NOT ALL TEACHERS ARE CREATED EQUAL: CONFIDENCE-AWARE PREDICTION-POWERED LEARNING FROM HETEROGENEOUS TEACHERS

Anonymous authors
Paper under double-blind review

ABSTRACT

Semi-supervised learning often uses teacher predictions to exploit unlabeled data, but unreliable predictions can introduce harmful supervision. This problem is especially acute when teachers are reliable only on particular inputs, classes, or domains. Prediction-powered inference (PPI) is a general statistical framework that combines predictions with labeled-data corrections to construct unbiased estimators. Building on this principle, we introduce confidence-aware prediction-powered correction (C-PP), which adapts to instance-dependent teacher reliability, and covariance-aware multi-teacher correction (CM-PP), which combines complementary teachers while accounting for their dependence. We establish finite-sample variance bounds and convergence guarantees, and develop online weight updates that avoid explicit variance estimation. Across image and tabular classification, and semantic segmentation, C-PP and CM-PP improve upon confidence-agnostic prediction-powered learning and established semi-supervised baselines under teacher-reliability shifts, partial coverage, and heterogeneous specialization.

1 INTRODUCTION

Modern learning systems increasingly use auxiliary predictors to exploit abundant unlabeled data (Chapelle et al., 2009; Van Engelen & Hoos, 2020), yet these predictors are rarely uniformly reliable: a pretrained model may transfer well only to certain classes or domains, a specialist may cover only part of the label space, and multiple teachers may provide complementary but correlated information. A key question is, not simply whether to use pseudo-labels, but *where to trust each teacher and how to combine several imperfect teachers*. Uniformly using their predictions can reinforce errors through confirmation bias (Lee et al., 2013; Sohn et al., 2020; Xie et al., 2020b; Arazo et al., 2019), while discarding partially reliable teachers wastes useful information where labels are scarce.

Existing SSL methods mitigate unreliable pseudo-labels through consistency regularization, teacher-student averaging, and confidence-based selection or weighting (Tarvainen & Valpola, 2017; Berthelot et al., 2019; Sohn et al., 2020). Although effective, these methods generally treat the resulting pseudo-labels as training targets, potentially introducing bias and providing limited finite-sample control over teacher error (Singh et al., 2008; Wei et al., 2020). Prediction-powered inference (PPI) instead treats predictions as auxiliary random variables and corrects their contribution using labeled data (Angelopoulos et al., 2023a). Its learning extensions construct unbiased stochastic gradients that interpolate between labeled-only and prediction-powered (PP) objectives (Angelopoulos et al., 2023b; Zhu et al., 2023). However, many formulations rely on a single teacher and a global interpolation parameter, which cannot express instance-dependent expertise or dependence among multiple teachers. Under teacher heterogeneity, increasing this weight can amplify correction variance on mismatched examples, while decreasing it retreats toward supervised learning and forfeits useful predictions.

Existing finite-sample analyses do not jointly characterize instance-dependent reliability, complementary expertise, and cross-teacher correlations in a non-asymptotic variance framework (Mani et al., 2025; Zrnic, 2024). This leaves open when confidence-aware correction improves over confidence-agnostic prediction-powered learning, when heterogeneous teachers provide complementary gains, and how interpolation weights should adapt when variance quantities are unknown and vary during optimization (Angelopoulos et al., 2023b; Zhu et al., 2023).

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

We introduce a PP framework for heterogeneous teachers. C-PP incorporates instance-dependent confidence, adapting to local teacher quality while preserving unbiasedness. CM-PP combines teachers using covariance weighting, exploiting complementary information without overweighting redundant or correlated information. We derive non-asymptotic conditions under which C-PP improves over confidence-agnostic PP learning. We link these reductions to stochastic-optimization convergence and develop online procedures that adapt weights without explicit variance estimation.

Our contributions.

- **Confidence-aware learning from partial and heterogeneous teachers.** We introduce C-PP and CM-PP, which preserve unbiasedness while exploiting locally reliable and complementary partial-coverage teachers. We establish when confidence-aware correction reduces variance and derive covariance-aware aggregation, recovering inverse-variance weighting as a diagonal special case.
- **Flexible confidence modeling.** Our framework supports binary or continuous, shared or teacher-specific confidence signals without requiring cross-teacher calibration. Confidence may encode coverage, domain, task, modality, or preference-related applicability, allowing the same correction principle to accommodate different notions of teacher reliability.
- **Finite-sample guarantees and adaptive optimization.** We provide non-asymptotic variance and convergence guarantees with online aggregation-weight adaptation, without explicit variance estimation.

Together, these contributions enable principled use of partial and specialized teachers, including off-the-shelf models for related tasks, even when no teacher is uniformly reliable.

2 Related work

Semi-supervised learning and pseudo-labeling. Semi-supervised learning (SSL) leverages unlabeled data to improve prediction (Chapelle et al., 2009). Pseudo-labeling is a prevalent SSL approach that augments training with predictions from an auxiliary model (Lee et al., 2013), generated through self-training (Yarowsky, 1995) or teacher-student schemes (Tarvainen & Valpola, 2017). Its central challenge is confirmation bias, where erroneous pseudo-labels are reinforced during training (Arazo et al., 2020). Many SSL methods mitigate this. However, they treat selected or weighted pseudo-labels as training targets rather than incorporating them into a labeled-data-corrected, unbiased estimator.

Prediction-powered inference and learning. PPI combines predictions with labeled-data corrections to obtain unbiased, variance-reduced estimators (Angelopoulos et al., 2023a). Extensions improve weight tuning (Angelopoulos et al., 2023b), connect to debiased estimation through cross-fitting and data reuse (Zrnic & Candès, 2024; Chernozhukov et al., 2018), incorporate prior information while preserving frequentist validity (Cortinovis & Caron, 2025), and combine multiple predictors under cost constraints (Cowen-Breen et al., 2026). For learning, Zhu et al. (2023) use scheduled weights and provide asymptotic variance guarantees, Ruah et al. (2025) improve robustness to teacher errors using fixed, tuned weights, and Shoham et al. (2026) adapt a global weight online; further work also studies self-training with imperfect teachers (Xie et al., 2020b; Hendrycks et al., 2019). Existing PP-SSL formulations typically use a single teacher and global weight, and do not address instance-dependent reliability or aggregation of heterogeneous predictors.

Confidence-aware and multi-teacher SSL. Confidence-aware SSL selects or reweights pseudo-labels using uncertainty signals (Sohn et al., 2020; Grandvalet & Bengio, 2004; Chen et al., 2023), but still uses them as training targets. Ensemble and multi-teacher methods exploit predictor agreement or knowledge distillation (Blum & Mitchell, 1998; Hinton et al., 2015; Liu & Tan, 2021), but generally do not jointly model instance-dependent reliability and dependence among teacher errors (Qiao et al., 2018). Our framework addresses both limitations by combining instance-dependent confidence within unbiased PP corrections, with covariance-aware aggregation across heterogeneous teachers.

3 Problem Setup

We consider an iterative first-order optimization problem, where in each round t, we have access to K prediction models (teachers) $\{f_j : \mathcal{X} \to \mathcal{Y}\}_{j=1}^{K}$, n labeled data points $(x^i, y^i)_{i=1}^{n}$, sampled i.i.d

from some unknown distribution denoted by $P = P_X \times P_{Y|X}$, and N unlabeled data points $(\tilde{x}^i)_{i=1}^N$ sampled i.i.d from the same distribution, denoted by P_X. We focus on the regime where $N > n$.

We train a student predictor $f_w : \mathcal{X} \to \mathcal{Y}$, parameterized by $w \in \mathbb{R}^d$, in a teacher-student setting. Our objective is to minimize the expected loss $\min_{w\in\mathbb{R}^d} \mathcal{L}(w) := \mathbb{E}_{(x,y)\sim P}[\ell(w; x, y)]$ where $\ell(w; x, y)$ is the loss function for a given data point and P is the underlying distribution over labeled data.

3.1 Teacher and Confidence Signals

Prediction confidence. Each teacher f_j is paired with a confidence model that assigns a score $c_j \in [0, 1]$ to each prediction. The model may use the input sample, teacher parameters, and internal forward-pass features. It may also use external signals. We denote the resulting effective score by $c_j(x)$, suppressing these additional dependencies. We focus on binary confidence, $c \in \{0, 1\}$ for a transparent conditional-error interpretation; the continuous case is analogous.

Definition 3.1. For each confidence model c_j, let $\tau_j = \mathbb{E}_{X\sim P_X}[c_j(x)]$ denote its expected confidence.

At iteration t, teacher and confidence functions are fixed given the prior optimization history before the current batches are sampled; label-dependent updates take effect only on subsequent iterations, preserving conditional unbiasedness. Our guarantees remain valid for any confidence signal; informativeness is required only to obtain variance improvements. In practice, confidence may be shared across teachers or estimated separately; Section 8 discusses constructions and aggregation. Severe miscalibration may nevertheless diminish the finite-sample gains.

Teacher quality. We measure teacher quality by its prediction error and, given confidence, by its conditional prediction error:

$$\varepsilon^{f_j} = \mathbb{E}_{(x,y)\sim P}[\|y - f_j(x)\|^2], \quad \varepsilon^{f_j,c} = \mathbb{E}_{(x,y)\sim P}[\|y - f_j(x)\|^2 | c_j(x) = c] \tag{1}$$

3.2 Assumptions

Loss properties. We first state the loss conditions used in our variance and convergence analysis.

Definition 3.2 (Natural loss)**.** A differentiable loss ℓ is *Natural* if for all $w \in \mathbb{R}^d$, $x \in \mathcal{X}$, $\nabla_w \ell(w; x, y)|_{y=f_w(x)} = 0$.

Naturality holds for many standard losses; see Appendix B.

Assumption 3.3 (β-smoothness)**.** The objective $\mathcal{L} : \mathbb{R}^d \to \mathbb{R}$ is β-smooth, i.e., for all $u, v \in \mathbb{R}^d$,

$$\|\nabla\mathcal{L}(u) - \nabla\mathcal{L}(v)\| \le \beta\|u - v\|.$$

Assumption 3.4 (Label-Lipschitz gradient)**.** The gradient of the loss is β_Y-Lipschitz with respect to the label argument, i.e., for all $w \in \mathbb{R}^d$, $\forall x$, $\forall y_1, y_2 : \|\nabla\ell(w; x, y_1) - \nabla\ell(w; x, y_2)\| \le \beta_Y\|y_1 - y_2\|$.

This condition holds for many losses, including squared and logistic losses; see Appendix A.1.

Variance and covariance assumptions.

Assumption 3.5 (Bounded stochastic-gradient variance)**.** There exists $\sigma^2 < \infty$ such that

$$\mathrm{Var}(\nabla\ell(w; x, y)) = \mathbb{E}\big[\|\nabla\ell(w; x, y) - \nabla\mathcal{L}(w)\|^2\big] \le \sigma^2. \tag{2}$$

For each teacher f_j, we define the instance-dependent loss error

$$\ell_e^{f_j}(w; x, y) := \ell(w; x, y) - \ell(w; x, f_j(x)), \qquad \nabla\mathcal{L}_e^{f_j}(w) := \mathbb{E}_{(x,y)\sim P}[\nabla\ell_e^{f_j}(w; x, y)]$$

Assumption 3.6 (Bounded error variance)**.** For each teacher j, there exists $\sigma_{e,j}^2 < \infty$ such that

$$\mathrm{Var}\big(\nabla\ell_e^{f_j}(w; x, y)\big) = \mathbb{E}\big[\|\nabla\ell_e^{f_j}(w; x, y) - \nabla\mathcal{L}_e^{f_j}(w)\|^2\big] \le \sigma_{e,j}^2. \tag{3}$$

With a confidence model, we use the following conditional bound.

Assumption 3.7 (Bounded high-confidence error variance)**.** For each teacher j, there exists $\sigma_{e,j,1}^2 < \infty$ such that

$$\mathrm{Var}\big(\nabla\ell_e^{f_j}(w; x, y) \,\big|\, c_j(x) = 1\big) = \mathbb{E}_{(x,y)\sim P}\big[\|\nabla\ell_e^{f_j}(w; x, y) - \nabla\mathcal{L}_e^{f_j,1}(w)\|^2 \mid c_j(x) = 1\big] \le \sigma_{e,j,1}^2. \tag{4}$$

We impose no corresponding assumption on low-confidence teacher errors. We use $\sigma^2_{j,0}$ and $\sigma^2_{j,1}$ for the conditional counterparts of Assumption 3.5 under $P(\cdot \mid c_j(x) = c)$, for $c \in \{0, 1\}$. In the single-teacher setting, we omit the index j.

Assumption 3.8 (Bounded confidence-weighted error covariance)**.** There exists a symmetric positive semi-definite matrix $\bar{\Sigma}_{\mathrm{CM}} \in \mathbb{R}^{K\times K}$ such that for all $w \in \mathbb{R}^d$, $\Sigma_{\mathrm{CM}}(w) \preceq \bar{\Sigma}_{\mathrm{CM}}$, where $[\Sigma_{\mathrm{CM}}(w)]_{jk} = \mathrm{Cov}(\nabla\ell^{f_j}_{e,c}, \nabla\ell^{f_k}_{e,c}) = \mathbb{E}\Big[\Big\langle \nabla\ell^{f_j}_{e,c} - \nabla\mathcal{L}^{f_j}_{e,c}, \nabla\ell^{f_k}_{e,c} - \nabla\mathcal{L}^{f_k}_{e,c}\Big\rangle\Big]$, and $\ell^{f_j}_{e,c}(w; x, y) = \ell(w; x, y) - c_j(x)\ell(w; x, f_j(x))$.

4 Confidence-Aware Prediction-Powered Learning

We first introduce confidence-aware PP for a single-teacher ($K = 1$), applying instance-dependent confidence symmetrically to labeled and unlabeled teacher terms, preserving unbiasedness while adapting to local teacher reliability.

Definition 4.1 (C-PP gradient)**.** For a parameter $\lambda \in [0, 1]$, the *confidence-aware prediction-powered gradient* is defined as $g^{\lambda}_{\mathrm{CPP}} = g^n + \lambda\big(g^{N,f} - g^{n,f}\big)$.

Here, we denote $g^n(w) = \frac{1}{n}\sum_{i=1}^n \nabla\ell(w; x^i, y^i)$, $g^{n,f}(w) = \frac{1}{n}\sum_{i=1}^n c(x^i)\nabla\ell(w; x^i, f(x^i))$, and $g^{N,f}(w) = \frac{1}{N}\sum_{i=1}^N c(\tilde{x}^i)\nabla\ell(w; \tilde{x}^i, f(\tilde{x}^i))$.

Since $\{x_i\}_{i=1}^n, \{\tilde{x}_i\}_{i=1}^N$ are i.i.d., $g^{\lambda}_{\mathrm{CPP}}$ is an unbiased estimator of $\nabla\mathcal{L}(w)$. Confidence weighting changes the estimator's variance without introducing bias.

Gradient variance analysis. Since PP gradients are unbiased, variance determines their efficiency in stochastic optimization. We next characterize how confidence, teacher quality, and the interpolation parameter λ determine the variance of the C-PP gradient.

Lemma 4.2. *For any Natural loss (Definition 3.2) and $\lambda \in [0, 1]$, the variance of the unbiased estimator $g^{\lambda}_{\mathrm{CPP}}$ satisfies*

$$\mathrm{Var}\big(g^{\lambda}_{\mathrm{CPP}}\big) \le 4\left((1-\lambda)^2\frac{\sigma^2}{n} + \lambda^2\left[\frac{\sigma^2}{N} + \left(\frac{1}{N} + \frac{1}{n}\right)s\right]\right), \quad s = \beta_Y^2(\tau\varepsilon^{f,1} + (1-\tau)\varepsilon^0) \tag{5}$$

Here, $\varepsilon^{f,1}$ is defined in equation 1, and ε^0 is the student error on low-confidence samples. Appendix D.1 gives a general-loss bound with the same asymptotic behavior. In the confidence-agnostic case, the analogous bound replaces s with $s_{PP} = \sigma_e^2 \le \beta_Y^2\varepsilon^f$. Thus, confidence weighting improves the bound whenever $s < s_{PP}$; Appendix H.2 gives the corresponding exact variance characterization.

We can tune $\lambda \in (0, 1]$ to minimize the variance bound in Lemma 4.2, yielding

$$V_{\mathrm{c}}^* = \mathrm{Var}(g^{\lambda^*}_{\mathrm{CPP}}) \le \frac{4\sigma^2}{n}\left(1 - \frac{1}{1+\frac{n}{N}}\frac{1}{1+\frac{s}{\sigma^2}}\right) \tag{6}$$

Effect of pseudo-label accuracy. Equation 6 implies $V_{\mathrm{c}}^* \le \frac{4\sigma^2}{n}\big(1 - \frac{1}{1+\frac{n}{N}}\frac{1}{1+\frac{s}{\sigma^2}}\big) \le \frac{4}{n}\big(\frac{n}{N}\sigma^2 + (1-\frac{n}{N})s\big)$; the second bound is tight when $n/N, s/\sigma^2$ are small. This is the supervised variance expression, up to a factor of 4, where the effective gradient variance interpolates between the nominal variance σ^2 and the correction error s.

At the favorable endpoint, $\tau \to 1$ and accurate retained pseudo-labels give $s/\sigma^2 \to (\beta_Y^2/\sigma^2)\varepsilon^{f,1} \approx 0$, and the bound in equation 6 approaches $4\sigma^2/(n+N)$, recovering the favorable variance reduction. At the opposite endpoint, $c(x) = 0$ almost surely, the correction vanishes and C-PP recovers the supervised estimator, with variance at most σ^2/n. Thus, C-PP remains unbiased while adapting between PP variance reduction and the supervised fallback according to confidence quality.

Compared with confidence-agnostic PP-SSL, C-PP improves the variance bound whenever confidence reduces s below s_{PP}, with larger gains when unlabeled data are abundant. Appendix C shows that near-optimal confidence gives $\frac{s}{s_{PP}} \le \frac{\varepsilon^0}{2} \le 1$. This advantage is observed in Appendix I.2. The Appendix further illustrates this effect for a teacher reliable only on a subset of classes, consistent with the partial-teacher experiments.

Algorithm 1 Confidence-Aware Multi-Model Prediction-Powered SSL (CM-PP)

Require: Prediction models $\{f_j : \mathcal{X} \to \mathcal{Y}\}_{j=1}^K$, confidence models $\{c_j : \mathcal{X} \to \{0,1\}\}_{j=1}^K$, stepsizes $\{\eta_t\}_{t=1}^T$, optional MW stepsizes $\{\gamma_t\}_{t=1}^T$

Initialize $w_1 \in \mathbb{R}^d$ and $\lambda_1 \in \Delta_{K+1}$
for $t = 1, \dots, T$ **do**
  Receive labeled batch $\{(x_t^i, y_t^i)\}_{i=1}^n$ and unlabeled batch $\{\tilde{x}_t^i\}_{i=1}^N$
  $g_t \leftarrow \frac{1}{n}\sum_{i=1}^n \nabla\ell(w_t; x_t^i, y_t^i)$ ▷ Labeled gradient estimate
  for $j = 1, \dots, K$ **do**
    $\hat{g}_{t,j}^u \leftarrow \frac{1}{N}\sum_{i=1}^N c_j(\tilde{x}_t^i)\nabla\ell\big(w_t; \tilde{x}_t^i, f_j(\tilde{x}_t^i)\big)$, $\hat{g}_{t,j}^\ell \leftarrow \frac{1}{n}\sum_{i=1}^n c_j(x_t^i)\nabla\ell\big(w_t; x_t^i, f_j(x_t^i)\big)$
  end for
  $d_{t,j} \leftarrow \hat{g}_{t,j}^u - \hat{g}_{t,j}^\ell$ for $j = 1, \dots, K$, and $g_t^{\lambda_t} \leftarrow g_t + \sum_{j=1}^K \lambda_{t,j} d_{t,j}$
  $w_{t+1} \leftarrow w_t - \eta_t g_t^{\lambda_t}$ ▷ Update parameters
  Set $\eta_t = \eta_0\big(\sum_{s=1}^t \|g_s^{\lambda_s}\|^2\big)^{-1/2}$. ▷ Optional AdaGrad update
  $q_{t,j} \leftarrow 2\langle d_{t,j}, g_t^{\lambda_t}\rangle,\ j = 0, 1, \dots, K,\ d_{t,0} := 0,$ ▷ Optional exponentiated-gradient update
  $\lambda_{t+1,j} \leftarrow \dfrac{\lambda_{t,j}\exp(-\gamma_t q_{t,j})}{\sum_{k=0}^K \lambda_{t,k}\exp(-\gamma_t q_{t,k})},\ j = 0, \dots, K$
end for
return w_{T+1}

In practice, lowering a binary confidence threshold increases coverage τ but also potentially increases $\varepsilon^{f,1}$; it is beneficial only while the newly admitted teacher predictions outperform the student. The resulting teacher-student error crossover, analyzed in Appendix H.2, shifts as the student improves. Online λ-updates can track this shift efficiently because the aggregation problem is low-dimensional.

Effective variance bound. When $\lambda = 0$, C-PP reduces to the labeled gradient, whose variance is bounded by σ^2/n. Combining this endpoint with equation 6 gives

$$V_c^\star \le \min\left\{\frac{\sigma^2}{n}, \frac{4\sigma^2}{n}\left(1 - \frac{1}{1+\frac{n}{N}}\frac{1}{1+\frac{s}{\sigma^2}}\right)\right\}. \tag{7}$$

Thus, C-PP exploits the teacher when its correction improves the variance bound and otherwise retains supervised learning as a feasible fallback. $\lambda = 0$ belongs to the feasible set, online adaptation competes with this fallback up to its regret term.

5 Confidence-Aware Multi-Teacher Prediction-Powered Learning

We extend C-PP to K heterogeneous teachers, each with a confidence signal. Confidence determines where each teacher contributes, while aggregation weights combine their corrections, exploiting complementary coverage even when no teacher covers the full task.

Let $\{f_j : \mathcal{X} \to \mathcal{Y}\}_{j=1}^K$ denote the teacher models. We assign nonnegative weights $\{\lambda_j\}_{j=0}^K$, where λ_0 corresponds to a zero correction and provides a supervised fallback. We denote the teacher weights by $\lambda_{1:K} := (\lambda_1, \dots, \lambda_K)$ and the full weight vector by $\lambda := (\lambda_0, \lambda_1, \dots, \lambda_K)$, with $\sum_{j=0}^K \lambda_j = 1$.

Definition 5.1 (CM-PP gradient)**.** The *confidence-aware multi-model prediction-powered gradient* is defined as $g_{CMPP}^\lambda(w) = g^n(w) + \sum_{j=1}^K \lambda_j\Big(g^{N,f_j}(w) - g^{n,f_j}(w)\Big)$, where for $j \ge 1$:
$g^{n,f_j}(w) = \frac{1}{n}\sum_{i=1}^n c_j(x^i)\nabla\ell\left(w; x^i, f_j(x^i)\right)$, $g^{N,f_j}(w) = \frac{1}{N}\sum_{i=1}^N c_j(\tilde{x}^i)\nabla\ell\left(w; \tilde{x}^i, f_j(\tilde{x}^i)\right)$.

Setting $c_j(x) \equiv 1$ for every teacher recovers the confidence-agnostic Multi-Model Prediction-Powered method (MM-PP). Since the labeled and unlabeled samples have the same input marginal, each teacher correction has zero expectation. Therefore, $g_{\mathrm{CMPP}}^\lambda$ remains an unbiased estimator of $\nabla\mathcal{L}(w)$, regardless of teacher or confidence accuracy.

Gradient variance analysis. The variance of CM-PP depends not only on the quality of each confidence-weighted teacher correction, but also on the covariance among teachers.

Lemma 5.2. *For any $\lambda \in \Delta^{K+1}$, the variance of the unbiased estimator $g^{\lambda}_{\mathrm{CMPP}}$ satisfies*

$$\mathrm{Var}(g^{\lambda}_{\mathrm{CMPP}}) \le 4\Big(\lambda_0^2 \tfrac{\sigma^2}{n} + (1-\lambda_0)^2 \tfrac{\sigma^2}{N} + \Big(\tfrac{1}{n} + \tfrac{1}{N}\Big)\lambda_{1:K}^T \bar{\Sigma}_{\mathrm{CM}} \lambda_{1:K}\Big). \tag{8}$$

The first two terms capture the sampling variance, while the quadratic term captures the joint covariance of the confidence-weighted teacher corrections. Consequently, CM-PP can exploit complementary teachers while limiting the contribution of inaccurate or redundant corrections.

Optimizing the covariance-aware bound in Lemma 5.2 over $\lambda \in \Delta^{K+1}$ yields

$$V^*_{\text{multi}} \le \frac{4\sigma^2}{n}\left(1 - \frac{1}{1+\frac{n}{N}} \frac{1}{1+\frac{1}{\sigma^2}\big(\mathbf{1}^\top \bar{\Sigma}^{-1}_{\mathrm{CM}} \mathbf{1}\big)^{-1}}\right) \tag{9}$$

where $\bar{\Sigma}_{\mathrm{CM}}$ is defined in Assumption 3.8. This bound is governed by the joint precision $\mathbf{1}^\top \bar{\Sigma}^{-1}_{\mathrm{CM}} \mathbf{1}$, which captures both high-confidence teacher quality and cross-teacher redundancy.

If the inverse-covariance solution lies outside the simplex, or if $\bar{\Sigma}_{\mathrm{CM}}$ is singular, the optimal teacher weights are instead obtained from the corresponding simplex-constrained quadratic program.

Diagonal envelope as a special case. As shown in Appendix E, there exists $\kappa_{\mathrm{CM}} \in [1, K]$ such that $\Sigma_{\mathrm{CM}}(w) \preceq \kappa_{\mathrm{CM}} \,\mathrm{diag}(s_1, \ldots, s_K)$ for all w. Using this diagonal envelope reduces equation 9 to

$$V^*_{\text{multi}} \le \frac{4\sigma^2}{n}\left(1 - \frac{1}{1+\frac{n}{N}} \frac{1}{1+\frac{\kappa_{\mathrm{CM}}}{\sum_{j=1}^K \left(\frac{s_j}{\sigma^2}\right)^{-1}}}\right). \tag{10}$$

In particular, setting $c_j(x) \equiv 1$ recovers the confidence-agnostic multi-model case, with $s_j = \sigma^2_{e,j}$. Setting $\kappa_{\mathrm{CM}} = K$ gives the worst-case diagonal envelope without additional cross-teacher covariance assumptions, for which the effective teacher error is the harmonic mean of $\{s_j\}_{j=1}^K$.

Effect of pseudo-label accuracy. Under the diagonal envelope, an accurate teacher contributes large normalized precision σ^2/s_j, and the bound approaches $V^\star_{\text{multi}} \le \frac{4\sigma^2}{n+N}$. Conversely, when $\sigma^2_{e,j} \gg \sigma^2$, a confidence-agnostic teacher contributes normalized precision $\sigma^2/\sigma^2_{e,j}$, which vanishes. If all teachers are unreliable, the aggregation approaches the supervised fallback. Informative confidence can preserve the benefit of unlabeled data even when teacher performance is poor, since the effective precision of a mostly unreliable teacher becomes $s_j \approx (1-\tau_j)\beta_Y^2 \varepsilon^{0,j}$, where $\varepsilon^{0,j}$ is the student error on low-confidence samples, preserving the reduction. At the exact endpoint where every $c_j(X) = 0$ almost surely, all corrections vanish and CM-PP recovers the supervised variance bound σ^2/n.

Under the full covariance-aware bound, the gain additionally depends on joint precision. Teachers with similar marginal accuracy and coverage can yield different improvements: complementary corrections increase effective precision, whereas correlated corrections provide less additional information. Appendix C.2.2 illustrates this with two partially overlapping teachers accurate on $\{0, \ldots, 4\}$ and $\{1, 3, 5, 7, 9\}$ in a 10-class problem. Without confidence, $s = \beta_Y^2$ per teacher and $0.5\beta_Y^2$ after diagonal aggregation. Near-optimal confidence gives $0.25\beta_Y^2\varepsilon^0$ for one teacher and $0.125\beta_Y^2\varepsilon^0$ for two, reduced further by covariance-aware aggregation to $0.10\beta_Y^2\varepsilon^0$. Covariance awareness adds a 20% reduction, and $\varepsilon^0 \le 2$ implies improvement by a factor of at least 2.5 over confidence-agnostic multi-teacher aggregation.

6 Practical Adaptation of Aggregation Weights

The variance bounds above characterize the optimal weights in terms of population variances and covariances, which are unknown in practice. We therefore consider strategies for adapting λ.

Batch-wise control of λ. λ can be estimated from mini-batch variances and covariances and applied to the next batch, preserving conditional unbiasedness but potentially noisy for small batches.

Online optimization of λ. Alternatively, we avoid explicit estimation by optimizing λ jointly with the model parameters. Tuning λ can be formulated as an online variance-minimization problem, where λ_t is treated as a decision variable. At iteration t, $h_t(\lambda) := \|g^\lambda_t\|^2$ is used to update λ_{t+1}. Conditional unbiasedness makes minimizing h_t equivalent in expectation to minimizing the conditional variance. We use projected AdaGrad for $\lambda_t \in [0, 1]$ in the single-teacher setting and exponentiated gradient for $\lambda_t \in \Delta^{K+1}$ in the multi-teacher setting. Full details are provided in Appendix G.

7 CONVERGENCE RATES

We show that CM-PP with online aggregation achieves the same leading convergence rate as the best fixed weights in hindsight minimizing the uniform variance bound, up to lower-order regret terms.

Assumption 7.1 (Bounded objective and gradients)**.** The objective $L : \mathbb{R}^d \to \mathbb{R}$ is M-bounded and the stochastic gradients are G-bounded, i.e., $|L(w)| \le M,\ \|\nabla \ell(w; x, y)\| \le G$ for all $w, (x, y)$.

Theorem 7.2 (Convergence rate of CM-PP)**.** *Suppose that Assumption 3.3 and Assumption 7.1 hold. Let $\{w_t\}_{t=1}^T$ be generated by AdaGrad, and $\eta_0 = \sqrt{2M/\beta}$ using the CM-PP gradient with online optimization of $\{\lambda_j\}_{j=0}^K$ subject to $\lambda_j \ge 0,\ \sum_{j=0}^K \lambda_j = 1$. Then*

$$\mathbb{E}\Big[\tfrac{1}{T}\sum_{t=1}^{T} \|\nabla \mathcal{L}(w_t)\|^2\Big] = \mathcal{O}\left(\sqrt{\tfrac{M\beta V^*_{\text{multi}}}{T}} + \tfrac{M\beta}{T} + \tfrac{\sqrt{M\beta}\, G\sqrt{\log(K+1)}}{T}\right) \tag{11}$$

Here, V^*_{multi} denotes the minimum achievable variance bound of the CM-PP gradient.

The first term matches the rate obtained using the best fixed aggregation weights and links CM-PP's variance reduction to faster convergence. The second is the standard optimization term, the third is the cost of online weight adaptation. The adaptation cost is $\mathcal{O}(\sqrt{\log(K+1)}/T)$, so it does not affect the leading $\mathcal{O}(1/\sqrt{T})$ rate, and depends only logarithmically on K. Because the supervised vertex is feasible, the optimal fixed aggregation can fall back to supervised learning, while the online scheme competes with this optimum up to its regret term. A complete proof is provided in the appendix.

The guarantee compares against the best fixed weights, although the iterates λ_t remain adaptive throughout training. The C-PP rate follows analogously by replacing $V^\star_{\text{multi}}$ with $V_c^\star$ and using projected AdaGrad for λ_t. A complete proof follows directly from the analysis of Shoham et al. (2026)

8 THE FLEXIBILITY OF THE CONFIDENCE MODEL

Confidence controls how strongly a teacher's label signal enters the PP correction, with minimal assumptions on its construction. It may encode coverage, uncertainty, agreement, alignment, or other task-specific cues for teacher use. Confidence signals can be **prior-informed**, using known teacher coverage or metadata, or **model-derived** from predictive entropy, logit margins, teacher disagreement, student-teacher agreement, or representation alignment. They may serve as soft weights or thresholded binary indicators. Appendix H provides practical construction, validation guidance, improvement criteria, and computational considerations.

In the multi-teacher setting, confidence may be shared or teacher-specific, allowing reliability to be estimated globally or separately for each teacher. If no single confidence signal is sufficiently informative, **confidence aggregation** can combine several heuristic or moderately informative signals into a cleaner teacher-level estimate. Because this aggregation is low-dimensional, it can be fitted efficiently using calibration data or updated online, providing an optional third adaptive process alongside the student and aggregation-weight updates. This allows CM-PP to handle weak, noisy, or correlated confidence estimates while preserving conditional unbiasedness.

The two CM-PP weighting mechanisms play distinct roles: c_j controls sample contributions within teacher j's correction, whereas λ_j weights the teacher-confidence pair (f_j, c_j) as a whole. The λ-update adds a protective layer by downweighting an unhelpful pair in favor of other teachers or the supervised fallback. Cross-teacher calibration is not required for validity, while online λ-adaptation can compensate for differences in effective scale and utility of the resulting corrections.

9 EXPERIMENTS

We evaluate C-PP and CM-PP under three forms of heterogeneous teacher reliability: teacher-reliability shift, subpopulation-dependent expertise, and partial label-space coverage. Controlled experiments isolate the predicted mechanisms, while larger-scale experiments test performance

Table 1: Tiny-ImageNet results under class-dependent teacher-reliability shift with labeled-to-unlabeled ratio $r = \frac{n}{N} = 1/9$, with labeled budget of $10,000$ samples.

Method	Top-1 (%)	Top-5 (%)	Peak memory (GB)
Soft knowledge distillation	25.86 ± 2.11	52.67 ± 2.75	18.04
UDA	37.99 ± 0.40	64.78 ± 0.70	21.26
FreeMatch	37.82 ± 0.20	62.45 ± 0.10	18.78
PP-SSL	29.05 ± 1.35	56.26 ± 2.09	21.20
Soft margin C-PP	31.01 ± 1.38	58.65 ± 1.39	17.15
Soft entropy C-PP	31.69 ± 0.93	59.28 ± 0.81	**11.25**
Binary entropy C-PP	$\mathbf{39.73 \pm 0.19}$	$\mathbf{66.49 \pm 0.44}$	13.25
Oracle C-PP	49.39 ± 0.24	74.46 ± 0.09	25.56

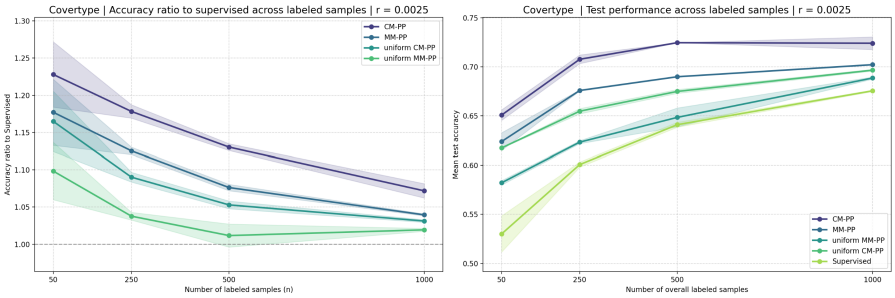


Figure 1: Covertype performance across overall labeled-sample budgets with $r = 0.0025$. Left: accuracy relative to supervised learning. Right: mean test accuracy.

across modalities, confidence signals, labeled-to-unlabeled ratios, and label budgets. Fixed teacher predictions and confidence scores can be cached, avoiding repeated teacher inference.

Teacher reliability shift. We evaluate C-PP when teacher reliability varies across the input distribution while the teacher retains coverage of the full label space. On Tiny-ImageNet (Le & Yang, 2015), we use off-the-shelf pretrained weights, and induce class-dependent corruption on the training data to create regions in which its reliability substantially degrades. As shown in Table 1, all confidence-aware variants improve over confidence-agnostic PP-SSL. Binary entropy C-PP performs best among practical variants, outperforming both UDA (Xie et al., 2020a) and FreeMatch (Wang et al., 2023). Oracle C-PP reaches higher accuracy results, highlighting the additional gains possible with more accurate teacher-reliability estimates. Practical C-PP variants add no peak GPU memory over PP-SSL or SSL baselines.

Subpopulation-reliable teacher. We first evaluate CM-PP on Covertype (Blackard, 1998) with topographic, hydrological, and land-cover specialists, and a weaker generalist, yielding complementary subpopulation-dependent reliability. A teacher analysis is provided in Appendix I. We use the softmax-margin as a lightweight local confidence signal.

Figure 1 shows that all PP variants improve over supervised learning across budgets, indicating that the heterogeneous teacher pool remains useful despite covariate mismatch. CM-PP consistently achieves the highest accuracy, with the largest gains in the low-label regime. Comparing each uniform variant with its vector-weighted counterpart shows the benefit of teacher-specific aggregation, while the MM-PP–CM-PP gap highlights the benefit of confidence-aware correction, consistent with Table 4. Gains narrow as the labeled budget grows, while CM-PP remains strongest across budgets.

We further evaluate CM-PP in a larger-scale dense-prediction setting on PASCAL VOC (Everingham et al., 2010). We use three architecturally distinct segmentation teachers with off-the-shelf pretrained weights. Each teacher is weakened on a different semantically related class subset using the same class-conditional corruption mechanism, producing partially overlapping regions of teacher reliability. We consider three confidence signals from the teachers' predictive distributions: normalized margin, entropy, and a spatial extension of normalized margin. The confidence type is shared across teachers. To stress-test CM-PP, we use $n/N = 1$, where unlabeled-data leverage is minimal.

Table 2: PASCAL VOC under class-dependent multi-teacher reliability shifts with $r = 1$, overall labeled budget of 92 and 4 labeled samples every update, for moderate and elevated teacher corruption.

Method	**mIoU (%)**		**Peak memory (GB)**
	Moderate	**Elevated**	
Supervised	48.39 ± 0.21	47.85 ± 0.21	4.9594
Uniform Knowledge distillation (KD)	55.47 ± 0.54	54.59 ± 0.37	5.0004
Margin-based KD	55.41 ± 0.37	54.75 ± 0.37	5.0004
Entropy KD	55.71 ± 0.40	55.21 ± 0.40	5.0004
Spatial-margin KD	55.82 ± 0.39	54.75 ± 0.39	5.0079
MM-PP	52.53 ± 1.43	56.09 ± 1.43	5.0004
Margin CM-PP	$\mathbf{55.94 \pm 0.28}$	$\mathbf{56.21 \pm 0.28}$	5.0004
Entropy CM-PP	55.74 ± 0.55	55.75 ± 0.55	5.0072
Spatial-margin CM-PP	55.22 ± 0.58	55.66 ± 0.58	5.0004

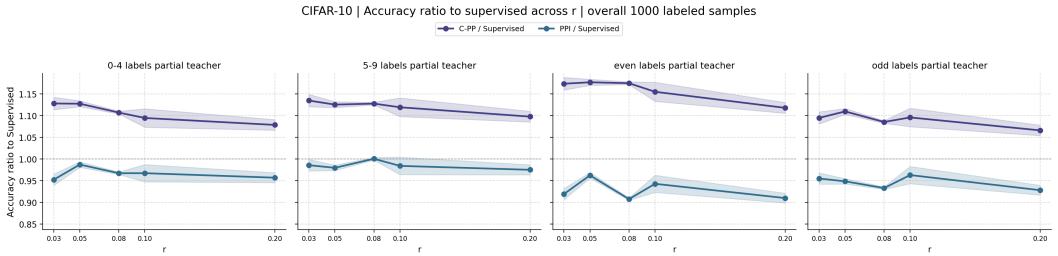


Figure 2: CIFAR-10 performance relative to supervised learning under partial teacher coverage across labeled-to-unlabeled ratios. Values above 1 indicate improvement over supervised learning.

Table 2 reports results under moderate and elevated teacher corruption. Margin-based CM-PP achieves the highest mIoU in both regimes, outperforming MM-PP and the corresponding knowledge-distillation baselines. Across matched confidence signals, CM-PP outperforms its distillation counterpart in five of six comparisons, including all three confidence choices under elevated corruption. The advantage is maintained as teacher corruption increases, showing that CM-PP can exploit complementary, class-dependent teacher reliability in a dense, architecture-heterogeneous setting.

Partial-coverage teachers. We evaluate C-PP with teachers covering only part of the student's label space. On CIFAR-10 (Figure 2), all teachers have equal nominal coverage but different label subsets, so differences reflect teacher quality and the learnability of covered classes, rather than coverage alone. With λ selected only by offline fine grid search, confidence-agnostic PPI can fall below supervised learning, whereas C-PP remains stable across teacher subsets and labeled-to-unlabeled ratios, demonstrating reduced sensitivity to imperfect aggregation.

Controlled MNIST (LeCun et al., 1998), Covertype, and SVHN (Netzer et al., 2011) experiments further support the behavior predicted in Appendices C, H: C-PP benefits from identifying reliable teacher regions; SVHN highlights the tradeoff between confidence quality and computational cost. Full results in Appendix I.

10 Conclusions and Future Work

We introduced C-PP and CM-PP, an unbiased PP framework for heterogeneous, partially reliable teachers. Confidence-aware correction adapts to local reliability, covariance-aware aggregation exploits complementary teachers while accounting for dependence; both reduce gradient variance, improve optimization, and yield empirical gains. Practical gains require informative confidence, and our guarantees assume shared labeled and unlabeled input marginals. Multi-teacher or dynamic-confidence settings may incur additional storage or inference, although fixed teacher outputs can be cached. Future work includes alignment-aware fine-tuning, reinforcement learning, and scaling to foundation models.

AI USE STATEMENT

In this work, we used generative AI tools for tasks requiring disclosure related to method implementation. In particular, these tools occasionally assisted with debugging implementations and with checking that our implementations were consistent with the algorithms described in this paper and that implementations of baseline methods were consistent with their corresponding papers and, when available, released reference code.

We have not used generative AI tools to generate synthetic datasets, develop theoretical models or conceptual frameworks, formulate mathematical claims, provide critical ingredients for or assist in writing mathematical proofs, propose or refine hypotheses, design or provide feedback on the research methodology or experiments, clean or reformat datasets, or interpret experimental results. Translation and qualitative or thematic data analysis were not applicable to this work.

Additionally, we used generative AI tools for editing software code during debugging and editing the manuscript to improve readability. In particular, these tools were used to help rephrase and shorten existing text to meet the page limit, identify typographical or presentation issues, and flag potential inconsistencies or problems in late-stage versions of the manuscript.

We reviewed all AI-assisted work. AI-assisted code changes and debugging suggestions were inspected and tested by the authors, and method and baseline implementations were checked against the corresponding algorithm descriptions, publications, and available reference implementations. All AI-assisted manuscript edits were reviewed to ensure that they preserved the intended technical meaning and claims.

We take responsibility for the final content of this work, including text, claims, code, and other artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

Algorithmic, implementation, and experimental details, including datasets, model architectures, optimization settings, confidence constructions, and baselines, are provided in the main paper and appendix, together with complete proofs of the theoretical results. We will release the code used to reproduce the reported experiments with the final version of the paper.

REFERENCES

Anastasios N Angelopoulos, Stephen Bates, Clara Fannjiang, Michael I Jordan, and Tijana Zrnic. Prediction-powered inference. *Science*, 382(6671):669–674, 2023a.

Anastasios N. Angelopoulos, John C. Duchi, and Tijana Zrnic. PPI++: Efficient prediction-powered inference, 2023b. URL `https://arxiv.org/abs/2311.01453`.

Eric Arazo, Diego Ortego, Paul Albert, Noel O'Connor, and Kevin McGuinness. Unsupervised label noise modeling and loss correction. In *International conference on machine learning*, pp. 312–321. PMLR, 2019.

Eric Arazo, Diego Ortego, Paul Albert, Noel E O'Connor, and Kevin McGuinness. Pseudo-labeling and confirmation bias in deep semi-supervised learning. In *2020 International joint conference on neural networks (IJCNN)*, pp. 1–8. IEEE, 2020.

David Berthelot, Nicholas Carlini, Ian Goodfellow, Nicolas Papernot, Avital Oliver, and Colin A Raffel. Mixmatch: A holistic approach to semi-supervised learning. *Advances in neural information processing systems*, 32, 2019.

Jock Blackard. Covertype. UCI Machine Learning Repository, 1998.

Jock A. Blackard and Denis J. Dean. Comparative accuracies of artificial neural networks and discriminant analysis in predicting forest cover types from cartographic variables. *Computers and Electronics in Agriculture*, 24(3):131–151, 1999. doi: 10.1016/S0168-1699(99)00046-0.

Avrim Blum and Tom Mitchell. Combining labeled and unlabeled data with co-training. In *Proceedings of the eleventh annual conference on Computational learning theory*, pp. 92–100, 1998.

Nicolo Cesa-Bianchi and Gábor Lugosi. *Prediction, learning, and games*. Cambridge university press, 2006.

Olivier Chapelle, Bernhard Schölkopf, and Alexander Zien. *Semi-Supervised Learning*. MIT Press, 2009.

Hao Chen, Ran Tao, Yue Fan, Yidong Wang, Jindong Wang, Bernt Schiele, Xing Xie, Bhiksha Raj, and Marios Savvides. Softmatch: Addressing the quantity-quality trade-off in semi-supervised learning. In *International Conference on Learning Representations*, 2023.

Victor Chernozhukov, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen, Whitney Newey, and James Robins. Double/debiased machine learning for treatment and structural parameters, 2018.

Stefano Cortinovis and François Caron. Fab-ppi: Frequentist, assisted by bayes, prediction-powered inference. *arXiv preprint arXiv:2502.02363*, 2025.

Charlie Cowen-Breen, Alekh Agarwal, Stephen Bates, William W Cohen, Jacob Eisenstein, Amir Globerson, and Adam Fisch. Multiple-prediction-powered inference. *arXiv preprint arXiv:2603.27414*, 2026.

Ron Dorfman, Naseem Yehya, and Kfir Y Levy. Dynamic byzantine-robust learning: adapting to switching byzantine workers. In *Proceedings of the 41st International Conference on Machine Learning*, pp. 11501–11543, 2024.

Mark Everingham, Luc Van Gool, Christopher K. I. Williams, John Winn, and Andrew Zisserman. The PASCAL visual object classes (VOC) challenge. *International Journal of Computer Vision*, 88(2):303–338, 2010. doi: 10.1007/s11263-009-0275-4.

Yves Grandvalet and Yoshua Bengio. Semi-supervised learning by entropy minimization. *Advances in neural information processing systems*, 17, 2004.

Bharath Hariharan, Pablo Arbeláez, Lubomir Bourdev, Subhransu Maji, and Jitendra Malik. Semantic contours from inverse detectors. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 991–998, 2011. doi: 10.1109/ICCV.2011.6126343.

Dan Hendrycks, Kimin Lee, and Mantas Mazeika. Using pre-training can improve model robustness and uncertainty. In *International conference on machine learning*, pp. 2712–2721. PMLR, 2019.

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

Jyrki Kivinen and Manfred K. Warmuth. Exponentiated gradient versus gradient descent for linear predictors. *Information and Computation*, 132(1):1–63, 1997.

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

Yann Le and Xuan Yang. Tiny imagenet visual recognition challenge. *CS231N*, 2015.

Yann LeCun, Leon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998. doi: 10.1109/5.726791.

Dong-Hyun Lee et al. Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks. In *Workshop on challenges in representation learning, ICML*, volume 3, pp. 896. Atlanta, 2013.

Lu Liu and Robby T. Tan. Certainty driven consistency loss on multi-teacher networks for semi-supervised learning. *Pattern Recognition*, 120:108140, 2021.

Pranav Mani, Peng Xu, Zachary C Lipton, and Michael Oberst. No free lunch: Non-asymptotic analysis of prediction-powered inference. *arXiv preprint arXiv:2505.20178*, 2025.

Yuval Netzer, Tao Wang, Adam Coates, Alessandro Bissacco, Bo Wu, and Andrew Y. Ng. Reading digits in natural images with unsupervised feature learning. In *NIPS Workshop on Deep Learning and Unsupervised Feature Learning*, 2011.

Francesco Orabona. A modern introduction to online learning. *arXiv preprint arXiv:1912.13213*, 2019.

Siyuan Qiao, Wei Shen, Zhishuai Zhang, Bo Wang, and Alan Yuille. Deep co-training for semi-supervised image recognition. In *Proceedings of the european conference on computer vision (eccv)*, pp. 135–152, 2018.

Clement Ruah, Houssem Sifaou, Osvaldo Simeone, and Bashir Al-Hashimi. Context-aware doubly-robust semi-supervised learning. *IEEE Signal Processing Letters*, 2025.

Shai Shalev-Shwartz. Online learning and online convex optimization. *Foundations and Trends in Machine Learning*, 4(2):107–194, 2011.

Noa Shoham, Ron Dorfman, Shalev Shaer, Kfir Y Levy, and Yaniv Romano. Prediction-powered semi-supervised learning with online power tuning. *Advances in Neural Information Processing Systems*, 38:108900–108937, 2026.

Aarti Singh, Robert Nowak, and Jerry Zhu. Unlabeled data: Now it helps, now it doesn't. *Advances in neural information processing systems*, 21, 2008.

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. *Advances in neural information processing systems*, 33:596–608, 2020.

Antti Tarvainen and Harri Valpola. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. *Advances in neural information processing systems*, 30, 2017.

Jesper E Van Engelen and Holger H Hoos. A survey on semi-supervised learning. *Machine learning*, 109(2):373–440, 2020.

Yidong Wang, Hao Chen, Qiang Heng, Wenxin Hou, Yue Fan, Zhen Wu, Jindong Wang, Marios Savvides, Takahiro Shinozaki, Bhiksha Raj, Bernt Schiele, and Xing Xie. Freematch: Self-adaptive thresholding for semi-supervised learning. In *International Conference on Learning Representations*, 2023.

Colin Wei, Kendrick Shen, Yining Chen, and Tengyu Ma. Theoretical analysis of self-training with deep networks on unlabeled data. *arXiv preprint arXiv:2010.03622*, 2020.

Qizhe Xie, Zihang Dai, Eduard Hovy, Thang Luong, and Quoc V. Le. Unsupervised data augmentation for consistency training. In *Advances in Neural Information Processing Systems*, volume 33, pp. 6256–6268, 2020a.

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*, pp. 10687–10698, 2020b.

David Yarowsky. Unsupervised word sense disambiguation rivaling supervised methods. In *33rd annual meeting of the association for computational linguistics*, pp. 189–196, 1995.

Banghua Zhu, Mingyu Ding, Philip Jacobson, Ming Wu, Wei Zhan, Michael Jordan, and Jiantao Jiao. Doubly-robust self-training. *Advances in Neural Information Processing Systems*, 36:41413–41431, 2023.

Tijana Zrnic. A note on the prediction-powered bootstrap. *arXiv preprint arXiv:2405.18379*, 2024.

Tijana Zrnic and Emmanuel J Candès. Cross-prediction-powered inference. *Proceedings of the National Academy of Sciences*, 121(15):e2322083121, 2024.

APPENDIX CONTENTS

A Teacher quality **14**

A.1 Label-Lipschitz gradient for standard losses 14

A.2 Gradient variance and teacher quality 15

A.3 Teacher–student distribution mismatch 16

B Examples of natural losses **16**

C Illustrative examples: confidence turns heterogeneous teacher reliability into useful signal **17**

C.1 Single-model confidence setting 18

C.2 Two-model confidence setting 19

D Theoretical analysis of prediction-powered gradients **22**

D.1 Proof of lemma 4.2 22

D.2 Proof of Lemma 5.2 24

D.3 Proof of Theorem 7.2 27

D.4 Proof of Lemma D.2 29

D.5 Proof of Lemma D.3 31

D.6 Optimal λ-Parameter Calculations 31

E Diagonal-scaled PSD envelopes and recovery of the worst-case bound **34**

E.1 Covariance-aware MM-PP bound 34

E.2 Retrieving the optimal weights from the diagonal-scaled special case 35

E.3 Proof of Lemma 5.2 for the Multi-Model Diagonal Case. 36

F Comparison of PPI++, PP-SSL, C-PP, MM-PP, and CM-PP **38**

F.1 Settings and limitations 38

F.2 Main-theory comparison of optimal interpolation weights 38

F.3 Linear-MSE closed forms 39

G Online Optimization of Interpolation Weights **40**

G.1 Single-Teacher Setting: Projected AdaGrad 41

G.2 Multi-Teacher Setting: Exponentiated Gradient Descent / Multiplicative Weights . 41

H Confidence model analysis and selection **42**

H.1 Empirical confidence selection 42

H.2 Exact variance and the variance-bound criterion 43

H.3 Confidence model types 44

H.4 Teacher, student, and data behavior 46

H.5 Runtime and memory considerations 47

I Experimental details **48**

I.1 Common experimental protocol 48

I.2 Controlled experiments 49

I.3 Heterogeneous multi-teacher experiments 54

I.4 Off-the-shelf teacher under reliability shift 57

A TEACHER QUALITY

In this appendix, we relate teacher prediction quality to the variance of the prediction-powered correction term. Throughout, let f denote a teacher model and let $c(x) \in \{0, 1\}$ be its binary confidence indicator.

Since the optimization and variance bounds are evaluated under the student distribution, we first define teacher quality under P:

$$\varepsilon^f := \mathbb{E}_{(x,y)\sim P}\left[\|y - f(x)\|^2\right], \qquad \varepsilon^{f,c} := \mathbb{E}_{(x,y)\sim P|c(x)=c}\left[\|y - f(x)\|^2\right], \quad c \in \{0, 1\}.$$

We also define the pointwise gradient error and its conditional mean

$$\nabla\ell_e^f(w; x, y) := \nabla\ell(w; x, y) - \nabla\ell(w; x, f(x)), \qquad \nabla\mathcal{L}_e^{f,1}(w) := \mathbb{E}_{(x,y)\sim P|c(x)=1}\left[\nabla\ell_e^f(w; x, y)\right].$$

The next assumption formalizes that teacher label errors translate smoothly into errors in the gradient.

Assumption A.1 (β_Y-smoothness to error)**.** Assume that the gradient $\nabla\ell(w; x, y)$ is β_Y-Lipschitz in y, i.e., for any $w \in \mathbb{R}^d$, $x \in \mathcal{X}$, and $y_1, y_2 \in \mathcal{Y}$, we have $\|\nabla\ell(w; x, y_1) - \nabla\ell(w; x, y_2)\| \leq \beta_Y|y_1 - y_2|$. Then the instance-dependent gradient error variance σ_e^2, satisfies $\sigma_e^2 \leq \beta_Y^2 \cdot \mathcal{E}^f$.

A.1 LABEL-LIPSCHITZ GRADIENT FOR STANDARD LOSSES

Several standard losses satisfy Assumptions 3.4 and A.1. The first two examples use scalar labels, while the third covers multiclass cross-entropy.

A.1.1 SQUARED LOSS

Proposition A.2. *Let*

$$\ell_{L_2}(w; x, y) = \frac{1}{2}\big(\phi(w; x) - y\big)^2.$$

If $\|\nabla_w\phi(w; x)\| \leq R$ uniformly, then $\nabla_w\ell_{L_2}(w; x, y)$ is R-Lipschitz in y.

Proof. We have

$$\nabla_w\ell_{L_2}(w; x, y) = \big(\phi(w; x) - y\big)\nabla_w\phi(w; x).$$

Hence, for any y_1, y_2,

$$\begin{aligned}\|\nabla_w\ell_{L_2}(w; x, y_1) - \nabla_w\ell_{L_2}(w; x, y_2)\| &= \|(y_2 - y_1)\nabla_w\phi(w; x)\| \\ &= |y_1 - y_2|\,\|\nabla_w\phi(w; x)\| \\ &\leq R|y_1 - y_2|.\end{aligned}$$

Therefore, the squared-loss gradient is R-Lipschitz with respect to the label argument. □

A.1.2 LOGISTIC LOSS

Proposition A.3. *Let $\sigma(t) = (1 + e^{-t})^{-1}$, and consider the binary logistic loss*

$$\ell_{\log}(w; x, y) = -y\log\big(\sigma(\phi(w; x))\big) - (1 - y)\log\big(1 - \sigma(\phi(w; x))\big).$$

If $\|\nabla_w\phi(w; x)\| \leq R$ uniformly, then $\nabla_w\ell_{\log}(w; x, y)$ is R-Lipschitz in y.

Proof. The gradient is

$$\nabla_w \ell_{\log}(w; x, y) = \big(\sigma(\phi(w; x)) - y\big)\nabla_w \phi(w; x).$$

Therefore, for any y_1, y_2,

$$\begin{aligned}\|\nabla_w \ell_{\log}(w; x, y_1) - \nabla_w \ell_{\log}(w; x, y_2)\| &= \|(y_2 - y_1)\nabla_w \phi(w; x)\| \\ &= |y_1 - y_2|\,\|\nabla_w \phi(w; x)\| \\ &\le R|y_1 - y_2|.\end{aligned}$$

Thus, the logistic-loss gradient is R-Lipschitz with respect to the label argument. □

A.1.3 Multiclass cross-entropy

Proposition A.4. *Let $y \in \Delta^C$ be a one-hot or soft label vector, and consider the multiclass cross-entropy loss*

$$\ell_{\mathrm{CE}}(w; x, y) = -\sum_{k=1}^{C} y_k \log\big(\textit{softmax}(\phi(w; x))_k\big),$$

where $\phi(w; x) \in \mathbb{R}^C$ denotes the logit vector. If the Jacobian $J_w\phi(w; x)$ satisfies

$$\|J_w \phi(w; x)\| \le R,$$

uniformly, where $\|\cdot\|$ denotes the spectral norm,

$$\|A\| := \sup_{\|v\|=1} \|Av\|,$$

then $\nabla_w \ell_{\mathrm{CE}}(w; x, y)$ is R-Lipschitz in y.

Proof. Let $p(w; x) = softmax(\phi(w; x))$. The gradient of the multiclass cross-entropy loss is

$$\nabla_w \ell_{\mathrm{CE}}(w; x, y) = J_w \phi(w; x)^\top \big(p(w; x) - y\big).$$

Therefore, for any $y_1, y_2 \in \Delta^C$,

$$\begin{aligned}\|\nabla_w \ell_{\mathrm{CE}}(w; x, y_1) - \nabla_w \ell_{\mathrm{CE}}(w; x, y_2)\| &= \big\|J_w \phi(w; x)^\top \big((p(w; x) - y_1) - (p(w; x) - y_2)\big)\big\| \\ &= \big\|J_w \phi(w; x)^\top (y_2 - y_1)\big\| \\ &\le \|J_w \phi(w; x)\|\,\|y_1 - y_2\| \\ &\le R\|y_1 - y_2\|.\end{aligned}$$

Hence, the multiclass cross-entropy gradient is R-Lipschitz with respect to the label argument. □

A.2 Gradient variance and teacher quality

We now use Assumption A.1 to show that teacher label error controls the second moment, and hence the variance, of the correction gradient under the student distribution.

Proposition A.5 (From label error to gradient-error variance)**.** *Under Assumption A.1,*

$$\mathbb{E}_{(x,y)\sim P|c(x)=1}\big[\|\nabla \ell_e^f(w; x, y)\|^2\big] \le \beta_Y^2 \varepsilon^{f,1},$$

and consequently

$$\mathbb{E}_{(x,y)\sim P|c(x)=1}\big[\|\nabla \ell_e^f(w; x, y) - \nabla \mathcal{L}_e^{f,1}(w)\|^2\big] \le \beta_Y^2 \varepsilon^{f,1}.$$

Proof. By definition of $\nabla \ell_e^f(w; x, y)$,

$$\mathbb{E}_{(x,y)\sim P|c(x)=1}\big[\|\nabla \ell_e^f(w; x, y)\|^2\big] = \mathbb{E}_{(x,y)\sim P|c(x)=1}\big[\|\nabla \ell(w; x, y) - \nabla \ell(w; x, f(x))\|^2\big].$$

Applying Assumption A.1,

$$\begin{aligned}\mathbb{E}_{(x,y)\sim P|c(x)=1}\big[\|\nabla \ell_e^f(w; x, y)\|^2\big] &\le \mathbb{E}_{(x,y)\sim P|c(x)=1}\big[\beta_Y^2 \|y - f(x)\|^2\big] \\ &= \beta_Y^2 \mathbb{E}_{(x,y)\sim P|c(x)=1}\big[\|y - f(x)\|^2\big] \\ &= \beta_Y^2 \varepsilon^{f,1}.\end{aligned}$$

Since variance is upper bounded by the second moment,

$$\mathbb{E}_{(x,y)\sim P|c(x)=1}\big[\|\nabla \ell_e^f(w; x, y) - \nabla \mathcal{L}_e^{f,1}(w)\|^2\big] \le \mathbb{E}_{(x,y)\sim P|c(x)=1}\big[\|\nabla \ell_e^f(w; x, y)\|^2\big],$$

which proves the result. □

Corollary A.6. *If the high-confidence error variance parameter is defined by*

$$\sigma_{e,1}^2 \;\geq\; \sup_{w\in\mathbb{R}^d} \mathbb{E}_{(x,y)\sim P|c(x)=1}\left[\|\nabla \ell_e^f(w;x,y) - \nabla \mathcal{L}_e^{f,1}(w)\|^2\right],$$

then taking $\sigma_{e,1}^2$ as the maximal conditional variance over w, Proposition A.5 implies that this variance is bounded

$$\sigma_{e,1}^2 \leq \beta_Y^2 \varepsilon^{f,1}.$$

A.3 Teacher–student distribution mismatch

Up to this point, all quantities were defined directly under the student distribution P. This is the most natural regime when the confidence signal is constructed or calibrated for the student task itself. In particular, if the confidence model is fitted to the student distribution, then the informative high-confidence region is defined where the student optimization is carried out, and no additional teacher–student mismatch term is needed.

In other settings, however, the teacher and its confidence model may be native to a related teacher distribution D. We then define

$$\varepsilon_D^f := \mathbb{E}_{(x,y)\sim D}\left[\|y - f(x)\|^2\right], \qquad \varepsilon_D^{f,c} := \mathbb{E}_{(x,y)\sim D|c(x)=c}\left[\|y - f(x)\|^2\right], \quad c \in \{0,1\}.$$

Assumption A.7 (Informative confidence under the teacher distribution)**.** The teacher confidence is informative under D, in the sense that

$$\varepsilon_D^{f,1} \leq \varepsilon_D^f \leq \varepsilon_D^{f,0}.$$

That is, high-confidence predictions are at least as accurate as average predictions, which are in turn at least as accurate as low-confidence predictions.

To transfer this property from D to P, we introduce a mismatch term $\varepsilon_{DP} \geq 0$ such that

$$\varepsilon^{f,1} \leq \varepsilon^f + \varepsilon_{DP}, \qquad \varepsilon^f - \varepsilon_{DP} \leq \varepsilon^{f,0}.$$

The quantity ε_{DP} captures the degradation induced by evaluating the teacher and its confidence signal on the student distribution rather than on the distribution on which they were originally fitted.

This term is expected to be small when the confidence model is adapted to the student distribution, or when the chosen teacher is relevant to the student task so that its native domain already overlaps substantially with P. In particular, teachers are typically selected because their training distribution shares meaningful support with the student distribution in the region where they are useful, making large mismatch less likely.

More concretely, ε_{DP} can be related to standard discrepancy measures between the conditional student and teacher distributions. Let P_c and D_c denote the distributions of (x,y) conditioned on $c(x) = c$ under P and D, respectively, and suppose $\|y - f(x)\|^2 \leq B$ almost surely. Then

$$|\varepsilon^{f,c} - \varepsilon_D^{f,c}| \leq 2B\,\mathrm{TV}(P_c, D_c),$$

where TV is the total variation distance. By Pinsker's inequality, this further implies

$$|\varepsilon^{f,c} - \varepsilon_D^{f,c}| \leq B\sqrt{2\,\mathrm{KL}(P_c\|D_c)}.$$

Hence, if the conditional teacher and student distributions are close, then the teacher error transfers from D to P with only a small degradation, so ε_{DP} is small.

Combining this discussion with Proposition A.5 shows that, under teacher–student mismatch,

$$\mathbb{E}_{(x,y)\sim P|c(x)=1}\left[\|\nabla \ell_e^f(w;x,y) - \nabla \mathcal{L}_e^{f,1}(w)\|^2\right] \leq \beta_Y^2\left(\varepsilon^f + \varepsilon_{DP}\right),$$

and therefore, taking $\sigma_{e,1}^2$ as the maximal conditional variance over w gives

$$\sigma_{e,1}^2 \leq \beta_Y^2\left(\varepsilon^f + \varepsilon_{DP}\right).$$

B Examples of natural losses

This appendix records several standard examples satisfying Definition 3.2. Recall that ℓ is natural if

$$\nabla_w \ell(w;x,y)|_{y=f_w(x)} = 0 \qquad \text{for all } w \in \mathbb{R}^d.$$

MSE loss

Proposition B.1. *Let*

$$\ell_{\mathrm{MSE}}(w; x, y) = \frac{1}{2}\|f_w(x) - y\|^2.$$

Then ℓ_{MSE} is natural.

Proof. The loss is non-negative. Substituting $y = f_w(x)$ gives

$$\ell_{\mathrm{MSE}}(w; x, f_w(x)) = \frac{1}{2}\|f_w(x) - f_w(x)\|^2 = 0.$$

□

Binary cross-entropy

Proposition B.2. *Let $p_w(x) \in (0, 1)$, and define*

$$\ell_{\mathrm{BCE}}(w; x, y) = -y \log p_w(x) - (1 - y) \log(1 - p_w(x)).$$

Then ℓ_{BCE} is Natural.

Proof. For fixed y,

$$\nabla_w \ell_{\mathrm{BCE}}(w; x, y) = \frac{p_w(x) - y}{p_w(x)(1 - p_w(x))} \nabla_w p_w(x).$$

Hence, at $y = p_w(x)$,

$$\nabla_w \ell_{\mathrm{BCE}}(w; x, p_w(x)) = 0.$$

□

Multiclass cross-entropy

Proposition B.3. *Let $y \in \Delta^C$ and $p_w(x) \in \Delta^C$, with $p_{w,k}(x) > 0$ for all k, and define*

$$\ell_{\mathrm{CE}}(w; x, y) = -\sum_{k=1}^{C} y_k \log p_{w,k}(x).$$

Then ℓ_{CE} is Natural.

Proof. For fixed y,

$$\nabla_w \ell_{\mathrm{CE}}(w; x, y) = -\sum_{k=1}^{C} \frac{y_k}{p_{w,k}(x)} \nabla_w p_{w,k}(x).$$

Hence, at $y = p_w(x)$,

$$\nabla_w \ell_{\mathrm{CE}}(w; x, p_w(x)) = -\sum_{k=1}^{C} \nabla_w p_{w,k}(x) = -\nabla_w \sum_{k=1}^{C} p_{w,k}(x) = 0.$$

□

C Illustrative examples: confidence turns heterogeneous teacher reliability into useful signal

We give compact multiclass toy examples that illustrate the qualitative mechanism behind our method in a transparent setting. The point of these examples is not to specialize the method to partial-label teachers, but to isolate a broader idea: confidence identifies where each teacher is reliable, so that heterogeneous teacher quality can be converted into useful variance reduction rather than uniformly trusted or discarded. In the single-teacher case, this lets a partial teacher help on the region where it is accurate without harming learning where it is weak. In the multi-teacher case, the gain can be larger when different teachers are reliable on different regions, while overlap reduces the gain through

Table 3: Summary of the toy examples. Smaller effective variance proxies indicate stronger variance reduction. Here $\pi_j = \Pr(Y \in \mathcal{G}_j)$ and $\tau_j = \pi_j(1-\rho) + (1-\pi_j)\rho$; under perfect routing, $\rho = 0$ and hence $\tau_j = \pi_j$.

Regime	**Reliable labels**	**Proxy comparison and implication**
Single partial teacher	General reliable set $\mathcal{G}$ with $\pi = \Pr(Y \in \mathcal{G})$ $\tau = \pi(1-\rho) + (1-\pi)\rho$.	No confidence: $s_{\text{no-conf}} = \beta^2 \varepsilon_f$. With confidence: $s_{\text{conf}} = \beta^2(\tau\varepsilon_{f,1} + (1-\tau)\varepsilon^0)$. Thus confidence helps iff $\varepsilon^0 \le \varepsilon_{f,0}$.
Perfect partial teacher	Teacher is perfect on $\mathcal{G}$ always wrong on $\mathcal{B} = \mathcal{Y} \setminus \mathcal{G}$; $\rho = 0, m_{\text{G}} = 0, m_{\text{B}} = 1$.	No confidence: $s_{\text{no-conf}} = 2\beta^2(1-\pi)$. With confidence: $s_{\text{conf}} = \beta^2(1-\pi)\varepsilon^0$. Since $\varepsilon^0 \le 2$, confidence makes the partial teacher useful.
Complementary teachers	$\mathcal{G}_1 = \{0,1,2\}$, $\mathcal{G}_2 = \{3,4,5,6,7,8,9\}$. $\pi_1 = \tau_1 = 0.3$ $\pi_2 = \tau_2 = 0.7$.	Best single no-confidence proxy: $0.6\,\beta^2$. Best single confidence proxy: $0.3\,\beta^2\varepsilon^0$. Multi-teacher confidence proxy: $0.21\,\beta^2\varepsilon^0$. Complementarity improves beyond the best single-teacher.
Overlapping teachers,	$\mathcal{G}_1 = \{0,1,2,3,4\}$, $\mathcal{G}_2 = \{1,3,5,7,9\}$. $\pi_1 = \pi_2 = 0.5$ under perfect routing, $\tau_1 = \tau_2 = 0.5$.	Single-teacher confidence proxy: $0.25\,\beta^2\varepsilon^0$. Multi-teacher covariance proxy: $0.10\,\beta^2\varepsilon^0$. Negative covariance strengthens the multi-teacher gain.

covariance. Partial-label structure is one clean way to visualize this mechanism, but the same logic applies more broadly under teacher–student mismatch and reliability shift.

Throughout, let $\mathcal{Y} = \{0, \dots, 9\}$, let $y = e_Y \in \mathbb{R}^{10}$ denote the one-hot true label, and let each teacher output a one-hot pseudo-label $f(x) = e_{\hat{Y}(x)}$. Moreover,

$$\|y - f(x)\|^2 = \begin{cases} 0, & \hat{Y}(x) = Y, \\ 2, & \hat{Y}(x) \neq Y, \end{cases}$$

C.1 Single-model confidence setting

Let $\mathcal{G} \subset \mathcal{Y}$ denote a set of labels on which the teacher is relatively reliable, and let $\mathcal{B} = \mathcal{Y} \setminus \mathcal{G}$ denote the remaining labels. We do not assume balanced mass across these two groups. Instead, we notate

$$\pi := \Pr(Y \in \mathcal{G}), \qquad 1 - \pi = \Pr(Y \in \mathcal{B}).$$

Let the teacher's mistake rates on these two regions be

$$m_{\text{G}} := \Pr(\hat{Y} \neq Y \mid Y \in \mathcal{G}), \qquad m_{\text{B}} := \Pr(\hat{Y} \neq Y \mid Y \in \mathcal{B}),$$

and let the binary confidence model $c(x) \in \{0, 1\}$ have routing error

$$\rho := \Pr(c = 0 \mid Y \in \mathcal{G}) = \Pr(c = 1 \mid Y \in \mathcal{B}).$$

Thus confidence aims to keep the teacher on $\mathcal{G}$ and suppress it on $\mathcal{B}$, but may make mistakes in either direction.

The corresponding teacher-quality quantities are

$$\varepsilon_f = 2\big(\pi m_{\text{G}} + (1-\pi)m_{\text{B}}\big),$$

$$\varepsilon_{f,1} = 2\left(\frac{\pi(1-\rho)m_{\text{G}} + (1-\pi)\rho m_{\text{B}}}{\pi(1-\rho) + (1-\pi)\rho}\right),$$

$$\varepsilon_{f,0} = 2\left(\frac{\pi\rho m_{\text{G}} + (1-\pi)(1-\rho)m_{\text{B}}}{\pi\rho + (1-\pi)(1-\rho)}\right),$$

with

$$\tau := \Pr(c = 1) = \pi(1-\rho) + (1-\pi)\rho.$$

Let ε^0 denote the student's squared-label error on the low-confidence region. Then the confidence-absent and confidence-aware single-model proxies are

$$s_{\text{no-conf}} = \beta^2 \varepsilon_f, \qquad s_{\text{conf}} = \beta^2 \big(\tau \varepsilon_{f,1} + (1-\tau)\varepsilon^0\big).$$

Using $\varepsilon_f = \tau \varepsilon_{f,1} + (1-\tau)\varepsilon_{f,0}$, we obtain

$$s_{\text{conf}} \le s_{\text{no-conf}} \quad \Longleftrightarrow \quad \varepsilon^0 \le \varepsilon_{f,0}.$$

The low-confidence region is exactly where the teacher is weak, so suppressing the teacher there helps once the student is at least competitive on that region. The teacher may become more harmful than helpful if kept indiscriminately. Confidence is therefore a general mechanism for protecting the estimator from unreliable teacher predictions.

Perfect partial teacher. In the stylized extreme case

$$\rho = 0, \qquad m_{\mathrm{G}} = 0, \qquad m_{\mathrm{B}} = 1,$$

the teacher is perfect on its reliable region, always wrong on its unreliable region, and confidence identifies that partition perfectly. Then

$$\varepsilon_{f,1} = 0, \qquad \varepsilon_{f,0} = 2, \qquad \varepsilon_f = 2(1-\pi).$$

Hence

$$s_{\text{conf}} = \beta^2(1-\pi)\varepsilon^0, \qquad s_{\text{no-conf}} = 2\beta^2(1-\pi),$$

so

$$s_{\text{conf}} \le s_{\text{no-conf}} \quad \Longleftrightarrow \quad \varepsilon^0 \le 2.$$

Since 2 is the maximum one-hot squared-label error, this holds automatically for simplex-valued student predictions. Thus even a teacher that is reliable on only part of the task becomes useful once confidence isolates the reliable part, $\frac{s_{\text{conf}}}{s_{\text{no-conf}}} = \frac{\varepsilon^0}{2} \le 1$.

C.2 Two-model confidence setting

We now consider two teachers f_1, f_2, each with its own reliable and unreliable regions. For teacher $j \in \{1,2\}$, let $\mathcal{G}_j$ denote its good-label set and $\mathcal{B}_j = \mathcal{Y} \setminus \mathcal{G}_j$ its bad-label set, and define $\pi_j := \Pr(Y \in \mathcal{G}_j)$. We allow the teachers to have different quality profiles:

$$m_{j,\mathrm{G}} := \Pr(\hat{Y}_j \ne Y \mid Y \in \mathcal{G}_j), \qquad m_{j,\mathrm{B}} := \Pr(\hat{Y}_j \ne Y \mid Y \in \mathcal{B}_j).$$

For simplicity, we keep a shared routing error

$$\rho := \Pr(c_j = 0 \mid Y \in \mathcal{G}_j) = \Pr(c_j = 1 \mid Y \in \mathcal{B}_j), \qquad j \in \{1,2\},$$

and let ε^0 denote the student's low-confidence squared-label error.

Then teacher j contributes the confidence-aware proxy $s_j = \beta^2\Big(\tau_j \varepsilon_{f_j,1} + (1-\tau_j)\varepsilon^0\Big)$, where

$$\tau_j = \pi_j(1-\rho) + (1-\pi_j)\rho, \qquad \varepsilon_{f_j} = 2\big(\pi_j m_{j,\mathrm{G}} + (1-\pi_j) m_{j,\mathrm{B}}\big).$$

The corresponding confidence-absent proxy is $\bar{s}_j := \beta^2 \varepsilon_{f_j}$.

The next formulas separate two effects: the individual usefulness of each teacher through s_1, s_2, and their redundancy through the covariance term.

General covariance-aware form. For a two-teacher confidence-aware covariance envelope

$$\bar{\Sigma}_{CM} = \begin{pmatrix} s_1 & c_{12} \\ c_{12} & s_2 \end{pmatrix}, \qquad 0 \le c_{12} < \sqrt{s_1 s_2}.$$

$$\bar{\Sigma}_{CM}^{-1} = \frac{1}{s_1 s_2 - c_{12}^2} \begin{pmatrix} s_2 & -c_{12} \\ -c_{12} & s_1 \end{pmatrix},$$

and the effective covariance-aware proxy is

$$s^{(2)}_{\text{conf,cov}} := \big(\mathbf{1}^\top \bar{\Sigma}_{CM}^{-1} \mathbf{1}\big)^{-1} = \frac{s_1 s_2 - c_{12}^2}{s_1 + s_2 - 2c_{12}}.$$

Thus multi-teacher gains depend not only on the per-teacher terms s_1, s_2, but also on the cross-teacher covariance c_{12}.

Relation to the diagonal envelope. Let $D = \mathrm{diag}(s_1, s_2)$. If $\bar{\Sigma}_{CM} \preceq \kappa_{CM} D$, then

$$s^{(2)}_{\mathrm{conf,cov}} \le \kappa_{CM} \left(\frac{1}{s_1} + \frac{1}{s_2} \right)^{-1}.$$

For the 2×2 positive-correlation case above,

$$D^{-1/2} \bar{\Sigma}_{CM} D^{-1/2} = \begin{pmatrix} 1 & \frac{c_{12}}{\sqrt{s_1 s_2}} \\ \frac{c_{12}}{\sqrt{s_1 s_2}} & 1 \end{pmatrix},$$

so

$$\kappa_{CM} = 1 + \frac{c_{12}}{\sqrt{s_1 s_2}} \in [1, 2).$$

Thus nearly complementary teachers have $\kappa_{CM} \approx 1$, while stronger overlap pushes κ_{CM} toward the worst-case value 2.

Comparison to the confidence-absent case. Under the diagonal-envelope view, the effective confidence-aware and confidence-absent two-teacher proxies are

$$s^{(2)}_{\mathrm{conf}} = \kappa_{CM} \left(\frac{1}{s_1} + \frac{1}{s_2} \right)^{-1}, \qquad s^{(2)}_{\text{no-conf}} = \kappa_{CM} \left(\frac{1}{\bar{s}_1} + \frac{1}{\bar{s}_2} \right)^{-1}.$$

If $s_j \le \bar{s}_j, \quad j \in \{1, 2\}$, then $s^{(2)}_{\mathrm{conf}} \le s^{(2)}_{\text{no-conf}}$. Thus if confidence improves each teacher individually, then it also improves the combined two-teacher correction.

More generally, for $s^{(2)}_{\text{no-conf,cov}} := \left(\mathbf{1}^\top \bar{\Sigma}^{-1}_{MM} \mathbf{1} \right)^{-1}$. If $\bar{\Sigma}_{CM} \preceq \bar{\Sigma}_{MM}$, then $s^{(2)}_{\mathrm{conf,cov}} \le s^{(2)}_{\text{no-conf,cov}}$. The same monotonicity extends to the full covariance-aware setting whenever confidence shrinks the multi-teacher covariance envelope.

C.2.1 Complementary teachers

Assume the two teachers are complementary:

$$\mathcal{G}_2 = \mathcal{B}_1, \qquad \mathcal{B}_2 = \mathcal{G}_1.$$

This is the favorable regime in which each teacher is reliable where the other is weak. Because the reliable regions are disjoint, the off-diagonal covariance is small, so aggregation is close to the ideal harmonic-mean gain.

In the stylized perfect partial-teacher case $\rho = 0,\ m_{1,\mathrm{G}} = m_{2,\mathrm{G}} = 0,\ m_{1,\mathrm{B}} = m_{2,\mathrm{B}} = 1$, each teacher is perfect on its own good labels and always wrong on its own bad labels.

For a concrete numerical example, assume the labels are uniformly distributed on $\{0, \ldots, 9\}$ and let

$$\mathcal{G}_1 = \{0, 1, 2\}, \qquad \mathcal{B}_1 = \{3, 4, 5, 6, 7, 8, 9\},$$

so that

$$\mathcal{G}_2 = \{3, 4, 5, 6, 7, 8, 9\}, \qquad \mathcal{B}_2 = \{0, 1, 2\}.$$

Then

$$\pi_1 = 0.3, \qquad \pi_2 = 0.7,$$

and hence

$$s_1 = \beta^2 (1 - \pi_1) \varepsilon^0 = 0.7\, \beta^2 \varepsilon^0, \qquad s_2 = \beta^2 (1 - \pi_2) \varepsilon^0 = 0.3\, \beta^2 \varepsilon^0.$$

The corresponding covariance-aware envelope is

$$\bar{\Sigma}_{CM} = \begin{pmatrix} 0.7\, \beta^2 \varepsilon^0 & c_{\mathrm{comp}} \\ c_{\mathrm{comp}} & 0.3\, \beta^2 \varepsilon^0 \end{pmatrix}.$$

Under the idealized complementary-support assumption, the two confidence-weighted correction terms are active on disjoint reliable regions, so $c_{\mathrm{comp}} = 0$. Therefore

$$\bar{\Sigma}_{CM} = \begin{pmatrix} 0.7\, \beta^2 \varepsilon^0 & 0 \\ 0 & 0.3\, \beta^2 \varepsilon^0 \end{pmatrix}, \qquad \kappa_{CM} = 1,$$

and

$$s^{(2)}_{\text{conf,cov}} = \frac{s_1 s_2}{s_1 + s_2} = \frac{(0.7)(0.3)}{0.7 + 0.3}\,\beta^2\varepsilon^0 = 0.21\,\beta^2\varepsilon^0.$$

For comparison, the corresponding confidence-absent terms are

$$\bar{s}_1 = 2\beta^2(1-\pi_1) = 1.4\,\beta^2, \qquad \bar{s}_2 = 2\beta^2(1-\pi_2) = 0.6\,\beta^2,$$

so

$$s^{(2)}_{\text{no-conf}} = \left(\frac{1}{\bar{s}_1} + \frac{1}{\bar{s}_2}\right)^{-1} = 0.42\,\beta^2.$$

Therefore

$$s^{(2)}_{\text{conf}} \le s^{(2)}_{\text{no-conf}} \quad \Longleftrightarrow \quad \varepsilon^0 \le 2,$$

and the confidence-aware two-teacher correction is exactly a factor of $\varepsilon^0/2$ smaller than the confidence-absent one. Relative to the corresponding single-teacher confidence-aware term $s_{\text{conf}} = \frac{\beta^2\varepsilon^0}{2}$, the complementary two-teacher construction reduces the effective correction term further:

$$s^{(2)}_{\text{conf,cov}} = 0.21\,\beta^2\varepsilon^0 < 0.5\,\beta^2\varepsilon^0 = s_{\text{conf}}.$$

Confidence does not only rescue a partial teacher; with complementary teachers it produces a stronger correction than any one partial teacher alone.

C.2.2 Partially overlapping teachers

As a partially overlapping example, take

$$\mathcal{G}_1 = \{0,1,2,3,4\}, \qquad \mathcal{G}_2 = \{1,3,5,7,9\}.$$

Then

$$\mathcal{G}_1 \cap \mathcal{G}_2 = \{1,3\}, \qquad \mathcal{G}_1 \cap \mathcal{B}_2 = \{0,2,4\},$$
$$\mathcal{B}_1 \cap \mathcal{G}_2 = \{5,7,9\}, \qquad \mathcal{B}_1 \cap \mathcal{B}_2 = \{6,8\}.$$

So the two teachers are simultaneously reliable on 2 labels, each has 3 labels on which it is reliable while the other is not, and they are jointly unreliable on 2 labels. Under the uniform label distribution,

$$\Pr(\mathcal{G}_1 \cap \mathcal{G}_2) = 0.2, \qquad \Pr(\mathcal{G}_1 \cap \mathcal{B}_2) = 0.3,$$
$$\Pr(\mathcal{B}_1 \cap \mathcal{G}_2) = 0.3, \qquad \Pr(\mathcal{B}_1 \cap \mathcal{B}_2) = 0.2.$$

Thus each teacher still has a reliable subset and an unreliable subset, but now those reliable regions overlap. This overlap introduces redundancy, although the sign of the resulting covariance is not determined by overlap alone. Overlap changes the covariance structure, and that covariance determines the additional gain available from combining teachers.

The optimized multi-teacher objective is still protected against harmful combinations, since the simplex of aggregation weights contains each single-teacher solution as a special case. Thus overlap does not imply that the multi-teacher method must underperform a single-teacher; rather, it reduces the *additional* gain available from combining teachers when the resulting covariance is less favorable.

To make this explicit in a minimal proxy model, consider the stylized scalar activation variables

$$X_1 = \mathbf{1}\{Y \in \mathcal{G}_1\}, \qquad X_2 = \mathbf{1}\{Y \in \mathcal{G}_2\}.$$

Since

$$\Pr(Y \in \mathcal{G}_1) = \Pr(Y \in \mathcal{G}_2) = 0.5, \qquad \Pr(Y \in \mathcal{G}_1 \cap \mathcal{G}_2) = 0.2,$$

we obtain

$$\mathbb{E}[X_1] = \mathbb{E}[X_2] = 0.5, \qquad \mathbb{E}[X_1 X_2] = 0.2,$$

and therefore

$$\mathrm{Var}(X_1) = \mathrm{Var}(X_2) = 0.25, \qquad \mathrm{Cov}(X_1, X_2) = 0.2 - (0.5)(0.5) = -0.05.$$

This yields the covariance matrix

$$\bar{\Sigma} = \begin{pmatrix} 0.25 & -0.05 \\ -0.05 & 0.25 \end{pmatrix},$$

which is positive definite since its eigenvalues are $0.25 \pm 0.05 = 0.30,\ 0.20$.

The corresponding covariance-aware effective proxy is

$$s^{(2)}_{\text{conf,cov}} = \left(\mathbf{1}^\top \bar{\Sigma}^{-1} \mathbf{1}\right)^{-1} = \frac{0.25 - 0.05}{2} = 0.10.$$

For comparison, the uncorrelated case would give $\frac{0.25}{2} = 0.125$.

So in this stylized activation model, the effective proxy is smaller than in the uncorrelated case, despite the geometric overlap in the reliable regions.

D THEORETICAL ANALYSIS OF PREDICTION-POWERED GRADIENTS

In this section, we establish the variance bounds of the prediction-powered gradients proposed, and their convergence rate bounds.

D.1 PROOF OF LEMMA 4.2

Proof. We compute a stochastic gradient estimate of the loss, while using the next notations

$$g^n = \frac{1}{n} \sum_{i=1}^{n} \nabla \ell(w; x^i, y^i)$$

$$g^{n,f} = \frac{1}{n} \sum_{i=1}^{n} c(x^i) \cdot \nabla \ell(w; x^i, f(x^i))$$

$$g^{N,f} = \frac{1}{N} \sum_{i=1}^{N} c(\tilde{x}^i) \cdot \nabla \ell(w; \tilde{x}^i, f(\tilde{x}^i))$$

g^n represents a standard mini batch gradient estimator based on n labeled samples.
$g^{n,f}$ replaces the true labels with predictions from the model f and weights them according to the prediction confidence.
$g^{N,f}$ is the confidence weighted gradient based on the unlabeled samples with pseudo-labels generated by f.

The gradient estimates defined as

$$g^\lambda_{CPP} = g^n + \lambda(g^{N,f} - g^{n,f}) = (1-\lambda) g^n + \lambda g^1_{CPP}, \quad g^1_{\text{CPP}} = g^n + g^{N,f} - g^{n,f}$$

for some weighting parameter $\lambda \in [0, 1]$. We can decompose g^1_{CPP} as follows

$$\begin{aligned} g^1_{\text{CPP}} &= g^N + (g^{N,f} - g^N) - (\nabla \mathcal{L}^f_c(w) - \nabla \mathcal{L}(w)) + (g^n - g^{n,f}) - (\nabla \mathcal{L}(w) - \nabla \mathcal{L}^f_c(w)) \\ &= g^N + (g^{N,f} - g^N) + \nabla \mathcal{L}^f_{e,c}(w) + (g^n - g^{n,f}) - \nabla \mathcal{L}^f_{e,c}(w) \end{aligned}$$

where $g^N = \frac{1}{N} \sum_{i=1}^{N} \nabla \ell(w; \tilde{x}^i, \tilde{y}^i)$ for $\tilde{y}^i$ the true unknown label of $\tilde{x}^i$, and $\mathcal{L}^f_c(w) = \mathbb{E}_{x \sim P_X}[c(x) \ell(w; x, f(x))]$, $\nabla \mathcal{L}^f_{e,c}(w) = \nabla \mathcal{L}(w) - \nabla \mathcal{L}^f_c(w)$.

The variance is

$$\begin{aligned} \text{Var}(g^\lambda_{\text{CPP}}) &= \mathbb{E} \left\| g^\lambda_{\text{CPP}} - \nabla \mathcal{L}(w) \right\|^2 \\ &= \mathbb{E} \left\| (1-\lambda)(g^n - \nabla \mathcal{L}(w)) + \lambda (g^1_{\text{CPP}} - \nabla \mathcal{L}(w)) \right\|^2 \\ &= \mathbb{E} \Big\| (1-\lambda)(g^n - \nabla \mathcal{L}(w)) + \lambda (g^N - \nabla \mathcal{L}(w)) \\ &\quad + \lambda (g^{N,f} - g^N + \nabla \mathcal{L}(w) - \nabla \mathcal{L}^f_c(w)) + \lambda (g^n - g^{n,f} - \nabla \mathcal{L}(w) + \nabla \mathcal{L}^f_c(w)) \Big\|^2 \end{aligned}$$

The inequality $\|u + v + w + z\|^2 \le 4(\|u\|^2 + \|v\|^2 + \|w\|^2 + \|z\|^2)$ gives

$$\begin{aligned} \text{Var}(g^\lambda_{\text{CPP}}) \le\ & 4(1-\lambda)^2 \, \mathbb{E} \|g^n - \nabla \mathcal{L}(w)\|^2 \\ & + 4\lambda^2 \, \mathbb{E} \|g^N - \nabla \mathcal{L}(w)\|^2 \\ & + 4\lambda^2 \, \mathbb{E} \|g^N - g^{N,f} - \nabla \mathcal{L}(w) + \nabla \mathcal{L}^f_c(w)\|^2 \\ & + 4\lambda^2 \, \mathbb{E} \|g^n - g^{n,f} - \nabla \mathcal{L}(w) + \nabla \mathcal{L}^f_c(w)\|^2. \end{aligned}$$

Based on the assumption on bounded variance of the instance-dependent loss gradient

$$\mathbb{E}\|g^n - \nabla\mathcal{L}(w)\|^2 \leq \tfrac{\sigma^2}{n}, \quad \mathbb{E}\|g^N - \nabla\mathcal{L}(w)\|^2 \leq \tfrac{\sigma^2}{N},$$

For the correction terms, we bound the centered second moments as follows:

$$\begin{aligned}
&\mathbb{E}[\|g^N - g^{N,f} - \nabla\mathcal{L}(w) + \nabla\mathcal{L}_c^f(w)\|^2] \\
&= \mathbb{E}\left[\left\|\frac{1}{N}\sum_{i=1}^N \nabla\ell(w;\tilde{x}^i,\tilde{y}^i) - \frac{1}{N}\sum_{i=1}^N c(\tilde{x}^i)\nabla\ell(w;\tilde{x}^i,f(\tilde{x}^i)) - \nabla\mathcal{L}(w) + \nabla\mathcal{L}_c^f(w)\right\|^2\right] \\
&= \frac{1}{N^2}\mathbb{E}\left[\left\|\sum_{i=1}^N \left(\nabla\ell(w;\tilde{x}^i,\tilde{y}^i) - c(\tilde{x}^i)\nabla\ell(w;\tilde{x}^i,f(\tilde{x}^i)) - \nabla\mathcal{L}(w) + \nabla\mathcal{L}_c^f(w)\right)\right\|^2\right]
\end{aligned}$$

The inequality $\|\sum_{j=1}^N u_j\|^2 \leq N\sum_{j=1}^N \|u_j\|^2$ gives

$$\leq \frac{1}{N}\mathbb{E}[\|\nabla\ell(w;\tilde{x}^i,\tilde{y}^i) - c(\tilde{x}^i)\nabla\ell(w;\tilde{x}^i,f(\tilde{x}^i)) - \nabla\mathcal{L}(w) + \nabla\mathcal{L}_c^f(w)\|^2] \tag{12}$$

The variance is smaller than the second moment

$$\leq \frac{1}{N}\mathbb{E}[\|\nabla\ell(w;\tilde{x}^i,\tilde{y}^i) - c(\tilde{x}^i)\nabla\ell(w;\tilde{x}^i,f(\tilde{x}^i))\|^2]$$

Suppose the loss function is Natural (Definition 3.2). Since $c(\tilde{x}^i) \in \{0,1\}$,

$$\begin{aligned}
c(\tilde{x}^i)\nabla\ell(w;\tilde{x}^i,f(\tilde{x}^i)) &= c(\tilde{x}^i)\nabla\ell(w;\tilde{x}^i,f(\tilde{x}^i)) + (1-c(\tilde{x}^i))\nabla\ell(w;\tilde{x}^i,f_w(\tilde{x}^i)) \\
&= \nabla\ell\big(w;\tilde{x}^i,\, c(\tilde{x}^i)f(\tilde{x}^i) + (1-c(\tilde{x}^i))f_w(\tilde{x}^i)\big)\,.
\end{aligned}$$

Here, f_w is the student training model with the weights w.

$$\mathbb{E}\big[\|g^N - g^{N,f} - \nabla\mathcal{L}(w) + \nabla\mathcal{L}_c^f(w)\|^2\big] \leq \frac{1}{N}\mathbb{E}\Big[\big\|\nabla\ell(w;\tilde{x}^i,\tilde{y}^i) - \nabla\ell\big(w;\tilde{x}^i,\, c(\tilde{x}^i)f(\tilde{x}^i) + (1-c(\tilde{x}^i))f_w(\tilde{x}^i)\big)\big\|^2\Big]\,.$$

By label-Lipschitzness,

$$\leq \frac{1}{N}\beta_Y^2\mathbb{E}[\|\tilde{y}^i - (c(\tilde{x}^i)f(\tilde{x}^i) + (1-c(\tilde{x}^i))f_w(\tilde{x}^i))\|^2]$$

From the law of total expectation

$$\begin{aligned}
&= \frac{1}{N}\beta_Y^2(\tau\mathbb{E}[\|\tilde{y}^i - f(\tilde{x}^i)\|^2|c(\tilde{x}^i)=1] + (1-\tau)\mathbb{E}[\|\tilde{y}^i - f_w(\tilde{x}^i)\|^2|c(\tilde{x}^i)=0]) \\
&= \frac{1}{N}\beta_Y^2(\tau\varepsilon^{f,1} + (1-\tau)\varepsilon^0)
\end{aligned}$$

Similarly, we get

$$\mathbb{E}\|g^n - g^{n,f} - \nabla\mathcal{L}_{e,c}^f(w)\|^2 \leq \beta_Y^2(\tau\varepsilon^{f,1} + (1-\tau)\varepsilon^0)$$

Substitution yields

$$\mathrm{Var}\big(g_{\mathrm{CPP}}^\lambda\big) \leq 4\left((1-\lambda)^2\frac{\sigma^2}{n} + \lambda^2\left[\frac{\sigma^2}{N} + \frac{\tau\beta_Y^2\varepsilon^{f,1} + (1-\tau)\beta_Y^2\varepsilon^0}{N} + \frac{\tau\beta_Y^2\varepsilon^{f,1} + (1-\tau)\beta_Y^2\varepsilon^0}{n}\right]\right).$$

□

Variance bound without loss assumptions. Another bound decomposition without an assumption on the loss is possible.

The right-hand side of equation 12 is a variance expression

$$= \frac{1}{N}\mathrm{Var}(\nabla\ell - c\nabla\ell^f)$$

From the law of total variance

$$
\begin{aligned}
&= \frac{1}{N}(\mathrm{Var}(\mathbb{E}[\nabla\ell - c\nabla\ell^f|c]) + \mathbb{E}[\mathrm{Var}(\nabla\ell - c\nabla\ell^f|c)]) \\
&= \frac{1}{N}(\mathrm{Var}(\mathbb{E}[\nabla\ell - c\nabla\ell^f|c]) + \tau\mathrm{Var}(\nabla\ell - \nabla\ell^f|c=1) + (1-\tau)\mathrm{Var}(\nabla\ell|c=0)) \\
&\leq \frac{1}{N}(\mathrm{Var}(\mathbb{E}[\nabla\ell - c\nabla\ell^f|c]) + \tau\sigma_{e,1}^2 + (1-\tau)\sigma_0^2) \\
&= \frac{1}{N}(\mathrm{Var}(\nabla\mathcal{L}^c - c\nabla\mathcal{L}^{f,c}) + \tau\sigma_{e,1}^2 + (1-\tau)\sigma_0^2) \\
&= \frac{1}{N}(\tau(1-\tau)\|\nabla\mathcal{L}^1 - \nabla\mathcal{L}^{f,1} - \nabla\mathcal{L}^0\|^2 + \tau\sigma_{e,1}^2 + (1-\tau)\sigma_0^2)
\end{aligned}
$$

Here, the last step is the expectation definition

$$
\begin{aligned}
\mathrm{Var}(\nabla\mathcal{L}^c - c\nabla\mathcal{L}^{f,c}) &= \mathbb{E}[\|\nabla\mathcal{L}^c - c\nabla\mathcal{L}^{f,c} - \nabla\mathcal{L} + \tau\nabla\mathcal{L}^{f,1}\|^2] \\
&= \tau\|\nabla\mathcal{L}^1 - \nabla\mathcal{L}^{f,1} - \nabla\mathcal{L} + \tau\nabla\mathcal{L}^{f,1}\|^2 + (1-\tau)\|\nabla\mathcal{L}^0 - \nabla\mathcal{L} + \tau\nabla\mathcal{L}^{f,1}\|^2 \\
&= \tau\|(1-\tau)\nabla\mathcal{L}^1 - (1-\tau)\nabla\mathcal{L}^{f,1} - (1-\tau)\nabla\mathcal{L}^0\|^2 + (1-\tau)\|\tau\nabla\mathcal{L}^1 - \tau\nabla\mathcal{L}^0 - \tau\nabla\mathcal{L}^{f,1}\|^2 \\
&= \tau(1-\tau)\|\nabla\mathcal{L}^1 - \nabla\mathcal{L}^{f,1} - \nabla\mathcal{L}^0\|^2 \\
&= \tau(1-\tau)\|\nabla\mathcal{L}_e^{f,1} - \nabla\mathcal{L}^0\|^2
\end{aligned}
$$

When the confidence is high, the pseudo-labels are mostly correct and the difference $\nabla\mathcal{L}_e^{f,1} = \nabla\mathcal{L}^1 - \nabla\mathcal{L}^{f,1}$ is close to zero.

We get the bound

$$
\mathbb{E}\|g^N - g^{N,f} - \nabla\mathcal{L}_{e,c}^f(w)\|^2 \leq \frac{\tau\sigma_{e,1}^2 + (1-\tau)\sigma_0^2 + \tau(1-\tau)\|\nabla\mathcal{L}_e^{f,1} - \nabla\mathcal{L}^0\|^2}{N}
$$

Similarly, we get

$$
\mathbb{E}\|g^n - g^{n,f} - \nabla\mathcal{L}_{e,c}^f(w)\|^2 \leq \frac{\tau\sigma_{e,1}^2 + (1-\tau)\sigma_0^2 + \tau(1-\tau)\|\nabla\mathcal{L}_e^{f,1} - \nabla\mathcal{L}^0\|^2}{n}
$$

Substitution yields

$$
\begin{aligned}
\mathrm{Var}\big(g_{\mathrm{CPP}}^{\lambda}\big) \leq 4\Bigg((1-\lambda)^2\frac{\sigma^2}{n} + \lambda^2\Bigg[&\frac{\sigma^2}{N} \\
&+ \frac{\tau\sigma_{e,1}^2 + (1-\tau)\sigma_0^2 + \tau(1-\tau)\left\|\nabla\mathcal{L}_e^{f,1} - \nabla\mathcal{L}^0\right\|^2}{N} \\
&+ \frac{\tau\sigma_{e,1}^2 + (1-\tau)\sigma_0^2 + \tau(1-\tau)\left\|\nabla\mathcal{L}_e^{f,1} - \nabla\mathcal{L}^0\right\|^2}{n}\Bigg]\Bigg).
\end{aligned}
$$

D.2 Proof of Lemma 5.2

Proof. We bound the variance of the multi-model prediction-powered gradient using the following notations:

$$
g^n = \frac{1}{n}\sum_{i=1}^n \nabla\ell(w; x^i, y^i), \qquad g^N = \frac{1}{N}\sum_{i=1}^N \nabla\ell(w; \tilde{x}^i, \tilde{y}^i),
$$

$$
g^{n,f} = \sum_{j=0}^K \lambda_j g^{n,f_j}, \qquad g^{N,f} = \sum_{j=0}^K \lambda_j g^{N,f_j},
$$

$$
g^{n,f_j} = \frac{1}{n}\sum_{i=1}^n c_j(x^i)\nabla\ell(w; x^i, f_j(x^i)), \qquad g^{N,f_j} = \frac{1}{N}\sum_{i=1}^N c_j(\tilde{x}^i)\nabla\ell(w; \tilde{x}^i, f_j(\tilde{x}^i)),
$$

with $g^{n,f_0} = g^{N,f_0} = 0$.

Here g^n is the labeled mini-batch gradient estimator and g^N is the corresponding hypothetical true-label estimator on the unlabeled batch. $g^{n,f}$, $g^{N,f}$ replace the true labels with predictions from the models f and weights them.

The CM-PP gradient estimator is

$$g^{\lambda}_{\mathrm{CMPP}} = g^n + g^{N,f} - g^{n,f} = g^n + \sum_{j=0}^{K} \lambda_j \big(g^{N,f_j} - g^{n,f_j}\big) = \lambda_0 g^n + \sum_{j=1}^{K} \lambda_j \big(g^n + g^{N,f_j} - g^{n,f_j}\big).$$

For each teacher $j \geq 1$, we define

$$\nabla \mathcal{L}^{f_j}_{e,c}(w) := \nabla \mathcal{L}(w) - \nabla \mathcal{L}^{f_j}_{c}(w).$$

Then

$$g^n + g^{N,f_j} - g^{n,f_j} - \nabla \mathcal{L}(w) = \big(g^N - \nabla \mathcal{L}(w)\big) + \big(g^{N,f_j} - g^N + \nabla \mathcal{L}^{f_j}_{e,c}(w)\big) + \big(g^n - g^{n,f_j} - \nabla \mathcal{L}^{f_j}_{e,c}(w)\big).$$

The variance is

$$\begin{aligned}
\mathrm{Var}(g^{\lambda}_{\mathrm{CMPP}}) &= \mathbb{E}[\|g^{\lambda}_{CMPP} - \nabla \mathcal{L}(w)\|^2] \\
&= \mathbb{E}\big\|\lambda_0(g^n - \nabla \mathcal{L}(w)) + \sum_{j=1}^{K} \lambda_j (g^n + g^{N,f_j} - g^{n,f_j} - \nabla \mathcal{L}(w)))\big\|^2 \\
&= \mathbb{E}\big\|\lambda_0(g^n - \nabla \mathcal{L}(w)) \\
&+ \sum_{j=1}^{K} \lambda_j (g^n - g^{n,f_j} - \nabla \mathcal{L}(w) + \nabla \mathcal{L}^{f_j}_{c}(w)) \\
&+ \sum_{j=1}^{K} \lambda_j (g^{N,f_j} - g^N + \nabla \mathcal{L}(w) - \nabla \mathcal{L}^{f_j}_{c}(w)) \\
&+ \sum_{j=1}^{K} \lambda_j (g^N - \nabla \mathcal{L}(w)))\big\|^2
\end{aligned}$$

The inequality $\|u+v+w+z\|^2 \leq 4(\|u\|^2 + \|v\|^2 + \|w\|^2 + \|z\|^2)$ gives

$$\begin{aligned}
\mathrm{Var}(g^{\lambda}_{\mathrm{CMPP}}) \leq\ & 4\lambda_0^2\, \mathbb{E}\|g^n - \nabla \mathcal{L}(w)\|^2 \\
&+ 4(1-\lambda_0)^2\, \mathbb{E}\|g^N - \nabla \mathcal{L}(w)\|^2 \\
&+ 4\, \mathbb{E}\Big\|\sum_{j=1}^{K} \lambda_j \big(g^{N,f_j} - g^N + \nabla \mathcal{L}^{f_j}_{e,c}(w)\big)\Big\|^2 \\
&+ 4\, \mathbb{E}\Big\|\sum_{j=1}^{K} \lambda_j \big(g^n - g^{n,f_j} - \nabla \mathcal{L}^{f_j}_{e,c}(w)\big)\Big\|^2.
\end{aligned}$$

Writing the final labeled correction term in vector form,

$$\mathbb{E}\left[\left\|\sum_{j=1}^{K} \lambda_j \big(g^n - g^{n,f_j} - \nabla \mathcal{L}^{f_j}_{e,c}(w)\big)\right\|^2\right] = \mathbb{E}\big[\lambda_{1:K}^{\top} A \lambda_{1:K}\big],$$

where, for $j, k \in [K]$,

$$\begin{aligned}
A_{jk} &= \Big\langle g^n - g^{n,f_j} - \nabla\mathcal{L}^{f_j}_{e,c}(w),\, g^n - g^{n,f_k} - \nabla\mathcal{L}^{f_k}_{e,c}(w)\Big\rangle \\
&= \left\langle \frac{1}{n}\sum_{a=1}^{n}\left(\nabla\ell^{f_j}_{e,c}(w;x^a,y^a) - \nabla\mathcal{L}^{f_j}_{e,c}(w)\right), \frac{1}{n}\sum_{b=1}^{n}\left(\nabla\ell^{f_k}_{e,c}(w;x^b,y^b) - \nabla\mathcal{L}^{f_k}_{e,c}(w)\right)\right\rangle \\
&= \frac{1}{n^2}\left\langle \sum_{a=1}^{n}\left(\nabla\ell^{f_j}_{e,c}(w;x^a,y^a) - \nabla\mathcal{L}^{f_j}_{e,c}(w)\right), \sum_{b=1}^{n}\left(\nabla\ell^{f_k}_{e,c}(w;x^b,y^b) - \nabla\mathcal{L}^{f_k}_{e,c}(w)\right)\right\rangle \\
&= \frac{1}{n^2}\sum_{a=1}^{n}\sum_{b=1}^{n}\Big\langle \nabla\ell^{f_j}_{e,c}(w;x^a,y^a) - \nabla\mathcal{L}^{f_j}_{e,c}(w), \nabla\ell^{f_k}_{e,c}(w;x^b,y^b) - \nabla\mathcal{L}^{f_k}_{e,c}(w)\Big\rangle.
\end{aligned}$$

Taking expectations gives

$$\begin{aligned}
\mathbb{E}[A_{jk}] &= \frac{1}{n^2}\sum_{a=1}^{n}\sum_{b=1}^{n}\mathbb{E}\Big[\Big\langle \nabla\ell^{f_j}_{e,c}(w;x^a,y^a) - \nabla\mathcal{L}^{f_j}_{e,c}(w), \nabla\ell^{f_k}_{e,c}(w;x^b,y^b) - \nabla\mathcal{L}^{f_k}_{e,c}(w)\Big\rangle\Big] \\
&= \frac{1}{n^2}\sum_{a=1}^{n}\mathbb{E}\Big[\Big\langle \nabla\ell^{f_j}_{e,c}(w;x^a,y^a) - \nabla\mathcal{L}^{f_j}_{e,c}(w), \nabla\ell^{f_k}_{e,c}(w;x^a,y^a) - \nabla\mathcal{L}^{f_k}_{e,c}(w)\Big\rangle\Big] \\
&= \frac{1}{n^2}\sum_{a=1}^{n}[\Sigma_{\mathrm{CM}}(w)]_{jk} \\
&= \frac{1}{n}[\Sigma_{\mathrm{CM}}(w)]_{jk},
\end{aligned}$$

where the second equality follows because the samples are independent and the centered terms have mean zero, so all terms with $a \neq b$ vanish. Therefore,

$$\begin{aligned}
\mathbb{E}\big[\lambda_{1:K}^\top A\lambda_{1:K}\big] &= \sum_{j=1}^{K}\sum_{k=1}^{K}\lambda_j\lambda_k\mathbb{E}[A_{jk}] \\
&= \frac{1}{n}\sum_{j=1}^{K}\sum_{k=1}^{K}\lambda_j\lambda_k[\Sigma_{\mathrm{CM}}(w)]_{jk} \\
&= \frac{1}{n}\lambda_{1:K}^\top\Sigma_{\mathrm{CM}}(w)\lambda_{1:K} \\
&\leq \frac{1}{n}\lambda_{1:K}^\top\overline{\Sigma}_{\mathrm{CM}}\lambda_{1:K}.
\end{aligned}$$

Similarly

$$\mathbb{E}\Big\|\sum_{j=1}^{K}\lambda_j\big(g^{N,f_j} - g^N + \nabla\mathcal{L}^{f_j}_{e,c}(w)\big)\Big\|^2 = \frac{1}{N}\lambda_{1:K}^\top\Sigma_{\mathrm{CM}}(w)\lambda_{1:K} \leq \frac{1}{N}\lambda_{1:K}^\top\overline{\Sigma}_{\mathrm{CM}}\lambda_{1:K}$$

By Assumption 3.5,

$$\mathbb{E}\|g^n - \nabla\mathcal{L}(w)\|^2 \leq \frac{\sigma^2}{n}, \qquad \mathbb{E}\|g^N - \nabla\mathcal{L}(w)\|^2 \leq \frac{\sigma^2}{N}.$$

Therefore,

$$\mathrm{Var}(g^\lambda_{\mathrm{CMPP}}) \leq 4\left(\lambda_0^2\frac{\sigma^2}{n} + (1-\lambda_0)^2\frac{\sigma^2}{N} + \Big(\frac{1}{n} + \frac{1}{N}\Big)\lambda_{1:K}^T\overline{\Sigma}_{\mathrm{CM}}\lambda_{1:K}\right),$$

which proves the claim. □

D.3 Proof of Theorem 7.2

The convergence analysis follows the same high-level structure as the proof of Shoham et al. (2026). The model parameters are updated by AdaGrad-Norm using the prediction-powered gradient estimator, while the interpolation weights are updated over the simplex using exponentiated gradient, equivalently multiplicative weights (Kivinen & Warmuth, 1997).

Proof. Recall that we defined

$$h_t(\lambda) := \|g_t^{\lambda}\|^2$$

And we use the AdaGrad step size to update the student's weights w_t

$$\eta_t = \eta_0 \Bigg(\sum_{s=1}^{t} \|g_s^{\lambda_s}\|^2 \Bigg)^{-1/2}, \; w_{t+1} = w_t - \eta_t g_t^{\lambda_t}$$

Given this AdaGrad use, we can employ lemma (see Lemma G.3 in Dorfman et al. (2024))

Lemma D.1. *For a β smooth and M-bounded function, the iterates $w_1, \cdots, w_T$ defined by the AdaGrad Norm (presented above) update rule satisfy:*

$$\sum_{t=1}^{T} \|\nabla \mathcal{L}(w_t)\|^2 \leq \left(\frac{2M}{\eta_0} + \eta_0 \beta \right) \sqrt{\sum_{t=1}^{T} \|g_t^{\lambda_t}\|^2} + \sum_{t=1}^{T} (\nabla \mathcal{L}(w_t) - g_t^{\lambda_t})^T \nabla \mathcal{L}(w_t)$$

Taking expectation of the bound above

$$\sum_{t=1}^{T} \mathbb{E}[\|\nabla \mathcal{L}(w_t)\|^2] \leq \left(\frac{2M}{\eta_0} + \eta_0 \beta \right) \mathbb{E}\left[\sqrt{\sum_{t=1}^{T} \|g_t^{\lambda_t}\|^2} \right] + \sum_{t=1}^{T} \mathbb{E}\left[(\nabla \mathcal{L}(w_t) - g_t^{\lambda_t})^\top \nabla \mathcal{L}(w_t) \right].$$

Since $g_t^{\lambda_t}$ is conditionally unbiased given the history,

$$\mathbb{E}_{t-1}[g_t^{\lambda_t}] = \nabla \mathcal{L}(w_t),$$

and $\nabla \mathcal{L}(w_t)$ is $\mathcal{F}_{t-1}$-measurable, the second term vanishes.

$$\begin{aligned} \sum_{t=1}^{T} \mathbb{E}[\|\nabla \mathcal{L}(w_t)\|^2] &\leq \left(\frac{2M}{\eta_0} + \eta_0 \beta \right) \mathbb{E}\left[\sqrt{\sum_{t=1}^{T} \|g_t^{\lambda_t}\|^2} \right] \\ &= \left(\frac{2M}{\eta_0} + \eta_0 \beta \right) \mathbb{E}\left[\sqrt{\sum_{t=1}^{T} h_t(\lambda_t)} \right] \end{aligned}$$

According to the definition of regret $\mathcal{R}_T^h(\lambda^\star) = \sum_{t=1}^{T} h_t(\lambda_t) - \sum_{t=1}^{T} h_t(\lambda^\star)$

$$\sum_{t=1}^{T} \mathbb{E}[\|\nabla \mathcal{L}(w_t)\|^2] \leq \left(\frac{2M}{\eta_0} + \eta_0 \beta \right) \mathbb{E}\left[\sqrt{\mathcal{R}_T^h(\lambda^\star) + \sum_{t=1}^{T} h_t(\lambda^\star)} \right]$$

We update λ_t using exponentiated gradient, so we can bound the regret $\mathcal{R}_T^h(\lambda^\star)$ as presented in lemma D.2 (proved in section D.4)

Lemma D.2 (Exponentiated-gradient regret bound)**.** *Assume the stochastic gradients are G-bounded; that is, for any $w \in \mathbb{R}^d$ and $(x, y) \in \mathcal{X} \times \mathcal{Y}$, we have $\|\nabla \ell(w; x, y)\| \leq G$. Consider the parameter vector $\lambda \in \mathcal{D} = \Delta_{K+1}$, where*

$$\Delta_{K+1} = \left\{ \lambda \in \mathbb{R}_+^{K+1} : \sum_{j=0}^{K} \lambda_j = 1 \right\}.$$

At each round t, define

$$h_t(\lambda) = \left\| g_t + \sum_{j=1}^{K} \lambda_j d_{t,j} \right\|^2,$$

where $d_{t,j}$ is the per-teacher correction term and $d_{t,0} = 0$. Let λ_t be updated by exponentiated gradient, equivalently multiplicative weights,

$$\lambda_{t+1,j} = \frac{\lambda_{t,j} \exp(-\gamma_t [\nabla h_t(\lambda_t)]_j)}{\sum_{k=0}^{K} \lambda_{t,k} \exp(-\gamma_t [\nabla h_t(\lambda_t)]_k)}, \qquad j = 0, \dots, K.$$

We use the adaptive stepsize

$$\gamma_t = \sqrt{\frac{2\log(K+1)}{\delta + \sum_{s=1}^{t} \|\nabla h_s(\lambda_s)\|_\infty^2}},$$

where $\delta > 0$ is a small numerical constant. Then, for any $\lambda^\star \in \Delta_{K+1}$,

$$R_T^h(\lambda^\star) := \sum_{t=1}^{T} h_t(\lambda_t) - \sum_{t=1}^{T} h_t(\lambda^\star)$$

satisfies

$$R_T^h(\lambda^\star) \le 128 G^2 \log(K+1) + 8\sqrt{2}\, G \sqrt{\log(K+1)} \sqrt{\sum_{t=1}^{T} h_t(\lambda^\star)}.$$

Implementing this regret bound, we get

$$\sum_{t=1}^{T} \mathbb{E}[\|\nabla \mathcal{L}(w_t)\|^2] \le \left(\frac{2M}{\eta_0} + \eta_0 \beta\right) \mathbb{E}\left[\sqrt{128 G^2 \log(K+1) + 8\sqrt{2}\, G \sqrt{\log(K+1)} \sqrt{\sum_{t=1}^{T} h_t(\lambda^\star)} + \sum_{t=1}^{T} h_t(\lambda^\star)}\right].$$

The two last terms under the square root can be written as $C\sqrt{A} + A$, where

$$A = \sum_{t=1}^{T} h_t(\lambda^\star), \qquad C = 8\sqrt{2}\, G \sqrt{\log(K+1)}.$$

Using the helper inequality

$$C\sqrt{A} + A \le 2C^2 + 2A,$$

we obtain

$$\sum_{t=1}^{T} \mathbb{E}[\|\nabla \mathcal{L}(w_t)\|^2] \le \left(\frac{2M}{\eta_0} + \eta_0 \beta\right) \mathbb{E}\left[\sqrt{384 G^2 \log(K+1) + 2\sum_{t=1}^{T} h_t(\lambda^\star)}\right].$$

Using $\sqrt{a+b} \le \sqrt{a} + \sqrt{b}$, we further obtain

$$\sum_{t=1}^{T} \mathbb{E}[\|\nabla \mathcal{L}(w_t)\|^2] \le \left(\frac{2M}{\eta_0} + \eta_0 \beta\right) 8\sqrt{6}\, G \sqrt{\log(K+1)} + \left(\frac{2M}{\eta_0} + \eta_0 \beta\right) \mathbb{E}\left[\sqrt{2\sum_{t=1}^{T} h_t(\lambda^\star)}\right].$$

By Jensen's inequality,

$$\sum_{t=1}^{T} \mathbb{E}[\|\nabla \mathcal{L}(w_t)\|^2] \le \left(\frac{2M}{\eta_0} + \eta_0 \beta\right) 8\sqrt{6}\, G \sqrt{\log(K+1)} + \left(\frac{2M}{\eta_0} + \eta_0 \beta\right) \sqrt{2\sum_{t=1}^{T} \mathbb{E}[h_t(\lambda^\star)]}.$$

This expectation can be bounded using Lemma D.3.

Lemma D.3 (Oracle squared-norm bound)**.** *For the optimal comparator $\lambda^\star$, and $h_t(\lambda) := \|g_t^\lambda\|^2$,*

$$\mathbb{E}[h_t(\lambda^\star)] \le V^\star + \mathbb{E}[\|\nabla \mathcal{L}(w_t)\|^2],$$

where $V^\star$ is the optimal multi-model variance bound.

We get

$$\begin{aligned}\sum_{t=1}^{T} \mathbb{E}[\|\nabla \mathcal{L}(w_t)\|^2] &\le \left(\frac{2M}{\eta_0} + \eta_0\beta\right) 8\sqrt{6}\, G\sqrt{\log(K+1)} \\ &\quad + \left(\frac{2M}{\eta_0} + \eta_0\beta\right)\sqrt{2V^\star T + 2\sum_{t=1}^{T} \mathbb{E}[\|\nabla \mathcal{L}(w_t)\|^2]} \\ &\le \left(\frac{2M}{\eta_0} + \eta_0\beta\right) 8\sqrt{6}\, G\sqrt{\log(K+1)} \\ &\quad + \left(\frac{2M}{\eta_0} + \eta_0\beta\right)\left(\sqrt{2V^\star T} + \sqrt{2\sum_{t=1}^{T} \mathbb{E}[\|\nabla \mathcal{L}(w_t)\|^2]}\right).\end{aligned}$$

This expression can be simplified using Lemma D.4.

Lemma D.4. *Let $A, B, C > 0$. If $A \le B\sqrt{A} + C$, then $A \le 4B^2 + 2C$.*

Let

$$S_T := \sum_{t=1}^{T} \mathbb{E}[\|\nabla \mathcal{L}(w_t)\|^2], \qquad C_\eta := \frac{2M}{\eta_0} + \eta_0\beta.$$

Applying Lemma D.4 with

$$A = S_T, \qquad B = \sqrt{2}\, C_\eta, \qquad C = 8\sqrt{6}\, C_\eta G\sqrt{\log(K+1)} + C_\eta\sqrt{2V^\star T},$$

gives

$$\sum_{t=1}^{T} \mathbb{E}[\|\nabla \mathcal{L}(w_t)\|^2] \le 8C_\eta^2 + 16\sqrt{6}\, C_\eta G\sqrt{\log(K+1)} + 2\sqrt{2}\, C_\eta\sqrt{V^\star T}.$$

Setting $\eta_0 = \sqrt{2M/\beta}$ gives

$$C_\eta = \frac{2M}{\eta_0} + \eta_0\beta = 2\sqrt{2M\beta}.$$

Therefore,

$$\sum_{t=1}^{T} \mathbb{E}[\|\nabla \mathcal{L}(w_t)\|^2] \le 64M\beta + 64\sqrt{3}\sqrt{M\beta}\, G\sqrt{\log(K+1)} + 8\sqrt{M\beta V^\star T}.$$

Dividing by T concludes the proof:

$$\frac{1}{T}\sum_{t=1}^{T} \mathbb{E}[\|\nabla \mathcal{L}(w_t)\|^2] \le 8\sqrt{\frac{M\beta\, V^\star}{T}} + 64\frac{M\beta}{T} + 64\sqrt{3}\frac{\sqrt{M\beta}\, G\sqrt{\log(K+1)}}{T}.$$

□

D.4 Proof of Lemma D.2

Proof. Recall that, for $\lambda \in \Delta_{K+1}$,

$$g_t^\lambda = g_t + \sum_{j=1}^{K} \lambda_j d_{t,j}, \qquad h_t(\lambda) = \|g_t^\lambda\|^2,$$

where $d_{t,j} = \hat{g}^u_{t,j} - \hat{g}^\ell_{t,j}$ for $j = 1, \dots, K$, and $d_{t,0} = 0$. Since g_t^λ is affine in λ, the function h_t is convex over Δ_{K+1}.

The gradient of h_t with respect to λ is

$$[\nabla h_t(\lambda)]_0 = 0, \qquad [\nabla h_t(\lambda)]_j = 2\left\langle d_{t,j}, g_t^\lambda \right\rangle, \quad j = 1, \dots, K.$$

The exponentiated-gradient update is

$$\lambda_{t+1,j} = \frac{\lambda_{t,j} \exp(-\gamma_t [\nabla h_t(\lambda_t)]_j)}{\sum_{k=0}^K \lambda_{t,k} \exp(-\gamma_t [\nabla h_t(\lambda_t)]_k)}, \qquad j = 0, \dots, K.$$

With the stepsize choice stated in Lemma D.2, this is the standard entropy-regularized mirror-descent, equivalently exponentiated-gradient, update over the simplex. By the standard exponentiated-gradient regret bound for linearized losses over Δ_{K+1} with the negative-entropy mirror map (Cesa-Bianchi & Lugosi, 2006; Shalev-Shwartz, 2011; Orabona, 2019), for any $\lambda^\star \in \Delta_{K+1}$,

$$R_T^h(\lambda^\star) := \sum_{t=1}^T h_t(\lambda_t) - \sum_{t=1}^T h_t(\lambda^\star) \le \sqrt{2\log(K+1) \sum_{t=1}^T \|\nabla h_t(\lambda_t)\|_\infty^2}.$$

It remains to bound the ℓ_∞ norm of the gradients. Under the bounded-gradient assumption $\|\nabla \ell(w; x, y)\| \le G$, we have

$$\|\hat{g}^u_{t,j}\| \le G, \qquad \|\hat{g}^\ell_{t,j}\| \le G,$$

and therefore

$$\|d_{t,j}\| = \|\hat{g}^u_{t,j} - \hat{g}^\ell_{t,j}\| \le 2G.$$

Hence, for every $j = 1, \dots, K$,

$$|[\nabla h_t(\lambda_t)]_j| = 2\left|\left\langle d_{t,j}, g_t^{\lambda_t} \right\rangle\right| \le 2\|d_{t,j}\| \, \|g_t^{\lambda_t}\| \le 4G\sqrt{h_t(\lambda_t)}.$$

Since $[\nabla h_t(\lambda_t)]_0 = 0$, this implies

$$\|\nabla h_t(\lambda_t)\|_\infty^2 \le 16 G^2 h_t(\lambda_t).$$

Substituting into the exponentiated-gradient regret bound gives

$$R_T^h(\lambda^\star) \le 4\sqrt{2}\, G\sqrt{\log(K+1)} \sqrt{\sum_{t=1}^T h_t(\lambda_t)}.$$

Using the regret identity

$$\sum_{t=1}^T h_t(\lambda_t) = R_T^h(\lambda^\star) + \sum_{t=1}^T h_t(\lambda^\star),$$

we obtain

$$R_T^h(\lambda^\star) \le 4\sqrt{2}\, G\sqrt{\log(K+1)} \sqrt{R_T^h(\lambda^\star) + \sum_{t=1}^T h_t(\lambda^\star)}.$$

Using $\sqrt{a+b} \le \sqrt{a} + \sqrt{b}$, we further get

$$R_T^h(\lambda^\star) \le 4\sqrt{2}\, G\sqrt{\log(K+1)} \sqrt{R_T^h(\lambda^\star)} + 4\sqrt{2}\, G\sqrt{\log(K+1)} \sqrt{\sum_{t=1}^T h_t(\lambda^\star)}.$$

Finally, apply the helper inequality

$$A \le B\sqrt{A} + C \quad \Longrightarrow \quad A \le 4B^2 + 2C.$$

Taking

$$A = R_T^h(\lambda^\star), \qquad B = 4\sqrt{2}\, G\sqrt{\log(K+1)},$$

and

$$C = 4\sqrt{2}\,G\sqrt{\log(K+1)}\sqrt{\sum_{t=1}^{T} h_t(\lambda^\star)},$$

yields

$$R_T^h(\lambda^\star) \le 128G^2\log(K+1) + 8\sqrt{2}\,G\sqrt{\log(K+1)}\sqrt{\sum_{t=1}^{T} h_t(\lambda^\star)}.$$

This proves the claim. □

D.5 Proof of Lemma D.3

Proof. Let $\mathcal{F}_{t-1}$ denote the filtration generated by all randomness up to the beginning of round t, and write

$$\mathbb{E}_{t-1}[\cdot] := \mathbb{E}[\cdot \mid \mathcal{F}_{t-1}].$$

By the law of total expectation,

$$\mathbb{E}[h_t(\lambda^\star)] = \mathbb{E}[\|g_t^{\lambda^\star}\|^2] = \mathbb{E}\Big[\mathbb{E}_{t-1}\Big[\|g_t^{\lambda^\star}\|^2\Big]\Big].$$

Using the conditional second-moment decomposition,

$$\mathbb{E}_{t-1}\Big[\|g_t^{\lambda^\star}\|^2\Big] = \mathrm{Var}_{t-1}\Big(g_t^{\lambda^\star}\Big) + \Big\|\mathbb{E}_{t-1}\Big[g_t^{\lambda^\star}\Big]\Big\|^2.$$

Since $g_t^{\lambda^\star}$ is conditionally unbiased,

$$\mathbb{E}_{t-1}\Big[g_t^{\lambda^\star}\Big] = \nabla\mathcal{L}(w_t).$$

Therefore,

$$\mathbb{E}[h_t(\lambda^\star)] = \mathbb{E}\Big[\mathrm{Var}_{t-1}\Big(g_t^{\lambda^\star}\Big)\Big] + \mathbb{E}\big[\|\nabla\mathcal{L}(w_t)\|^2\big].$$

The conditional variance at the comparator $\lambda^\star$ is bounded by the optimal multi-model variance bound $V^\star$, explicitly presented in Section 5. Hence,

$$\mathbb{E}[h_t(\lambda^\star)] \le V^\star + \mathbb{E}\big[\|\nabla\mathcal{L}(w_t)\|^2\big].$$

□

D.6 Optimal λ-Parameter Calculations

The variance bounds in Lemmas 4.2, 5.2 can be minimized over λ.

D.6.1 Optimal λ for the confidence-aware single-teacher estimator

We derive the variance-minimizing interpolation parameter $\lambda^\star$ for the confidence-aware prediction-powered gradient with a single-teacher. From the variance bound (Lemma 4.2), minimizing $\mathrm{Var}(g_{\mathrm{CPP}}^{\lambda})$ is equivalent (dropping the constant factor 4) to

$$\min_{\lambda\in[0,1]} \phi(\lambda) \quad \text{where} \quad \phi(\lambda) := (1-\lambda)^2\frac{\sigma^2}{n} + \lambda^2\Big(\frac{\sigma^2}{N} + \Big(\frac{1}{n}+\frac{1}{N}\Big)s\Big).$$

Let

$$a := \frac{\sigma^2}{n}, \qquad b := \frac{\sigma^2}{N} + \Big(\frac{1}{n}+\frac{1}{N}\Big)s,$$

so that $\phi(\lambda) = a(1-\lambda)^2 + b\lambda^2$ with $a \ge 0$ and $b \ge 0$.

Derivative and stationary point. A direct derivative gives

$$\phi'(\lambda) = 2(\lambda-1)a + 2\lambda b = 2\big((a+b)\lambda - a\big).$$

Thus the unique stationary point (when $a+b>0$) is

$$\lambda^\star_{\mathrm{unc}} = \frac{a}{a+b} = \frac{\frac{\sigma^2}{n}}{\frac{\sigma^2}{n} + \frac{\sigma^2}{N} + \left(\frac{1}{n}+\frac{1}{N}\right)s} = \frac{1}{\left(1+\frac{n}{N}\right)\left(1+\frac{s}{\sigma^2}\right)}$$

Comparison to the endpoints. Since ϕ is a convex quadratic with $\phi''(\lambda) = 2(a+b) \geq 0$, the stationary point is the global minimizer on $\mathbb{R}$. Constraining to $[0,1]$ yields

$$\lambda^\star = \Pi_{[0,1]}\left(\frac{a}{a+b}\right),$$

where $\Pi_{[0,1]}(x) = \min\{1, \max\{0, x\}\}$. In the non-degenerate case $a > 0$ and $b > 0$, we have $0 < \frac{a}{a+b} < 1$, hence clipping is unnecessary and

$$\lambda^\star = \frac{1}{\left(1+\frac{n}{N}\right)\left(1+\frac{s}{\sigma^2}\right)}$$

Degenerate cases. If $a = 0$, then $\phi(\lambda) = b\lambda^2$ and $\lambda^\star = 0$. If $b = 0$, then $\phi(\lambda) = a(1-\lambda)^2$ and $\lambda^\star = 1$.

Optimal variance value. Plugging $\lambda^\star = \frac{a}{a+b}$ into $\phi(\lambda) = a(1-\lambda)^2 + b\lambda^2$ yields $\phi(\lambda^\star) = \frac{ab}{a+b}$, hence

$$\mathrm{Var}(g_{\mathrm{CPP}}^{\lambda^\star}) \leq 4 \cdot \frac{\frac{\sigma^2}{n} b}{\frac{\sigma^2}{n} + b}, \qquad b = \frac{\sigma^2}{N} + \left(\frac{1}{n} + \frac{1}{N}\right)s.$$

Rearranging,

$$\mathrm{Var}(g_{\mathrm{CPP}}^{\lambda^\star}) \leq \frac{4\sigma^2}{n}\left(1 - \frac{1}{1+\frac{n}{N}} \cdot \frac{\frac{\sigma^2}{s}}{1+\frac{\sigma^2}{s}}\right)$$

D.6.2 KKT SOLUTION FOR THE DIAGONAL MULTI-MODEL VARIANCE MINIMIZATION

We solve the simplex-constrained quadratic program

$$\min_{\lambda \in \Delta_{K+1}} J(\lambda) = \frac{\sigma^2}{n}\lambda_0^2 + \frac{\sigma^2}{N}(1-\lambda_0)^2 + K\left(\frac{1}{n} + \frac{1}{N}\right)\sum_{j=1}^{K} \sigma_{e,j}^2 \lambda_j^2, \tag{13}$$

where $\Delta_{K+1} = \{\lambda \geq 0 : \sum_{j=0}^{K} \lambda_j = 1\}$. Let $A := \frac{\sigma^2}{n},\ B := \frac{\sigma^2}{N},\ C := \frac{1}{n} + \frac{1}{N},\ q_j := \sigma_{e,j}^2$. Since $A, B, C > 0$ and $q_j > 0$, the objective is strictly convex.

Lagrangian. With multiplier ν for $\sum_j \lambda_j = 1$ and $\mu_j \geq 0$ for $\lambda_j \geq 0$,

$$\mathcal{L} = A\lambda_0^2 + B(1-\lambda_0)^2 + KC\sum_{j=1}^{K} q_j\lambda_j^2 + \nu\left(\sum_{j=0}^{K}\lambda_j - 1\right) - \sum_{j=0}^{K}\mu_j\lambda_j.$$

Stationarity gives

$$2KCq_j\lambda_j + \nu - \mu_j = 0, \quad j = 1, \ldots, K, \tag{14}$$

$$2A\lambda_0 + 2B(\lambda_0 - 1) + \nu - \mu_0 = 0. \tag{15}$$

Complementary slackness: $\mu_j\lambda_j = 0$.

Interior solution. Assume $\lambda_j > 0$ for all j, so $\mu_j = 0$. From equation 14,

$$\lambda_j = -\frac{\nu}{2KCq_j},$$

hence $\lambda_j \propto q_j^{-1} = \sigma_{e,j}^{-2}$. Let $S := \sum_{k=1}^{K} \sigma_{e,k}^{-2}$. Using $\sum_{j=1}^{K}\lambda_j = 1 - \lambda_0$,

$$-\frac{\nu}{2KC}S = 1 - \lambda_0 \quad \Rightarrow \quad \nu = -\frac{2KC(1-\lambda_0)}{S}.$$

Substitution yields

$$\lambda_j^\star = (1-\lambda_0^\star)\frac{\sigma_{e,j}^{-2}}{\sum_{k=1}^{K}\sigma_{e,k}^{-2}}, \qquad j = 1, \ldots, K. \tag{16}$$

Solving for $\lambda_0^\star$. With $\mu_0 = 0$, equation 15 reduces to $2(A+B)\lambda_0 - 2B + \nu = 0$, so $\lambda_0 = (B - \nu/2)/(A+B)$. Substituting $\nu = -2KC(1-\lambda_0)/S$ gives

$$(A+B)\lambda_0 = B + \frac{KC}{S} - \frac{KC}{S}\lambda_0,$$

hence

$$\lambda_0^\star = \frac{B + \frac{KC}{S}}{A + B + \frac{KC}{S}} = \frac{\frac{\sigma^2}{N} + \frac{K\left(\frac{1}{n}+\frac{1}{N}\right)}{\sum_{k=1}^K \sigma_{e,k}^{-2}}}{\frac{\sigma^2}{n} + \frac{\sigma^2}{N} + \frac{K\left(\frac{1}{n}+\frac{1}{N}\right)}{\sum_{k=1}^K \sigma_{e,k}^{-2}}}. \tag{17}$$

Boundary case. If the interior solution gives $\lambda_j^\star < 0$ for some j, the constraint becomes active. Let $\mathcal{A} \subseteq \{1, \dots, K\}$ denote the active set and $S_\mathcal{A} = \sum_{k\in\mathcal{A}} \sigma_{e,k}^{-2}$. Repeating the derivation with $S_\mathcal{A}$ yields

$$\lambda_j^\star = (1-\lambda_0^\star)\frac{\sigma_{e,j}^{-2}}{S_\mathcal{A}}, \quad j \in \mathcal{A}, \qquad \lambda_j^\star = 0 \quad j \notin \mathcal{A}.$$

Strict convexity ensures uniqueness.

Optimal variance value. Substituting equation 16 into equation 13 reduces the objective to the one-dimensional problem equation 17, i.e.

$$J(\lambda_0) = \frac{\sigma^2}{n}\lambda_0^2 + (1-\lambda_0)^2\left(\frac{\sigma^2}{N} + \frac{KC}{\sum_{k=1}^K \sigma_{e,k}^{-2}}\right).$$

Let

$$A := \frac{\sigma^2}{N} + \frac{K\left(\frac{1}{n}+\frac{1}{N}\right)}{\sum_{k=1}^K \sigma_{e,k}^{-2}}.$$

Using $\lambda_0^\star = \frac{A}{\frac{\sigma^2}{n}+A}$ and $1-\lambda_0^\star = \frac{\frac{\sigma^2}{n}}{\frac{\sigma^2}{n}+A}$, the optimal value becomes

$$Var(g_{\text{MM-PP}}^{\lambda^\star}) \le 4 \cdot \frac{\frac{\sigma^2}{n}A}{\frac{\sigma^2}{n}+A}. \tag{18}$$

Equivalently,

$$\text{Var}(g_{\text{MM-PP}}^{\lambda^\star}) \le 4 \cdot \frac{\frac{\sigma^2}{n}\left(\frac{\sigma^2}{N} + \frac{K\left(\frac{1}{n}+\frac{1}{N}\right)}{\sum_{k=1}^K \sigma_{e,k}^{-2}}\right)}{\frac{\sigma^2}{n} + \frac{\sigma^2}{N} + \frac{K\left(\frac{1}{n}+\frac{1}{N}\right)}{\sum_{k=1}^K \sigma_{e,k}^{-2}}} = 4\frac{\sigma^2}{n}\left(1 - \frac{1}{\left(1+\frac{n}{N}\right)\left(1 + \frac{K}{\sigma^2\sum_{k=1}^K \sigma_{e,k}^{-2}}\right)}\right).$$

Thus the multi-model optimal variance has the same harmonic-mean structure as the confidence case, with the single-teacher error variance replaced by the effective inverse-variance aggregate $\left(\sum_{k=1}^K \sigma_{e,k}^{-2}\right)^{-1}$.

D.6.3 KKT SOLUTION FOR THE CONFIDENCE-AWARE MULTI-MODEL VARIANCE MINIMIZATION

For a fixed λ_0,

$$\lambda_{1:K} = (1-\lambda_0)\alpha, \qquad \alpha \in \Delta^K.$$

Under the general matrix bound 3.8, the teacher-dependent term becomes

$$(1-\lambda_0)^2\alpha^\top \overline{\Sigma}_{\text{CM}}\alpha.$$

Thus the optimal teacher mixture solves

$$\min_{\alpha\in\Delta^K} \alpha^\top \overline{\Sigma}_{\text{CM}}\alpha.$$

If $\overline{\Sigma}_{\mathrm{CM}} \succ 0$ and the optimum is interior, the KKT conditions yield

$$\alpha^\star_{\mathrm{CM}} = \frac{\overline{\Sigma}^{-1}_{\mathrm{CM}}\mathbf{1}}{\mathbf{1}^\top \overline{\Sigma}^{-1}_{\mathrm{CM}}\mathbf{1}},$$

where $\mathbf{1} \in \mathbb{R}^K$ is the all-ones vector. The minimum value is

$$\min_{\alpha \in \Delta^K} \alpha^\top \overline{\Sigma}_{\mathrm{CM}} \alpha = \left(\mathbf{1}^\top \overline{\Sigma}^{-1}_{\mathrm{CM}}\mathbf{1}\right)^{-1}.$$

These expressions apply when the resulting $\alpha^\star_{\mathrm{CM}}$ belongs to the simplex. Otherwise, the constrained optimum lies on the boundary and is obtained by projecting this solution onto the simplex under the metric induced by $\overline{\Sigma}_{\mathrm{CM}}$, equivalently by solving the same quadratic program on the appropriate simplex face.

We will use the notation

$$A_{\mathrm{CM}} := \frac{\sigma^2}{N} + \left(\frac{1}{n} + \frac{1}{N}\right)\left(\mathbf{1}^\top \overline{\Sigma}^{-1}_{\mathrm{CM}}\mathbf{1}\right)^{-1}.$$

Then

$$\lambda^\star_{0,\mathrm{CM}} = \frac{A_{\mathrm{CM}}}{\frac{\sigma^2}{n} + A_{\mathrm{CM}}}, \qquad \lambda^\star_{1:K,\mathrm{CM}} = (1 - \lambda^\star_{0,\mathrm{CM}})\alpha^\star_{\mathrm{CM}},$$

and the resulting optimal variance bound is

$$V^\star_{\mathrm{CM}} \le 4\,\frac{\frac{\sigma^2}{n} A_{\mathrm{CM}}}{\frac{\sigma^2}{n} + A_{\mathrm{CM}}}.$$

E Diagonal-scaled PSD envelopes and recovery of the worst-case bound

The variance bound from Lemma 5.2, proven in Appendix D.2, states a bound using a general positive semidefinite bounding matrix. Here we show that a diagonal-scaled envelope of the form κD always provides a valid choice for a positive semidefinite bounding matrix. In the worst case, this recovers the factor K appearing in the diagonal variance analysis and aligns with the second bound from Appendix D.2.

Notation. For a diagonal matrix $D \succeq 0$, we write $D^{-1/2}$ for the inverse square root on the support of D, equivalently the pseudoinverse square root. When all diagonal entries are strictly positive, this coincides with the usual inverse square root.

E.1 Covariance-aware MM-PP bound

The covariance-aware bound reveals that the multi-teacher aggregation problem is governed by a quadratic form rather than only by a weighted sum of marginal variances. The next lemma shows that a diagonal-scaled matrix always provides a valid envelope.

Lemma E.1 (Diagonal-scaled covariance envelope for MM-PP)**.** *Let*

$$D_{\mathrm{MM}} := \mathrm{diag}(\sigma^2_{e,1}, \dots, \sigma^2_{e,K}).$$

Under Assumption 3.6, for every $w \in \mathbb{R}^d$,

$$[\Sigma_{\mathrm{MM}}(w)]_{jj} \le \sigma^2_{e,j}, \qquad j \in [K].$$

Moreover, if

$$\kappa_{\mathrm{MM}} \ge \sup_{w \in \mathbb{R}^d} \lambda_{\max}\Big(D^{-1/2}_{\mathrm{MM}} \Sigma_{\mathrm{MM}}(w) D^{-1/2}_{\mathrm{MM}}\Big),$$

then

$$\Sigma_{\mathrm{MM}}(w) \preceq \kappa_{\mathrm{MM}} D_{\mathrm{MM}} \qquad \textit{for all } w \in \mathbb{R}^d.$$

Proof. By Assumption 3.6,

$$\mathbb{E}\|X_j(w;x,y)\|^2 \le \sigma_{e,j}^2, \qquad j \in [K].$$

Since

$$[\Sigma_{\mathrm{MM}}(w)]_{jj} = \mathbb{E}\|X_j(w;x,y)\|^2,$$

the diagonal bound follows immediately.

For the matrix inequality,

$$\Sigma_{\mathrm{MM}}(w) = D_{\mathrm{MM}}^{1/2}\Big(D_{\mathrm{MM}}^{-1/2}\Sigma_{\mathrm{MM}}(w)D_{\mathrm{MM}}^{-1/2}\Big)D_{\mathrm{MM}}^{1/2}.$$

If the middle matrix has largest eigenvalue at most κ_{MM}, then

$$D_{\mathrm{MM}}^{-1/2}\Sigma_{\mathrm{MM}}(w)D_{\mathrm{MM}}^{-1/2} \preceq \kappa_{\mathrm{MM}} I,$$

and therefore

$$\Sigma_{\mathrm{MM}}(w) \preceq \kappa_{\mathrm{MM}} D_{\mathrm{MM}}.$$

□

Combining Lemmas 5.2 and E.1, we obtain the explicit bound

$$\mathrm{Var}(g^{\lambda}_{\mathrm{MMPP}}) \le 4\left(\lambda_0^2\frac{\sigma^2}{n} + (1-\lambda_0)^2\frac{\sigma^2}{N} + \kappa_{\mathrm{MM}}\Big(\frac{1}{n}+\frac{1}{N}\Big)\lambda_{1:K}^{\top}D_{\mathrm{MM}}\lambda_{1:K}\right). \tag{19}$$

Remark E.2 (Worst-case diagonal envelope). Setting $\kappa_{\mathrm{MM}} = K$ recovers the worst-case envelope without additional cross-teacher covariance assumptions

$$\Sigma_{\mathrm{MM}}(w) \preceq K\, D_{\mathrm{MM}}.$$

Indeed, by Cauchy–Schwarz,

$$\big|[\Sigma_{\mathrm{MM}}(w)]_{jk}\big| = \big|\mathbb{E}\langle X_j, X_k\rangle\big| \le \sqrt{\mathbb{E}\|X_j\|^2}\sqrt{\mathbb{E}\|X_k\|^2} \le \sigma_{e,j}\sigma_{e,k}.$$

Hence, for every $a \in \mathbb{R}^K$,

$$a^{\top}\Sigma_{\mathrm{MM}}(w)a \le \sum_{j,k=1}^{K}|a_j||a_k|\sigma_{e,j}\sigma_{e,k} = \Big(\sum_{j=1}^{K}|a_j|\sigma_{e,j}\Big)^2 \le K\sum_{j=1}^{K}a_j^2\sigma_{e,j}^2 = a^{\top}(KD_{\mathrm{MM}})a.$$

The diagonal-scaled envelope always exists, and $\kappa_{\mathrm{MM}} \in [1, K]$ can be interpreted as a correlation inflation factor.

Confidence-aware diagonal envelope. For CM-PP

$$D_{\mathrm{CM}} := \mathrm{diag}(s_1,\dots,s_K), \qquad s_j = \beta_Y^2\left(\tau_j\varepsilon^{f_j,1} + (1-\tau_j)\varepsilon^{0,j}\right).$$

Consequently, the same diagonal-scaling argument gives a $\kappa_{\mathrm{CM}} \in [1, K]$ such that

$$\Sigma_{\mathrm{CM}}(w) \preceq \kappa_{\mathrm{CM}} D_{\mathrm{CM}} \qquad \text{for all } w \in \mathbb{R}^d.$$

Thus, $\bar{\Sigma}_{\mathrm{CM}} := \kappa_{\mathrm{CM}} D_{\mathrm{CM}}$ is a valid covariance envelope. Setting $\kappa_{\mathrm{CM}} = K$ gives the worst-case diagonal envelope without additional cross-teacher covariance assumptions.

E.2 Retrieving the optimal weights from the diagonal-scaled special case

We show that the optimal interpolation parameters from Sections D.6.2, D.6.3 coincide and recover the variance bound in equation 10.

If we use the explicit envelope

$$\overline{\Sigma}_{\mathrm{MM}} = \kappa_{\mathrm{MM}} D_{\mathrm{MM}},$$

then the optimal teacher mixture reduces to inverse-variance weighting:

$$\alpha^{\star}_{\mathrm{MM},j} = \frac{\sigma_{e,j}^{-2}}{\sum_{k=1}^{K}\sigma_{e,k}^{-2}},$$

with the convention that zero-variance entries receive all mass before any positive-variance teacher. In this case,

$$A_{\mathrm{MM}} := \frac{\sigma^2}{N} + \kappa_{\mathrm{MM}}\Big(\frac{1}{n}+\frac{1}{N}\Big)\Big(\sum_{k=1}^{K}\sigma_{e,k}^{-2}\Big)^{-1},$$

and therefore

$$\lambda^\star_{0,\mathrm{MM}} = \frac{A_{\mathrm{MM}}}{\frac{\sigma^2}{n}+A_{\mathrm{MM}}}, \qquad \lambda^\star_{1:K,\mathrm{MM}} = (1-\lambda^\star_{0,\mathrm{MM}})\alpha^\star_{\mathrm{MM}},$$

with

$$V^\star_{\mathrm{MM}} \le 4\,\frac{\frac{\sigma^2}{n}A_{\mathrm{MM}}}{\frac{\sigma^2}{n}+A_{\mathrm{MM}}}.$$

For the confidence-aware envelope $\bar{\Sigma}_{\mathrm{CM}} = \kappa_{\mathrm{CM}} D_{\mathrm{CM}}$, the optimal teacher mixture is $\alpha^\star_{\mathrm{CM},j} = \frac{s_j^{-1}}{\sum_{k=1}^{K} s_k^{-1}}$, $j \in [K]$. Accordingly, the effective confidence-aware teacher term is $q_{\mathrm{CM}} = \kappa_{\mathrm{CM}}\left(\sum_{j=1}^{K} s_j^{-1}\right)^{-1}$. Substituting this term into the covariance-aware bound yields Equation equation 10.

E.3 Proof of Lemma 5.2 for the Multi-Model Diagonal Case.

For the confidence-agnostic multi-model diagonal worst case, proof of Lemma 5.2 can be simplified:

Proof. We bound the variance of the multi-model prediction-powered gradient using the following notations:

$$g^n = \frac{1}{n}\sum_{i=1}^{n}\nabla\ell(w;x^i,y^i), \qquad g^N = \frac{1}{N}\sum_{i=1}^{N}\nabla\ell(w;\tilde{x}^i,\tilde{y}^i),$$

$$g^{n,f} = \sum_{j=0}^{K}\lambda_j g^{n,f_j}, \qquad g^{N,f} = \sum_{j=0}^{K}\lambda_j g^{N,f_j},$$

$$g^{n,f_j} = \frac{1}{n}\sum_{i=1}^{n}\nabla\ell(w;x^i,f_j(x^i)), \qquad g^{N,f_j} = \frac{1}{N}\sum_{i=1}^{N}\nabla\ell(w;\tilde{x}^i,f_j(\tilde{x}^i)),$$

with $g^{n,f,0} = g^{N,f,0} = 0$.

Here g^n is the labeled mini-batch gradient estimator and g^N is the corresponding hypothetical true-label estimator on the unlabeled batch. $g^{n,f}$, $g^{N,f}$ replace the true labels with predictions from the models f and weights them.

The MM-PP gradient estimator is

$$g^{\lambda}_{\mathrm{MMPP}} = g^n + g^{N,f} - g^{n,f} = g^n + \sum_{j=0}^{K}\lambda_j\big(g^{N,f_j} - g^{n,f_j}\big) = \lambda_0 g^n + \sum_{j=1}^{K}\lambda_j\big(g^n + g^{N,f_j} - g^{n,f_j}\big).$$

For each teacher $j \ge 1$, define

$$\nabla\mathcal{L}^{f_j}_e(w) := \nabla\mathcal{L}(w) - \nabla\mathcal{L}^{f_j}(w).$$

Then

$$g^n + g^{N,f_j} - g^{n,f_j} - \nabla\mathcal{L}(w) = \big(g^N - \nabla\mathcal{L}(w)\big) + \big(g^{N,f_j} - g^N + \nabla\mathcal{L}^{f_j}_e(w)\big) + \big(g^n - g^{n,f_j} - \nabla\mathcal{L}^{f_j}_e(w)\big).$$

The variance is

$$\begin{aligned}\mathrm{Var}(g^{\lambda}_{\mathrm{MMPP}}) &= \mathbb{E}[g^{\lambda}_{MMPP} - \nabla\mathcal{L}(w)] \\ &= \mathbb{E}\big\|\lambda_0(g^n - \nabla\mathcal{L}(w)) + \sum_{j=1}^{K}\lambda_j(g^n + g^{N,f_j} - g^{n,f_j} - \nabla\mathcal{L}(w)))\big\|^2 \\ &= \mathbb{E}\big\|\lambda_0(g^n - \nabla\mathcal{L}(w)) \\ &+ \sum_{j=1}^{K}\lambda_j(g^n - g^{n,f_j} - \nabla\mathcal{L}(w) + \nabla\mathcal{L}^{f_j}(w)) \\ &+ \sum_{j=1}^{K}\lambda_j(g^{N,f_j} - g^N + \nabla\mathcal{L}(w) - \nabla\mathcal{L}^{f_j}(w)) \\ &+ \sum_{j=1}^{K}\lambda_j(g^N - \nabla\mathcal{L}(w)))\big\|^2\end{aligned}$$

The inequality $\|u+v+w+z\|^2 \le 4(\|u\|^2+\|v\|^2+\|w\|^2+\|z\|^2)$ gives

$$\begin{aligned}\mathrm{Var}(g^{\lambda}_{\mathrm{MMPP}}) \le\ & 4\lambda_0^2\,\mathbb{E}\|g^n - \nabla\mathcal{L}(w)\|^2 \\ &+ 4(1-\lambda_0)^2\,\mathbb{E}\|g^N - \nabla\mathcal{L}(w)\|^2 \\ &+ 4\,\mathbb{E}\Big\|\sum_{j=1}^{K}\lambda_j\big(g^{N,f_j} - g^N + \nabla\mathcal{L}_e^{f_j}(w)\big)\Big\|^2 \\ &+ 4\,\mathbb{E}\Big\|\sum_{j=1}^{K}\lambda_j\big(g^{n} - g^{n,f_j} - \nabla\mathcal{L}_e^{f_j}(w)\big)\Big\|^2.\end{aligned}$$

Using $\|\sum_{j=1}^{K}u_j\|^2 \le K\sum_{j=1}^{K}\|u_j\|^2$, we obtain

$$\begin{aligned}\mathrm{Var}(g^{\lambda}_{\mathrm{MMPP}}) \le\ & 4\lambda_0^2\,\mathbb{E}\|g^n - \nabla\mathcal{L}(w)\|^2 + 4(1-\lambda_0)^2\,\mathbb{E}\|g^N - \nabla\mathcal{L}(w)\|^2 \\ &+ 4K\sum_{j=1}^{K}\lambda_j^2\,\mathbb{E}\|g^{N,f_j} - g^N + \nabla\mathcal{L}_e^{f_j}(w)\|^2 \\ &+ 4K\sum_{j=1}^{K}\lambda_j^2\,\mathbb{E}\|g^{n} - g^{n,f_j} - \nabla\mathcal{L}_e^{f_j}(w)\|^2.\end{aligned}$$

By Assumption 3.5,

$$\mathbb{E}\|g^n - \nabla\mathcal{L}(w)\|^2 \le \frac{\sigma^2}{n}, \qquad \mathbb{E}\|g^N - \nabla\mathcal{L}(w)\|^2 \le \frac{\sigma^2}{N}.$$

By Assumption 3.6,

$$\mathbb{E}\|g^{N,f_j} - g^N + \nabla\mathcal{L}_e^{f_j}(w)\|^2 \le \frac{\sigma_{e,j}^2}{N}, \qquad \mathbb{E}\|g^{n} - g^{n,f_j} - \nabla\mathcal{L}_e^{f_j}(w)\|^2 \le \frac{\sigma_{e,j}^2}{n}.$$

Therefore,

$$\begin{aligned}\mathrm{Var}(g^{\lambda}_{\mathrm{MMPP}}) &\le 4\left(\lambda_0^2\frac{\sigma^2}{n} + (1-\lambda_0)^2\frac{\sigma^2}{N} + \frac{K}{N}\sum_{j=1}^{K}\lambda_j^2\sigma_{e,j}^2 + \frac{K}{n}\sum_{j=1}^{K}\lambda_j^2\sigma_{e,j}^2\right) \\ &= 4\left(\lambda_0^2\frac{\sigma^2}{n} + (1-\lambda_0)^2\frac{\sigma^2}{N} + K\Big(\frac{1}{n}+\frac{1}{N}\Big)\sum_{j=1}^{K}\lambda_j^2\sigma_{e,j}^2\right),\end{aligned}$$

□

Table 4: Problem setting and main limitations.

Method	Setting	Main limitations
PP-SSL	Single teacher	Assumes one globally useful teacher. If pseudo-labels are poor or misaligned, the method tends to revert toward labeled-only behavior and cannot exploit complementary teachers.
C-PP	Single teacher with confidence	Still restricted to one teacher. Gains rely on an informative confidence signal, and low-confidence behavior remains tied to ε^0.
MM-PP	K teachers	Uses only global teacher weights. If all teachers are poor, the method again tends toward labeled-only behavior. Under covariance-aware analysis, gains also depend on teacher complementarity.
CM-PP	K teachers with confidence	Most flexible, but also the most complex. It requires per-teacher confidence signals and, under covariance-aware analysis, depends on the interaction between confidence modulation and cross-teacher covariance.

F Comparison of PPI++, PP-SSL, C-PP, MM-PP, and CM-PP

This appendix compares the five prediction-powered objectives in a unified framework. We first summarize the settings and limitations of each method, then give a compact table of the optimal interpolation weights and variance bounds. We derive linear-MSE specializations in the style of Appendix C of *Prediction-Powered Semi-Supervised Learning with Online Power Tuning* (Shoham et al., 2026), and finally connect these closed forms back to the general variance-based theory developed in the main text.

Throughout this appendix, we use ϵ and ϵ_f only in the linear-regression comparison below, where they denote residual quantities.

F.1 Settings and limitations

Table 4 highlights the two main axes along which the methods differ: whether they use a single-teacher or multiple teachers, and whether they incorporate confidence-aware weighting. The progression from PPI++ to PP-SSL to C-PP to MM-PP to CM-PP corresponds to increasing modeling flexibility, but also to greater dependence on confidence quality and aggregation structure.

C-PP addresses the single-teacher failure mode of PP-SSL under heterogeneous pseudo-label quality, while MM-PP addresses the inability of PP-SSL to exploit multiple complementary teachers. CM-PP combines both extensions. In particular, C-PP is recovered from CM-PP by restricting to one teacher, while MM-PP and PP-SSL are recovered by setting $c_j(x) \equiv 1$.

F.2 Main-theory comparison of optimal interpolation weights

We first summarize the variance-minimizing interpolation weights obtained from the main-theory variance bounds. For a generic effective teacher-variance term q, we use the notations

$$m(q) := \frac{1}{1+\frac{n}{N}}\,\frac{1}{1+\frac{q}{\sigma^2}}, \qquad \mathcal{V}_{\mathrm{main}}(q) := \frac{4\sigma^2}{n}\big(1-m(q)\big).$$

Here $m(q)$ is the optimal total mass assigned to prediction-powered correction terms, and $\mathcal{V}_{\mathrm{main}}(q)$ is the corresponding optimized variance bound.

For the multi-teacher methods, the remaining mass is allocated according to the covariance-aware precision vector. In particular, when the covariance envelope is invertible and the optimum is interior,

$$\lambda^\star_{1:K,\text{MM-PP}} = \big(1-\lambda^\star_{0,\text{MM-PP}}\big)\frac{\overline{\Sigma}_{\mathrm{MM}}^{-1}\mathbf{1}}{\mathbf{1}^\top\overline{\Sigma}_{\mathrm{MM}}^{-1}\mathbf{1}}, \quad \lambda^\star_{1:K,\text{CM-PP}} = \big(1-\lambda^\star_{0,\text{CM-PP}}\big)\frac{\overline{\Sigma}_{\mathrm{CM}}^{-1}\mathbf{1}}{\mathbf{1}^\top\overline{\Sigma}_{\mathrm{CM}}^{-1}\mathbf{1}}.$$

Thus PP-SSL and C-PP differ only through the single-teacher effective variance term, while MM-PP and CM-PP replace the single-teacher variance by a covariance-aware multi-teacher effective variance.

Table 5: Main-theory comparison of optimal interpolation weights and optimized variance bounds. Smaller effective teacher variance q yields larger prediction-powered mass and a smaller variance bound.

Method	**Effective teacher variance** q	**Optimal interpolation weight**	**Optimized bound**
PP-SSL	$q_{\mathrm{PP}} = \sigma_e^2$	$\lambda^\star_{\text{PP-SSL}} = m(q_{\mathrm{PP}})$	$\mathcal{V}_{\mathrm{main}}(q_{\mathrm{PP}})$
C-PP	$q_{\mathrm{C}} = s = \beta_Y^2\left(\tau\varepsilon_{f,1}+(1-\tau)\varepsilon^0\right)$	$\lambda^\star_{\text{C-PP}} = m(q_{\mathrm{C}})$	$\mathcal{V}_{\mathrm{main}}(q_{\mathrm{C}})$
MM-PP	$q_{\mathrm{MM}} = \left(\mathbf{1}^\top\overline{\Sigma}_{\mathrm{MM}}^{-1}\mathbf{1}\right)^{-1}$	$1-\lambda^\star_{0,\text{MM-PP}} = m(q_{\mathrm{MM}})$	$\mathcal{V}_{\mathrm{main}}(q_{\mathrm{MM}})$
CM-PP	$q_{\mathrm{CM}} = \left(\mathbf{1}^\top\overline{\Sigma}_{\mathrm{CM}}^{-1}\mathbf{1}\right)^{-1}$	$1-\lambda^\star_{0,\text{CM-PP}} = m(q_{\mathrm{CM}})$	$\mathcal{V}_{\mathrm{main}}(q_{\mathrm{CM}})$

Table 6: Optimal interpolation weights and variance bounds in the linear-MSE comparison. The quantity v denotes the effective teacher variance entering $\mathcal{V}(v)$.

Method	**Effective teacher variance**	**Optimal teacher mass**	**Bound**
PP-SSL	$v_{\mathrm{PP}} = \mathrm{Var}(\epsilon_f)$	$\lambda^\star_{\text{PP-SSL}} = \frac{1}{1+r}\frac{\mathrm{Var}(\epsilon)}{\mathrm{Var}(\epsilon)+v_{\mathrm{PP}}}$	$\mathcal{V}(v_{\mathrm{PP}})$
C-PP	$v_{\mathrm{C}} = \mathrm{Var}(c(x)\epsilon_f)$	$\lambda^\star_{\text{C-PP}} = \frac{1}{1+r}\frac{\mathrm{Var}(\epsilon)}{\mathrm{Var}(\epsilon)+v_{\mathrm{C}}}$	$\mathcal{V}(v_{\mathrm{C}})$
MM-PP	$v_{\mathrm{MM}} = \left(\mathbf{1}^\top\Gamma_f^{-1}\mathbf{1}\right)^{-1}$	$1 - \lambda^\star_{0,\text{MM-PP}} = \frac{1}{1+r}\frac{\mathrm{Var}(\epsilon)}{\mathrm{Var}(\epsilon)+v_{\mathrm{MM}}}$	$= \mathcal{V}(v_{\mathrm{MM}})$
CM-PP	$v_{\mathrm{CM}} = \left(\mathbf{1}^\top\Gamma_{f,c}^{-1}\mathbf{1}\right)^{-1}$	$1 - \lambda^\star_{0,\text{CM-PP}} = \frac{1}{1+r}\frac{\mathrm{Var}(\epsilon)}{\mathrm{Var}(\epsilon)+v_{\mathrm{CM}}}$	$= \mathcal{V}(v_{\mathrm{CM}})$

In the diagonal special case, $q_{\mathrm{MM}} = \kappa_{\mathrm{MM}}\left(\sum_{j=1}^K \sigma_{e,j}^{-2}\right)^{-1}$, $q_{\mathrm{CM}} = \kappa_{\mathrm{CM}}\left(\sum_{j=1}^K s_j^{-1}\right)^{-1}$, where $s_j = \beta_Y^2\left(\tau_j\varepsilon_{f_j,1} + (1-\tau_j)\varepsilon^0\right)$. This recovers inverse-variance weighting as the diagonal-envelope special case.

F.3 Linear-MSE closed forms

We follow the linear-regression comparison from Appendix C of (Shoham et al., 2026). Consider squared loss

$$\ell(w;(x,y)) = \frac{1}{2}(y - x^\top w)^2, \qquad r = \frac{n}{N},$$

and let $w^\star$ be the population minimizer. We define the residuals $\epsilon := y - x^\top w^\star$, $\epsilon_f := f(x) - x^\top w^\star$. For an effective correction-residual variance v, define the generic optimized bound

$$\mathcal{V}(v) := \frac{4\,\mathrm{Var}(\epsilon)\,\mathrm{Tr}(\Sigma_x)}{n}\left(1 - \frac{1}{(1+r)\left(1+\frac{v}{\mathrm{Var}(\epsilon)}\right)}\right). \tag{20}$$

The table summarizes the resulting variance-ratio weights. Smaller v corresponds to a more useful prediction-powered correction.

The expressions in Table 6 follow by substituting the linear-MSE variance identities into the general variance-minimizing formulas from Appendix D.6. In the single-teacher case, the abstract labeled-gradient variance becomes $\sigma^2 = \mathrm{Var}(\epsilon)\,\mathrm{Tr}(\Sigma_x)$, and the teacher-error variance becomes $\sigma_e^2 = \mathrm{Var}(\epsilon_f)\,\mathrm{Tr}(\Sigma_x)$.

Thus the PP-SSL variance-ratio expression is

$$\lambda^\star_{\text{PP-SSL}} = \frac{1}{1+r}\frac{\mathrm{Var}(\epsilon)}{\mathrm{Var}(\epsilon)+\mathrm{Var}(\epsilon_f)}.$$

For completeness, the exact linear PP-SSL solution coincides with the PPI++ covariance-form expression

$$\lambda^\star_{\text{PPI++}} = \lambda^\star_{\text{PP-SSL,linear}} = \frac{1}{1+r}\frac{\text{Cov}(\epsilon, \epsilon_f)}{\text{Var}(\epsilon_f)}.$$

Hence PPI++ and the exact linear PP-SSL solution are covariance-driven, whereas the unified comparison in Table 6 uses the corollary-based variance-ratio form.

Confidence-aware single-teacher. For C-PP, confidence replaces the teacher residual ϵ_f by $c(x)\epsilon_f$. Hence the teacher-error variance becomes $\sigma^2_{e,c} = \text{Var}(c(x)\epsilon_f)\,\text{Tr}(\Sigma_x)$.

If $c(x) \in \{0,1\}$, $\tau = \Pr(c(x)=1)$, and $\epsilon_{f|1} \sim (\epsilon_f \mid c(x)=1)$, then

$$\text{Var}(c(x)\epsilon_f) = \tau\text{Var}(\epsilon_{f|1}) + \tau(1-\tau)\big(\mathbb{E}[\epsilon_{f|1}]\big)^2.$$

Thus informative confidence reduces the effective teacher variance by preserving the high-confidence, low-error region. In the favorable case $\mathbb{E}[\epsilon_{f|1}] \approx 0$ and $\text{Var}(\epsilon_{f|1}) \to 0$, the C-PP weight approaches $(1+r)^{-1}$, the largest possible prediction-powered mass in this variance-ratio form.

Covariance-aware multiple teachers. For K teachers, define residuals $\epsilon_{f_j} := f_j(x) - x^\top w^\star$ and the covariance matrix $\Gamma_f = \big[\text{Cov}(\epsilon_{f_j}, \epsilon_{f_k})\big]_{j,k=1}^K$. The linear-MSE identity $\Sigma_{MM} = \text{Tr}(\Sigma_x)\Gamma_f$ turns the main-theory covariance precision term into $v_{\text{MM}} = \left(\mathbf{1}^\top\Gamma_f^{-1}\mathbf{1}\right)^{-1}$. The total teacher mass is therefore

$$1 - \lambda^\star_{0,\text{MM-PP}} = \frac{1}{1+r}\frac{\text{Var}(\epsilon)}{\text{Var}(\epsilon) + v_{\text{MM}}},$$

and the mass is allocated across teachers according to

$$\lambda^\star_{1:K,\text{MM-PP}} = \big(1 - \lambda^\star_{0,\text{MM-PP}}\big)\frac{\Gamma_f^{-1}\mathbf{1}}{\mathbf{1}^\top\Gamma_f^{-1}\mathbf{1}}.$$

Thus MM-PP weights teachers by covariance-aware precision rather than by independent marginal quality alone.

For CM-PP, confidence modifies the teacher covariance before aggregation. $\Gamma_{f,c} = \big[\text{Cov}(c_j(x)\epsilon_{f_j}, c_k(x)\epsilon_{f_k})\big]_{j,k=1}^K$. Then $\Sigma_{CM} = \text{Tr}(\Sigma_x)\Gamma_{f,c}$, $v_{\text{CM}} = \left(\mathbf{1}^\top\Gamma_{f,c}^{-1}\mathbf{1}\right)^{-1}$, and the optimal allocation is

$$\lambda^\star_{1:K,\text{CM-PP}} = \big(1 - \lambda^\star_{0,\text{CM-PP}}\big)\frac{\Gamma_{f,c}^{-1}\mathbf{1}}{\mathbf{1}^\top\Gamma_{f,c}^{-1}\mathbf{1}}.$$

If $c_j(x) \equiv 1$ for all teachers, then $\Gamma_{f,c} = \Gamma_f$, so CM-PP reduces to MM-PP. In the diagonal case, the effective variances reduce to inverse-variance aggregates:

$$v_{\text{MM}} = \left(\sum_{j=1}^K \text{Var}(\epsilon_{f_j})^{-1}\right)^{-1}, \qquad v_{\text{CM}} = \left(\sum_{j=1}^K \text{Var}(c_j(x)\epsilon_{f_j})^{-1}\right)^{-1}.$$

Interpretation. All variance-ratio expressions share the same structure: smaller effective teacher variance v yields larger teacher mass and a smaller optimized bound. The linear-MSE quantities $\text{Var}(\epsilon_f)$, $\text{Var}(c(x)\epsilon_f)$, v_{MM}, and v_{CM} should therefore be viewed as interpretable special cases of the abstract variance terms in the main theory, such as σ_e^2, Σ_{MM}, and Σ_{CM}.

G Online Optimization of Interpolation Weights

This appendix describes how the interpolation weights are adapted online during training. The goal is to avoid explicit estimation of the variance and covariance terms that determine the offline optimal interpolation weights, while still reducing the variance of the prediction-powered gradient estimator.

At iteration t, let g_t^λ denote the prediction-powered gradient estimator formed with interpolation weights λ. We define the online objective

$$h_t(\lambda) := \|g_t^\lambda\|^2.$$

At round t, λ_t is selected from the history $\mathcal{F}_{t-1}$, before observing the current mini-batch. The resulting gradient $g_t^{\lambda_t}$ is used to update w_{t+1}, after which the same batch may be used to update λ_{t+1}. Consequently,

$$\mathbb{E}\Big[g_t^{\lambda_t} \mid \mathcal{F}_{t-1}\Big] = \nabla\mathcal{L}(w_t).$$

For any fixed $\lambda \in \Delta_{K+1}$, the same conditional unbiasedness yields

$$\mathbb{E}[g_t^\lambda \mid w_t] = \nabla\mathcal{L}(w_t),$$

and therefore

$$\mathbb{E}\big[\|g_t^\lambda\|^2 \mid w_t\big] = \|\nabla\mathcal{L}(w_t)\|^2 + \mathbb{E}\big[\|g_t^\lambda - \nabla\mathcal{L}(w_t)\|^2 \mid w_t\big].$$

Thus, for fixed w_t, minimizing $h_t(\lambda)$ is equivalent in expectation to minimizing the conditional variance of the gradient estimator up to the deterministic term $\|\nabla\mathcal{L}(w_t)\|^2$. This motivates treating λ_t as an online decision variable and updating it by online convex optimization.

G.1 Single-Teacher Setting: Projected AdaGrad

In the single-teacher setting, the interpolation parameter is scalar and the feasible set is

$$D = [0, 1].$$

We update $\lambda_t \in D$ by projected AdaGrad:

$$\lambda_{t+1} = \Pi_D(\lambda_t - \gamma_t \nabla h_t(\lambda_t)),$$

where Π_D denotes Euclidean projection onto D. The adaptive stepsize is

$$\gamma_t = \mathcal{R}_D \left(2\sum_{s=1}^{t} |\nabla h_s(\lambda_s)|^2\right)^{-1/2}, \qquad \mathcal{R}_D := \max_{u,v\in D} |u - v| = 1.$$

The model parameters are updated using the corresponding prediction-powered gradient:

$$w_{t+1} = w_t - \eta_t g_t^{\lambda_t}, \qquad \eta_t = \eta_0 \left(\sum_{s=1}^{t} \|g_s^{\lambda_s}\|^2\right)^{-1/2}.$$

G.2 Multi-Teacher Setting: Exponentiated Gradient Descent / Multiplicative Weights

In the multi-teacher setting, the interpolation weights lie on the simplex

$$\lambda_t = (\lambda_{t,0}, \lambda_{t,1}, \ldots, \lambda_{t,K}) \in \Delta_{K+1},$$

where $\lambda_{t,0}$ is the weight assigned to the zero teacher, which corresponds to the labeled-only component.

For each teacher $j = 1, \ldots, K$, define the prediction-powered correction

$$d_{t,j} := \hat{g}_{t,j}^{u} - \hat{g}_{t,j}^{\ell},$$

and set $d_{t,0} = 0$. Then

$$g_t^\lambda = g_t + \sum_{j=1}^{K} \lambda_j d_{t,j}, \qquad h_t(\lambda) = \left\| g_t + \sum_{j=1}^{K} \lambda_j d_{t,j} \right\|^2.$$

Since g_t^λ is affine in λ, the function h_t is convex over Δ_{K+1}.

To respect the simplex geometry, we update λ_t by mirror descent with the negative-entropy mirror map

$$\psi(\lambda) = \sum_{j=0}^{K} \lambda_j \log \lambda_j.$$

The update is

$$\lambda_{t+1} = \arg\min_{\lambda \in \Delta_{K+1}} \left\{ \langle \nabla h_t(\lambda_t), \lambda \rangle + \frac{1}{\gamma_t} D_{\mathrm{KL}}(\lambda \| \lambda_t) \right\}.$$

The gradient coordinates are

$$[\nabla h_t(\lambda_t)]_0 = 0, \qquad [\nabla h_t(\lambda_t)]_j = 2 \left\langle d_{t,j}, g_t^{\lambda_t} \right\rangle, \quad j = 1, \ldots, K.$$

Equivalently, defining

$$q_{t,0} = 0, \qquad q_{t,j} = 2 \left\langle d_{t,j}, g_t^{\lambda_t} \right\rangle, \quad j = 1, \ldots, K,$$

the mirror-descent update takes the multiplicative-weights form

$$\lambda_{t+1,j} = \frac{\lambda_{t,j} \exp(-\gamma_t q_{t,j})}{\sum_{k=0}^{K} \lambda_{t,k} \exp(-\gamma_t q_{t,k})}, \qquad j = 0, \ldots, K.$$

We use the adaptive stepsize

$$\gamma_t = \sqrt{\frac{2 \log(K+1)}{\delta + \sum_{s=1}^{t} \|\nabla h_s(\lambda_s)\|_\infty^2}},$$

where $\delta > 0$ is a small numerical constant. The model parameters are updated as

$$w_{t+1} = w_t - \eta_t g_t^{\lambda_t}, \qquad \eta_t = \eta_0 \left(\sum_{s=1}^{t} \|g_s^{\lambda_s}\|^2 \right)^{-1/2}.$$

This entropy-regularized update preserves feasibility of λ_t automatically and yields a multiplicative-weights rule over the zero teacher and the K predictive teachers. Consequently, the online procedure adapts the teacher weights according to their observed contribution to the squared norm of the prediction-powered gradient, without explicitly estimating teacher variances or cross-teacher covariances.

H Confidence model analysis and selection

This appendix discusses the selection of confidence models for C-PP and CM-PP. The appropriate signal depends on the source of teacher unreliability. Output-based confidence is suitable when errors are accompanied by predictive uncertainty, whereas structural confidence or a learned router is preferable when a teacher remains confident outside its region of expertise.

The effective term s in Lemma 4.2 provides a tractable interpretation of this choice: a favorable confidence model assigns greater weight where the teacher is locally useful and suppresses its correction where the student is preferable. Since s controls a variance upper bound rather than the exact variance, pseudo-label accuracy and calibration are intermediate diagnostics rather than selection objectives. The exact criterion depends on gradient covariance and variance, as formalized in equation 21. Candidate confidence models are therefore compared using validation performance or empirical gradient variance. Tables 7 and 8 summarize suitable candidates and empirical diagnostics, while Table 9 compares their runtime and memory requirements.

H.1 Empirical confidence selection

The selection procedure proceeds in four stages and does not require an additional labeled split beyond the validation samples already used for model selection.

Table 7: Confidence signals for common teacher reliability regimes. The table describes the signals by their observable behavior; formal definitions of the output-based scores are given in Section H.3.

Teacher behavior	Typical setting	Observable reliability signal
Its output becomes uncertain when it is wrong	A pretrained teacher used on a similar task or domain	The largest predicted probability, separation between the two leading classes, or concentration of the full predictive distribution
It supports only known labels	A specialist trained on a known subset of classes	A support indicator or available label metadata
It supports only an unknown subset of labels	A specialist whose coverage is not documented	A learned router, optionally combined with output uncertainty
It works only on part of the input distribution	A site-, domain-, or subpopulation-specific teacher	Domain membership, representation similarity, or a density-ratio score
Its reliability decreases with input quality	Blur, noise, compression, or augmentation	Output uncertainty, corruption metadata, or an input-quality model
It remains confident when it is wrong	Confident but inaccurate out-of-domain predictions	Domain membership, representation similarity, or teacher-student disagreement rather than output confidence
Different teachers are useful on different inputs	Class-, domain-, or subgroup-specific experts	A router or a separate reliability signal for each teacher

First, the relation between the teacher, student, and data is used to restrict the relevant confidence families. Output-based scores are natural when teacher errors are reflected by predictive uncertainty, whereas coverage indicators, routers, representation similarity, or learned evaluators are more appropriate when reliability is determined by label support, domain, or another structural property.

Second, candidates that are incompatible with the available runtime, memory, or storage budget are removed using the considerations in Section H.5.

The remaining candidates are then run with the student for several initial epochs. During these runs, the validation samples are used to record the reliability separation, coverage-quality tradeoff, and empirical gradient variance in Table 8; for multiple teachers, the dependence between their correction terms is also recorded. Evaluating the confidence signal after student training has begun is informative even when enough labeled data are available for an offline comparison, since some classes, attributes, or subpopulations may be learned faster than others, changing the relative advantage of the teacher during optimization.

Finally, the confidence model and its threshold or softness are selected from this reduced set using the validation trajectory and the corresponding variance diagnostics; calibration error alone does not determine the choice. An offline threshold estimate can be particularly useful under substantial domain shift or severe input corruption. In these regimes, a binary confidence rule or a cutoff that excludes unreliable regions is often preferable to a smoothly varying weight. The available small validation set can then be used to locate the teacher-student error crossover before the short training comparison, without introducing a separate calibration partition.

H.2 Exact variance and the variance-bound criterion

Lemma 4.2 uses s to characterize a tractable variance upper bound. The exact variance instead follows the identity below, which also motivates selection by empirical gradient variance or validation performance.

Let $G_Y = \nabla_w \ell(w; X, Y)$ and $G_f^c = c(X)\nabla_w \ell(w; X, f(X))$. For any measurable $c(X) \in [0, 1]$ with finite second moments,

$$\mathrm{Var}\big(g_{\mathrm{CPP}}^{\lambda}(w)\big) = \frac{1}{n}\mathrm{Var}(G_Y) - \frac{2\lambda}{n}\mathrm{Cov}(G_Y, G_f^c) + \lambda^2\left(\frac{1}{n} + \frac{1}{N}\right)\mathrm{Var}(G_f^c). \tag{21}$$

Thus, before projection to the feasible interval, the variance-minimizing correction strength is determined by the covariance between G_Y and G_f^c relative to the variance of G_f^c.

Table 8: Empirical diagnostics for confidence selection.

Diagnostic	What to measure	Desired behavior
Reliability separation	Teacher–student error gap across confidence bins	The teacher has lower error in trusted regions
Threshold crossover	Conditional teacher and student errors across the score	A stable crossing near the selected threshold
Coverage–quality	Accepted fraction and accepted-region teacher error across thresholds	A smooth and useful quality–coverage tradeoff
C-PP efficiency	Empirical gradient variance or covariance-to-variance gain	Lower variance than unweighted PP and labeled-only training
CM-PP redundancy	Covariance or correlation between residual teacher corrections	Complementary rather than fully redundant directions
Cached-confidence stability	Teacher–student ordering at several training checkpoints	Few high-margin reversals; otherwise refresh confidence

Derivation. The labeled and unlabeled teacher terms have the same expectation. Expanding the variance of the resulting mean-zero correction gives equation 21; minimizing this quadratic gives the stated covariance-to-variance criterion.

This result does not prescribe a universal probability cutoff: the useful threshold depends on the teacher, student, task, and stage of training. It also need not minimize the exact variance in equation 21.

The confidence rule used in the current C-PP or CM-PP gradient must be fixed before any label-dependent update from that mini-batch. The same labeled samples may update the confidence model afterward, with the new rule used only in later iterations. Moreover, the labeled and unlabeled correction terms must use the same measurable correction map and augmentation distribution. Independent calibration data and cross-fitting are alternative ways to enforce these conditions.

H.3 Confidence model types

Confidence may be binary or continuous. Binary confidence provides a clear accept-or-reject decision, whereas soft confidence is often preferable when teacher quality changes gradually. A per label threshold can be found in the binary case. Class-adaptive thresholding can also be applied to cached scores for uneven class learning, unknown coverage, or partial-teacher routing.

In either case, $c(x)$ controls local teacher use, while λ_t controls the overall strength of the correction. For CM-PP, each teacher has its own $c_j(x)$ and aggregation weight $\lambda_{t,j}$; the confidence scores need not be jointly calibrated across teachers.

Fixed confidence with online adaptation. A fixed or cached confidence model does not imply a fixed teacher contribution. The confidence $c_j(x)$ specifies where teacher j is locally reliable, while $\lambda_{t,j}$ adapts the overall contribution of that teacher as the student changes. Since optimizing λ_t is one-dimensional for C-PP and involves only $K+1$ weights for CM-PP, it can often respond on a shorter empirical time scale than the high-dimensional student optimization. Thus, even with fixed confidence, the method adapts the correction strength during training rather than selecting a single fixed λ only in hindsight.

Output-based confidence. For a predictive distribution $p_f(x)$, common scores include maximum softmax probability, top-two margin, normalized top-two margin, and normalized inverse entropy:

$$p_{(1)}(x), \qquad p_{(1)}(x) - p_{(2)}(x), \qquad 1 - \frac{p_{(2)}(x)}{p_{(1)}(x)}, \qquad 1 - \frac{H(p_f(x))}{\log |\mathcal{Y}|}.$$

Maximum softmax probability (MSP). MSP is a simple default when lower peak probability accompanies errors. In the s-based criterion, it is favorable when conditional teacher error decreases as $p_{(1)}(x)$ increases, so high-MSP samples receive the larger teacher contribution. Empirically, teacher accuracy should increase across MSP bins and the teacher–student error gap should favor the teacher above the selected threshold.

Normalized top-two margin. The normalized margin is particularly useful when failures arise from competition between a few similar classes. Its benefit follows when increasing separation between the two leading classes corresponds to decreasing conditional teacher error and therefore a smaller contribution to s on the selected region. Empirically, teacher error should decrease with the margin, with a stable teacher-student crossover across nearby thresholds.

Entropy. Entropy is useful when uncertainty becomes diffuse across several classes, as can happen under broad corruption or shift. Normalized inverse entropy is favorable when concentrated predictive distributions have lower conditional teacher error, reducing the corresponding term in s. Empirically, teacher error should increase with entropy, and entropy should separate useful from unreliable predictions more clearly than MSP or margin when probability mass is distributed across several alternatives.

All three output-based scores can fail when a teacher is confidently wrong. Their interpretation therefore requires the score to be meaningfully related to conditional teacher error; poor calibration weakens this connection.

Hard thresholds versus soft weights. A hard threshold is appropriate when teacher usefulness changes sharply at a score boundary. The useful threshold is where the teacher and student have similar conditional quality, rather than a universal confidence value. Soft weighting is preferable when medium-confidence predictions remain informative and teacher quality deteriorates gradually. The coverage-quality diagnostic in Table 8 helps distinguish these cases. Under the s-based criterion, binary confidence approximates the pointwise choice between the teacher and student, whereas soft confidence interpolates between them. Empirically, a sharp sign change in the binned teacher-student error gap favors a threshold; a gradual change with useful intermediate-confidence samples favors soft weighting.

Structural confidence and routers. Output uncertainty is unnecessary when applicability is known from the task. A support indicator can encode known label coverage, and metadata can encode site, domain, sensor, modality, or subpopulation membership. When coverage is unknown, a learned router may produce either a hard assignment to one teacher or soft teacher-specific scores. Soft routing permits several teachers to contribute to the same example. This is useful for partial-label, domain-specific, hierarchical, and overlapping specialists, especially when they remain confident outside their expertise. In the s-based interpretation, routing errors are weighted by the local teacher-student performance gap and are most costly when the wrong source is selected far from a tie. Empirically, router coverage, teacher error within each routed region, and the corresponding teacher-student gap should be reported; for multiple teachers, residual correlations reveal whether overlapping routes are redundant or complementary.

Learned evaluators, agreement, and reward signals. A learned confidence model may combine the input, teacher output, representations, metadata, teacher-student disagreement, or outputs from other teachers. This is useful when no single signal has a monotone relationship with teacher error. Correctness predictors, quality evaluators, reward models, and alignment scores can all be used after mapping their outputs to $[0, 1]$. An arbitrary quality score should not automatically be interpreted as calibrated confidence: it remains useful only if it separates regions in which the teacher correction is beneficial. Mathematically, improved calibration alone does not guarantee a lower exact variance; the learned score should improve the covariance-to-variance relation in equation 21 or reduce its empirical counterpart. Empirically, the evaluator is assessed through teacher-student error separation across score bins and the variance of the resulting C-PP or CM-PP gradient.

Augmentation-based confidence. When teacher reliability varies across transformations, confidence may be computed from a weak view and used to train the student on a paired strong view, as in Match-style methods. Weak transformations should preserve teacher reliability, while strong transformations should improve student invariance without destroying label information. Typical randomness includes crop or translation offsets, flips, photometric or affine transformations, Cutout masks, and RandAugment operation sequences. The confidence signal may use output uncertainty, augmentation metadata, or cross-view agreement. Under the s-based interpretation, it should suppress views where teacher error exceeds paired-view student error; this can be evaluated through error and

coverage across augmentation families and severities. Sample-mixing augmentations additionally require the corresponding transformed target.

Because the augmented view is resampled, augmentation-dependent confidence requires a new forward pass for each active teacher and view at every iteration. This captures view-specific reliability but increases runtime and prevents confidence from being cached per original example. When reliability varies little across augmentations, or the same results are visible on the clean input's output, confidence from a cached teacher softmax vector may retain similar, though sometimes weaker, separation without repeated inference. Stability of teacher-student separation across augmentations and training checkpoints can guide the choice between cached and online confidence.

Teacher-student disagreement. Disagreement provides a dynamic confidence signal when the relative usefulness of a fixed teacher changes as the student improves. It is informative when larger disagreement identifies regions with a less useful teacher correction, so suppressing those regions reduces s or the exact correction variance. Empirically, teacher and student errors should be compared across disagreement bins at several training checkpoints; a changing crossover indicates that the signal should be refreshed rather than cached.

H.4 Teacher, student, and data behavior

Partial label coverage. When the supported classes are known, a support indicator or label metadata is the most direct confidence model. If support is unknown, a router can learn it from labeled calibration data. Output confidence is suitable only when unsupported classes also reduce MSP or margin or increase entropy. Routing errors matter most when the teacher and student have a large performance gap; mistakes near a tie have little effect on the bound represented by s.

Partial domain or covariate coverage. A teacher may cover all labels but remain reliable only on part of the input space. Domain routers, representation similarity, and density-ratio scores are natural choices. A confidence rule learned on the source domain transfers only when the selected target region remains sufficiently similar and the relevant labeling mechanism is stable. Output confidence is appropriate when leaving the native domain also changes the teacher's predictive distribution.

Noisy or corrupted inputs. MSP, margin, and entropy work well when corruption makes the teacher uncertain, this is a common case; entropy is particularly natural when probability mass becomes diffuse. If the teacher remains confidently wrong, augment-based methods can highlight the differences and test stability. If runtime is scarce, use a domain router, or disagreement instead. For a threshold based on corruption severity, the useful boundary is where the teacher and student have comparable conditional quality.

An improving student. A fixed confidence model does not imply a fixed teacher contribution. The local confidence $c(x)$ describes where the teacher is trusted, while the online weight λ_t can reduce or increase the overall correction as the student changes. Cached confidence becomes stale only where student improvement reverses the teacher-student ordering. Such reversals are important when the performance gap is large and negligible near a tie. If this ordering changes materially, refresh a dynamic signal such as disagreement using held-out, cross-fitted, or past labeled data.

Complementary and overlapping teachers. Teacher-specific confidence identifies where each source is locally useful; CM-PP then accounts for dependence between their residual corrections. Hard exclusive routing prevents simultaneous activation but does not, by itself, guarantee zero centered covariance. Soft routers may activate overlapping teachers whose corrections are positively or negatively correlated. The learned $\lambda_{t,j}$ values and residual-covariance diagnostic reveal whether the teachers are redundant or complementary. This setting is most useful when no single teacher dominates across the entire distribution.

Hierarchical teachers. When expertise follows a known coarse-to-fine taxonomy, a coarse router can select the relevant specialist before fine-label prediction. Performance then depends on both routing errors between superclasses and prediction errors within the selected specialist. Reporting these two components separately makes clear whether confidence or the specialist itself is the limiting factor.

Table 9: Runtime and memory cost of common confidence implementations. Online methods may additionally require the listed transient mini-batch memory.

Implementation	Extra forward cost	Persistent / transient storage	Use
Known support / metadata	None	None or $O(N)$ persistent	Fixed structural applicability
Cached scalar score + pseudo-label	None after preprocessing	$O(NK)$ persistent	Cached confidence and pseudo-label per teacher and sample
Cached full teacher outputs	None after preprocessing	$O(NKC)$ persistent	Supports MSP, margin, entropy, disagreement, and post-hoc thresholds
Cached router / route	None after preprocessing	$O(N)$–$O(NK)$ persistent	Structural or hierarchical teacher selection
Match-style augmentation confidence	About one teacher forward per active teacher and view	$O(bK)$–$O(bKC)$ transient	Recompute confidence and pseudo-labels on the current view
Precomputed augmented outputs	None during training	Up to $O(VNKC)$ persistent	Trades teacher inference for storage
Frozen learned evaluator	None after preprocessing	Typically $O(NK)$ persistent	Cached evaluator scores from fixed inputs
Online learned evaluator	Evaluator plus required teacher or student forwards	Batch-dependent transient memory	Confidence depends on changing inputs or student state
Jointly trained evaluator	Evaluator forward and backward	Parameters, optimizer state, and activations	Confidence model itself evolves during training
Teacher-student disagreement	Zero or one teacher forward if teacher outputs are cached	$O(NK)$–$O(NKC)$ if cached	Tracks changing local teacher–student reliability
All teachers online	K teacher forwards per required view	Up to $O(bKC)$ transient	Reevaluate every teacher on the current view
Router + selected specialist	About one router and one specialist forward	Optional	Evaluate only one or a few specialists

Routing over teacher subsets. When the oracle routing structure selects one teacher or a subgroup $\mathcal{A}(x)$ whose maximum size is smaller than K, routing can precede aggregation. Each route then applies MM-PP or CM-PP only to its active subgroup, excluding irrelevant teachers and reducing the dimension of the corresponding correction. Hard routing activates and updates only one subgroup per iteration. With soft routing, however, several subgroups may be active simultaneously and require separate online λ updates, increasing the per-iteration runtime. This creates a tradeoff between more targeted aggregation and the cost of maintaining multiple route-specific corrections.

H.5 Runtime and memory considerations

The main computational distinction is whether confidence can be cached or must be recomputed during training. MSP, normalized margin, entropy, frozen routers, and frozen evaluators are cacheable when their inputs are fixed. Confidence becomes online when it depends on a fresh augmentation, the current student, an updated teacher, or a changing evaluator. Network inference generally dominates the cost of computing the scalar confidence score.

For example, in the Match-style augmentation setting, confidence and pseudo-labels are computed from a stochastic weak view, while the student is trained on a paired strong view. Because the weak view changes across iterations, its confidence generally cannot be cached per original example. One active teacher requires approximately one forward pass per unlabeled mini-batch and confidence view; evaluating all K teachers requires K forwards, while routing can activate only a subset.

The storage tradeoff can be stated directly. Let b denote the unlabeled mini-batch size, N the number of unlabeled examples, K the number of teachers, C the number of classes, and V the number of precomputed augmented views. Caching complete teacher distributions requires $O(NKC)$ storage, whereas caching only pseudo-labels and confidence scores requires $O(NK)$. Online evaluation removes this persistent cache and stores only $O(bKC)$ transient probabilities, or $O(bK)$ hard pseudo-labels and confidence values, but repeats the teacher forwards whenever samples are revisited - which require runtime and memory as well. Precomputing V augmented outputs per teacher and sample instead requires VNK teacher evaluations once and up to $O(VNKC)$ persistent storage. This eliminates teacher inference during student training, but provides less augmentation diversity than fresh online sampling and may require substantially more storage.

For binary confidence derived from cached teacher scores, the thresholds can still be updated online using running statistics. This requires only $O(bK)$ scalar operations per mini-batch and no additional teacher forward passes; class-specific thresholds require $O(KC)$ stored values.

Teacher outputs, the confidence rule, and the aggregation weights therefore form three distinct levels of adaptation. Caching teacher outputs does not prevent the confidence thresholds or λ_t and $\lambda_{t,j}$ from adapting during training. Confidence must be recomputed from current model outputs only when the underlying reliability signal changes with the augmentation, teacher, student, or evaluator state. The stability diagnostic in Table 8 can indicate whether this additional cost is warranted.

I EXPERIMENTAL DETAILS

Overview. We organize the experiments around three complementary settings. First, controlled single-teacher experiments on MNIST, Covertype, CIFAR-10, and SVHN study partial teacher coverage and confidence quality. Second, multi-teacher experiments on Covertype and PASCAL VOC evaluate confidence-aware aggregation under overlapping regions of teacher reliability. Third, Tiny-ImageNet evaluates a frozen, off-the-shelf teacher under corruption-induced teacher-reliability shift. Together, these settings test partial coverage, model-derived confidence, heterogeneous teacher specialization, reliability shift, and dense prediction. All experiments were run on NVIDIA L40S GPUs.

I.1 COMMON EXPERIMENTAL PROTOCOL

Labeled and unlabeled sampling. Throughout this section, n and N denote the numbers of labeled and unlabeled examples in a mini-batch, respectively, and $r = \frac{n}{N}$ denotes the labeled-to-unlabeled ratio. For every experiment, the labeled-to-unlabeled ratio within each mini-batch is chosen to equal the corresponding ratio between the overall labeled and unlabeled sample budgets. Consequently, one epoch processes every labeled and unlabeled training example once. When we vary r, we keep the overall labeled budget and labeled mini-batch fixed and vary only the overall unlabeled budget and the unlabeled mini-batch size.

Comparison design. Labeled-only supervised learning is the common reference in all settings. Each experiment also includes the prediction-powered baseline corresponding to its teacher configuration and its proposed confidence-aware counterpart. We add setting-specific baselines when they isolate the question of interest: standard semi-supervised methods for single-teacher image classification, knowledge-distillation methods for pretrained teachers, and uniform or confidence-agnostic aggregation for multi-teacher settings. The precise methods and confidence construction are specified with each experiment.

Hyperparameter search and selection. For each method and experimental configuration, we search over the relevant hyperparameters and refine the grid around the strongest-performing region. Hyperparameters are selected according to mean validation performance across student seeds. For each dataset, the validation set is disjoint from the student-training pool and has the same number of examples as the corresponding test set. The reported labeled budget counts only trusted examples used for student training. All methods use the same validation set, and the test set is used only for final evaluation.

Architectures and data splits. On MNIST (LeCun et al., 1998), the teacher and student are two-layer convolutional networks. On SVHN (Netzer et al., 2011), both are ResNet-18 models, and on CIFAR-10 (Krizhevsky, 2009), both are ResNet-9 models. The Covertype (Blackard, 1998; Blackard & Dean, 1999) experiments use TabResNet-18 teachers and students. On Tiny-ImageNet (Le & Yang, 2015), the teacher is an off-the-shelf pretrained ResNet-50 and the student is a ResNet-18. For PASCAL VOC 2012 and SBD (Everingham et al., 2010; Hariharan et al., 2011), the student is DeepLabV3+ with a ResNet-101 backbone, and the frozen teachers are DeepLabV3–ResNet-101, FCN–ResNet-50, and LR-ASPP–MobileNetV3. We use the default train/test partitions provided by `torchvision`. Labeled and unlabeled sets are sampled from the training split, and final accuracy is reported on the corresponding test split. All compared methods are evaluated under the same data splits, student architecture, training budget, and model-selection protocol.

Fairness and random seeds. Within each dataset, all methods use the same labeled and unlabeled splits, student architecture, and training budget. When a teacher is trained, its training examples are removed from the student's training data. All methods in a corruption experiment receive the same corrupted inputs or cached teacher outputs. The smaller controlled experiments use two independent student seeds; PASCAL VOC, Tiny-ImageNet and CIFAR-100 use four. Seeds affect the data split, model initialization, data order, and student optimization. Teachers are trained once per configuration and reused across student runs. We report means and one empirical standard deviation across student seeds.

I.2 Controlled experiments

I.2.1 MNIST and single-teacher Covertype

MNIST partial teachers and confidence. Each MNIST teacher is trained only on a restricted subset of digit labels, whereas the student is trained and evaluated on the full ten-class task. We use teachers supported on labels `0-2` and `0-7`, corresponding to approximately 30% and 80% coverage. The teachers achieve 97–99% accuracy on their respective training distributions. For a teacher supported on S_t, oracle confidence is $c_t(x, y) = \mathbf{1}\{y \in S_t\}$. This controlled signal isolates coverage-aware weighting from the separate problem of estimating confidence. The setting corresponds to the construction in Appendix C.

Single-teacher Covertype construction. The single-teacher Covertype experiment is the tabular counterpart of the MNIST partial-teacher setting. A TabResNet-18 teacher is trained on a restricted portion of the data and covers 43% of the student distribution, whereas the TabResNet-18 student is trained and evaluated on the full task. This experiment tests whether confidence-aware correction remains useful when partial coverage is defined over a tabular data distribution rather than over a subset of image labels.

Setup. The overall MNIST labeled budget is 300. We compare supervised learning, pseudo-labeled supervised learning, PP-SSL, C-PP, FixMatch (Sohn et al., 2020), and MixMatch (Berthelot et al., 2019). Figures 3 and 4 show the results for the two partial teachers. Table 10 additionally reports the corresponding single-partial-teacher Covertype experiments.

Oracle confidence is used only in controlled experiments to isolate the effect of correctly identifying the teacher-reliable region and to quantify the remaining opportunity for improved practical confidence estimation.

Results. Table 10 examines two complementary sources of difficulty. On MNIST, we fix the overall labeled budget and vary teacher coverage from 30% to 80%. C-PP improves substantially as coverage increases, whereas confidence-agnostic PPI remains close to the supervised baseline under limited coverage and becomes beneficial mainly in the higher-coverage setting. C-PP also outperforms FixMatch and MixMatch in both reported MNIST settings, showing that its gain is not explained solely by access to unlabeled data.

On Covertype, we instead fix teacher coverage at 43% and vary the overall labeled budget from 50 to 250. C-PP improves over supervised learning, confidence-agnostic PPI, and FixMatch at both budgets. The improvement in the 50-sample regime shows that confidence-aware correction remains useful when trusted labels are particularly scarce, while the 250-sample result shows that the benefit persists as the supervised baseline becomes stronger.

Together, these experiments test complementary axes: MNIST isolates sensitivity to teacher coverage at a fixed labeled budget, whereas Covertype isolates sensitivity to the labeled budget at fixed teacher coverage. In both cases, confidence enables prediction-powered learning to exploit the teacher's reliable region without being dominated by predictions outside that region.

I.2.2 MNIST interpolation-weight analysis

We compare supervised learning, confidence-agnostic PPI, and C-PP over $\lambda \in \{0, 0.05, \ldots, 1\}$ for four partial teachers on labels `0-4`, `5-9`, `even`, `odd`. We also compare the multi-teacher variants MM-PP and CM-PP over a fixed grid of feasible aggregation vectors.

Table 10: Single-partial-teacher test accuracy on Covertype and MNIST.

Dataset	r	Labeled budget	Coverage	Supervised	C-PP	PPI	FixMatch
Covertype	0.003	50	43%	0.492 ± 0.016	$\mathbf{0.514 \pm 0.016}$	0.492 ± 0.017	0.491 ± 0.020
Covertype	0.003	250	43%	0.605 ± 0.007	$\mathbf{0.627 \pm 0.004}$	0.625 ± 0.003	0.604 ± 0.011
MNIST	0.05	300	30%	0.7355 ± 0.0131	$\mathbf{0.7490 \pm 0.0164}$	0.7269 ± 0.0101	0.7132 ± 0.0371
MNIST	0.05	300	80%	0.7371 ± 0.0109	$\mathbf{0.8218 \pm 0.0205}$	0.7462 ± 0.0083	0.7488 ± 0.0039

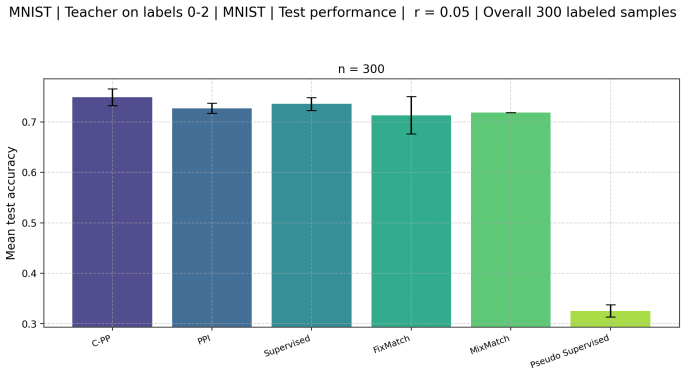


Figure 3: MNIST test accuracy for the `0-2` partial-label teacher with an overall labeled budget of 300. Bars report mean test accuracy for C-PP, PPI, supervised learning, FixMatch, MixMatch, and pseudo-labeled supervised learning.

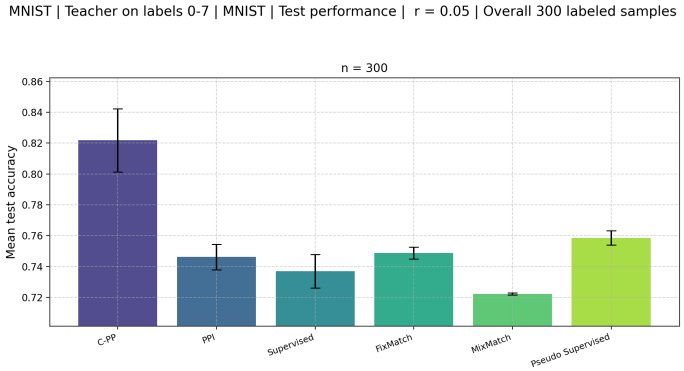


Figure 4: MNIST test accuracy for the `0-7` partial-label teacher with an overall labeled budget of 300. Bars report mean test accuracy for C-PP, PPI, supervised learning, FixMatch, MixMatch, and pseudo-labeled supervised learning.

Table 11 compares single- and multi-teacher prediction-powered methods when the interpolation weights are selected using only a fixed grid. Because the four partial teachers have the same nominal coverage, this comparison isolates how the identity of the supported labels and teacher quality affect both student accuracy and the selected balance between supervised and teacher-based information. Here, λ_0 denotes the weight assigned to the supervised gradient, while the remaining weight is assigned to the prediction-powered teacher corrections. In the single-teacher experiments, $\lambda_0 = 1 - \lambda$.

Results. For every partial teacher, confidence-agnostic PPI selects a larger λ_0 than C-PP. Thus, without confidence, the selected solution relies more strongly on supervised learning to limit the effect of unreliable pseudo-labels. Confidence-aware correction permits greater use of the teacher signal while achieving higher accuracy.

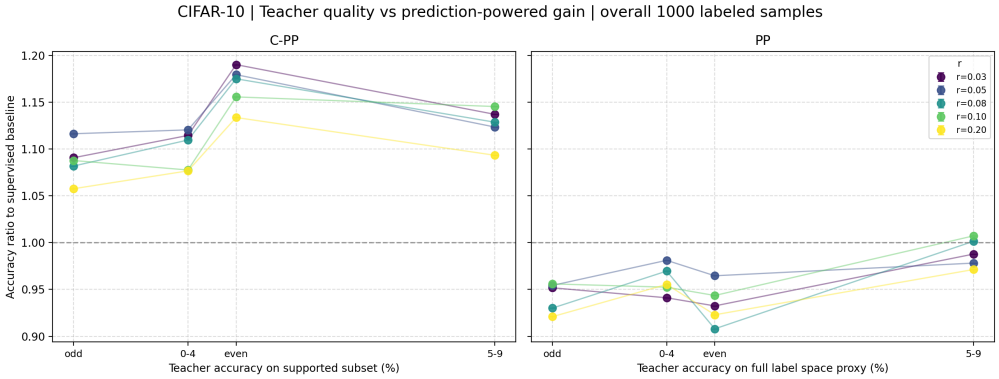


Figure 5: Teacher quality versus prediction-powered gain on CIFAR-10 across labeled-to-unlabeled ratios. Values above 1 indicate improvement over supervised learning.

The same effect is more pronounced in the multi-teacher comparison. MM-PP selects a large supervised weight, because the aggregate noise across the partial teachers weakens the utility of the confidence-agnostic correction. CM-PP instead selects the smallest supervised weight and achieves the strongest result. Confidence therefore makes the complementary multi-teacher signal beneficial even when confidence-agnostic aggregation is strongly damped. Notably, even when the confidence-agnostic multi-teacher correction is strongly damped by a large supervised weight, combining the partial teachers still outperforms every corresponding single-teacher PPI variant, demonstrating the benefit of aggregating complementary teacher information.

Table 11: Accuracy results by teacher and method for fixed-grid analysis of interpolation weights on MNIST using four partial-label teachers with identical nominal coverage. The label below each bar denotes the selected supervised weight λ_0, with $r = 0.05$.

Teacher	Method	λ_0	Accuracy $\pm$ std
Supervised	Supervised	1.00	0.975 ± 0.003
Full	CM-PP	0.00	$\mathbf{0.979 \pm 0.002}$
	MM-PP	0.90	$\mathbf{0.977 \pm 0.003}$
0-4 labels	C-PP	0.50	0.967 ± 0.002
	PP	0.75	0.961 ± 0.006
5-9 labels	C-PP	0.50	0.966 ± 0.002
	PP	0.75	0.962 ± 0.000
Even labels	C-PP	0.50	0.968 ± 0.004
	PP	0.75	0.960 ± 0.002
Odd labels	C-PP	0.50	0.971 ± 0.003
	PP	0.75	0.962 ± 0.003

I.2.3 CIFAR-10

Partial teachers and confidence. The CIFAR-10 teachers are restricted to the label subsets `even`, `odd`, `0-4`, and `5-9`. Table 12 shows that they are accurate on their supported subsets but substantially weaker on the full student distribution. C-PP uses oracle support confidence, whereas PPI treats the teacher as confidence-agnostic.

Setup and results. We compare supervised learning, confidence-agnostic PPI, and C-PP over $\lambda \in \{0, 0.05, \ldots, 1\}$ and several labeled-to-unlabeled ratios. Figure 2 reports accuracy relative to supervised learning. As the unlabeled budget grows, C-PP consistently remains above the supervised baseline, whereas PPI is more sensitive to incomplete teacher coverage and the selected interpolation weight. Figure 5 further shows that the gain is related to teacher quality but is not determined by it alone: confidence coverage, error structure, and the selected interpolation weight also contribute.

Table 12: CIFAR-10 partial-teacher performance on the supported label subset and on the full student training distribution.

Teacher	Partial labels	Supported-subset acc.	Full-data acc.
$T_{0\text{-}4}$	$\{0,1,2,3,4\}$	90.5%	45.8%
$T_{5\text{-}9}$	$\{5,6,7,8,9\}$	95.6%	47.8%
T_{odd}	$\{1,3,5,7,9\}$	88.6%	44.3%
T_{even}	$\{0,2,4,6,8\}$	91.3%	45.6%

I.2.4 SVHN Confidence-Mechanism Analysis

Setup. We evaluate the coverage-quality diagnostic of Appendix H.1 on SVHN using seven partial-label teachers and a full-label teacher under homogeneous input corruption. The partial teachers instantiate Appendix C: each is trained on a subset of the ten labels but evaluated on the full distribution. We distinguish *label support*, the fraction of labels used to train a teacher, from *sample mass* τ, the fraction of evaluation examples retained by a confidence rule.

We evaluate the teachers on the SVHN training split, disjoint from the *extra* split used to train them. Binary thresholds trace the risk–coverage curve, while soft weighting applies each score directly. Predicted-label-specific policies use a separate threshold for each predicted class, with retained mass measured over the full evaluation population. The reported $\varepsilon^{f,1} \leq 0.15$ operating points are retrospective diagnostics and are not used to tune the student experiments. Homogeneous corruption adds Gaussian noise in $[0,1]$ pixel space before normalization; corruption results average two independent draws.

We compare MSP, raw and normalized margin, inverse entropy, and augmentation stability. Augmentation stability averages the probability assigned to the original prediction across two weak crop-and-color views. These views differ from the Gaussian pixel corruption used to induce distribution shift. The first four use cached teacher outputs, whereas augmentation stability requires additional teacher forwards. We report the complete confidence-coverage curves and summarize them at

$$\tau_{.15} = \max_{q} \{\tau(q) : \varepsilon^{f,1}(q) \leq 0.15\},$$

corresponding to at least 85% teacher accuracy among accepted examples. Thus, the conclusions do not depend on a single threshold q. We consider a global rule with one threshold q for all predictions and a per-label rule with a separate threshold q_k for each teacher-predicted class k. Each q_k maximizes retained mass within its class subject to class-specific accepted error at most 0.15. We use AUROC to compare the complete threshold sweeps. For continuous weighting, soft mass is the mean confidence weight and soft error is the corresponding weighted teacher error.

Partial teachers. Table 13 separates nominal label support from its prevalence in the evaluation distribution. Although the 0–4, 5–9, even, and odd teachers each support five labels, their oracle-region masses range from 0.381 to 0.619 because SVHN is class-imbalanced. Their accuracies within these regions are nearly identical (0.945–0.948), so the large difference in full-distribution accuracy primarily reflects which labels are supported, rather than teacher quality within the supported region.

We report inverse entropy as the representative cached score; MSP and the raw and normalized margins produce nearly identical confidence–coverage curves, with within-teacher AUROC differences of at most 0.004. With separate thresholds for each predicted label, the retained sample mass is 62%-90% of the corresponding oracle-region mass, while teacher accuracy remains similar to that within the oracle region. Inverse entropy and augmentation stability are nearly indistinguishable: their retained masses differ by at most 0.001, except for the 0–7 teacher, where augmentation retains 0.019 additional mass at lower accepted accuracy. For the 0–6 and 0–7 teachers, global thresholds extend beyond the oracle-supported region while maintaining the target accepted accuracy, but no global threshold is feasible for the more restricted teachers. Thus, the available reliable mass is governed mainly by the prevalence of the teacher-supported region, and cached inverse entropy recovers essentially the same mass as augmentation stability in this structured shift.

Table 13: Partial-teacher operating points at $\varepsilon^{f,1} \leq 0.15$. Entries report sample mass / teacher accuracy; "–" denotes no admissible global threshold.

Teacher	Label support	Full acc.	Oracle mass / acc.	Global mass / acc.		Per-label mass / acc.	
				Inv. entropy	Aug. stability	Inv. entropy	Aug. stability
0–2	3/10	.390	.401 / .971	–	–	.251 / .971	.251 / .971
0–4	5/10	.587	.619 / .948	–	–	.448 / .948	.448 / .948
0–6	7/10	.742	.791 / .938	.830 / .852	.837 / .852	.655 / .938	.654 / .939
0–7	8/10	.809	.868 / .933	.937 / .852	.939 / .851	.766 / .933	.785 / .910
5–9	5/10	.361	.381 / .947	–	–	.235 / .947	.235 / .948
Even	5/10	.437	.461 / .948	–	–	.299 / .948	.299 / .949
Odd	5/10	.509	.539 / .945	–	–	.369 / .945	.369 / .946

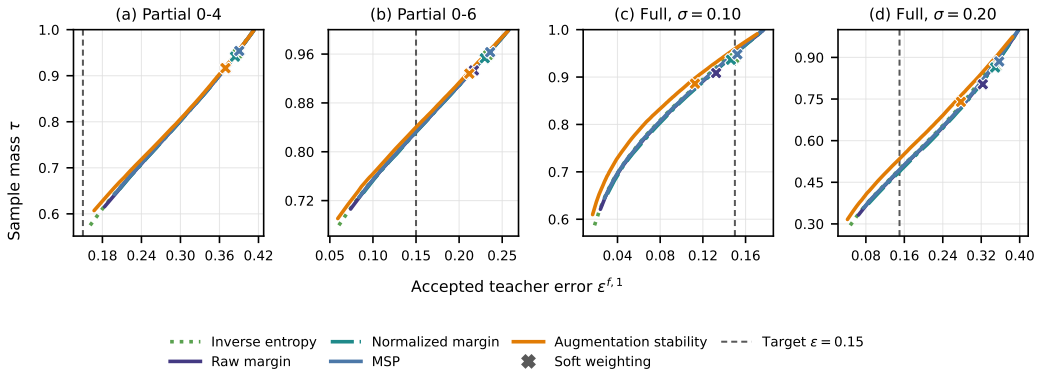


Figure 6: Risk-coverage curves under partial support and homogeneous Gaussian corruption. Curves vary the binary threshold, the dashed line marks $\varepsilon^{f,1} = 0.15$, and crosses denote soft weighting. Augmentation stability averages the probability of the original prediction across two weak views.

Corruption and label-specific thresholds. Corruption reveals a regime-dependent difference between the confidence signals. The full teacher's accuracy decreases from 0.925 without corruption to 0.823 and 0.602 at $\sigma = 0.10$ and 0.20, respectively. Figure 6 shows representative threshold sweeps, while Table 14 reports the $\varepsilon^{f,1} \leq 0.15$ operating points. Under mild corruption, the binary scores behave similarly, each retaining approximately 95% of the sample. The cacheable soft scores differ mainly in selectivity, with raw margin trading mass for lower error. Under stronger corruption, augmentation stability shifts the risk-coverage frontier: relative to the best cacheable score, it retains 0.039 more global mass and 0.031 more mass under predicted-label-specific thresholds. All ten labels remain feasible, so this gain is not driven by a single easy class. Under soft weighting, augmentation stability reduces weighted error from 0.323 to 0.278, but assigns less effective mass. Thus, its advantage under stronger corruption is better concentration of teacher influence rather than uniformly greater coverage.

Overall, cacheable confidence and augmentation stability behave similarly under partial support and mild corruption, but diverge as corruption increases. This supports the mechanism of Appendix H: confidence is useful through the joint $(\tau, \varepsilon^{f,1})$ tradeoff rather than through coverage or error alone.

I.2.5 SVHN Student Performance

Setup. We evaluate downstream student performance with the partial 0–4 teacher using 400 labeled and 3600 unlabeled examples, with 4 labeled and 36 teacher-labeled samples per update. We compare confidence-agnostic PP-SSL, C-PP with cached raw-margin confidence, C-PP with online augmentation stability, and oracle-confidence C-PP. Because augmentation stability requires additional teacher forwards, we report both equal-epoch and time-matched performance.

Table 15 separates the benefit of confidence quality from the cost of obtaining it. Cached raw-margin confidence does not improve over PP-SSL, despite exhibiting a useful teacher-side risk-coverage

Table 14: Confidence under homogeneous corruption. Binary entries report sample mass / accepted accuracy at $\varepsilon^{f,1} \leq 0.15$; soft columns report effective mass and weighted error.

σ (full acc.)	Confidence	Global mass / acc.	Per-label mass / acc.	Soft mass	Soft error
.10 (.823)	Inverse entropy	.952 / .851	.903 / .870	.940	.149
	Raw margin	.949 / .850	.901 / .870	.909	.132
	Normalized margin	.947 / .851	.901 / .870	.937	.146
	MSP	.949 / .851	.902 / .870	.948	.152
	Aug. stability	.958 / .851	.901 / .872	.886	.113
.20 (.602)	Inverse entropy	.492 / .853	.503 / .855	.867	.351
	Raw margin	.491 / .851	.498 / .855	.803	.323
	Normalized margin	.479 / .855	.494 / .856	.864	.349
	MSP	.483 / .857	.494 / .859	.885	.357
	Aug. stability	.531 / .852	.534 / .854	.740	.278

Table 15: SVHN student accuracy with the partial 0–4 teacher. Time-matched augmentation agreement is evaluated at the training-time budget of the cached confidence reference.

Method	**Evaluation**	**Accuracy (%)**
PP-SSL	Equal epochs	44.17 ± 9.52
C-PP (raw margin)	Equal epochs	36.97 ± 3.96
C-PP (augmentation agreement)	Time matched	43.87 ± 6.92
C-PP (augmentation agreement)	Equal epochs	47.99 ± 5.07
C-PP (oracle)	Equal epochs	84.24 ± 0.07

tradeoff in Figure 6. In contrast, online augmentation stability improves over both cached confidence and PP-SSL at equal epochs. After matching the training-time budget, however, this advantage is reduced, illustrating that stronger online confidence can sometimes come with a non-negligible inference cost.

This behavior is consistent with Appendix H: teacher-side coverage and accepted error alone do not determine downstream utility; the confidence signal must induce a useful prediction-powered correction. Oracle C-PP substantially exceeds all practical methods, showing that the partial teacher contains useful signal and that stronger practical confidence could unlock substantially larger gains.

I.3 Heterogeneous multi-teacher experiments

I.3.1 Covertype

Partial-teacher construction. Covertype serves as our tabular benchmark for heterogeneous predictive sources. Unlike the label-restricted image teachers, this setting uses semantically biased tabular teachers.

The teacher pool consists of four TabResNet-18 models trained on differently biased subsets of the Covertype data. These subsets emphasize distinct sources of predictive structure: topographic features, hydrology and distance-to-water features, land-category indicators, and a weaker generalist teacher trained on a smaller, broadly sampled subset. This construction induces teacher-specific covariate misalignment: a teacher may be accurate on subpopulations aligned with its training bias but unreliable elsewhere. Table 16 summarizes the construction and intended role of each teacher. Figure 7 further shows the aggregate teacher signal by class for the canonical margin-vector setting. The high-confidence teacher signal is clearly class-dependent: teacher accuracy is high for Classes 1, 2, and 7, but weaker for Classes 3–6. High-confidence assignments are also unevenly distributed across labels. This motivates confidence-aware weighting rather than treating all teacher predictions uniformly.

Confidence model. For Covertype, we use a soft margin-based confidence score computed from the teacher softmax probabilities. Given teacher probabilities $p_t(y \mid x)$, let $p_{(1)}(x)$ and $p_{(2)}(x)$ denote

Table 16: Covertype partial-teacher construction. Each teacher is trained on a biased subset of the data, yielding heterogeneous expertise across regions of the input and label space.

Teacher	Subset bias	Expected expertise	Purpose in evaluation
Topography	High-elevation and topographically distinctive regions.	Stronger on examples whose class structure is correlated with elevation and terrain features.	Tests whether prediction-powered learning can exploit localized geographic and topographic expertise.
Hydrology	Regions near hydrological features, such as streams and water-related terrain attributes.	Stronger on examples for which distance to water or hydrological proximity is predictive.	Tests whether the method can use expertise specialized to a different feature subspace.
Land-cover	Wilderness- and land-cover-biased regions.	Stronger on classes associated with wilderness areas and land-cover patterns.	Tests robustness when teacher coverage follows semantic and environmental structure.
Generalist-small	A small, broadly sampled subset of the training data.	Weaker but less specialized coverage across the full data distribution.	Provides a weak general-purpose teacher and contrasts it with the specialized partial experts.

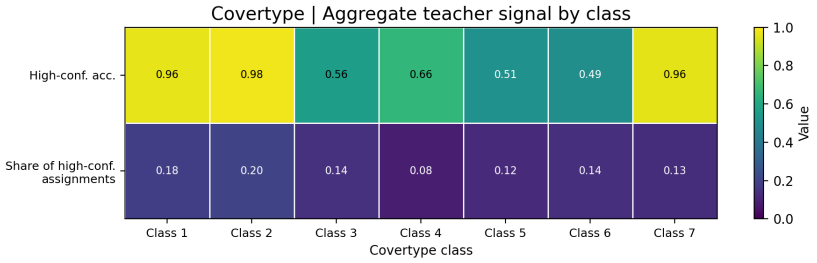


Figure 7: Aggregate teacher signal by Covertype class for the canonical margin-vector setting. The first row reports high-confidence teacher accuracy, and the second reports the share of high-confidence teacher assignments assigned to each class. Some classes receive highly accurate teacher predictions, whereas others receive weaker teacher signal, and high-confidence assignments are unevenly distributed across labels.

the largest and second-largest predicted class probabilities. We define

$$c_t(x) = p_{(1)}(x) - p_{(2)}(x),$$

Thus, a teacher has high confidence when its top prediction is well separated from the runner-up class and low confidence when its predictive distribution is ambiguous. This confidence model is lightweight, requires no auxiliary calibration model, and provides a local reliability signal for weighting teacher predictions.

Experimental setup. We use a TabResNet-18 student trained on the full Covertype feature space. The overall labeled budget is varied over $\{50, 250, 500, 1000\}$ samples. The number of unlabeled examples is set using $r = 0.0025$, where r controls the labeled-to-unlabeled sample ratio. All methods use the same four-teacher pool from Table 16. We compare supervised learning, Uniform MM-PP and Uniform CM-PP with a single optimized interpolation parameter, and MM-PP and CM-PP with optimized vector-valued interpolation parameters. The MM-PP variants are confidence-agnostic, whereas the CM-PP variants use margin-based teacher confidence. All interpolation parameters are optimized online.

Results. Figure 1 reports Covertype performance as the labeled budget varies. With online optimization of the interpolation parameters, all prediction-powered variants improve over the supervised baseline across the evaluated budgets, indicating that the teacher signal remains useful despite

teacher heterogeneity. Among these methods, CM-PP achieves the highest test accuracy for every labeled-sample budget, with the largest gains in the low-label regime.

The ratio plot further separates the contributions of the different design choices. CM-PP gives the largest improvement relative to supervised learning, followed by MM-PP and the uniform variants. This suggests that both components are beneficial: vector-valued interpolation improves over a single uniform interpolation parameter, and margin-based confidence further improves over confidence-agnostic aggregation. As expected, the relative benefit of prediction-powered learning decreases as n increases: with more labeled data, the supervised baseline becomes stronger and the marginal utility of pseudo-labeled teacher information is reduced. Nevertheless, all prediction-powered variants remain above the supervised baseline, and CM-PP remains consistently strongest.

I.3.2 PASCAL VOC

Teachers and structured reliability. PASCAL VOC is our dense-prediction benchmark. The three frozen teachers are DeepLabV3–ResNet-101, FCN–ResNet-50, and LR-ASPP–MobileNetV3; the student is DeepLabV3+–ResNet-101 and predicts all 21 classes. DeepLabV3 is weakened on {`bird`, `boat`, `cat`, `cow`, `dog`, `horse`, `sheep`}; FCN on {`aeroplane`, `bicycle`, `boat`, `bus`, `car`, `chair`, `motorbike`, `train`}; and LR-ASPP on {`bottle`, `chair`, `diningtable`, `pottedplant`, `sofa`, `tvmonitor`}. The overlaps on `boat` and `chair`, while leaving no class deliberately weak for all three teachers, create correlated but complementary regions of reliability.

Teacher corruption. For pixels whose ground-truth class belongs to a teacher's weak-class set, we apply Gaussian corruption

$$\tilde{x} = \mathrm{clip}_{[0,1]}(x + \sigma\epsilon), \qquad \epsilon \sim \mathcal{N}(0, I),$$

with $\sigma = 0.10$ for moderate corruption and $\sigma = 0.15$ for elevated corruption; other pixels are unchanged. Noise realizations are fixed across methods.

Confidence and compared methods. From each corrupted teacher distribution, we compute the normalized margin, raw margin, and normalized inverse entropy

$$c_{\text{marginal normed}}(x) = 1 - \frac{p_{(2)}(x)}{p_{(1)}(x)},\ c_{\text{marginal raw}}(x) = p_{(1)}(x) - p_{(2)}(x),\ c_{\text{entrophy}}(x) = 1 - \frac{H(p(x))}{\log 21}.$$

Spatial-margin confidence multiplies the normalized margin by the fraction of pixels in a 5×5 neighborhood that agree with the center pseudo-label.

$$c_{\text{spatial}}(u) = c_{\text{norm}}(u) \frac{1}{|\mathcal{N}_{5\times5}(u)|} \sum_{v \in \mathcal{N}_{5\times5}(u)} \mathbf{1}\{\widehat{y}(v) = \widehat{y}(u)\}.$$

All confidence maps are computed after corruption. We compare supervised learning, uniform knowledge distillation, confidence-weighted knowledge distillation, MM-PP, and CM-PP. The primary metric is mean intersection-over-union (mIoU), with pixel accuracy and semantic-group IoU as additional diagnostics.

Results. As shown in Table 2, all confidence-aware CM-PP variants outperform supervised learning and confidence-agnostic MM-PP, demonstrating the benefit of confidence-aware correction under class-dependent teacher reliability. Margin CM-PP achieves the highest mIoU overall. CM-PP also outperforms its matched knowledge-distillation baseline with margin and entropy confidence, while spatial-margin KD remains stronger than spatial-margin CM-PP in this setting.

I.3.3 CIFAR-100

Teacher construction and confidence. CIFAR-100 (Krizhevsky, 2009) provides a natural-transfer setting in which teacher heterogeneity arises from distinct pretrained architectures rather than constructed regions of specialization. The teacher pool consists of ImageNet-pretrained ResNet-50, ConvNeXt-Tiny, and ViT-B/16 backbones, which remain frozen; only linear CIFAR-100 prediction heads are trained. The heads are trained by two-fold cross-fitting using 1,000 trusted examples in total, so each head is fit on 500 examples (5 per class). For every trusted example used in the

labeled correction, the corresponding teacher prediction is produced by a head that was not trained on that example. The resulting out-of-fold head accuracies are 51.1%, 55.1%, and 60.4% for ResNet-50, ConvNeXt-Tiny, and ViT-B/16, respectively. Thus, the teachers provide useful but imperfect predictive signals of differing quality under transfer mismatch.

We use soft normalized-margin confidence. Oracle CM-PP uses teacher correctness only as a diagnostic upper bound.

Natural-transfer mechanism diagnostic. We use $n = 1{,}000$, $N = 39{,}000$, batch size 160 (4 trusted and 156 unlabeled), and 40 training epochs, corresponding to $n/N = 1/39$. This unlabeled-dominated regime probes whether confidence-aware correction remains useful with naturally heterogeneous pretrained teachers. CM-PP reaches $19.35 \pm 0.42\%$ Top-1 accuracy, compared with $15.88 \pm 0.95\%$ for MM-PP, a gain of 3.48 percentage points; Oracle CM-PP reaches $19.56 \pm 0.55\%$. Using the same confidence signal, confidence-weighted KD improves from $16.74 \pm 0.71\%$ to $17.23 \pm 0.46\%$, a gain of 0.49 points. Thus, in this setting, confidence yields a substantially larger improvement when incorporated into the prediction-powered correction than when used only to reweight distillation, while practical CM-PP nearly matches its oracle-confidence counterpart.

The improvement requires essentially no additional memory: standard and confidence-weighted KD use approximately 3.00 GB of peak allocated CUDA memory (4.46 GB reserved), while MM-PP, CM-PP, and Oracle CM-PP use approximately 3.02 GB (4.47 GB reserved).

I.4 Off-the-shelf teacher under reliability shift

I.4.1 Tiny-ImageNet

Teacher construction. The teacher is an off-the-shelf ImageNet-pretrained ResNet-50 from `torchvision`, used without fine-tuning; the student is a ResNet-18. We retain the pretrained 1000-way output head, select the logits corresponding to the 200 Tiny-ImageNet WordNet classes, and normalize over these classes. The overall training sample contains $10{,}000$ trusted labeled examples and $90{,}000$ unlabeled examples. We split the 200 Tiny-ImageNet classes into two equal groups and apply additive Gaussian noise with $\sigma = 0.01$ to one group and $\sigma = 0.15$ to the other. For class y, $\widetilde{x} = \mathrm{clip}_{[0,1]}(x + \sigma_y \epsilon)$, $\epsilon \sim \mathcal{N}(0, I)$, where $\sigma_y \in \{0.01, 0.15\}$. The same corruption is used across methods.

Confidence and baselines. We compare soft knowledge distillation, UDA, FreeMatch, PP-SSL, C-PP with soft normalized-margin confidence, C-PP with soft entropy confidence, C-PP with binary entropy confidence, and Oracle C-PP. Oracle confidence uses teacher correctness and is included only as a diagnostic upper bound. The practical confidence variants use only the teacher's predictive distribution.

Training and caching. The student is trained at 64×64 resolution. Weak augmentation uses a random crop with reflection padding and random horizontal flipping; strong augmentation additionally uses RandAugment. Gaussian corruption is applied before normalization. For methods based on a fixed input view, teacher targets and confidence scores are cached. For matched external-teacher baselines, UDA and FreeMatch evaluate the frozen teacher on each weak view and train the student on the paired strong view.

Results. As shown in Table 1, all practical confidence-aware variants improve over PP-SSL. Soft normalized margin and entropy confidence reach 31.01% and 31.69% Top-1 accuracy, respectively, while binary entropy confidence gives the strongest non-oracle result at 39.73%. Oracle C-PP reaches 49.39%, indicating that the practical signals capture useful but incomplete information about local teacher reliability. The cached C-PP variants require no recurring teacher inference; UDA and FreeMatch evaluate the teacher on the $90{,}000$ unlabeled examples each epoch. Binary-entropy C-PP therefore exceeds both online baselines while avoiding their recurring teacher-forward cost.