefit established for nonzero shrinkage or selection.

5.3 ARC language-model W2S experiment

We evaluate option-permutation correction on the Easy and Challenge subsets of the AI2 Reasoning Challenge (ARC) [6], using Qwen3-1.7B as the weak teacher and Qwen3-4B as the student [20]. A structural audit found no unresolved position-dependent references. Teacher probabilities under option permutations are mapped back to canonical answer identities before averaging and student training. On this audited corpus, option permutation preserves the correct answer by construction, giving $\beta = 0$ independently of teacher behavior and downstream evaluation.

Table 3: ARC student improvements over Raw W2S across five paired training seeds. Accuracy differences are in percentage points (pp); negative Brier differences indicate improvement. We report mean $\pm$ standard deviation and paired 95% confidence intervals.

Subset	Accuracy diff. (pp)	95% CI	Brier diff.	95% CI
Easy	$+1.48 \pm 0.30$	$[1.11, 1.86]$	-0.0222 ± 0.0042	$[-0.0274, -0.0169]$
Challenge	$+1.86 \pm 0.65$	$[1.06, 2.66]$	-0.0481 ± 0.0099	$[-0.0604, -0.0358]$
Combined	$+1.61 \pm 0.27$	$[1.28, 1.94]$	-0.0307 ± 0.0057	$[-0.0378, -0.0236]$

Table 3 shows gains on both subsets. A single permutation yields no stable accuracy benefit, while an invalid semantic-changing control reduces combined accuracy by 0.33 percentage points and increases Brier by 0.0101 relative to Raw W2S. These results demonstrate that an independently valid constraint can identify a correctable weak-label component that transfers to a stronger language-model student.

6 Conclusion

This work identified sufficient conditions under which independently specified semantic constraints yield label-free weak-label correction for weak-to-strong generalization. The central observation is that a weak predictor's violation of an independently valid semantic constraint yields an observable lower bound on the risk reduction from correcting weak labels. For reversible Markov constraints, trajectory inconsistency equals squared-risk gain under exact validity and remains a conservative lower bound under approximate validity. Under a Laplacian validity budget, the minimax analysis gives a resolvent smoothing correction when a positive worst-case gain is possible and the identity correction when validity uncertainty is too large. The stochastic-teacher extension has the same spectral dependence with additional attenuation in high-variance modes. The ERM and split-sample transfer bounds specify when a population correction lower bound becomes an end-to-end W2S risk-improvement bound after student training. Controlled graph-signal experiments, split-sample CIFAR-10 classification, and ARC language-model experiments support this analysis: valid constraints enable weak-label correction and student gain, while invalid controls show why raw inconsistency alone is insufficient. The results therefore identify a mathematically specified regime in which weak labels can be improved using label-free constraints, without asserting a general explanation of all W2S behavior.

Acknowledgements

This work was partially supported by NSF IIS 2347592, 2348169, DBI 2405416, CCF 2348306, CNS 2347617, RISE 2536663.

Competing Interests Disclosure

The authors declare no competing interests.

References

[1] Mikhail Belkin, Partha Niyogi, and Vikas Sindhwani. Manifold regularization: A geometric framework for learning from labeled and unlabeled examples. *Journal of Machine Learning Research*, 7(85):2399–2434, 2006. URL https://jmlr.org/papers/v7/belkin06a.html.

[2] David Berthelot, Nicholas Carlini, Ian Goodfellow, Nicolas Papernot, Avital Oliver, and Colin A. Raffel. MixMatch: A holistic approach to semi-supervised learning. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL https://papers.nips.cc/paper/8749-mixmatch-a-holistic-approach-to-semi-supervised-learning.

[3] Collin Burns, Pavel Izmailov, Jan Hendrik Kirchner, Bowen Baker, Leo Gao, Leopold Aschenbrenner, Yining Chen, Adrien Ecoffet, Manas Joglekar, Jan Leike, Ilya Sutskever, and Jeffrey Wu. Weak-to-strong generalization: Eliciting strong capabilities with weak supervision. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 4971–5012. PMLR, 2024. URL https://proceedings.mlr.press/v235/burns24b.html.

[4] Moses Charikar, Chirag Pabbaraju, and Kirankumar Shiragur. Quantifying the gain in weak-to-strong generalization. In *Advances in Neural Information Processing Systems*, volume 37, pages 126474–126499. Curran Associates, Inc., 2024. doi: 10.52202/079017-4017. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/e4a0d8aef3567f742b0794844d9b5847-Abstract-Conference.html.

[5] Shuxiao Chen, Edgar Dobriban, and Jane H. Lee. A group-theoretic framework for data augmentation. *Journal of Machine Learning Research*, 21(245):1–71, 2020. URL https://jmlr.org/papers/v21/20-163.html.

[6] Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. Think you have solved question answering? try ARC, the AI2 reasoning challenge, 2018. URL https://arxiv.org/abs/1803.05457.

[7] Yijun Dong, Yicheng Li, Yunai Li, Jason D. Lee, and Qi Lei. Discrepancies are virtue: Weak-to-strong generalization through lens of intrinsic dimension. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pages 14079–14113. PMLR, 2025. URL https://proceedings.mlr.press/v267/dong25g.html.

[8] Heinz W. Engl, Martin Hanke, and Andreas Neubauer. *Regularization of Inverse Problems*, volume 375 of *Mathematics and Its Applications*. Kluwer Academic Publishers, Dordrecht, 1996. ISBN 978-0-7923-4157-4. URL https://link.springer.com/book/9780792341574.

[9] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network, 2015. URL https://arxiv.org/abs/1503.02531.

[10] Muhammed Emrullah Ildiz, Halil Alperen Gozeten, Ege Onur Taga, Marco Mondelli, and Samet Oymak. High-dimensional analysis of knowledge distillation: Weak-to-strong generalization and scaling laws. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=1xzqz73hvL.

[11] Xiaoye Jiang, Lek-Heng Lim, Yuan Yao, and Yinyu Ye. Statistical ranking and combinatorial hodge theory. *Mathematical Programming*, 127(1):203–244, 2011. doi: 10.1007/s10107-010-0419-x. URL https://link.springer.com/article/10.1007/s10107-010-0419-x.

[12] Alex Krizhevsky. Learning multiple layers of features from tiny images. Technical report, University of Toronto, 2009. URL https://www.cs.toronto.edu/~kriz/learning-features-2009-TR.pdf.

[13] Samuli Laine and Timo Aila. Temporal ensembling for semi-supervised learning. In *International Conference on Learning Representations*, 2017. URL https://openreview.net/forum?id=BJ6oOfqge.

[14] Hunter Lang, David A. Sontag, and Aravindan Vijayaraghavan. Theoretical analysis of weak-to-strong generalization. In *Advances in Neural Information Processing Systems*, volume 37, pages 46837–46880. Curran Associates, Inc., 2024. doi: 10.52202/079017-1486. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/5358d1e6138dff5718c5e9790f5fa593-Abstract-Conference.html.

[15] Andreas Maurer. A note on the PAC bayesian theorem, 2004. URL https://arxiv.org/abs/cs/0411099.

[16] Marko Medvedev, Kaifeng Lyu, Dingli Yu, Sanjeev Arora, Zhiyuan Li, and Nathan Srebro. Weak-to-strong generalization even in random feature networks, provably. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pages 43519–43556. PMLR, 2025. URL https://proceedings.mlr.press/v267/medvedev25a.html.

[17] Takeru Miyato, Shin-Ichi Maeda, Masanori Koyama, and Shin Ishii. Virtual adversarial training: A regularization method for supervised and semi-supervised learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(8):1979–1993, 2019. doi: 10.1109/TPAMI.2018.2858821. URL https://pubmed.ncbi.nlm.nih.gov/30040630/.

[18] Abhijeet Mulgund and Chirag Pabbaraju. Relating misfit to gain in weak-to-strong generalization beyond the squared loss. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pages 45129–45150. PMLR, 2025. URL https://proceedings.mlr.press/v267/mulgund25a.html.

[19] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F. Christiano, Jan Leike, and Ryan Lowe. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems*, volume 35, pages 27730–27744. Curran Associates, Inc., 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/b1efde53be364a73914f58805a001731-Abstract.html.

[20] Qwen Team. Qwen3 technical report, 2025. URL https://arxiv.org/abs/2505.09388.

[21] Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D. Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. In *Advances in Neural Information Processing Systems*, volume 36, pages 53728–53741. Curran Associates, Inc., 2023. doi: 10.52202/075280-2338. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/a85b405ed65c6477a4fe8302b5e06ce7-Abstract-Conference.html.

[22] Changho Shin, John Cooper, and Frederic Sala. Weak-to-strong generalization through the data-centric lens. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=uogG8BfLs2.

[23] Kihyuk Sohn, David Berthelot, Nicholas Carlini, Zizhao Zhang, Han Zhang, Colin A. Raffel, Ekin Dogus Cubuk, Alexey Kurakin, and Chun-Liang Li. FixMatch: Simplifying semi-supervised learning with consistency and confidence. In *Advances in Neural Information Processing Systems*, volume 33, pages 596–608. Curran Associates, Inc., 2020. URL https://papers.nips.cc/paper_files/paper/2020/hash/06964dce9addb1c5cb5d6e3d9838f733-Abstract.html.

[24] Antti Tarvainen and Harri Valpola. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://neurips.cc/virtual/2017/poster/8912.

[25] Qizhe Xie, Minh-Thang Luong, Eduard Hovy, and Quoc V. Le. Self-training with noisy student improves imagenet classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10687–10698, 2020. doi: 10.1109/CVPR42600.2020.01070. URL https://openaccess.thecvf.com/content_CVPR_2020/html/Xie_Self-Training_With_Noisy_Student_Improves_ImageNet_Classification_CVPR_2020_paper.html.

[26] Yihao Xue, Jiping Li, and Baharan Mirzasoleiman. Representations shape weak-to-strong generalization: Theoretical insights and empirical predictions. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pages 70120–70159. PMLR, 2025. URL https://proceedings.mlr.press/v267/xue25d.html.

[27] Wei Yao, Wenkai Yang, Ziqiao Wang, Yankai Lin, and Yong Liu. Revisiting weak-to-strong generalization in theory and practice: Reverse KL vs. forward KL. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 2860–2888, Vienna, Austria, 2025. Association for Computational Linguistics. doi: 10.18653/v1/2025.findings-acl.148. URL https://aclanthology.org/2025.findings-acl.148/.

[28] Dengyong Zhou, Olivier Bousquet, Thomas Navin Lal, Jason Weston, and Bernhard Schölkopf. Learning with local and global consistency. In *Advances in Neural Information Processing Systems*, volume 16, pages 321–328. MIT Press, 2003. URL https://proceedings.neurips.cc/paper_files/paper/2003/hash/87682805257e619d49b8e0dfdc14affa-Abstract.html.

[29] Xiaojin Zhu, Zoubin Ghahramani, and John D. Lafferty. Semi-supervised learning using gaussian fields and harmonic functions. In *Proceedings of the 20th International Conference on Machine Learning*, pages 912–919, 2003. URL https://mlanthology.org/icml/2003/zhu2003icml-semi/.

A Proof for Theorem 1

The proof uses stationarity, reversibility, and the spectral calculus for the self-adjoint contraction $0 \preceq T \preceq I$. The first step identifies the observable trajectory quantity with the quadratic form $C_k(t)$. Since the chain is stationary, both X_0 and X_{2k} have marginal distribution P. Since the transition from X_0 to X_{2k} is K^{2k}, the cross term satisfies

$$\mathbb{E}[t(X_0)t(X_{2k})] = \int t(x)(T^{2k}t)(x)\, dP(x) = \langle t, T^{2k}t \rangle .$$

Therefore

$$\frac{1}{2}\mathbb{E}\big(t(X_0) - t(X_{2k})\big)^2 = \|t\|^2 - \langle t, T^{2k}t \rangle = \langle t, (I - T^{2k})t \rangle = C_k(t),$$

which proves the identity in (6). Now let $u_k = T^k t$. Expanding the two squared risks gives

$$\begin{aligned}\mathcal{R}(t) - \mathcal{R}(u_k) &= \|t - f^\star\|^2 - \left\|T^k t - f^\star\right\|^2 \\ &= \|t\|^2 - \left\|T^k t\right\|^2 - 2\langle t - T^k t, f^\star \rangle .\end{aligned}$$

Because T is self-adjoint, $\left\|T^k t\right\|^2 = \langle T^k t, T^k t \rangle = \langle t, T^{2k} t \rangle$. Therefore the first two terms in the last line equal $C_k(t)$, and

$$\mathcal{R}(t) - \mathcal{R}(u_k) = C_k(t) - 2\langle (I - T^k)t, f^\star \rangle .$$

This proves (8). If $\Lambda f^\star = 0$, then $T f^\star = f^\star$ and $T^k f^\star = f^\star$. Self-adjointness then gives

$$\langle (I - T^k)t, f^\star \rangle = \langle t, (I - T^k) f^\star \rangle = 0,$$

so (9) follows. Under approximate validity, set $c_k = (I - T^k)t$. The vector c_k has no fixed-point component. Indeed, if $g \in \ker(\Lambda)$, then $Tg = g$ and

$$\langle c_k, g \rangle = \langle (I - T^k)t, g \rangle = \langle t, (I - T^k)g \rangle = 0.$$

In fact, $c_k = \Lambda \sum_{r=0}^{k-1} T^r t \in \mathrm{range}(\Lambda)$, so

$$\langle c_k, f^\star \rangle = \left\langle (\Lambda^\dagger)^{1/2} c_k, \Lambda^{1/2} f^\star \right\rangle$$

is valid. Cauchy–Schwarz and $\left\|\Lambda^{1/2} f^\star\right\| \le \beta$ give

$$|\langle c_k, f^\star \rangle| \le \left\|(\Lambda^\dagger)^{1/2} c_k\right\| \left\|\Lambda^{1/2} f^\star\right\| \le \beta A_k(t).$$

Together with (8), this proves

$$\mathcal{R}(t) - \mathcal{R}(u_k) \ge C_k(t) - 2\beta A_k(t).$$

It remains to derive the auxiliary bound $A_k(t) \le \sqrt{k C_k(t)}$. Suppose first that T has an orthonormal eigenbasis $T u_j = \mu_j u_j$ with $\mu_j \in [0, 1]$, and write $t = \sum_j t_j u_j$. Then

$$A_k(t)^2 = \sum_j \frac{(1 - \mu_j^k)^2}{1 - \mu_j} t_j^2, \qquad C_k(t) = \sum_j (1 - \mu_j^{2k}) t_j^2.$$

Terms with $\mu_j = 1$ are interpreted as zero, as prescribed by the Moore–Penrose inverse. For $\mu \in [0, 1)$,

$$\frac{(1 - \mu^k)^2}{1 - \mu} = (1 + \mu + \cdots + \mu^{k-1})(1 - \mu^k) \le k(1 - \mu^k) \le k(1 - \mu^{2k}),$$

so $A_k(t)^2 \le k C_k(t)$. The general self-adjoint case follows by applying the same scalar inequality under the spectral measure of T. This completes the proof of (10).

The finite-query estimate (11) is a direct concentration consequence. Each

$$Z_i = \frac{1}{2}\big(t(X_0^{(i)}) - t(X_{2k}^{(i)})\big)^2$$

has mean $C_k(t)$ and lies in $[0, 2B^2]$ when $|t| \le B$. Hoeffding's inequality yields

$$\Pr\{\widehat{C}_k - C_k(t) \ge r\} \le \exp\left(-\frac{2Nr^2}{(2B^2)^2}\right),$$

and the stated radius follows by solving for probability α.

B Proof for Proposition 1

We work conditionally on any data used to construct or select the corrected label $\bar{t}$. Thus $\bar{t}$ is fixed relative to the student-training sample. All norms and inner products are in $L^2(P)$. Define

$$G_{\bar{t}} := \mathcal{R}(t) - \mathcal{R}(\bar{t}) = \|t - f^\star\|^2 - \|\bar{t} - f^\star\|^2 .$$

First we prove the deterministic transfer identity. Let

$$e = \widehat{h} - \bar{t}.$$

Then

$$\mathcal{R}(\widehat{h}) = \left\|\widehat{h} - f^\star\right\|^2 = \|\bar{t} - f^\star + e\|^2 = \mathcal{R}(\bar{t}) + 2\langle \bar{t} - f^\star, e\rangle + \|e\|^2 .$$

Subtracting from $\mathcal{R}(t)$ gives the exact identity

$$\mathcal{R}(t) - \mathcal{R}(\widehat{h}) = G_{\bar{t}} - 2\langle \bar{t} - f^\star, e\rangle - \|e\|^2 .$$

Consequently, if $G_{\bar{t}} \geq \Gamma$, $\|e\| \leq \Delta$, and

$$M := \|\bar{t} - f^\star\| ,$$

then Cauchy–Schwarz gives

$$\mathcal{R}(t) - \mathcal{R}(\widehat{h}) \geq \Gamma - 2M\Delta - \Delta^2.$$

This radius-only lower bound is sharp: if one only knows $\|e\| \leq \Delta$, equality can be attained when e has norm Δ and is aligned with $\bar{t} - f^\star$.

If $\Gamma > 0$, the right-hand side is positive exactly when

$$\Delta < \tau(\Gamma, M) := \sqrt{M^2 + \Gamma} - M.$$

Indeed,

$$\Gamma - 2M\Delta - \Delta^2 > 0 \quad \Longleftrightarrow \quad \Delta < \sqrt{M^2 + \Gamma} - M.$$

Moreover, since $2M\tau + \tau^2 = \Gamma$, if $\Delta \leq \rho\tau$ for some $\rho \in [0, 1]$, then

$$2M\Delta + \Delta^2 \leq 2M\rho\tau + \rho^2\tau^2 \leq \rho(2M\tau + \tau^2) = \rho\Gamma.$$

Therefore

$$\mathcal{R}(t) - \mathcal{R}(\widehat{h}) \geq (1 - \rho)\Gamma.$$

The radius condition follows from approximate empirical risk minimization. Conditional on the construction and selection data, let $\mathcal{S} \subset L^2(P)$ be a fixed nonempty student hypothesis class, and let $(X_i, Y_i)_{i=1}^n$ be independent and identically distributed with law Q, X-marginal P, $\mathbb{E}_Q[Y \mid X] = \bar{t}(X)$, and $\mathbb{E}_Q Y^2 < \infty$. For a function g of (X, Y), write $Qg = \mathbb{E}_Q g(X, Y)$ and $Q_n g = n^{-1}\sum_{i=1}^n g(X_i, Y_i)$. Define the pseudo-label risks

$$\widehat{\mathcal{L}}_n(h) = \frac{1}{n}\sum_{i=1}^n (h(X_i) - Y_i)^2, \qquad \mathcal{L}(h) = \mathbb{E}[(h(X) - Y)^2].$$

Because

$$Y = \bar{t}(X) + \xi, \qquad \mathbb{E}[\xi \mid X] = 0,$$

we have

$$\begin{aligned}\mathcal{L}(h) &= \mathbb{E}[(h(X) - \bar{t}(X) - \xi)^2] \\ &= \|h - \bar{t}\|^2 + \mathbb{E}[\xi^2] - 2\mathbb{E}[(h(X) - \bar{t}(X))\xi] \\ &= \|h - \bar{t}\|^2 + \mathbb{E}[\xi^2].\end{aligned}$$

The cross term vanishes because

$$\mathbb{E}[(h(X) - \bar{t}(X))\xi] = \mathbb{E}[(h(X) - \bar{t}(X))\mathbb{E}(\xi \mid X)] = 0.$$

Assume the minimum below is attained, and set

$$h_{\mathcal{S}} \in \arg\min_{h \in \mathcal{S}} \|h - \bar{t}\|^2, \qquad a_{\mathcal{S}} := \|h_{\mathcal{S}} - \bar{t}\|^2.$$

Then the decomposition above implies

$$\mathcal{L}(h) - \mathcal{L}(h_{\mathcal{S}}) = \|h - \bar{t}\|^2 - a_{\mathcal{S}}.$$

Let $\widehat{h} \in \mathcal{S}$ be an approximate ERM with optimization tolerance $\eta_{\text{opt}} \geq 0$, so that

$$\widehat{\mathcal{L}}_n(\widehat{h}) \leq \inf_{h \in \mathcal{S}} \widehat{\mathcal{L}}_n(h) + \eta_{\text{opt}}.$$

Comparing the infimum with $h_{\mathcal{S}}$, we obtain

$$Q_n\left[(\widehat{h} - Y)^2 - (h_{\mathcal{S}} - Y)^2\right] \leq \eta_{\text{opt}}.$$

Define the excess empirical-process error

$$\epsilon_n = \sup_{h \in \mathcal{S}} \left| (Q - Q_n)\left[(h - Y)^2 - (h_{\mathcal{S}} - Y)^2\right] \right|.$$

Applying this definition to $h = \widehat{h}$ and using the empirical comparison,

$$\begin{aligned}
\mathcal{L}(\widehat{h}) - \mathcal{L}(h_{\mathcal{S}}) &= (Q - Q_n)\left[(\widehat{h} - Y)^2 - (h_{\mathcal{S}} - Y)^2\right] \\
&\quad + Q_n\left[(\widehat{h} - Y)^2 - (h_{\mathcal{S}} - Y)^2\right] \\
&\leq \epsilon_n + \eta_{\text{opt}}.
\end{aligned}$$

Combining this with

$$\mathcal{L}(h) - \mathcal{L}(h_{\mathcal{S}}) = \|h - \bar{t}\|^2 - a_{\mathcal{S}}$$

gives

$$\left\|\widehat{h} - \bar{t}\right\|^2 \leq a_{\mathcal{S}} + \epsilon_n + \eta_{\text{opt}} =: \Delta^2.$$

Substituting this radius into the deterministic transfer bound proves

$$\mathcal{R}(t) - \mathcal{R}(\widehat{h}) \geq \Gamma - 2\|\bar{t} - f^\star\|\Delta - \Delta^2.$$

Finally, if $|\bar{t}| \leq B$ and $|f^\star| \leq B$, then

$$M = \|\bar{t} - f^\star\| \leq 2B.$$

Thus the lower bound is positive whenever

$$\Gamma > 4B\Delta + \Delta^2.$$

Equivalently, the bounded-output transfer radius is

$$\Delta < \sqrt{4B^2 + \Gamma} - 2B.$$

C Proof for Theorem 2

The proof is a minimax calculation over the correction vector. For any corrected label h, write $c = t - h$, so optimizing over h is equivalent to optimizing over the removed component c. Then for every candidate target f,

$$\|t - f\|^2 - \|h - f\|^2 = \|t - f\|^2 - \|t - c - f\|^2 = 2\langle t, c\rangle - \|c\|^2 - 2\langle c, f\rangle.$$

Therefore

$$\Psi_\Lambda(h; t, \beta) = 2\langle t, c\rangle - \|c\|^2 - 2 \sup_{\|\Lambda^{1/2} f\| \leq \beta} \langle c, f\rangle. \tag{26}$$

The constraint does not restrict $\ker(\Lambda)$, so the nullspace component must be preserved. If $\Pi_0 c \neq 0$, then one can take $f = M\Pi_0 c$ with arbitrarily large M. This target remains feasible because

$\Lambda^{1/2}\Pi_0 c = 0$, while $\langle c, f\rangle = M\|\Pi_0 c\|^2$ diverges. Hence the supremum in (26) is infinite and $\Psi_\Lambda(h;t,\beta) = -\infty$. Thus every rule with finite worst-case gain must satisfy $\Pi_0 c = 0$, equivalently $\Pi_0 h = \Pi_0 t$.

Once $c \perp \ker(\Lambda)$, the adversary acts only on the positive subspace of Λ. Writing $\langle c, f\rangle = \left\langle (\Lambda^\dagger)^{1/2} c, \Lambda^{1/2} f\right\rangle$ on this subspace and applying Cauchy–Schwarz gives

$$\sup_{\|\Lambda^{1/2} f\| \le \beta} \langle c, f\rangle = \beta \left\|(\Lambda^\dagger)^{1/2} c\right\|,$$

with equality for $f_c = \beta \Lambda^\dagger c / \left\|(\Lambda^\dagger)^{1/2} c\right\|$ when $c \neq 0$. Thus the support-function calculation identifies the worst-case target in the validity class. Hence every correction with finite worst-case gain has

$$\Psi_\Lambda(h;t,\beta) = J(c) := 2\langle t, c\rangle - \|c\|^2 - 2\beta \left\|(\Lambda^\dagger)^{1/2} c\right\|. \tag{27}$$

This also proves the minimax attainment statement: the infimum is attained by an explicit target in $\mathcal{F}_\beta(\Lambda)$.

It remains to maximize $J(c)$ over $c \perp \ker(\Lambda)$. The boundary cases explain the assumption $0 < \beta < \left\|\Lambda^{1/2} t\right\|$. If $\beta \ge \left\|\Lambda^{1/2} t\right\|$, then

$$\langle t, c\rangle = \left\langle \Lambda^{1/2} t, (\Lambda^\dagger)^{1/2} c\right\rangle \le \left\|\Lambda^{1/2} t\right\| \left\|(\Lambda^\dagger)^{1/2} c\right\| \le \beta \left\|(\Lambda^\dagger)^{1/2} c\right\|,$$

so $J(c) \le -\|c\|^2 \le 0$, with equality at $c = 0$, i.e. $h = t$. If $\beta = 0$, then $J(c) = 2\langle t, c\rangle - \|c\|^2$ and is uniquely maximized by $c = \Pi_+ t$, giving $h^\star = \Pi_0 t$ and value $\|\Pi_+ t\|^2$.

Assume now that $0 < \beta < \left\|\Lambda^{1/2} t\right\|$. The function J is strictly concave, since it is a linear term plus $-\|c\|^2$ and the concave term $-2\beta \left\|(\Lambda^\dagger)^{1/2} c\right\|$. The origin is not optimal in this regime: moving in a small positive multiple of Λt gives a first-order gain proportional to $\left\|\Lambda^{1/2} t\right\| - \beta > 0$. Therefore the maximizer is nonzero, and any nonzero stationary point is the unique global maximizer. For $c \neq 0$,

$$\nabla_{\Pi_+ \mathcal{H}} J(c) = 2\Pi_+ t - 2c - 2\beta \frac{\Lambda^\dagger c}{\left\|(\Lambda^\dagger)^{1/2} c\right\|},$$

where the gradient is taken within $\Pi_+ \mathcal{H}$. Setting it to zero and defining $\theta = \beta / \left\|(\Lambda^\dagger)^{1/2} c\right\| > 0$ gives

$$\Pi_+ t = c + \theta \Lambda^\dagger c, \qquad \Lambda t = (\Lambda + \theta I) c, \qquad c = (\Lambda + \theta I)^{-1} \Lambda t.$$

Thus the optimal correction is $c = (\Lambda + \theta I)^{-1} \Lambda t$. Therefore, in spectral coordinates,

$$c_j = \frac{\lambda_j}{\lambda_j + \theta} t_j, \qquad h_j = t_j - c_j = \frac{\theta}{\lambda_j + \theta} t_j,$$

which is (14). It remains to determine θ. The definition $\theta = \beta / \left\|(\Lambda^\dagger)^{1/2} c\right\|$ yields

$$\beta^2 = \theta^2 \left\|(\Lambda^\dagger)^{1/2} c\right\|^2 = \theta^2 \sum_{\lambda_j > 0} \frac{\lambda_j t_j^2}{(\lambda_j + \theta)^2}.$$

This scalar equation has a unique solution. Indeed,

$$\phi(\theta) = \theta^2 \sum_{\lambda_j > 0} \frac{\lambda_j t_j^2}{(\lambda_j + \theta)^2}$$

is continuous, strictly increasing on every nonzero teacher component, satisfies $\phi(0+) = 0$, and satisfies $\phi(\infty) = \left\|\Lambda^{1/2} t\right\|^2$. Since $0 < \beta < \left\|\Lambda^{1/2} t\right\|$, the equation $\phi(\theta) = \beta^2$ has exactly one solution. This proves existence and uniqueness of θ. Finally, substituting the optimizer back into J gives

$$\begin{aligned} J(c^\star) &= 2\langle t, c^\star\rangle - \|c^\star\|^2 - 2\beta \left\|(\Lambda^\dagger)^{1/2} c^\star\right\| \\ &= \sum_{\lambda_j > 0} \left[\frac{2\lambda_j}{\lambda_j + \theta} - \frac{\lambda_j^2}{(\lambda_j + \theta)^2} - \frac{2\theta\lambda_j}{(\lambda_j + \theta)^2}\right] t_j^2 \\ &= \sum_{\lambda_j > 0} \left(\frac{\lambda_j}{\lambda_j + \theta}\right)^2 t_j^2, \end{aligned}$$

where $\theta = \beta / \left\|(\Lambda^\dagger)^{1/2} c^\star\right\|$. This proves (16). The fixed-correction formula (17) is (27); taking $c = \Pi_+ t$ gives the hard-projection identity (18). If that expression is negative, the adversarial target above satisfies $\|\Pi_0 t - f_c\|^2 > \|t - f_c\|^2$.

D Proof for Proposition 2

The proof of Proposition 2 is carried out on the positive subspace of Λ. The nullspace mean is preserved as in the deterministic minimax theorem because the validity class imposes no restriction on target variation in $\ker(\Lambda)$. Thus the only coordinates to optimize are those with $\lambda_j > 0$. Let $S_s u_j = s_j u_j$ and write $\zeta_j = \langle \zeta, u_j \rangle$, so $\nu_j = \mathbb{E}\zeta_j^2 = \langle u_j, \Sigma u_j \rangle$.

For a fixed spectral smoother, the expected gain satisfies the following identity. No simultaneous diagonalization of Σ and Λ is needed:

$$\mathbb{E}\,\|Y - f\|^2 = \|t - f\|^2 + \operatorname{tr}(\Sigma), \qquad \mathbb{E}\,\|S_s Y - f\|^2 = \|S_s t - f\|^2 + \operatorname{tr}(S_s^2 \Sigma),$$

and, since S_s is diagonal in the Λ-eigenbasis,

$$\operatorname{tr}\{(I - S_s^2)\Sigma\} = \sum_j (1 - s_j^2)\,\langle u_j, \Sigma u_j \rangle = \sum_j (1 - s_j^2)\nu_j.$$

The mean part is deterministic after averaging over ζ:

$$\|t - f\|^2 - \|S_s t - f\|^2 = \sum_j (1 - s_j^2) t_j^2 - 2 \sum_j (1 - s_j) t_j f_j.$$

Taking the infimum over targets with $\sum_j \lambda_j f_j^2 \le \beta^2$ subtracts the supremum of the last linear term. This is the support function of the validity ellipsoid, and Cauchy–Schwarz gives

$$\sup_{\sum_j \lambda_j f_j^2 \le \beta^2} \sum_j (1 - s_j) t_j f_j = \beta \left(\sum_j \frac{(1 - s_j)^2 t_j^2}{\lambda_j} \right)^{1/2}.$$

This proves (19).

It remains to maximize the gain. If $t_j = 0$, the coordinate does not enter the adversarial penalty because it carries no target-dependent mean. Its only contribution is $(1 - s_j^2)\nu_j$, so every maximizer sets $s_j = 0$ when $\nu_j > 0$; if $\nu_j = 0$, the coordinate is irrelevant. On the remaining mean-carrying set $J = \{j : t_j \neq 0\}$, set $q_j = 1 - s_j \in [0, 1]$, where q_j is the coordinate-wise correction coefficient, and define

$$D(q) = \sum_{j \in J} \frac{q_j^2 t_j^2}{\lambda_j}.$$

Ignoring constants independent of q, the objective on J becomes

$$F(q) = \sum_{j \in J} a_j (2q_j - q_j^2) - 2\beta \sqrt{D(q)}, \qquad a_j = t_j^2 + \nu_j > 0. \tag{28}$$

The function F is concave on $[0, 1]^J$, since the first term is a strictly concave quadratic and the second term is the negative of a norm. The threshold β_c is obtained by comparing the linear gain term with the adversarial penalty. For every q,

$$\sum_{j \in J} a_j q_j \le \left(\sum_{j \in J} \frac{\lambda_j a_j^2}{t_j^2} \right)^{1/2} \left(\sum_{j \in J} \frac{q_j^2 t_j^2}{\lambda_j} \right)^{1/2} = \beta_c \sqrt{D(q)}.$$

If $\beta \ge \beta_c$, then

$$F(q) = 2 \sum_j a_j q_j - \sum_j a_j q_j^2 - 2\beta \sqrt{D(q)} \le - \sum_j a_j q_j^2 \le 0,$$

and the value zero is attained by $q = 0$, equivalently $s_j = 1$ for all $j \in J$. Thus when the validity budget is too large, no correction of mean-carrying modes has positive minimax value. At the other

boundary, if $\beta = 0$, then (28) becomes $\sum_j a_j(2q_j - q_j^2)$, which is maximized at $q_j = 1$, equivalently $s_j = 0$, for every $j \in J$.

Assume now that $0 < \beta < \beta_c$. The point $q = 0$ is not optimal because the directional derivative along $v_j = \lambda_j a_j / t_j^2$ is positive at the origin. At any optimum with $D(q) > 0$, no coordinate can lie on the boundary of $[0, 1]$. If $q_j = 0$, then $\partial F/\partial q_j = 2a_j > 0$; if $q_j = 1$, then $\partial F/\partial q_j = -2\beta t_j^2/(\lambda_j\sqrt{D(q)}) < 0$. Hence the unique optimum is interior, and it is characterized by first-order conditions. Differentiating (28) gives

$$0 = \frac{\partial F}{\partial q_j} = 2a_j(1 - q_j) - 2\beta\frac{q_j t_j^2}{\lambda_j\sqrt{D(q)}}.$$

Since $s_j = 1 - q_j$ and $\eta = \sqrt{D(q)}/\beta > 0$, this becomes

$$a_j s_j = \frac{q_j t_j^2}{\eta\lambda_j}, \qquad s_j = \frac{t_j^2}{t_j^2 + \eta\lambda_j a_j}.$$

Substituting $q_j = 1 - s_j = \eta\lambda_j a_j/(t_j^2 + \eta\lambda_j a_j)$ into $D(q) = \eta^2\beta^2$ gives

$$\beta^2 = \sum_{j\in J} \frac{\lambda_j t_j^2 a_j^2}{(t_j^2 + \eta\lambda_j a_j)^2},$$

which is (20). The right-hand side is continuous and strictly decreasing in $\eta > 0$, with limit β_c^2 as $\eta \downarrow 0$ and limit zero as $\eta \to \infty$. Thus the scalar equation has exactly one solution whenever $0 < \beta < \beta_c$, and the corresponding interior point is the unique maximizer by concavity.

E Proof for Theorem 3 and Proposition 3

We first prove the confidence-score selection theorem. Work on the event in (22) and the event $\beta_a \le \bar{\beta}_a \le \beta_a + \varepsilon_\beta$ for all a. Since $\underline{H}_a \le H_a$ and $\overline{S}_a \ge S_a$,

$$\mathcal{L}_a = \underline{H}_a - 2\bar{\beta}_a\overline{S}_a \le H_a - 2\bar{\beta}_a S_a \le H_a - 2\beta_a S_a = \Gamma_a.$$

Thus each selected score is a valid lower bound. The same event also gives

$$\mathcal{L}_a \ge H_a - \varepsilon_H - 2\bar{\beta}_a(S_a + \varepsilon_S) = H_a - 2\bar{\beta}_a S_a - \varepsilon_H - 2\bar{\beta}_a\varepsilon_S.$$

Let $a^\star \in \arg\max_a \Gamma_a$. Since $\widehat{a}$ maximizes $\mathcal{L}_a$ and $\Gamma_{\widehat{a}} \ge \mathcal{L}_{\widehat{a}}$,

$$\Gamma_{\widehat{a}} \ge \mathcal{L}_{\widehat{a}} \ge \mathcal{L}_{a^\star} \ge \Gamma_{a^\star} - \varepsilon_H - 2\bar{\beta}_{\max}\varepsilon_S - 2(\bar{\beta}_{a^\star} - \beta_{a^\star})S_{a^\star}.$$

Using $\bar{\beta}_a - \beta_a \le \varepsilon_\beta$ and $S_a \le S_{\max}$ proves (24). For exact budgets, $\varepsilon_\beta = 0$. A union bound gives the stated probability when score and budget events are obtained separately.

For the bounded empirical instantiation, condition on the candidate construction and let $Z_1, \dots, Z_N$ be independent and identically distributed scoring observations. Assume fixed estimators $U_a(Z), V_a(Z)$ satisfy $\mathbb{E}U_a = H_a$, $\mathbb{E}V_a = D_a$, $|U_a| \le B_U$, and $0 \le V_a \le B_V$, for constants $B_U, B_V > 0$. Set

$$\widehat{H}_a = \frac{1}{N}\sum_{i=1}^N U_a(Z_i), \qquad \widehat{D}_a = \frac{1}{N}\sum_{i=1}^N V_a(Z_i).$$

Hoeffding and a union bound give $|\widehat{H}_a - H_a| \le r_H$ and $|\widehat{D}_a - D_a| \le r_D$ for all a, where

$$r_H = B_U\sqrt{\frac{2\log(4M/\delta)}{N}}, \qquad r_D = B_V\sqrt{\frac{\log(4M/\delta)}{2N}}.$$

Then $\underline{H}_a = \widehat{H}_a - r_H$ satisfies $H_a - 2r_H \le \underline{H}_a \le H_a$, and $\overline{S}_a = \sqrt{\widehat{D}_a + r_D}$ satisfies

$$S_a \le \overline{S}_a \le \sqrt{D_a + 2r_D} \le S_a + \sqrt{2r_D}.$$

Thus one may take $\varepsilon_H = 2r_H$ and $\varepsilon_S = \sqrt{2r_D}$.

For Proposition 3, sample splitting makes the selected correction fixed relative to the data used for student training. Condition first on D_{build}, so the candidate family and correction rules are fixed. By the assumed scoring guarantee on the independent sample D_{score}, with probability at least $1-\delta_{\text{score}}$ the selected candidate satisfies $\Gamma_{\widehat{a}} \geq \mathcal{L}_{\widehat{a}}$. On this event, condition further on D_{score}. The selected corrected-label mean $\bar{t}_{\widehat{a}} = h_{\widehat{a}}$ is now fixed relative to the independent student-training data D_{train}, so Proposition 1 applies conditionally and gives, with probability at least $1-\delta_{\text{train}}$,

$$\mathcal{R}(t) - \mathcal{R}(\widehat{h}_{\widehat{a}}) \geq \Gamma_{\widehat{a}} - 2M_{\widehat{a}}\Delta_{\widehat{a}} - \Delta_{\widehat{a}}^2.$$

Combining the two inequalities and taking a union bound proves (25). If the same sample is used both to select $\widehat{a}$ and to train $\widehat{h}$, this conditioning argument no longer applies; one must instead control a joint empirical-process term over pairs $(a,h) \in \mathcal{A} \times \mathcal{S}$, or include the appropriate finite-M, covering-number, or PAC-Bayes complexity penalty.

F PAC-Bayes extension for label-free selection

For a large or continuous candidate family, a finite union bound can be too crude or inapplicable. Condition on the candidate construction, and fix a prior π independently of the scoring sample. Let $Z_1, \ldots, Z_N$ be independent and identically distributed scoring observations, with $N \geq 1$, and let $U_a(Z), V_a(Z)$ be unbiased estimators of H_a, D_a. For a data-dependent posterior ρ, write $\mathbb{E}_\rho$ for expectation over $a \sim \rho$, and let $\mathrm{KL}(\rho\|\pi)$ denote Kullback–Leibler divergence, infinite when ρ is not absolutely continuous with respect to π. Define

$$\widehat{H}(\rho) = \mathbb{E}_\rho \frac{1}{N}\sum_{i=1}^N U_a(Z_i), \qquad \widehat{D}(\rho) = \mathbb{E}_\rho \frac{1}{N}\sum_{i=1}^N V_a(Z_i).$$

Assume $\beta_{\max} := \sup_{a\in\mathcal{A}} \beta_a < \infty$, and set $\Gamma(\rho) = \mathbb{E}_\rho H_a - 2\beta_{\max}\sqrt{\mathbb{E}_\rho D_a}$. The following bound replaces the finite $\log M$ term by posterior complexity.

Proposition 4 (PAC-Bayes lower bound for label-free selection)**.** *Assume $|U_a| \leq B_U$ and $0 \leq V_a \leq B_V$ for fixed constants $B_U, B_V > 0$. For $\delta \in (0,1)$, set*

$$\kappa_N(\rho,\delta) = \mathrm{KL}(\rho\|\pi) + \log(4\sqrt{N}/\delta).$$

With probability at least $1-\delta$, uniformly for all posteriors ρ with finite $\mathrm{KL}(\rho\|\pi)$,

$$\Gamma(\rho) \geq \widehat{H}(\rho) - B_U\sqrt{\frac{2\kappa_N(\rho,\delta)}{N}} - 2\beta_{\max}\sqrt{\widehat{D}(\rho) + B_V\sqrt{\frac{\kappa_N(\rho,\delta)}{2N}}}. \tag{29}$$

If the algorithm outputs the well-defined mean predictor $h_\rho = \mathbb{E}_\rho h_a$ and the validity budgets hold simultaneously for candidates in the support of ρ, then convexity of squared risk implies $\mathcal{R}(t) - \mathcal{R}(h_\rho) \geq \Gamma(\rho)$.

We apply the PAC-Bayes-kl argument to bounded observations. Let $W_a(Z) \in [0,1]$, $m_a = \mathbb{E}W_a(Z)$, and $\widehat{m}_a = N^{-1}\sum_i W_a(Z_i)$. Write $\mathrm{kl}(p\|q) = p\log(p/q) + (1-p)\log((1-p)/(1-q))$ for binary relative entropy, with the usual boundary conventions. The bounded-variable exponential-moment inequality gives

$$\mathbb{E}\exp\{N\,\mathrm{kl}(\widehat{m}_a\|m_a)\} \leq 2\sqrt{N}.$$

This is the bound of Maurer [15] for $N \geq 8$. For $1 \leq N < 8$, the same bound follows by evaluating the Bernoulli upper bound $\sum_{j=0}^N \binom{N}{j}(j/N)^j(1-j/N)^{N-j}$, with $0^0 = 1$. Integrating over π and applying Markov's inequality gives, with probability at least $1-\delta/2$,

$$\mathbb{E}_\pi \exp\{N\,\mathrm{kl}(\widehat{m}_a\|m_a)\} \leq 4\sqrt{N}/\delta.$$

On this event, the change-of-measure inequality and joint convexity of binary relative entropy yield, simultaneously for all posteriors with finite divergence,

$$N\,\mathrm{kl}(\mathbb{E}_\rho\widehat{m}_a\|\mathbb{E}_\rho m_a) \leq N\mathbb{E}_\rho\,\mathrm{kl}(\widehat{m}_a\|m_a) \leq \kappa_N(\rho,\delta).$$

Pinsker's inequality, $\mathrm{kl}(p\|q) \geq 2(p-q)^2$, therefore gives

$$\left|\mathbb{E}_\rho\widehat{m}_a - \mathbb{E}_\rho m_a\right| \leq \sqrt{\frac{\kappa_N(\rho,\delta)}{2N}}.$$

Apply this argument first to $W_a = (U_a + B_U)/(2B_U)$ and then to $W_a = V_a/B_V$. A union bound gives, on a common event of probability at least $1-\delta$ and for every posterior,

$$\mathbb{E}_\rho H_a \geq \widehat{H}(\rho) - B_U\sqrt{\frac{2\kappa_N(\rho,\delta)}{N}}, \qquad \mathbb{E}_\rho D_a \leq \widehat{D}(\rho) + B_V\sqrt{\frac{\kappa_N(\rho,\delta)}{2N}}.$$

Substituting these inequalities into the definition of $\Gamma(\rho)$ proves (29).

It remains to justify the posterior-averaged predictor statement. Assume each candidate correction h_a has a simultaneous validity lower bound

$$\mathcal{R}(t) - \mathcal{R}(h_a) \geq H_a - 2\beta_a\sqrt{D_a}.$$

Let $h_\rho = \mathbb{E}_\rho h_a$. By convexity of squared risk, averaging predictors cannot increase risk beyond the posterior average of their risks:

$$\mathcal{R}(h_\rho) = \|\mathbb{E}_\rho h_a - f^\star\|^2 \leq \mathbb{E}_\rho \|h_a - f^\star\|^2 = \mathbb{E}_\rho \mathcal{R}(h_a).$$

Therefore

$$\mathcal{R}(t) - \mathcal{R}(h_\rho) \geq \mathbb{E}_\rho[\mathcal{R}(t) - \mathcal{R}(h_a)] \geq \mathbb{E}_\rho H_a - 2\mathbb{E}_\rho[\beta_a\sqrt{D_a}].$$

With a uniform validity budget, Jensen's inequality gives $\mathbb{E}_\rho[\beta_a\sqrt{D_a}] \leq \beta_{\max}\sqrt{\mathbb{E}_\rho D_a}$, so $\mathcal{R}(t) - \mathcal{R}(h_\rho) \geq \Gamma(\rho)$.

G Preference-cycle lower bound

Let $G = (V, E, \omega)$ be a weighted comparison graph and let $r \in \mathbb{R}^E$ be an antisymmetric logit flow, with weighted inner product $\langle a, b\rangle_\omega = \sum_{e\in E}\omega_e a_e b_e$. Let $d : \mathbb{R}^V \to \mathbb{R}^E$ be the graph-gradient map, $(du)_{ij} = u_i - u_j$, let $\mathcal{G} = \mathrm{range}(d)$ be the subspace of utility-consistent flows, let P_∇ be the orthogonal projection onto $\mathcal{G}$, and let $Q_\nabla = I - P_\nabla$ be the cyclic projection. If the target flow is utility-based, $r^\star \in \mathcal{G}$, then

$$\|r - r^\star\|_\omega^2 - \|P_\nabla r - r^\star\|_\omega^2 = \|Q_\nabla r\|_\omega^2 . \tag{30}$$

If the target is only approximately utility-based, $\|Q_\nabla r^\star\|_\omega \leq \delta$, the same calculation gives

$$\|r - r^\star\|_\omega^2 - \|P_\nabla r - r^\star\|_\omega^2 \geq \|Q_\nabla r\|_\omega^2 - 2\delta\|Q_\nabla r\|_\omega .$$

Thus weak-judge cycle energy yields a label-free lower bound only when target cyclic energy is independently bounded.

The exact-validity identity follows directly from orthogonal projection. Because P_∇ is the orthogonal projection onto $\mathcal{G}$ and $Q_\nabla = I - P_\nabla$, one has

$$r - r^\star = (P_\nabla r - r^\star) + Q_\nabla r$$

whenever $r^\star \in \mathcal{G}$, with the two terms orthogonal. Pythagoras gives

$$\|r - r^\star\|_\omega^2 = \|P_\nabla r - r^\star\|_\omega^2 + \|Q_\nabla r\|_\omega^2 ,$$

which proves (30). If $\|Q_\nabla r^\star\|_\omega \leq \delta$, then

$$r - r^\star = (P_\nabla r - P_\nabla r^\star) + (Q_\nabla r - Q_\nabla r^\star),$$

and projection removes $Q_\nabla r$ but not $Q_\nabla r^\star$. Expanding squared risks gives

$$\|r - r^\star\|_\omega^2 - \|P_\nabla r - r^\star\|_\omega^2 = \|Q_\nabla r\|_\omega^2 - 2\langle Q_\nabla r, Q_\nabla r^\star\rangle_\omega \geq \|Q_\nabla r\|_\omega^2 - 2\delta\|Q_\nabla r\|_\omega .$$

For logistic transfer, assume all logits under comparison lie in $[-M, M]$. Let $\sigma(z) = (1+e^{-z})^{-1}$, let $p_e = \sigma(r_e^\star)$, define $\ell_e(a_e) = -p_e\log\sigma(a_e) - (1-p_e)\log(1-\sigma(a_e))$, and set $\mathcal{L}(a) = \sum_{e\in E}\omega_e\ell_e(a_e)$. Let $m_M = \min_{|z|\leq M}\sigma(z)(1-\sigma(z))$. Since $\ell_e''(z) = \sigma(z)(1-\sigma(z))$, ℓ_e is m_M-strongly convex and $1/4$-smooth on $[-M, M]$. Applying strong convexity around $r^\star$ to r and smoothness around $r^\star$ to $P_\nabla r$, summing with weights, and using the approximate-cycle inequality gives

$$\mathcal{L}(r) - \mathcal{L}(P_\nabla r) \geq \frac{m_M}{2}\left(\|Q_\nabla r\|_\omega^2 - 2\delta\|Q_\nabla r\|_\omega\right) - \left(\frac{1}{8} - \frac{m_M}{2}\right)\|P_\nabla r - P_\nabla r^\star\|_\omega^2 - \frac{1}{8}\delta^2.$$

This is a sufficient condition, not a necessary one; it is conservative because DPO-style objectives are nonlinear and projection identities for squared risk do not automatically imply logistic-risk improvement.

H Additional experimental details

H.1 Controlled graph-signal regression

The synthetic experiment uses the normalized Laplacian of a paired-copy cycle graph with $m = 1024$ nodes and the first $k = 128$ nonconstant eigenvectors. The experiment uses twenty seeds starting at 20260504, validity budgets $\beta \in \{0, 0.05, 0.1, 0.2, 0.4, 0.8\}$, and weak-label noise $\sigma = 0.1$. The candidate family is fixed in advance: raw labels, Markov smoothers with depths $1, 2, 4$, and 8, hard projection, the minimax resolvent smoother, weak-label cross-validated graph smoothing, and an invalid pair-averaging control. Finite-sample selection uses label-free budgets $N \in \{16, 32, 64, 128, 256, 512\}$ with 400 repetitions per graph instance, comparing unadjusted-inconsistency selection with VAC selection. Ground truth is used only to report true risk reduction relative to the raw teacher; the complete population and selection results are reported in Table 1. Two additional ablations evaluate the minimax interpretation: when the teacher is already consistent, the resolvent gain is 0.000 ± 0.000 at $\beta = 0.2$; when β is misspecified as zero, excessive smoothing yields gain -11.579 ± 2.774, whereas the correctly specified β yields 0.449 ± 0.056 and a conservative overestimate yields 0.000 ± 0.000.

H.2 CIFAR-10 protocols

The neural experiment uses the cats-versus-dogs subset of CIFAR-10 [12]. For each of the ten seeds 20260504 through 20260513, the split contains 500 labeled teacher-training images per class, 100 reserved validation images per class, and 8,800 weak-labeled pool images. The main split-sample run uses 2,200 pool examples for label-free scoring and the remaining 6,600 examples for student training. The reported main run uses only predeclared validity budgets for selection. The weak teacher is a four-layer ConvNet trained for 100 epochs, and the student is a CIFAR-adapted ResNet-18 trained for 200 epochs by KL distillation at temperature 2.0. Both models use SGD with momentum 0.9, weight decay 5×10^{-4}, cosine learning-rate decay, and initial learning rates 0.05 for the teacher and 0.1 for the student.

The CIFAR candidate families are horizontal flip, random crop, color jitter, their combined transformation, cutout, strong random resized crop, class swap, and random image replacement. The default number of transformed views is four; horizontal flip and the invalid controls use one view, and the combined transformation uses eight views. The probability-space validity budgets are $0.02, 0.03, 0.03, 0.06, 0.18, 0.25, 0.70$, and 0.70 in the same order. For the predeclared-budget split-sample run, VAC selects the combined transformation in 9 of 10 seeds and random crop in 1 of 10 seeds; the graph-resolvent variant selects crop in all 10 seeds.

For the matched comparison in Table 2, all four methods share the same teacher checkpoint, student initialization, data order, student-side augmentation, and 6,600 training examples within each seed. Fixed TTA averages eight transformed views without the original view. Fixed shrinkage uses the same average with $\beta = 0.06$, while full VAC selects on the independent scoring split. Student batch size is 128 and evaluation batch size is 256. The paired confidence intervals use 10,000 percentile-bootstrap resamples of the ten seed-level differences. The supplementary results below retain their original-view-inclusive averaging and augmentation settings.

H.3 ARC language-model experiment

The ARC experiment combines the official ARC-Easy and ARC-Challenge training partitions: 3,370 training questions (2,251 Easy and 1,119 Challenge), 869 validation questions (570 and 299), and 3,548 test questions (2,376 and 1,172). A structural audit of question and option text found no unresolved position-dependent references; all 7,787 examples were retained. Qwen3-1.7B supplies fixed soft labels, and Qwen3-4B is trained using low-rank adaptation (LoRA). Both models use the official chat template with thinking disabled. Answer probabilities are normalized next-token logits over the displayed option labels, mapped back to canonical answer identities before averaging. The frozen view set uses all six permutations for three-option questions and eight unique permutations, including the identity, for four- and five-option questions, with permutation seed 20260728.

Each of the four methods (raw, permutation average, single permutation, and invalid replacement average) uses five paired student-training seeds, 20260728–20260732, and cached teacher targets. Students train for three epochs with AdamW, learning rate 2×10^{-4}, weight decay 0.01, batch size two, 16 gradient-accumulation steps, and maximum sequence length 512. LoRA uses rank 16, scaling

parameter 32, dropout 0.05, and the attention and feed-forward projection layers. The learning rate warms up linearly over 5% of updates and then decays linearly. The reported student confidence intervals are paired t-intervals across the five seeds, using sample standard deviations; they reflect student-training randomness conditional on the fixed teacher and targets. Dataset and model revisions, prompts, permutation manifests, and per-run settings are included in the code supplement.

H.4 Supplementary CIFAR comparisons

Table 4: Supplementary predeclared-budget split-sample CIFAR-10 results with the original view included in the local VAC average. Candidate selection uses 2,200 scoring examples and student training uses a disjoint set of 6,600 examples. Table 2 reports the matched reruns averaging transformed views only.

Method	Aug.	Test acc.	W2S gain	Err. red.	Brier ↓	NLL ↓
Weak teacher	–	0.7275 ± 0.0109	–	–	0.4190	0.8002
Invalid class swap	no	0.7275 ± 0.0117	0.0001 ± 0.0117	-0.0004	0.4265	0.8282
Invalid random image	no	0.7375 ± 0.0107	0.0101 ± 0.0085	0.0366	0.4187	0.8248
Raw W2S	no	0.7337 ± 0.0103	0.0063 ± 0.0077	0.0227	0.4205	0.8289
VAC	no	0.7387 ± 0.0151	0.0113 ± 0.0107	0.0415	0.3848	**0.6599**
Raw W2S + aug.	yes	0.7465 ± 0.0129	0.0191 ± 0.0081	0.0701	0.3989	0.8581
VAC + aug.	yes	$\mathbf{0.7559 \pm 0.0113}$	$\mathbf{0.0285 \pm 0.0047}$	**0.1044**	**0.3679**	0.6929

Table 5: Graph-resolvent comparison under the supplementary split-sample protocol of Table 4. The local VAC baseline includes the original view in its average; the graph-resolvent variant has lower W2S gain in this comparison.

Method	Test acc.	W2S gain	Err. red.	Brier ↓	NLL ↓
Raw W2S + aug.	0.7465 ± 0.0129	0.0191 ± 0.0081	0.0701	0.3989	0.8581
Graph-resolvent VAC + aug.	0.7488 ± 0.0168	0.0213 ± 0.0080	0.0790	0.3931	0.8231
VAC + aug.	$\mathbf{0.7559 \pm 0.0113}$	$\mathbf{0.0285 \pm 0.0047}$	**0.1044**	**0.3679**	**0.6929**

Table 6: Validation-budget split-sample ablation based on upper confidence bounds. The validation-based upper confidence budgets are conservative with only 100 validation images per class. The resulting selected correction is effectively the identity correction, so its performance is close to raw W2S with augmentation rather than replacing the predeclared-budget main result.

Method	Test acc.	W2S gain	Err. red.	Brier ↓	NLL ↓
Raw W2S + aug.	0.7497 ± 0.0149	0.0202 ± 0.0102	0.0747	0.3987	0.8755
Validation-budget VAC + aug.	$\mathbf{0.7501 \pm 0.0148}$	$\mathbf{0.0206 \pm 0.0073}$	**0.0763**	**0.3983**	0.8673
Invalid class swap	0.7322 ± 0.0115	0.0027 ± 0.0079	0.0092	0.4192	**0.8337**
Invalid random image	0.7287 ± 0.0181	-0.0008 ± 0.0109	-0.0029	0.4310	0.8702

Validation-budget outcome. The reserved validation split is used only in the validation-budget ablation based on upper confidence bounds, not in the main predeclared-budget run. In the code-proxy validation, the label-preserving transformations have zero failures out of 200 validation trials, but the one-sided upper confidence bound still gives a conservative failure-rate upper bound of about 0.036, corresponding to a probability-space budget about 0.268. Class swap fails all 200 validation trials, giving budget 1.414; random image replacement fails on average 102.2 of 200 trials, giving budget about 1.100. These upper confidence budgets make all non-identity correction lower-bound values nonpositive, so the validation-based rule selects an effectively identity correction. This outcome follows from using a small validation set and is reported as a robustness analysis; the resulting upper bounds are insufficiently tight to replace the predeclared budgets in the main table.

Table 7: Non-split full-pool CIFAR run from the previous main table. This run uses all 8,800 weak-labeled pool examples for both scoring and training and is therefore reported as an appendix ablation rather than the main split-sample result.

Method	Aug.	Test acc.	W2S gain	Err. red.	Brier ↓	NLL ↓
Weak teacher	–	0.7274 ± 0.0112	–	–	0.4191	0.8058
Invalid class swap	no	0.7370 ± 0.0122	0.0097 ± 0.0069	0.0355	0.4238	0.8613
Invalid random image	no	0.7336 ± 0.0127	0.0062 ± 0.0106	0.0223	0.4294	0.8837
Raw W2S	no	0.7401 ± 0.0133	0.0127 ± 0.0070	0.0467	0.4229	0.8721
Raw W2S + aug.	yes	0.7497 ± 0.0138	0.0223 ± 0.0069	0.0822	0.3927	0.8532
VAC	no	0.7466 ± 0.0123	0.0192 ± 0.0054	0.0707	0.3789	**0.6571**
VAC + aug.	yes	$\mathbf{0.7584 \pm 0.0128}$	$\mathbf{0.0311 \pm 0.0051}$	**0.1141**	**0.3651**	0.6903

I Additional implementation notes

For each CIFAR candidate, VAC constructs a corrected soft label by moving the original teacher probability $q_0(x)$ toward the average transformed-view probability $\bar{q}(x)$ with a budget-dependent shrinkage weight. The implementation uses $r(x) = \|q_0(x) - \bar{q}(x)\|_2$ and $\alpha(x) = \max\{0, 1 - \beta/(r(x) + 10^{-8})\}$, clipped at one, so the corrected label is $q_{\text{corr}}(x) = (1 - \alpha(x))q_0(x) + \alpha(x)\bar{q}(x)$. Selection is performed before student training with $\mathcal{L} = H - 2\beta\sqrt{D} - 1.64\sqrt{D/N}$, where $H = \mathbb{E}[\|q_0\|_2^2 - \|q_{\text{corr}}\|_2^2]$, $D = \mathbb{E}[\|q_0 - q_{\text{corr}}\|_2^2]$, and N is the scoring-split size for the main split-sample run. Student-side augmentation is controlled separately: all four methods in Table 2 use random crop and horizontal flip during distillation; in the supplementary comparisons (Tables 4–7), augmentation is indicated by the "+ aug." suffix.

The scripts write the outputs required for reproducibility to disk: synthetic oracle, student, selection, and ablation CSV files with JSON summaries, and CIFAR per-seed candidate statistics, per-method student results, metadata, and aggregate summaries. These files record the selected candidate, validity budgets, number of views, teacher and student metrics, training-example counts, architecture, epochs, device, runtime, split protocol, scoring-split size, and training-split size. Runs that change the candidate list, validity budgets, split seeds, teacher temperature, distillation temperature, student augmentation flag, or split protocol should therefore be treated as separate experimental configurations.

Supplementary code and recorded results. The accompanying `w2s_label_free_code_release.zip` contains the synthetic, CIFAR, and ARC runners, a table-to-command guide, tests, and recorded CIFAR/ARC metrics and configurations. Its README separates the matched CIFAR comparison from the supplementary protocols; `docs/REPRODUCIBILITY.md` gives the ARC training commands and fixed revisions. The scripts write new runs to separate output directories, and `scripts/verify_paper_results.py` checks the main CIFAR and ARC tables from the included seed-level metrics without model training. Third-party datasets and pretrained weights are obtained from their original sources.

Computational resources. Experiments ran primarily on the Nexus cluster. The matched CIFAR runs used one NVIDIA GeForce RTX 2080 Ti (11 GB GPU memory), four CPU cores, and 32 GB host memory per seed; ARC used one NVIDIA RTX A5000 (24 GB GPU memory), eight CPU cores, and 32 GB host memory per run. CIFAR student training and evaluation took 9.7–48.8 minutes per method/seed (40 runs, 7.74 hours in total); ARC training, adapter saving, and evaluation took 38.3–47.7 minutes per method/seed (20 runs, 13.74 hours). These program timings exclude teacher preparation and, for ARC, initial model loading. Recovered project allocations total 46.54 GPU-hours, including preliminary and unsuccessful jobs. The supplement's `docs/COMPUTE.md` gives allocation-time totals, timing coverage, software versions, and hardware records.

J Additional related work

Consistency regularization and augmentation invariance. Consistency has long been used as a regularizer in semi-supervised learning. Temporal ensembling and Mean Teacher encourage predictions to remain stable across stochastic perturbations or training trajectories [13, 24]; virtual

adversarial training enforces local smoothness around inputs [17]; MixMatch and FixMatch combine augmentation consistency with pseudo-labeling and confidence thresholding [2, 23]. Group-theoretic data augmentation similarly views valid transformations as label-preserving orbits [5]. The present analysis differs from consistency regularization: under exact validity, weak-teacher trajectory inconsistency equals a squared-risk reduction from smoothing, whereas under approximate validity the same quantity is reduced by a validity-budget penalty. Thus consistency defines a label-free lower bound and supplies the input to a minimax correction problem.

Graph smoothing, inverse problems, and preference consistency. Graph-based semi-supervised learning and manifold regularization use Laplacian smoothness to propagate labels over data graphs [1, 28, 29]. Classical inverse-problem theory studies spectral filters and Tikhonov regularization [8]. These tools are mathematically close to the Markov and resolvent operators studied here, but their usual role is to regularize supervised or semi-supervised estimators rather than to lower-bound weak-label error from label-free queries. In preference learning, HodgeRank decomposes pairwise comparisons into globally ranked and cyclic components [11], while RLHF and DPO use pairwise preferences to align language models [19, 21]. This decomposition provides the orthogonal structure needed to analyze cyclic weak-preference components under an explicit approximate-utility validity assumption.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes].

   Justification: The abstract and introduction state the paper as a sufficient label-free correction mechanism for W2S, define validity-adjusted correction (VAC), and limit the claim to independently specified semantic constraints rather than a universal explanation of W2S. The theoretical claims are supported by the Markov lower-bound, minimax correction, finite-query selection, and split-sample transfer results, and the experiments are presented as mechanism checks.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes].

   Justification: The manuscript limits its claims to independently specified validity constraints, explains why invalid or weakly validated constraints do not provide lower-bound evidence, uses invalid controls, reports the non-split full-pool CIFAR run only as an appendix ablation because it lacks the split-sample independence used by the theorem, and reports the validation-budget ablation as conservative with the available validation size.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best

judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes].

   Justification: The assumptions are stated in Section 3 and the theorem or proposition statements specify the relevant conditions. Each main result points to its proof, and the appendix gives complete proofs for the Markov lower bound, student-transfer identity, minimax correction, stochastic-teacher correction, finite-query selection, split-sample transfer, PAC-Bayes extension, and preference-cycle result.

   Guidelines:

   - The answer [N/A] means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes].

   Justification: Appendix H and Appendix I specify the controlled graph construction, seeds, validity budgets, candidate corrections, selection protocol, CIFAR splits, model architectures, training hyperparameters, candidate transformations, number of transformed views, scoring rule, and recorded outputs. Appendix H.3 gives the ARC splits, view construction, training settings, and uncertainty estimates. The accompanying code artifact provides executable scripts and README instructions for the synthetic, CIFAR, and ARC experiments.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example

(a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
(b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
(c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
(d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes].

   Justification: A camera-ready code artifact is included. It contains the controlled graph-signal, CIFAR, and ARC runners, requirements, tests, local reproduction commands, archived CIFAR/ARC metrics, and a README mapping paper tables to commands and result files. Datasets and pretrained weights are obtained from their original sources rather than redistributed.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes].

   Justification: The main text and appendices report the synthetic graph construction, seeds, validity budgets, weak-label noise, candidate family, finite-sample selection sizes, CIFAR train/validation/scoring/student-training splits, weak-teacher and student architectures, epochs, optimizers, learning rates, distillation temperature, candidate transformations, transformed-view counts, and invalid controls. Appendix H.3 also specifies the ARC split sizes, model settings, permutations, LoRA training, and confidence intervals. The released README provides corresponding commands and environment setup.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes].

   Justification: The controlled graph-signal, CIFAR, and ARC tables report mean $\pm$ standard deviation where applicable. The matched CIFAR and ARC comparisons report paired 95% confidence intervals, and the finite-sample selection table reports selection frequencies across repeated trials.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.
   - It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
   - For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
   - If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes].

   Justification: The appendix and supplementary compute documentation report neural-experiment hardware, memory, per-run timings, and recovered GPU allocation totals, including preliminary and unsuccessful jobs.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics https://neurips.cc/public/EthicsGuidelines?

   Answer: [Yes].

   Justification: The work is theoretical and empirical on synthetic data and the public CIFAR-10 and ARC benchmarks. It does not collect human-subject data, release high-risk datasets or models, or deploy a system affecting individuals, and it explicitly warns that invalid constraints do not provide label-free lower-bound evidence.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [N/A].

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
    - If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A].

    Justification: The released assets are experimental code, numerical records, and small CIFAR classifiers. They do not include general-purpose language models, image generators, scraped datasets, or other high-risk assets.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.

- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes].

    Justification: CIFAR-10, ARC, and Qwen3 are cited. The supplement's `docs/ASSETS.md` identifies their original sources and terms: ARC is CC BY-SA 4.0 and the Qwen3 checkpoints are Apache 2.0; the CIFAR-10 source page does not specify a standard license. Third-party datasets and pretrained weights are not redistributed. The authors' code and CIFAR teacher checkpoints are released under the MIT license.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
    - If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes].

    Justification: The experimental artifact includes README documentation, requirements, local reproduction commands, tests, archived metrics, and a license. The ten frozen CIFAR teacher checkpoints are documented and verified against the hashes in the original experiment records. The work does not introduce a new dataset or benchmark.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [N/A].

Justification: The work does not include crowdsourcing or human-subject experiments. No participant instructions, screenshots, or compensation information are applicable.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [N/A].

Justification: The paper does not involve human-subject research or crowdsourcing. IRB approval or equivalent review is therefore not applicable.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

Answer: [Yes].

Justification: Section 5.3 and Appendix H.3 describe Qwen3-1.7B as the weak teacher and Qwen3-4B as the student in the ARC experiments, including option-permutation correction and student training. The accompanying code artifact records the model revisions, prompts, and training settings.

Guidelines:

- The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.