Prediction-Powered Adaptive Shrinkage Estimation

Sida Li* Nikolaos Ignatiadis†

February 1, 2025

Abstract

Prediction-Powered Inference (PPI) is a powerful framework for enhancing statistical estimates by combining limited gold-standard data with machine learning (ML) predictions. While prior work has demonstrated PPI's benefits for individual statistical tasks, modern applications require answering numerous parallel statistical questions. We introduce Prediction-Powered Adaptive Shrinkage (`PAS`), a method that bridges PPI with empirical Bayes shrinkage to improve estimation of multiple means. `PAS` debiases noisy ML predictions *within* each task and then borrows strength *across* tasks by using those same predictions as a reference point for shrinkage. The amount of shrinkage is determined by minimizing an unbiased estimate of risk, and we prove that this tuning strategy is asymptotically optimal. Experiments on both synthetic and real-world datasets show that `PAS` adapts to the reliability of the ML predictions and outperforms traditional and modern baselines in large-scale applications.

1 Introduction

A major obstacle in answering modern scientific questions is the scarcity of gold-standard data [Miao et al., 2024b]. While advancements in data collection, such as large-scale astronomical surveys [York et al., 2000] and web crawling [Penedo et al., 2024], have led to an abundance of covariates (or features), scientific inferences often rely on outcomes (or labels), which are often expensive and labor-intensive to obtain. The rapid development of machine learning (ML) algorithms has offered a path forward, with ML predictions increasingly used to supplement the gold-standard outcomes and increase the statistical efficiency of subsequent analyses [Liang et al., 2007, Wang et al., 2020].

Prediction-Powered Inference (PPI) [Angelopoulos et al., 2023] addresses the scarcity issue by providing a framework for valid statistical analysis using predictions from black-box ML models. By combining ML-predicted and gold-standard outcomes, PPI and its variants [Angelopoulos et al., 2024, Zrnic and Candès, 2024, Zrnic, 2025] use the abundance of predictions to reduce variance while relying on the accuracy of labeled[1] data to control bias.

In this work, we adapt PPI to the estimation of multiple outcome means in compound estimation settings. Many applications of PPI naturally involve parallel statistical problems that can be solved simultaneously. For instance, several PPI methods [Angelopoulos et al., 2024, Fisch et al., 2024] have shown improvements in estimating the fraction of spiral galaxies using predictions on images from the Galaxy Zoo 2 dataset [Willett et al., 2013]. While these methods focus on estimating a single overall fraction, a richer analysis emerges from partitioning galaxies based on metadata (such as celestial coordinates or pre-defined bins) and estimating the fraction of galaxies within each partition. This compound estimation approach enables more granular scientific inquiries that account for heterogeneity across galaxy clusters and spatial locations [Nair and Abraham, 2010].

We demonstrate, both theoretically and empirically, the benefits of solving multiple mean estimation problems simultaneously. Our approach builds on the empirical Bayes principle of sharing information *across problems* [Robbins, 1956, Efron, 2010] as exemplified by James-Stein shrinkage [James and Stein, 1961, Xie et al., 2012]. The connection between modern and classical statistical ideas allows us to perform *within problem* PPI estimation in the first place, followed by a shrinkage process reusing the ML predictions in an adaptive manner, which becomes possible through borrowing information *across problems*. Our contributions are as follows:

*Data Science Institute, The University of Chicago. Email: listar2000@uchicago.edu.
†Department of Statistics and Data Science Institute, The University of Chicago. E-mail: ignat@uchicago.edu.
[1]Throughout the paper, we will use the terms "labeled" and "gold-standard" interchangeably.

1. We propose Prediction-Powered Adaptive Shrinkage (`PAS`) for compound mean estimation. `PAS` inherits the flexibility of PPI in working with *any* black-box predictive model and makes minimal distributional assumptions about the data. Its two-stage estimation process makes efficient use of the ML predictions as both a variance-reduction device and a shrinkage target.

2. We develop a *correlation-aware unbiased risk estimate* (CURE) for optimizing the `PAS` estimator, establish its asymptotic consistency, and derive an interpretation in terms of a Bayes oracle risk upper bound.

3. We conduct extensive experiments on both synthetic and real-world datasets. Our experiments demonstrate `PAS`'s applicability to large-scale problems with state-of-the-art predictors, showing improved estimation accuracy compared to other estimators (e.g., classical, PPI++).

2 Preliminaries and Notations

2.1 Prediction-Powered Inference (PPI)

The PPI framework considers a setting where we have access to a small number of labeled data points $(X_i, Y_i)_{i=1}^n \in (\mathcal{X} \times \mathcal{Y})^n$ and a large number of unlabeled covariates $(\tilde{X}_i)_{i=1}^N \in (\mathcal{X})^N$, where $\mathcal{X}$ and $\mathcal{Y}$ represent the covariate and outcome space, respectively. The data points are drawn iid from a joint distribution $\mathbb{P}_{XY}$.[2] We are also given a black-box predictive model $f : \mathcal{X} \to \mathcal{Y}$ that is independent of the datasets (e.g. pre-trained on similar but unseen data). For mean estimation and with $\mathcal{Y} \subset \mathbb{R}$, the goal is to leverage the predicted outcomes $f(X_i)$ to improve the estimation of $\theta := \mathbb{E}[Y_i]$. Some simple estimators take the form of the following *aggregated statistics*

$$\begin{aligned} \bar{Y} &:= \frac{1}{n}\sum_{i=1}^{n} Y_i, & \tilde{Y} &:= \frac{1}{N}\sum_{i=1}^{N} \tilde{Y}_i, \\ \bar{Z}^f &:= \frac{1}{n}\sum_{i=1}^{n} f(X_i), & \tilde{Z}^f &:= \frac{1}{N}\sum_{i=1}^{N} f(\tilde{X}_i), \end{aligned} \tag{1}$$

where $\bar{Y}$ is the classical estimator,[3] $\bar{Z}^f, \tilde{Z}^f$ are the prediction means on the labeled and unlabeled data, and $\tilde{Y}$ (greyed out) is unobserved. The vanilla PPI estimator is defined as,

$$\hat{\theta}^{\text{PPI}} := \underbrace{\bar{Y}}_{\text{Baseline}} + \underbrace{(\tilde{Z}^f - \bar{Z}^f)}_{\text{Variance Reduction}} = \underbrace{\tilde{Z}^f}_{\text{Baseline}} + \underbrace{(\bar{Y} - \bar{Z}^f)}_{\text{Debiasing}}. \tag{2}$$

Both definitions represent $\hat{\theta}^{\text{PPI}}$ in the form of a baseline estimator plus a correction term. In the first representation, the baseline estimator is the unbiased classical estimator $\bar{Y}$, while the correction term has expectation 0 and helps reduce the variance of $\bar{Y}$. In the second representation, the baseline estimator is the prediction mean on unlabeled data $\tilde{Z}^f$ (which in general may be biased for θ), while the correction term removes the bias of $\tilde{Z}^f$ by estimating the bias of the ML model f on the labeled dataset. Writing $\hat{\theta}^{\text{PPI}} = \frac{1}{N}\sum_{i=1}^{N} f(\tilde{X}_i) + \frac{1}{n}\sum_{i=1}^{n}(Y_i - f(X_i))$, we find, $\mathbb{E}[\hat{\theta}^{\text{PPI}}] = \mathbb{E}[Y_i] = \theta$ and

$$\text{Var}[\hat{\theta}^{\text{PPI}}] = \frac{1}{N}\text{Var}[f(\tilde{X}_i)] + \frac{1}{n}\text{Var}[Y_i - f(X_i)], \tag{3}$$

that is, $\hat{\theta}^{\text{PPI}}$ is unbiased for θ and its variance becomes smaller when the model predicts the true outcomes well. The mean squared error (MSE) of $\hat{\theta}^{\text{PPI}}$ is also equal to $\text{Var}[\hat{\theta}^{\text{PPI}}]$. Although we motivated $\hat{\theta}^{\text{PPI}}$ in (2) as implementing a correction step on two possible baseline estimators ($\bar{Y}$ and $\tilde{Z}^f$), $\hat{\theta}^{\text{PPI}}$ may have MSE for estimating θ that is arbitrarily worse than either of these baselines.

Comparison to classical estimator $\bar{Y}$. The classical estimator $\bar{Y}$ which only uses labeled data is unbiased for θ and has variance (and MSE) equal to $n^{-1}\text{Var}[Y_i]$, while the MSE of $\hat{\theta}^{\text{PPI}}$ in (3) may be arbitrarily large when the predictive model is inaccurate.

[2] To be concrete: $(X_i, Y_i) \overset{\text{iid}}{\sim} \mathbb{P}_{XY}$ and $(\tilde{X}_i, \tilde{Y}_i) \overset{\text{iid}}{\sim} \mathbb{P}_{XY}$, but $\tilde{Y}_i$ is unobserved.

[3] From now on, we will use the term "classical estimator" to refer to the sample average of the labeled outcomes.

Power-Tuned PPI (PPI++). To overcome the above limitation, Angelopoulos et al. [2024] introduce a power-tuning parameter λ and define

$$\hat{\theta}^{\text{PPI}}_{\lambda} := \bar{Y} + \lambda \left(\tilde{Z}^f - \bar{Z}^f \right), \tag{4}$$

which recovers the classical estimator when $\lambda = 0$ and the vanilla PPI estimator when $\lambda = 1$. For all values of λ, $\hat{\theta}^{\text{PPI}}_{\lambda}$ is unbiased, so if we select the λ that minimizes $\text{Var}[\hat{\theta}^{\text{PPI}}_{\lambda}]$, we can improve our estimator over both the classical estimator and vanilla PPI. Such an estimator is defined as the *Power-Tuned PPI* (PT) estimator $\hat{\theta}^{\text{PT}} := \hat{\theta}^{\text{PPI}}_{\lambda^*}$, where

$$\lambda^* := \underset{\lambda \in [0,1]}{\arg\min}\, \text{Var}[\hat{\theta}^{\text{PPI}}_{\lambda}],$$

which will become one pillar of `PAS` in Section 4.

Comparison to $\tilde{Z}^f$. Consider the ideal scenario for PPI with $N = \infty$ (that is, the unlabeled dataset is much larger than the labeled dataset) so that $\tilde{Z}^f \equiv \mathbb{E}[f(\tilde{X}_i)]$. Even then, the MSE of $\hat{\theta}^{\text{PPI}}$ in (3) is always lower bounded[4] by $\mathbb{E}[\text{Var}[Y_i \mid X_i]]/n$, which is attained for the perfect ML predictor $f(\cdot) \equiv \mathbb{E}[Y_i \mid X_i = \cdot]$. In words, if Y_i is not perfectly predictable from X_i, then PPI applied to a labeled dataset of fixed size n must have nonnegligible MSE. By contrast, for $N = \infty$, the prediction mean of unlabeled data $\tilde{Z}^f$ has zero variance and MSE equal to the squared bias $(\mathbb{E}[f(X_i)] - \theta_i)^2$. Thus if the predictor satisfies a calibration-type property that $\mathbb{E}[f(X_i)] \approx \mathbb{E}[Y_i]$ (which is implied by, but much weaker than the requirement $f(X_i) \approx Y_i$), then the MSE of $\tilde{Z}^f$ could be nearly 0. By contrast, PPI (and PPI++) can only partially capitalize on such a predictor $f(\cdot)$.

While PPI and PPI++ are constrained by their reliance on unbiased estimators, we show that the compound estimation setting (Section 2.2) enables a different approach. By carefully navigating the bias-variance tradeoff through information sharing *across* parallel estimation problems, we can provably match the performance of both $\bar{Y}$ and $\tilde{Z}^f$.

2.2 The Compound Estimation Setting

In this section, we introduce the problem setting that `PAS` is designed to address—estimating the mean of $m > 1$ parallel problems with a single black-box predictive model f.[5] We start with modeling heterogeneity across problems via

$$\eta_j \overset{\text{iid}}{\sim} \mathbb{P}_{\eta},\ j \in [m],\ \text{and } \boldsymbol{\eta} := (\eta_1, ..., \eta_m)^{\intercal} \tag{5}$$

with $[m] := \{1, \ldots, m\}$. We do not place any restriction over the unknown prior $\mathbb{P}_{\eta}$.[6] Here, η_j is an unobserved latent variable that fully specifies the distribution of the j-th labeled dataset $(X_{ij}, Y_{ij})_{i=1}^{n_j}$ and the j-th unlabeled dataset $(\tilde{X}_{ij})_{i=1}^{N_j}$ for $n_j, N_j \in \mathbb{N}$. In our setting, we are specifically interested in the means

$$\theta_j := \mathbb{E}_{\eta_j}[Y_{ij}],\ j \in [m],\ \text{and } \boldsymbol{\theta} := (\theta_1, \ldots, \theta_m)^{\intercal}. \tag{6}$$

For the sake of genericity, the exact form of the observation distribution is neither assumed nor required in our arguments. Since f is fixed and the covariate space $\mathcal{X}$ can be high-dimensional, we directly model the joint distribution between the outcomes and the predictions.

Assumption 2.1. For each problem $j \in [m]$, we assume that the joint distribution of $(f(X_{ij}), Y_{ij})$ has *finite first and second moments* conditioning on η_j. We then write,

$$\begin{bmatrix} f(X_{ij}) \\ Y_{ij} \end{bmatrix} \overset{\text{iid}}{\sim} \mathbb{F}_j \left(\begin{bmatrix} \mu_j \\ \theta_j \end{bmatrix}, \begin{bmatrix} \tau_j^2 & \rho_j \tau_j \sigma_j \\ \rho_j \tau_j \sigma_j & \sigma_j^2 \end{bmatrix} \right), \tag{7}$$

[4] The same lower bound also applies to power-tuned PPI $\hat{\theta}^{\text{PT}}$.

[5] Our proposal also accommodates using separate predictors $\{f_j\}_{j=1}^m$ for each problem. To streamline exposition, we focus on the practical scenario where a single (large) model (e.g., an LLM or vision model) can handle multiple tasks simultaneously [Radford et al., 2019, He et al., 2022].

[6] Henceforth we will use the notation $\mathbb{E}_{\eta_j}[\cdot]$ (resp. $\mathbb{E}_{\boldsymbol{\eta}}[\cdot]$) to denote the expectation conditional on η_j (resp. $\boldsymbol{\eta}$), while $\mathbb{E}_{\mathbb{P}_{\eta}}[\cdot]$ denotes an expectation also integrating out $\mathbb{P}_{\eta}$ in (5).

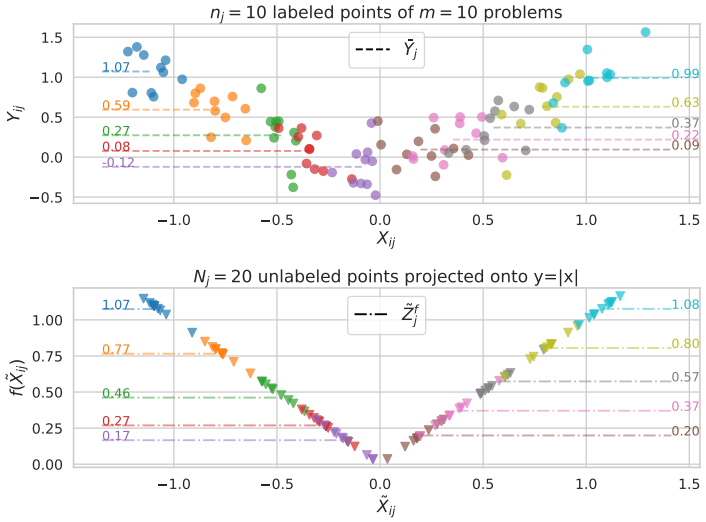

Figure 1: Example 2.2 with $m = 10$ problems in different colors. $n_j = 10, N_j = 20$ for all j. (Top) labeled data with the outcome mean $\bar{Y}_j$. (Bottom) predicted outcomes on unlabeled data by $f(x) = |x|$ with the prediction mean $\tilde{Z}_j$.

where $\mathbb{F}_j, \mu_j, \theta_j, \rho_j, \sigma_j^2, \tau_j^2$ are functions of η_j. The notation in (7) is to be interpreted as follows: $\mathbb{F}_j$ is *any distribution* satisfying the moment constraints in (6) and

$$\mathbb{E}_{\eta_j}[f(X_{ij})] = \mu_j, \qquad \mathrm{Var}_{\eta_j}[f(X_{ij})] = \tau_j^2,$$
$$\mathrm{Corr}_{\eta_j}[f(X_{ij}), Y_{ij}] = \rho_j, \qquad \mathrm{Var}_{\eta_j}[Y_{ij}] = \sigma_j^2.$$

We further denote $\gamma_j := \mathrm{Cov}_{\eta_j}[f(X_{ij}), Y_{ij}] = \rho_j \tau_j \sigma_j$.

Similar to Eq. (1), we define the *aggregated statistics* $\bar{Y}_j, \bar{Z}_j^f, \tilde{Z}_j^f$ for each $j \in [m]$. Following prior work,[7] we treat τ_j, σ_j, ρ_j as known in our theoretical development. In our numerical implementation, we use sample-based estimates thereof. The rationale is that estimation of τ_j, σ_j, ρ_j is a second-order concern in our setup of mean estimation and treating these as known facilitates exposition.

We next introduce a synthetic model that we will also use in our numerical study.

Example 2.2 (Synthetic model). For each problem j, let $\eta_j \sim \mathcal{U}[-1, 1]$. We think of η_j as both indexing the problems, and generating heterogeneity across problems. The j-th dataset then follows (with constant $c = 0.05, \psi = 0.1$),

$$X_{ij} \overset{\text{iid}}{\sim} \mathcal{N}(\eta_j, \psi^2), \quad Y_{ij}|X_{ij} \overset{\text{ind}}{\sim} \mathcal{N}(2\eta_j X_{ij} - \eta_j^2, c). \tag{8}$$

In Figure 1, we visualize realizations from this model with $m = 10$ problems, $n_j = 10$ labeled observations, and $N_j = 20$ unlabeled observations for each problem. We apply a flawed predictor $f(x) = |x|$. The classical estimator $\bar{Y}_j$ and the prediction mean $\tilde{Z}_j^f$ deviate from each other. Nevertheless, $\tilde{Z}_j^f$ contains information that can help us improve upon $\bar{Y}_j$ as an estimator of θ_j by learning from within problem (PPI, PPI++, this work) and across problem (this work) structure. We note that as specified in (7), our approach formally only models first and second moments of the joint distribution of $(f(X_{ij}), Y_{ij})$. For instance, in this synthetic model, $\theta_j = \eta_j^2$ and $\sigma_j^2 = 4\eta_j^2\psi^2 + c$, while μ_j, τ_j^2 and ρ_j also admit closed-form expressions in terms of η_j (see Appendix C.1).

To conclude this section, we define the *compound risk* [Robbins, 1951, Jiang and Zhang, 2009] for any estimator $\hat{\boldsymbol{\theta}} = (\hat{\theta}_1, \ldots, \hat{\theta}_m)^\intercal$ as the expected squared-error loss

$$\mathcal{R}_m(\hat{\boldsymbol{\theta}}, \boldsymbol{\theta}) := \mathbb{E}_{\boldsymbol{\eta}}\left[\ell_m(\hat{\boldsymbol{\theta}}, \boldsymbol{\theta})\right], \tag{9}$$

$$\text{where} \quad \ell_m(\hat{\boldsymbol{\theta}}, \boldsymbol{\theta}) := \frac{1}{m}\sum_{j=1}^{m}(\hat{\theta}_j - \theta_j)^2. \tag{10}$$

[7] For EB, examples include Ignatiadis and Wager [2019], Xie et al. [2012]; For PPI, see recent works like Fisch et al. [2024].

Table 1: Estimator comparison in the stylized example of Section 3.

Estimator	MSE	VR	P	CP
$\hat{\theta}^{\mathrm{cl}}$	$\sigma_\xi^2 + \sigma_\varepsilon^2$	✗	✗	✗
$\hat{\theta}^{\mathrm{cl}} - \xi$	σ_ε^2	✓	✗	✗
$\mathbb{E}[\theta \mid \hat{\theta}^{\mathrm{cl}}]$	$\frac{(\sigma_\xi^2+\sigma_\varepsilon^2)\sigma_\theta^2}{(\sigma_\xi^2+\sigma_\varepsilon^2)+\sigma_\theta^2}$	✗	✓	✗
$\mathbb{E}[\theta \mid \hat{\theta}^{\mathrm{cl}}, \phi]$	$\frac{(\sigma_\xi^2+\sigma_\varepsilon^2)\sigma_{\theta\mid\phi}^2}{(\sigma_\xi^2+\sigma_\varepsilon^2)+\sigma_{\theta\mid\phi}^2}$	✗	✓	✓
$\mathbb{E}[\theta \mid \hat{\theta}^{\mathrm{cl}} - \xi, \phi]$	$\frac{\sigma_\varepsilon^2\sigma_{\theta\mid\phi}^2}{\sigma_\varepsilon^2+\sigma_{\theta\mid\phi}^2}$	✓	✓	✓

VR: Variance Reduction, P: Prior Information, CP: Contextual Prior Information.

The *Bayes risk*, which we also refer to simply as mean squared error (MSE), further integrates over randomness in the unknown prior $\mathbb{P}_\eta$ in (5):

$$\mathcal{B}_m^{\mathbb{P}_\eta}(\hat{\boldsymbol{\theta}}) := \mathbb{E}_{\mathbb{P}_\eta}\left[\mathcal{R}_m(\hat{\boldsymbol{\theta}}, \boldsymbol{\theta})\right]. \tag{11}$$

3 A Stylized Example and Related Work

We explain how our proposal relates to existing literature through the lens of the following stylized Gaussian model.

$$\text{Sampling: } \hat{\theta}^{\mathrm{cl}} = \theta + (\xi + \varepsilon),\ \xi \sim \mathcal{N}(0, \sigma_\xi^2),\ \varepsilon \sim \mathcal{N}(0, \sigma_\varepsilon^2).$$
$$\text{Prior: } \quad \theta \sim \mathcal{N}(0, \sigma_\theta^2),\ \phi \sim \mathcal{N}(0, \sigma_\phi^2),\ \mathrm{Corr}[\theta, \phi] = \rho.$$

For the stylized example, we assume that $(\theta, \phi, \varepsilon, \xi)$ are jointly normal and that all the pairwise correlations among them are zero with the exception of $\mathrm{Corr}[\theta, \phi] = \rho \neq 0$. We also write $\sigma_{\theta|\phi}^2 := \mathrm{Var}[\theta \mid \phi] = (1-\rho^2)\sigma_\theta^2 < \sigma_\theta^2$.

We think of $\hat{\theta}^{\mathrm{cl}}$ as the baseline classical statistical estimator of a quantity θ that we seek to estimate in MSE. In this example, $\hat{\theta}^{\mathrm{cl}}$ is unbiased for θ and has noise term $\xi+\varepsilon$, so that $\mathbb{E}[(\hat{\theta}^{\mathrm{cl}} - \theta)^2] = \mathrm{Var}_\theta[\hat{\theta}^{\mathrm{cl}}] = \sigma_\xi^2 + \sigma_\varepsilon^2$. We describe three high-level strategies used to improve the MSE of $\hat{\theta}^{\mathrm{cl}}$. These strategies are not tied in any way to the stylized model; nevertheless, the stylized model enables us to give precise expressions for the risk reductions possible, see Table 1.

Variance reduction (VR). An important statistical idea is to improve $\hat{\theta}^{\mathrm{cl}}$ via obtaining further information to intercept some of its noise, say ξ, and replacing $\hat{\theta}^{\mathrm{cl}}$ by $\hat{\theta}^{\mathrm{cl}} - \xi$ which has MSE σ_ε^2 and remains unbiased for θ. This idea lies at the heart of approaches such as control variates in simulation [Lavenberg and Welch, 1981, Hickernell et al., 2005], variance reduction in randomized controlled experiments via covariate adjustment [Lin, 2013] and by utilizing pre-experiment data [Deng et al., 2013, CUPED], as well as model-assisted estimation in survey sampling [Cochran, 1977, Breidt and Opsomer, 2017]. It is also the idea powering PPI and related methods: the unlabebeled dataset and the predictive model are used to intercept some of the noise in the classical statistical estimator $\hat{\theta}^{\mathrm{cl}} \triangleq \bar{Y}$; compare to Eq. (2) with $\xi \triangleq \bar{Z}^f - \tilde{Z}^f$. We refer to Ji et al. [2025] for an informative discussion of how PPI relates to traditional ideas in semiparametric inference [Robins et al., 1994].

Prior information (P) via empirical Bayes (EB). In the Bayesian approach we seek to improve upon $\hat{\theta}^{\mathrm{cl}}$ by using the prior information that $\theta \sim \mathcal{N}(0, \sigma_\theta^2)$. The Bayes estimator $\mathbb{E}[\theta \mid \hat{\theta}^{\mathrm{cl}}] = \{\sigma_\theta^2/(\sigma_\xi^2 + \sigma_\varepsilon^2 + \sigma_\theta^2)\}\hat{\theta}^{\mathrm{cl}}$ reduces variance by shrinking $\hat{\theta}^{\mathrm{cl}}$ toward 0 (at the cost of introducing some bias). When σ_θ^2 is small, the MSE of $\mathbb{E}[\theta \mid \hat{\theta}^{\mathrm{cl}}]$ can be substantially smaller than that of $\hat{\theta}^{\mathrm{cl}}$.

Now suppose that the variance of the prior, σ_θ^2, is unknown but we observe data from multiple related problems generated from the same model and indexed by $j = 1, \ldots, m$, say, $\theta_j \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma_\theta^2)$ and $\hat{\theta}_j^{\mathrm{cl}} \overset{\text{ind}}{\sim} \mathcal{N}(\theta_j, \sigma_\xi^2 + \sigma_\varepsilon^2)$. Then an EB analysis can mimic the MSE of the oracle Bayesian that has full

knowledge of the prior. To wit, we can estimate σ_θ^2 via $\hat{\sigma}_\theta^2 = \{\frac{1}{m-2}\sum_{j=1}^m (\hat{\theta}_j^{\text{cl}})^2\} - (\sigma_\xi^2 + \sigma_\varepsilon^2)$, and then consider a plug-in approximation of the Bayes rule, $\hat{\theta}_j^{\text{JS}} = \hat{\mathbb{E}}[\theta_j \mid \hat{\theta}_j^{\text{cl}}] := \{\hat{\sigma}_\theta^2/(\sigma_\xi^2 + \sigma_\varepsilon^2 + \hat{\sigma}_\theta^2)\}\hat{\theta}^{\text{cl}}$. The resulting estimator is the celebrated James-Stein estimator [James and Stein, 1961, Efron and Morris, 1973], whose risk is very close to the Bayes risk under the hierarchical model. The James-Stein estimator also always dominates the classical estimator under a frequentist evaluation of compound risk in (9): $\mathcal{R}_m(\hat{\boldsymbol{\theta}}^{\text{JS}}, \boldsymbol{\theta}) < \mathcal{R}_m(\hat{\boldsymbol{\theta}}^{\text{cl}}, \boldsymbol{\theta})$ for all $\boldsymbol{\theta} \in \mathbb{R}^m$.

Contextual prior information (CP) via EB. Instead of using the same prior for each problem, we may try to sharpen the prior and increase its relevance [Efron, 2011] by using further information ϕ. In the stylized example, as seen in Table 1, such an approach reduces the variance of the prior from σ_θ^2 to $\sigma_{\theta|\phi}^2 < \sigma_\theta^2$ with corresponding MSE reduction of the Bayes estimator. With multiple related problems, such a strategy can be instantiated via EB shrinkage toward an informative but biased predictor [Fay III and Herriot, 1979, Green and Strawderman, 1991, Mukhopadhyay and Maiti, 2004, Kou and Yang, 2017, Ignatiadis and Wager, 2019, Rosenman et al., 2023].

Putting everything together. We can get even smaller MSE (last row of Table 1) by using both variance reduction, shrinkage, and a contextual prior. In that case, the Bayes estimator $\mathbb{E}[\theta \mid \hat{\theta}^{\text{cl}} - \xi, \phi]$ takes the form,

$$\frac{\sigma_{\theta|\phi}^2}{\sigma_\varepsilon^2 + \sigma_{\theta|\phi}^2}(\hat{\theta}^{\text{cl}} - \xi) + \frac{\sigma_\varepsilon^2}{\sigma_\varepsilon^2 + \sigma_{\theta|\phi}^2}\mathbb{E}[\theta \mid \phi]\,. \tag{12}$$

4 Prediction-Powered Adaptive Shrinkage

On a high level, `PAS` aims to provide a lightweight approach that outperforms both baselines in (2) and PPI/PPI++ in terms of MSE when estimating multiple means. `PAS` also aims at minimal modeling requirements and assumptions.

The stylized example from Section 3 serves as a guiding analogy. We seek to benefit from ML predictions in two ways: first by variance reduction (acting akin to ξ in the stylized example), and second by increasing prior relevance (acting as a proxy for ϕ). We implement both steps to adapt to the unknown data-generating process in an assumption-lean way using *within*-problem information for the first step (Section 4.1) and *across*-problem information for the second step (Section 4.2), drawing on ideas from the EB literature.

4.1 The *Within* Problem Power-Tuning Stage

Extending the notation from (4) to each problem j, we have a class of unbiased estimators $\hat{\theta}_{j,\lambda}^{\text{PPI}} := \bar{Y}_j + \lambda(\tilde{Z}_j^f - \bar{Z}_j^f)$, for every $\lambda \in \mathbb{R}$. Explicitly calculating the variance gives

$$\text{Var}_{\eta_j}\left[\hat{\theta}_{j,\lambda}^{\text{PPI}}\right] = \frac{\sigma_j^2}{n_j} + \overbrace{\frac{n_j + N_j}{n_j N_j}\lambda^2\tau_j^2 - \frac{2}{n_j}\lambda\gamma_j}^{\delta_j(\lambda)}\,.$$

Note that the classical estimator has risk σ_j^2/n_j and so is outperformed whenever $\delta_j(\lambda) < 0$. We can analytically solve for the optimal λ, which yields

$$\lambda_j^* := \arg\min_\lambda \delta_j(\lambda) = \left(\frac{N_j}{n_j + N_j}\right)\frac{\gamma_j}{\tau_j^2}, \tag{13}$$

and the *Power-Tuned* (PT) estimator $\hat{\theta}_j^{\text{PT}} := \hat{\theta}_{j,\lambda_j^*}^{\text{PP}}$ with

$$\tilde{\sigma}_j^2 := \text{Var}_{\eta_j}\left[\hat{\theta}_j^{\text{PT}}\right] = \frac{\sigma_j^2}{n_j} - \frac{N_j}{n_j(n_j + N_j)}\frac{\gamma_j^2}{\tau_j^2}. \tag{14}$$

The formulation of the above PT estimators has been well understood in the single problem setting Angelopoulos et al. [2024], Miao et al. [2024a]. In `PAS`, we execute this stage separately for each problem, as the optimal power-tuning parameter is data-dependent and varies case by case.

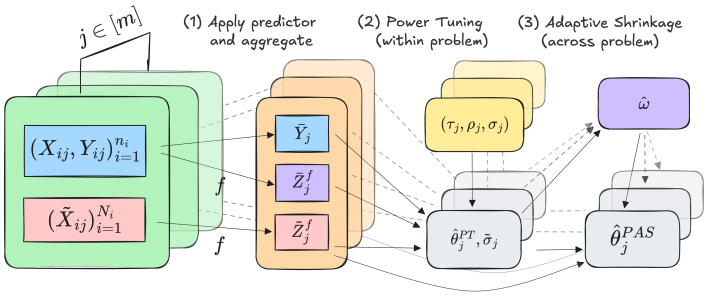

Figure 2: A flowchart illustration (top) of the pseudo-code implementation (down) of the full `PAS` method.

Algorithm 1 Prediction-Powered Adaptive Shrinkage

Require: $(X_{ij}, Y_{ij})_{i=1}^{n_j}$, $(\tilde{X}_{ij})_{i=1}^{N_j}$, ρ_j, τ_j, σ_j for $j \in [m]$, predictive model f

for $j = 1$ to m **do**
  {Step 1: Apply predictor (Eq. (1))}
  $\bar{Y}_j, \bar{Z}_j^f, \tilde{Z}_j^f = \texttt{get\_means}((X_{ij}, Y_{ij})_{i=1}^{n_j}, (\tilde{X}_{ij})_{i=1}^{N_j}, f)$
  {Step 2: Power tuning (Eq. (13))}
  $\lambda_j^* = \texttt{get\_pt\_param}(\rho_j, \tau_j, n_j, N_j)$
  $\hat{\theta}_j^{\mathrm{PT}} = \bar{Y}_j + \lambda_j^*(\tilde{Z}_j^f - \bar{Z}_j^f)$
  $\tilde{\sigma}_j^2 = \texttt{get\_pt\_var}(\hat{\theta}_j^{\mathrm{PT}})$ {(Eq. (14))}
end for
{Step 3: Adaptive shrinkage (Eq. (18))}
$\hat{\omega} = \texttt{get\_shrink\_param}(\{\hat{\theta}_j^{\mathrm{PT}}\}_{j=1}^m, \{\tilde{Z}_j^f\}_{j=1}^m, \{\tilde{\sigma}_j^2\}_{j=1}^m)$
for $j = 1$ to m **do**
  $\hat{\omega}_j = \hat{\omega}/(\hat{\omega} + \tilde{\sigma}_j^2)$
  $\hat{\theta}_j^{\mathrm{PAS}} = \hat{\omega}_j \hat{\theta}_j^{\mathrm{PT}} + (1 - \hat{\omega}_j)\tilde{Z}_j^f$
end for
return $\{\hat{\theta}_j^{\mathrm{PAS}}\}_{j=1}^m$

4.2 The *Across* Problem Adaptive Shrinkage Stage

The PT estimator derived in Section 4.1 already possesses many appealing properties: it is unbiased and has lower variance than both the classical estimator and vanilla PPI. However, as our setting involves working with many parallel problems together, there is the possibility of further risk reduction. Concretely, based on the PT estimator obtained in Section 4.1, we consider a class of shrinkage estimators $\hat{\boldsymbol{\theta}}_{\omega}^{\mathrm{PAS}} := (\hat{\theta}_{1,\omega}^{\mathrm{PAS}}, \ldots, \hat{\theta}_{m,\omega}^{\mathrm{PAS}})^{\intercal}$, where for any $\omega \geq 0$

$$\begin{aligned} \hat{\theta}_{j,\omega}^{\mathrm{PAS}} &:= \omega_j \hat{\theta}_j^{\mathrm{PT}} + (1 - \omega_j)\tilde{Z}_j^f, \\ \omega_j &:= \omega_j(\omega) = \omega/(\omega + \tilde{\sigma}_j^2), \end{aligned} \tag{15}$$

The motivation is to match the form of the Bayes estimator with variance reduction and contextual prior information in (12) with the following analogies:

$$\begin{aligned} \hat{\theta}^{\mathrm{cl}} - \xi &\longleftrightarrow \hat{\theta}_j^{\mathrm{PT}}, \qquad \mathbb{E}[\theta \mid \phi] \longleftrightarrow \tilde{Z}_j^f, \\ \sigma_\varepsilon^2 &\longleftrightarrow \tilde{\sigma}_j^2, \qquad \sigma_{\theta|\phi}^2 \longleftrightarrow \omega. \end{aligned} \tag{16}$$

The highlighted ω is a *global shrinkage parameter* that acts as follows:

(i) Fixing ω, any problem whose PT estimator has higher variance possesses smaller ω_j and shrinks more towards $\tilde{Z}_j^f$; a smaller variance increases ω_j and makes the final estimator closer to $\hat{\theta}_j^{\mathrm{PT}}$.

(ii) Fixing all the problems, increasing ω has an overall effect of recovering $\hat{\theta}_j^{\mathrm{PT}}$ (full recovery when $\omega \to \infty$), and setting $\omega = 0$ recovers $\tilde{Z}^f$.

The above establishes the importance of ω, though it needs to be selected in a data-driven way. Our strategy is to minimize an unbiased estimate of risk [Stein, 1981] as a surrogate of the unknown risk $\mathcal{R}_m(\hat{\boldsymbol{\theta}}^{\text{PAS}}_{\omega}, \boldsymbol{\theta})$, following e.g., Xie et al. [2012].[8] One challenge here is that

$$\tilde{\gamma}_j := \text{Cov}_{\eta_j}[\hat{\theta}^{\text{PT}}_j, \tilde{Z}^f_j] = \lambda^*_j \text{Var}_{\eta_j}[\tilde{Z}^f_j] = \frac{\gamma_j}{n_j + N_j} \tag{17}$$

is not necessarily zero. This implies that we must account for the correlation between shrinkage source and target.

Theorem 4.1. *Under Assumption 2.1, define the "correlation-aware unbiased risk estimate" (*CURE*) as*

$$\begin{aligned}\text{CURE}\left(\hat{\theta}^{\text{PAS}}_{j,\omega}\right) &:= (2\omega_j - 1)\tilde{\sigma}^2_j + 2(1-\omega_j)\tilde{\gamma}_j \\ &\quad + \left[(1-\omega_j)(\hat{\theta}^{PT}_{j,\omega_j} - \tilde{Z}^f_j)\right]^2, \\ \text{CURE}\left(\hat{\boldsymbol{\theta}}^{\text{PAS}}_{\omega}\right) &:= \frac{1}{m}\sum_{j=1}^{m} \text{CURE}\left(\hat{\theta}^{\text{PAS}}_{j,\omega_j}\right).\end{aligned}$$

Then CURE is an unbiased estimator of the compound risk defined in (9), that is,

$$\mathbb{E}_{\boldsymbol{\eta}}\left[\text{CURE}\left(\hat{\boldsymbol{\theta}}^{\text{PAS}}_{\omega}\right)\right] = \mathcal{R}_m(\hat{\boldsymbol{\theta}}^{\text{PAS}}_{\omega}, \boldsymbol{\theta}).$$

(See Appendix A.2 for the proof and motivation.) With Theorem 4.1, we now have a systematic strategy to pick ω by minimizing CURE

$$\hat{\omega} := \arg\min_{\omega \geq 0} \text{CURE}\left(\hat{\boldsymbol{\theta}}^{\text{PAS}}_{\omega}\right). \tag{18}$$

Even though $\hat{\omega}$ does not admit a closed-form expression, the one-dimensional minimization can be efficiently carried out by numerical methods like grid search. The final `PAS` estimator is then given by

$$\hat{\theta}^{\text{PAS}}_j := \hat{\theta}^{\text{PAS}}_{j,\hat{\omega}} = \frac{\hat{\omega}}{\hat{\omega} + \tilde{\sigma}^2_j}\hat{\theta}^{\text{PT}}_j + \frac{\tilde{\sigma}^2_j}{\hat{\omega} + \tilde{\sigma}^2_j}\tilde{Z}^f_j.$$

To appreciate how flexible and adaptive `PAS` is, we briefly revisit the synthetic model in Example 2.2, whose special structure allows us to visualize how the power-tuned and adaptive shrinkage parameters vary across problems and different predictors. In Figure 3, we consider $m = 200$ problems and two predictors: a good predictor $f_1(x) = x^2$ and a flawed predictor $f_2(x) = |x|$. The model setup in (8) is such that the magnitude of $\text{Cov}_{\eta_j}[X_j, Y_j]$ relative to $\text{Var}_{\eta_j}[Y_j]$ is much larger for problems with η_j closer to the origin. Therefore, for both predictors, we see a dip in λ^*_j near the middle (top panel), which shows that `PAS` adapts to the *innate difficulties of each problem* when deciding how much power-tuning to apply. On the other hand (bottom panel), the overall shrinkage effect is much stronger (smaller $\hat{\omega}_j$ for all j) with f_1 than with f_2, which demonstrates `PAS`'s ability to adapt to the *predictor's quality across problems*—while still allowing each problem to have its own shrinkage level. Numerical results are postponed to Section 6.

5 Theoretical Results

In (18), we proposed selecting $\hat{\omega}$ by optimizing an unbiased surrogate of true risk. Our theoretical results justify this procedure. As we consider more and more problems ($m \to \infty$), CURE approximates the true loss uniformly in ω.

[8] The connection to empirical Bayes is the following. As explained by Xie et al. [2012] and Tibshirani and Rosset [2019], the James-Stein estimator may be derived by tuning σ^2_θ (in Section 3) via minimization of Stein's [1981] unbiased risk estimate.

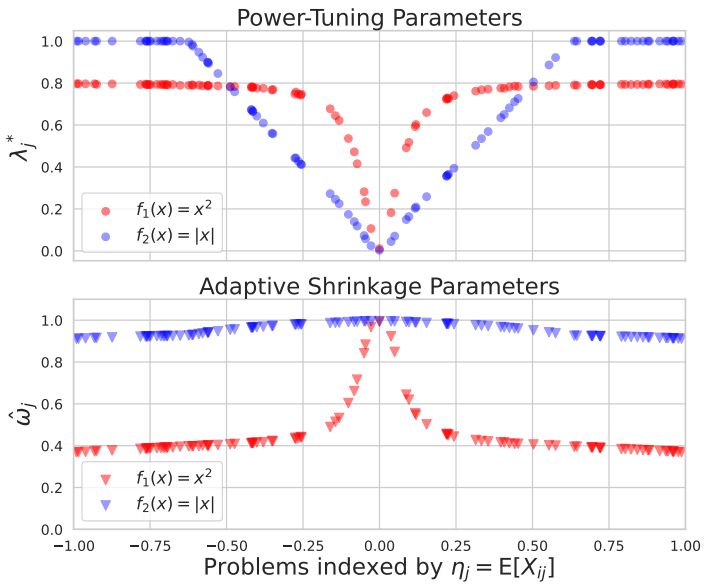

Figure 3: The power-tuned and adaptive shrinkage parameters, $\hat{\lambda}_j$ and $\hat{\omega}_j$ across $m = 200$ problems in Example 2.2. On the x-axis, we identify the problem by its η_j so the trend is more visible.

Proposition 5.1. *Suppose the datasets are generated according to (5) and Assumption 2.1 and further assume that* $\mathbb{E}_{\mathbb{P}_\eta}[\theta_j^4] < \infty$, $\mathbb{E}_{\mathbb{P}_\eta}[\mu_j^4] < \infty$, $\sup_j \mathbb{E}_{\mathbb{P}_\eta}[(\hat{\theta}_j^{PT})^4] < \infty$ *and* $\sup_j \mathbb{E}_{\mathbb{P}_\eta}[(\tilde{Z}_j^f)^4] < \infty$. *Then*

$$\mathbb{E}_{\mathbb{P}_\eta}\left[\sup_{\omega \geq 0}\left|\mathrm{CURE}(\hat{\boldsymbol{\theta}}_\omega^{PAS}) - \ell_m(\hat{\boldsymbol{\theta}}_\omega^{PAS}, \boldsymbol{\theta})\right|\right] = o(1),$$

where $o(1)$ *denotes a term that converges to* 0 *as* $m \to \infty$.

A principal consequence of Proposition 5.1 is that `PAS` with the data-driven choice of $\hat{\omega}$ in (18) has asymptotically smaller MSE (11) than any of the estimators in (15), and so has smaller risk than both baselines in (2) and also than PPI and PT estimators.

Theorem 5.2. *Under the assumptions in Proposition 5.1,*

$$\mathcal{B}_m^{\mathbb{P}_\eta}(\hat{\boldsymbol{\theta}}_{\hat{\omega}}^{PAS}) \leq \inf_{\omega \geq 0}\left\{\mathcal{B}_m^{\mathbb{P}_\eta}(\hat{\boldsymbol{\theta}}_\omega^{PAS})\right\} + o(1) \;\; \textit{as} \;\; m \to \infty,$$

and so $\mathcal{B}_m^{\mathbb{P}_\eta}(\hat{\boldsymbol{\theta}}_{\hat{\omega}}^{PAS}) \leq \min\{\mathcal{B}_m^{\mathbb{P}_\eta}(\tilde{\boldsymbol{Z}}^{\boldsymbol{f}}), \mathcal{B}_m^{\mathbb{P}_\eta}(\hat{\boldsymbol{\theta}}^{PT})\} + o(1)$.

Our next proposition makes strong assumptions (that are untenable in practice), but connects Theorem 5.2 with the possible risk reductions in the stylized example of Section 3.

Proposition 5.3. *In addition to existing assumptions, further assume that* $N_j = \infty$ *and that there exist* $n \in \mathbb{N}$, $\tilde{\sigma}^2 > 0$ *such that* $n_j = n$ *and* $\tilde{\sigma}_j^2 = \tilde{\sigma}^2$ *for all* j *almost surely. Let* $\beta^2 := \mathbb{E}_{\mathbb{P}_\eta}[(\tilde{Z}_j^f - \theta_j)^2]$ *(which does not depend on* j *as we are integrating over* $\mathbb{P}_\eta$*). Then, as* $m \to \infty$,

$$\mathcal{B}_m^{\mathbb{P}_\eta}(\hat{\boldsymbol{\theta}}^{PAS}) \leq \frac{\tilde{\sigma}^2\beta^2}{\tilde{\sigma}^2 + \beta^2} + o(1).$$

To interpret the result, it is instructive to compare the asymptotic upper bound on the MSE of `PAS` with the MSE in the last line of Table 1, i.e., with $(\sigma_\varepsilon^2\sigma_{\theta|\phi}^2)/(\sigma_\varepsilon^2 + \sigma_{\theta|\phi}^2)$. Observe that $\tilde{\sigma}^2$ plays the role of σ_ε^2 (as already anticipated in (16)) which is smaller than the variance of the classical estimator (due to power tuning). Meanwhile, β^2 plays the role of $\sigma_{\theta|\phi}^2$. If the baseline $\tilde{Z}_j^f$ that is the mean of the ML predictions on the unlabeled datasets is doing a good job of predicting θ_j, then β^2 will be small, and so `PAS` may potentially have MSE substantially smaller than that of PPI++. On the other hand, even if β^2 is large (that is, even if the ML model is very biased), `PAS` asymptotically still has MSE less or equal than $\tilde{\sigma}^2$, the MSE of PPI++. We emphasize that the assumptions of Proposition 5.3 are not needed for Theorem 5.2, whose statement does not restrict heterogeneity (e.g., heteroscedasticity) across problems and allows for varying, finite unlabeled and labeled sample sizes.

6 Experiments

We apply the proposed `PAS` estimator and conduct extensive experiments in both the synthetic model proposed in Example 2.2 and two real-world datasets.

Baselines. We compare the `PAS` estimator against both classical and modern baseline estimators: (1) the classical estimator, (2) prediction mean on unlabeled data, (3) the vanilla PPI estimator [Angelopoulos et al., 2023], (4) the PT estimator [Angelopoulos et al., 2024, PPI++], and (5) the shrinkage-only estimator that directly shrinks the classical estimator towards the prediction mean (see Appendix B). The comparisons with (4) and (5) also directly serve as ablation studies of the two stages of the `PAS` estimator.

Metrics. We report the mean squared error (MSE) ($\pm$ 1 standard error) of each estimator $\hat{\boldsymbol{\theta}}$ by averaging $\frac{1}{m}\sum_{j=1}^{m}(\hat{\theta}_j - \theta_j)^2$ across $K = 200$ Monte Carlo replicates. In the synthetic model, we sample η_j (and thus θ_j) from the known prior $\mathbb{P}_\eta$. For the real-world datasets, since $\mathbb{P}_\eta$ is unknown, we follow the standard evaluation strategy in the PPI literature: we start with a large labeled dataset and use it to compute a pseudo-ground truth for each mean θ_j. Then in each Monte Carlo replicate, we randomly split the data points of each problem into the labeled/unlabeled partitions with a 20-80 ratio. We provide more details on the benchmarking procedure for the MSE in Appendix C.

A common concern over shrinkage estimators is that there is only "marginal" guarantee on the compound risk reduction: it is possible to hugely improve a few noisy estimates at the expense of a large number of clean ones. For this reason, we also measure the "percentage of problems improved" (denoted in table as "% **Improved** ↑), defined as

$$\frac{1}{m}\sum_{j=1}^{m}\mathbf{1}\left[(\hat{\theta}_j - \theta_j)^2 < (\hat{\theta}_j^{\text{Classical}} - \theta_j)^2\right] \times (100\%)$$

with respect to the classical estimator. We report this metric for real-world datasets where varying sample sizes across problems results in "outlier" problems that are much harder to improve than otherwise. Larger values of this metric are preferable.

6.1 Synthetic Model

Estimator	MSE f_1 ($\times 10^{-3}$)	MSE f_2 ($\times 10^{-3}$)
Classical	3.142 $\pm$ 0.033	3.142 $\pm$ 0.033
Prediction Avg	0.273 $\pm$ 0.004	34.335 $\pm$ 0.147
PPI	2.689 $\pm$ 0.027	2.756 $\pm$ 0.027
PT	2.642 $\pm$ 0.027	2.659 $\pm$ 0.026
Shrinkage	0.273 $\pm$ 0.003	3.817 $\pm$ 0.042
`PAS` **(ours)**	**0.272 $\pm$ 0.003**	**2.496 $\pm$ 0.025**

Table 2: MSE ($\pm$ standard error) of different estimators under the synthetic model with predictor $f_1(x) = x^2$ and $f_2(x) = |x|$.

This is the synthetic model from Example 2.2, where we choose $m = 200$, $n_j = 20$, and $N_j = 80$ for all j. Since we have already visualized the model and the parameters of the `PAS` estimator in previous sections, we simply report the numerical results (for the good predictor f_1 and the flawed predictor f_2) in Table 2.

For both predictors, we see that `PAS` outperforms all the baselines. With a good estimator f_1, both the prediction mean and the shrinkage estimator closely track `PAS`; in contrast, the PPI and PT estimators fail to fully leverage the accurate predictions, as their design enforces unbiasedness. The situation reverses for the less reliable predictor f_2: the prediction mean and the shrinkage estimator have high MSE, while estimators with built-in de-biasing mechanisms demonstrate greater resilience. `PAS` adapts effectively across these extremes, making it a robust choice for a wide range of problems and predictors.

6.2 Real-World Datasets

Table 3: Results aggregated over $K = 200$ replicates on the Amazon review dataset with `BERT-base` and `BERT-tuned` predictors. Metrics are reported with ± 1 standard error.

	Amazon (base f)		Amazon (tuned f)	
Estimator	**MSE** ($\times 10^{-3}$)	**% Improved** ↑	**MSE** ($\times 10^{-3}$)	**% Improved** ↑
Classical	24.305 ± 0.189	baseline	24.305 ± 0.189	baseline
Prediction Avg	41.332 ± 0.050	30.7 ± 0.2	3.945 ± 0.011	75.4 ± 0.2
PPI	11.063 ± 0.085	62.4 ± 0.2	7.565 ± 0.066	70.4 ± 0.2
PT	10.633 ± 0.089	70.3 ± 0.2	6.289 ± 0.050	76.0 ± 0.2
Shrinkage	15.995 ± 0.121	56.4 ± 0.3	3.828 ± 0.039	78.9 ± 0.2
`PAS` (ours)	**8.517 ± 0.071**	**71.4 ± 0.2**	**3.287 ± 0.024**	**80.8 ± 0.2**

Table 4: Results aggregated over $K = 200$ replicates on the Galaxy dataset with `ResNet50` predictor. Metrics are reported with ± 1 standard error.

	Galaxy	
Estimator	**MSE** ($\times 10^{-3}$)	**% Improved** ↑
Classical	2.073 ± 0.028	baseline
Prediction Avg	7.195 ± 0.008	17.0 ± 0.2
PPI	1.149 ± 0.017	59.4 ± 0.3
PT	1.026 ± 0.015	**67.7 ± 0.3**
Shrinkage	1.522 ± 0.016	48.8 ± 0.4
`PAS` (ours)	**0.893 ± 0.011**	67.3 ± 0.4

We next evaluate `PAS` on two large-scale real-world datasets, highlighting its ability to leverage state-of-the-art deep learning models in different settings. We include only the essential setup below and defer additional data and model details (e.g., hyper-parameters, preprocessing) to Appendix C.

Amazon Review Ratings [SNAP, 2014]. Many commercial and scientific studies involve collecting a large corpus of text and estimating an average score (rating, polarity, etc.) from it. A practitioner would often combine limited human annotations with massive automatic evaluations from ML models [Tyser et al., 2024, Baly et al., 2020]. To replicate this setup, we consider *mean rating estimation* problems using the *Amazon Fine Food Review* dataset from Kaggle, where we artificially hide the labels in a random subset of the full data to serve as the unlabeled partition. Concretely, we estimate the average rating for the top $m = 200$ products with the most reviews (from ∼200 to ∼900). For the i-th review of the j-th product, the covariate X_{ij} consists of the review's title and text concatenated, while the outcome Y_{ij} is the star rating in $\{1, \ldots, 5\}$. We employ two black-box predictors: (1) `BERT-base`, a language model without fine-tuning [Devlin, 2018] and (2) `BERT-tuned` which is the same model but fine-tuned on a held-out set of reviews from other products. Neither of these models have seen the reviews for the 200 products during training.

Spiral Galaxy Fractions [Willett et al., 2013]. The *Galaxy Zoo 2* project contains the classification results of galaxy images from the the Sloan Digital Sky Survey [York et al., 2000, SDSS]. We are interested in estimating the fraction of galaxies that are classified as "spiral," i.e., have at least one spiral arm. The covariates X_{ij} in this applications are images (we provide some examples in Figure 4 of the appendix). Existing PPI papers have focused on estimating the overall fraction; we demonstrate how this dataset's metadata structure enables compound estimation of spiral fractions across distinct galaxy subgroups. We first use a pre-defined partition of the galaxies into 122 subgroups that is based on `REDSHIFT`. Second, we estimate the fraction of spiral galaxies in all of the galaxy subgroups simultaneously. For the predictor, we train a `ResNet50` convolutional neural network on a held-out set with around 50k images.

Results & Discussion. For both datasets, each problem (a food product or galaxy subgroup) has its data randomly split into a labeled and unlabeled partition with a 20-80 ratio—a procedure that is repeated for $K = 200$ trials. The results are displayed in Table 3, and here is a brief summary:

- **Amazon Review:** similar to the trend in the synthetic model, the more accurate `BERT-tuned` model enables stronger shrinkage for `PAS` while the biased `BERT-base` predictions necessitate less shrinkage. Our `PAS` estimator adapts to both predictors and outperforms other baselines. `PAS` has the lowest MSE without sacrificing performance on our second metric.
- **Galaxy Zoo 2:** the predictions from `ResNet50` are suboptimal, so the variance-reduction from power tuning dominates any benefit from shrinkage. `PAS` achieves the lowest MSE among all estimators, and improves individual estimates at a level on par with the PT estimator.

7 Conclusion

This paper introduces `PAS`, a novel method for compound mean estimation that effectively combines PPI and EB principles. We motivate the problem through the lens of variance reduction and contextual prior information—then demonstrate how `PAS` could achieve both goals, in theory and in practice. Our paper differs from many other PPI-related works in its focus on estimation, so a natural next step is to develop average coverage controlling intervals for the means that are centered around `PAS`. To this end, it may be fruitful to build on the robust empirical Bayes confidence intervals of Armstrong et al. [2022]. Modern scientific inquiries increasingly demand the simultaneous analysis of multiple related problems. The framework developed in this paper—combining empirical Bayes principles with machine learning predictions—represents a promising direction for such settings.

Impact Statement

By developing a method to improve the efficiency and accuracy of mean estimation using machine learning predictions, this research has the potential to enhance data analysis across various domains where labeled data is limited but predictive models are available.

We acknowledge that advancements in machine learning can have broader societal consequences. However, the ethical considerations directly arising from this methodological contribution are those typically associated with the general advancement of statistical methods and machine learning. We do not foresee specific negative ethical impacts unique to this work that require further detailed discussion.

References

A. N. Angelopoulos, S. Bates, C. Fannjiang, M. I. Jordan, and T. Zrnic. Prediction-powered inference. *Science*, 382(6671):669–674, 2023.

A. N. Angelopoulos, J. C. Duchi, and T. Zrnic. PPI++: Efficient Prediction-Powered Inference. *arXiv preprint*, arXiv:2311.01453, 2024.

T. B. Armstrong, M. Kolesár, and M. Plagborg-Møller. Robust empirical Bayes confidence intervals. *Econometrica*, 90(6):2567–2602, 2022.

R. Baly, G. Da San Martino, J. Glass, and P. Nakov. We can detect your bias: Predicting the political ideology of news articles. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4982–4991. Association for Computational Linguistics, 2020.

F. J. Breidt and J. D. Opsomer. Model-assisted survey estimation with modern prediction techniques. *Statistical Science*, 32(2), 2017.

W. G. Cochran. *Sampling Techniques*. Wiley Series in Probability and Mathematical Statistics. John Wiley & Sons, New York, 3d ed edition, 1977.

A. Deng, Y. Xu, R. Kohavi, and T. Walker. Improving the sensitivity of online controlled experiments by utilizing pre-experiment data. In *Proceedings of the Sixth ACM International Conference on Web Search and Data Mining*, pages 123–132, Rome Italy, 2013. ACM.

J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255. Ieee, 2009.

J. Devlin. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint*, arXiv:1810.04805, 2018.

R. Durrett. *Probability: Theory and Examples*, volume 49. Cambridge University Press, 2019.

B. Efron. *Large-Scale Inference: Empirical Bayes Methods for Estimation, Testing, and Prediction*. Institute of Mathematical Statistics Monographs. Cambridge University Press, Cambridge, 2010.

B. Efron. Tweedie's formula and selection bias. *Journal of the American Statistical Association*, 106 (496):1602–1614, 2011.

B. Efron and C. Morris. Stein's estimation rule and its competitors—an empirical Bayes approach. *Journal of the American Statistical Association*, 68(341):117–130, 1973.

R. E. Fay III and R. A. Herriot. Estimates of income for small places: An application of James-Stein procedures to census data. *Journal of the American Statistical Association*, 74(366a):269–277, 1979.

A. Fisch, J. Maynez, R. A. Hofer, B. Dhingra, A. Globerson, and W. W. Cohen. Stratified prediction-powered inference for hybrid language model evaluation. *arXiv preprint*, arXiv:2406.04291, 2024.

E. J. Green and W. E. Strawderman. A James-Stein type estimator for combining unbiased and possibly biased estimators. *Journal of the American Statistical Association*, 86(416):1001–1006, 1991.

R. E. Hart, S. P. Bamford, K. W. Willett, K. L. Masters, C. Cardamone, C. J. Lintott, R. J. Mackay, R. C. Nichol, C. K. Rosslowe, B. D. Simmons, and R. J. Smethurst. Galaxy Zoo: comparing the demographics of spiral arm number and a new method for correcting redshift bias. *Monthly Notices of the Royal Astronomical Society*, 461(4):3663–3682, Oct. 2016.

K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016.

K. He, X. Chen, S. Xie, Y. Li, P. Dollár, and R. Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16000–16009, 2022.

F. J. Hickernell, C. Lemieux, and A. B. Owen. Control variates for Quasi-Monte Carlo. *Statistical Science*, 20(1):1–31, 2005.

N. Ignatiadis and S. Wager. Covariate-powered empirical Bayes estimation. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

W. James and C. Stein. Estimation with quadratic loss. In *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability*, volume 1, pages 361–379, 1961.

W. Ji, L. Lei, and T. Zrnic. Predictions as surrogates: Revisiting surrogate outcomes in the age of AI. *arXiv preprint*, arXiv:2501.09731, 2025.

W. Jiang and C.-H. Zhang. General maximum likelihood empirical Bayes estimation of normal means. *The Annals of Statistics*, 37(4):1647–1684, 2009.

D. P. Kingma. Adam: A method for stochastic optimization. *arXiv preprint*, arXiv:1412.6980, 2014.

S. C. Kou and J. J. Yang. Optimal shrinkage estimation in heteroscedastic hierarchical linear models. In S. E. Ahmed, editor, *Big and Complex Data Analysis*, pages 249–284. Springer International Publishing, Cham, 2017.

S. S. Lavenberg and P. D. Welch. A perspective on the use of control variables to increase the efficiency of Monte Carlo simulations. *Management Science*, 27(3):322–335, 1981.

F. Liang, S. Mukherjee, and M. West. The use of unlabeled data in predictive modeling. *Statistical Science*, 22(2):189–205, 2007.

J. Y.-Y. Lin, S.-M. Liao, H.-J. Huang, W.-T. Kuo, and O. H.-M. Ou. Galaxy morphological classification with efficient vision transformer. *arXiv preprint*, arXiv:2110.01024, 2021.

W. Lin. Agnostic notes on regression adjustments to experimental data: Reexamining Freedman's critique. *The Annals of Applied Statistics*, 7(1):295–318, 2013.

J. Miao, X. Miao, Y. Wu, J. Zhao, and Q. Lu. Assumption-lean and data-adaptive post-prediction inference, 2024a.

J. Miao, Y. Wu, Z. Sun, X. Miao, T. Lu, J. Zhao, and Q. Lu. Valid inference for machine learning-assisted genome-wide association studies. *Nature Genetics*, pages 1–9, 2024b.

P. Mukhopadhyay and T. Maiti. Two stage non-parametric approach for small area estimation. *Proceedings of ASA Section on Survey Research Methods, hal*, pages 4058–4065, 2004.

P. B. Nair and R. G. Abraham. On the fraction of barred spiral galaxies. *The Astrophysical Journal Letters*, 714(2):L260, 2010.

NLP Town. bert-base-multilingual-uncased-sentiment (revision edd66ab), 2023. URL https://huggingface.co/nlptown/bert-base-multilingual-uncased-sentiment.

G. Penedo, H. Kydlíček, L. B. allal, A. Lozhkov, M. Mitchell, C. Raffel, L. V. Werra, and T. Wolf. The fineweb datasets: Decanting the web for the finest text data at scale. In *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2024.

A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, I. Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.

H. Robbins. Asymptotically subminimax solutions of compound statistical decision problems. In *Proceedings of the Second Berkeley Symposium on Mathematical Statistics and Probability*, volume 2, pages 131–149. University of California Press, 1951.

H. Robbins. An empirical Bayes approach to statistics. In *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Contributions to the Theory of Statistics*, pages 157–163. The Regents of the University of California, 1956.

J. M. Robins, A. Rotnitzky, and L. P. Zhao. Estimation of regression coefficients when some regressors are not always observed. *Journal of the American Statistical Association*, 89(427):846–866, 1994.

E. T. Rosenman, G. Basse, A. B. Owen, and M. Baiocchi. Combining observational and experimental datasets using shrinkage estimators. *Biometrics*, page biom.13827, 2023.

SNAP. Amazon Fine Food Reviews (Stanford network analysis project). Kaggle Dataset, 2014. URL https://www.kaggle.com/datasets/snap/amazon-fine-food-reviews/data. Accessed: 2024-01-01.

C. M. Stein. Estimation of the mean of a multivariate normal distribution. *The Annals of Statistics*, pages 1135–1151, 1981.

R. J. Tibshirani and S. Rosset. Excess optimism: How biased is the apparent error of an estimator tuned by SURE? *Journal of the American Statistical Association*, 114(526):697–712, 2019.

K. Tyser, B. Segev, G. Longhitano, X.-Y. Zhang, Z. Meeks, J. Lee, U. Garg, N. Belsten, A. Shporer, M. Udell, et al. AI-driven review systems: evaluating LLMs in scalable and bias-aware academic reviews. *arXiv preprint arXiv:2408.10365*, 2024.

S. Wang, T. H. McCormick, and J. T. Leek. Methods for correcting inference based on outcomes predicted by machine learning. *Proceedings of the National Academy of Sciences*, 117(48):30266–30275, 2020.

K. W. Willett, C. J. Lintott, S. P. Bamford, K. L. Masters, B. D. Simmons, K. R. Casteels, E. M. Edmondson, L. F. Fortson, S. Kaviraj, W. C. Keel, et al. Galaxy Zoo 2: detailed morphological classifications for 304 122 galaxies from the Sloan Digital Sky Survey. *Monthly Notices of the Royal Astronomical Society*, 435(4):2835–2860, 2013.

T. Wolf. Huggingface's transformers: State-of-the-art natural language processing. *arXiv preprint*, arXiv:1910.03771, 2019.

X. Xie, S. Kou, and L. D. Brown. SURE estimates for a heteroscedastic hierarchical model. *Journal of the American Statistical Association*, 107(500):1465–1479, 2012.

D. G. York, J. Adelman, J. E. Anderson Jr, S. F. Anderson, J. Annis, N. A. Bahcall, J. Bakken, R. Barkhouser, S. Bastian, E. Berman, et al. The sloan digital sky survey: Technical summary. *The Astronomical Journal*, 120(3):1579, 2000.

T. Zrnic. A Note on the prediction-powered bootstrap. *arXiv preprint*, arXiv:2405.18379, 2025.

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

♣ Appendix: Table of Contents

A Proof of Theoretical Results.

A.1 The Correlation-Aware Unbiased Risk Estimate

Theorem A.1. *Let X, Y be two random variables satisfying $\mathbb{E}_\mu[X] = \mu$, $\mathrm{Var}_\mu[X] = \sigma^2$, $\mathrm{Cov}_\mu[X, Y] = \gamma$, and the second moment of Y exists. Consider estimating μ with the shrinkage estimator $\hat{\mu}_c = cX + (1-c)Y$ with $c \in [0, 1]$. Assuming that σ^2 and γ are known, the following estimator*

$$\mathrm{CURE}(\hat{\mu}_c) := (2c-1)\sigma^2 + 2(1-c)\gamma + \{(1-c)(X-Y)\}^2, \tag{19}$$

*defined as the **correlation-aware unbiased risk estimate**, is an unbiased estimator for the risk of $\hat{\mu}_c$ under the quadratic loss, i.e., $\mathbb{E}_\mu[\mathrm{CURE}(\hat{\mu}_c)] = R(\hat{\mu}_c, \mu) = \mathbb{E}_\mu\left[(\hat{\mu}_c - \mu)^2\right]$.*

Proof. First, expand the risk:

$$\begin{aligned} R(\hat{\mu}_c, \mu) &= \mathbb{E}_\mu[(\hat{\mu}_c - \mu)^2] = \mathbb{E}_\mu[(cX + (1-c)Y - \mu)^2] \\ &= \mathrm{Var}_\mu[cX + (1-c)Y] + (\mathbb{E}_\mu[cX + (1-c)Y] - \mu)^2 \\ &= c^2\sigma^2 + (1-c)^2\mathrm{Var}_\mu[Y] + 2c(1-c)\gamma + [(1-c)(\mathbb{E}_\mu[Y] - \mu)]^2. \end{aligned}$$

Then, taking the expectation of $\mathrm{CURE}(\hat{\mu}_c)$:

$$\mathbb{E}_\mu[\mathrm{CURE}(\hat{\mu}_c)] = \underbrace{(2c-1)\sigma^2 + 2(1-c)\gamma}_{\mathbb{I}} + \mathbb{E}_\mu\left[\{(1-c)(X-Y)\}^2\right], \tag{20}$$

where the last term is

$$\begin{aligned} \mathbb{E}_\mu\left[\{(1-c)(X-Y)\}^2\right] &= (1-c)^2\left[(\mathbb{E}_\mu[X-Y])^2 + \mathrm{Var}_\mu[X-Y]\right] \\ &= (1-c)^2\left[(\mathbb{E}_\mu[Y] - \mu)^2 + \sigma^2 + \mathrm{Var}_\mu[Y] - 2\gamma\right] \\ &= [(1-c)(\mathbb{E}_\mu[Y] - \mu)]^2 + \underbrace{(1-c)^2(\sigma^2 + \mathrm{Var}_\mu[Y] - 2\gamma)}_{\mathbb{II}}. \end{aligned}$$

With a little algebra, we observe

$$\begin{aligned} \mathbb{I} + \mathbb{II} &= (2c-1)\sigma^2 + 2(1-c)\gamma + (1-c)^2(\sigma^2 + \mathrm{Var}_\mu[Y] - 2\gamma) \\ &= c^2\sigma^2 + (1-c)^2\mathrm{Var}_\mu[Y] + 2c(1-c)\gamma. \end{aligned}$$

Thus, a term-by-term matching confirms $\mathbb{E}_\mu[\mathrm{CURE}(\hat{\mu}_c)] = R(\hat{\mu}_c, \mu)$. □

Corollary A.2 (Specializations for independence and/or deterministic Y)**.** *If X and Y are independent (i.e. $\gamma = 0$), then (19) simplifies to*

$$\mathrm{CURE}(\hat{\mu}_c) = (2c-1)\,\sigma^2 + \left[(1-c)(X-Y)\right]^2.$$

In particular, this remains an unbiased estimator of the risk

$$R(\hat{\mu}_c, \mu) = \mathbb{E}_\mu\left[\left(\hat{\mu}_c - \mu\right)^2\right],$$

under each of the following scenarios:

*(1) (**Conditioning on both μ and Y**) Even for $\mathbb{E}_\mu\left[(\hat{\mu}_c - \mu)^2 \mid Y\right]$, the expectation of $\mathrm{CURE}(\hat{\mu}_c)$ (conditioned on Y) matches that conditional risk.*

*(2) (**Conditioning only on** μ) The unconditional statement $\mathbb{E}_\mu[\mathrm{CURE}(\hat{\mu}_c)] = R(\hat{\mu}_c, \mu)$ continues to hold when $\gamma = 0$.*

*(3) (**Y is a scalar**) If Y is deterministic (or degenerate) so that $\mathrm{Var}_\mu[Y] = 0$, the expression simplifies accordingly, and the unbiasedness is preserved.*

Hence, the independence assumption $\gamma = 0$ subsumes all these specializations, and $\mathrm{CURE}(\hat{\mu}_c)$ remains an unbiased estimate of the quadratic risk.

Remark A.3. The original **Stein's unbiased risk estimate** (SURE) was proposed in Charles Stein's seminal work [Stein, 1981] to study the quadratic risk of estimating μ_i in Gaussian sequence models $Z_i \sim \mathcal{N}(\mu_i, \sigma^2)$, $i \in [n]$. Let $h : \mathbb{R} \to \mathbb{R}$ be an *absolutely continuous* function and $\mathbb{E}_\mu[|h'(Z)|] < \infty$, then SURE is defined as

$$\mathrm{SURE}(h) := (h(Z) - Z)^2 + 2\sigma^2 h'(Z) - \sigma^2,$$

with the property that $\mathbb{E}_\mu[\mathrm{SURE}(h)] = R(h(Z), \mu) = \mathbb{E}_\mu\big[(h(Z) - \mu)^2\big]$. The original proof relies on **Stein's lemma**, an identity specific to Gaussian random variables Stein [1981], but for the specific class of shrinkage estimators $h_c(Z) = cZ + (1-c)Y$, with $c \in [0,1]$ and $Y \in \mathbb{R}$ being fixed (think about either Y being a constant, or $Y \perp Z$ and we condition on Y), the result extend beyond normal models. We have

$$\begin{aligned}
\mathrm{SURE}(h_c) &= (h_c(Z) - Z)^2 + 2\sigma^2 h_c'(Z) - \sigma^2 \\
&= (cZ + (1-c)Y - Z)^2 + 2c\sigma^2 - \sigma^2 \\
&= [(1-c)(Y - Z)]^2 + (2c-1)\sigma^2 \\
&= \mathrm{CURE}(h_c(Z)),
\end{aligned}$$

which explains how CURE defined in A.1 is connected to SURE.

A.2 Proof of Theorem. 4.1

For each problem $j \in [m]$, we are shrinking the PT estimator $\hat{\theta}_j^{\mathrm{PT}}$ obtained from the first stage towards $\tilde{Z}_j^f$, the prediction mean on the unlabeled data. Conditioning on η_j, we denote

$$\begin{aligned}
\tilde{\sigma}_j^2 &:= \mathrm{Var}_{\eta_j}\left[\hat{\theta}_j^{\mathrm{PT}}\right] = \mathrm{Var}_{\eta_j}\left[\hat{\theta}_{j,\lambda_j^*}^{\mathrm{PPI}}\right] \\
\tilde{\gamma}_j &:= \mathrm{Cov}_{\eta_j}\left[\hat{\theta}_j^{\mathrm{PT}}, \tilde{Z}_j^f\right] = \lambda_j^* \mathrm{Var}_{\eta_j}\left[\tilde{Z}_j^f\right]
\end{aligned}$$

where all the first and second moments of $\hat{\theta}_j^{\mathrm{PT}}$ and $\tilde{Z}_j^f$ exist under the conditions of Assumption 2.1. For each "global" $\omega \geq 0$, the shrinkage parameter for the j-th problem is defined as $\omega_j := \omega/(\omega + \tilde{\sigma}_j^2)$. Then, following the result in Theorem A.1, CURE for $\hat{\theta}_{j,\omega_j}^{\mathrm{PAS}} := \omega_j \hat{\theta}_{j,\omega_j}^{\mathrm{PT}} + (1-\omega_j)\tilde{Z}_j^f$,

$$\mathrm{CURE}\left(\hat{\theta}_{j,\omega}^{\mathrm{PAS}}\right) = (2\omega_j - 1)\tilde{\sigma}_j^2 + 2(1-\omega_j)\tilde{\gamma}_j + \left[(1-\omega_j)(\hat{\theta}_{j,\omega_j}^{\mathrm{PT}} - \tilde{Z}_j^f)\right]^2,$$

is an unbiased estimator of the risk, i.e.,

$$\mathbb{E}_{\eta_j}\left[\mathrm{CURE}\left(\hat{\theta}_{j,\omega}^{\mathrm{PAS}}\right)\right] = R(\hat{\theta}_{j,\omega_j}^{\mathrm{PAS}}, \theta_j).$$

Finally, the CURE for the collection of estimators is $\hat{\boldsymbol{\theta}}_\omega^{\mathrm{PAS}} := (\hat{\theta}_{1,\omega}^{\mathrm{PAS}}, \ldots, \hat{\theta}_{m,\omega}^{\mathrm{PAS}})^\intercal$

$$\mathrm{CURE}\left(\hat{\boldsymbol{\theta}}_\omega^{\mathrm{PAS}}\right) := \frac{1}{m}\sum_{j=1}^m \mathrm{CURE}\left(\hat{\theta}_{j,\omega}^{\mathrm{PAS}}\right),$$

which is an unbiased estimator of the compound risk $\mathcal{R}_m(\hat{\boldsymbol{\theta}}_\omega^{\mathrm{PAS}}, \boldsymbol{\theta})$ by the linearity of the expectation. □

A.3 Formal conditions and proof of Proposition 5.1

We aim to prove that the correlation-aware unbiased risk estimate (CURE) converges uniformly to the true squared-error loss $\ell_m(\hat{\boldsymbol{\theta}}_\omega^{\mathrm{PAS}}, \boldsymbol{\theta})$ as $m \to \infty$. Specifically, our goal is to establish

$$\sup_{\omega \geq 0} \left| \mathrm{CURE}(\hat{\boldsymbol{\theta}}_\omega^{\mathrm{PAS}}) - \ell_m(\hat{\boldsymbol{\theta}}_\omega^{\mathrm{PAS}}, \boldsymbol{\theta}) \right| \xrightarrow[m\to\infty]{L^1} 0.$$

For this proposition, all the expectation and variance terms without subscript are conditioning on $\boldsymbol{\eta}$. We keep using the notations $\theta_j = \mathbb{E}[\hat{\theta}_j^{\mathrm{PT}}]$, $\mu_j = \mathbb{E}[\tilde{Z}_j^f]$, $\tilde{\sigma}_j^2 := \mathrm{Var}[\hat{\theta}_j^{\mathrm{PT}}]$ and $\tilde{\gamma}_j := \mathrm{Cov}[\hat{\theta}_j^{\mathrm{PT}}, \tilde{Z}_j^f]$.

Step 1: Decompose the difference. We first decompose both CURE and the loss separately as

$$\begin{aligned}
\mathrm{CURE}(\hat{\boldsymbol{\theta}}_\omega^{\mathrm{PAS}}) &= \frac{1}{m}\sum_{j=1}^m (2\omega_j - 1)\tilde{\sigma}_j^2 + \left[(1-\omega_j)(\hat{\theta}_j^{\mathrm{PT}} - \tilde{Z}_j^f)\right]^2 + 2(1-\omega_j)\tilde{\gamma}_j \\
&= \underbrace{\frac{1}{m}\sum_{j=1}^m (2\omega_j - 1)\tilde{\sigma}_j^2 + (1-\omega_j)^2(\hat{\theta}_j^{\mathrm{PT}} - \mu_j)^2}_{\mathbb{I}(\omega)} \\
&\quad + \underbrace{\frac{1}{m}\sum_{j=1}^m 2(1-\omega_j)\tilde{\gamma}_j + 2(1-\omega_j)^2(\tilde{Z}_j^f - \mu_j)(\hat{\theta}_j^{\mathrm{PT}} - \mu_j) + (1-\omega_j)^2(\mu_j - \tilde{Z}_j^f)^2}_{\mathbb{III}(\omega)} \\
\ell_m(\hat{\boldsymbol{\theta}}_\omega^{\mathrm{PAS}}, \boldsymbol{\theta}) &= \underbrace{\frac{1}{m}\sum_{j=1}^m (\omega_j\hat{\theta}_j^{\mathrm{PT}} + (1-\omega_j)\mu_j - \theta_j)^2}_{\mathbb{I}^*(\omega)} \\
&\quad + \underbrace{\frac{1}{m}\sum_{j=1}^m 2(1-\omega_j)(\tilde{Z}_j^f - \mu_j)(\hat{\theta}_j^{\mathrm{PT}} - \theta_j) + 2(1-\omega_j)^2(\tilde{Z}_j^f - \mu_j)(\hat{\theta}_j^{\mathrm{PT}} - \mu_j) + (1-\omega_j)^2(\mu_j - \tilde{Z}_j^f)^2}_{\mathbb{III}^*(\omega)}
\end{aligned}$$

and we are interested in bounding

$$\sup_{\omega \geq 0} \left| \mathrm{CURE}(\hat{\boldsymbol{\theta}}_\omega^{\mathrm{PAS}}) - \ell_m(\hat{\boldsymbol{\theta}}_\omega^{\mathrm{PAS}}, \boldsymbol{\theta}) \right| \leq \sup_{\omega \geq 0} |\mathbb{I}(\omega) - \mathbb{I}^*(\omega)| + \sup_{\omega \geq 0} |\mathbb{III}(\omega) - \mathbb{III}^*(\omega)| \tag{21}$$

Step 2: Bounding the first difference $\Delta_1(\omega) := \mathbb{I}(\omega) - \mathbb{I}^*(\omega)$. The proof in this step is directly adapted from **Theorem 5.1** in Xie et al. [2012] and generalizes to non-Gaussian data. With some algebraic manipulation, we can further decompose

$$\begin{aligned}
\Delta_1(\omega) &= \frac{1}{m}\sum_{j=1}^m (2\omega_j - 1)\tilde{\sigma}_j^2 + (1-\omega_j)^2(\hat{\theta}_j^{\mathrm{PT}} - \mu_j)^2 \\
&\quad - \frac{1}{m}\sum_{j=1}^m (\omega_j\hat{\theta}_j^{\mathrm{PT}} + (1-\omega_j)\mu_j - \theta_j)^2 \\
&= \mathrm{CURE}(\hat{\boldsymbol{\theta}}_\omega^0) - \ell_m(\hat{\boldsymbol{\theta}}_\omega^0, \boldsymbol{\theta}) - \frac{2}{m}\sum_{j=1}^m \mu_j(1-\omega_j)(\hat{\theta}_j^{\mathrm{PT}} - \theta_j)
\end{aligned}$$

$$\text{where} \quad \mathrm{CURE}(\hat{\boldsymbol{\theta}}_\omega^0) = \frac{1}{m}\sum_{j=1}^m (2\omega_j - 1)\tilde{\sigma}_j^2 + (1-\omega_j)^2\left(\hat{\theta}_j^{\mathrm{PT}}\right)^2, \quad \ell_m(\hat{\boldsymbol{\theta}}_\omega^0, \boldsymbol{\theta}) = \frac{1}{m}\sum_{j=1}^m (\omega_j\hat{\theta}_j^{\mathrm{PT}} - \theta_j)^2$$

are equal to CURE and the loss of the "shrink-towards-zero" estimator $\hat{\theta}_{j,\omega}^0 := \omega_j\hat{\theta}_j^{\mathrm{PT}}$. We thus have

$$\sup_{\omega \geq 0} |\Delta_1(\omega)| \leq \sup_{\omega \geq 0} \left| \mathrm{CURE}(\hat{\boldsymbol{\theta}}_\omega^0) - \ell_m(\hat{\boldsymbol{\theta}}_\omega^0, \boldsymbol{\theta}) \right| + \frac{2}{m}\sup_{\omega \geq 0} \left| \sum_{j=1}^m \mu_j(1-\omega_j)(\hat{\theta}_j^{\mathrm{PT}} - \theta_j) \right| \tag{22}$$

Now, rearrangements of terms gives that

$$\sup_{\omega\geq 0}\left|\mathrm{CURE}(\hat{\boldsymbol{\theta}}^0_\omega)-\ell_m(\hat{\boldsymbol{\theta}}^0_\omega,\boldsymbol{\theta})\right| = \sup_{\omega\geq 0}\left|\frac{1}{m}\sum_{j=1}^m \left(\hat{\theta}_j^{\mathrm{PT}}\right)^2-\tilde{\sigma}_j^2-\theta_j^2-2\omega_j(\left(\hat{\theta}_j^{\mathrm{PT}}\right)^2-\hat{\theta}_j^{\mathrm{PT}}\theta_j-\tilde{\sigma}_j^2)\right|$$
$$\leq \left|\frac{1}{m}\sum_{j=1}^m \left(\hat{\theta}_j^{\mathrm{PT}}\right)^2-\tilde{\sigma}_j^2-\theta_j^2\right| + \sup_{\omega\geq 0}\left|\frac{1}{m}\sum_{j=1}^m 2\omega_j\Big(\left(\hat{\theta}_j^{\mathrm{PT}}\right)^2-\hat{\theta}_j^{\mathrm{PT}}\theta_j-\tilde{\sigma}_j^2\Big)\right|$$

For the first term,

$$\mathbb{E}_{\mathbb{P}_\eta}\left[\mathbb{E}_{\boldsymbol{\eta}}\left[\left(\frac{1}{m}\sum_{j=1}^m \left(\hat{\theta}_j^{\mathrm{PT}}\right)^2-\tilde{\sigma}_j^2-\theta_j^2\right)^2\right]\right] = \frac{1}{m^2}\sum_{j=1}^m \mathbb{E}_{\mathbb{P}_\eta}\left[\mathrm{Var}_{\eta_j}\left[\left(\hat{\theta}_j^{\mathrm{PT}}\right)^2\right]\right] \leq \frac{1}{m}\sup_j \mathrm{Var}_{\mathbb{P}_\eta}\left[\left(\hat{\theta}_j^{\mathrm{PT}}\right)^2\right].$$

Thus by Jensen's inequality and iterated expectation:

$$\mathbb{E}_{\mathbb{P}_\eta}\left[\left|\frac{1}{m}\sum_{j=1}^m \left(\hat{\theta}_j^{\mathrm{PT}}\right)^2-\tilde{\sigma}_j^2-\theta_j^2\right|\right] \leq \left(\frac{1}{m}\sup_j \mathrm{Var}_{\mathbb{P}_\eta}\left[\left(\hat{\theta}_j^{\mathrm{PT}}\right)^2\right]\right)^{1/2}.$$

For the second term, we first assume that $\tilde{\sigma}_1^2 \leq ... \leq \tilde{\sigma}_m^2$ without loss of generality. Then, since ω_j is monotonic function of $\tilde{\sigma}_j^2$ and lies in $[0,1]$, we have (since $1\geq\omega_1,\geq ...\geq\omega_m\geq 0$ for all $\omega\geq 0$)

$$\sup_{\omega\geq 0}\left|\frac{1}{m}\sum_{j=1}^m 2\omega_j(\left(\hat{\theta}_j^{\mathrm{PT}}\right)^2-\hat{\theta}_j^{\mathrm{PT}}\theta_j-\tilde{\sigma}_j^2)\right| \leq \max_{1\geq c_1\geq\ldots\geq c_m\geq 0}\left|\frac{2}{m}\sum_{j=1}^m c_j(\left(\hat{\theta}_j^{\mathrm{PT}}\right)^2-\hat{\theta}_j^{\mathrm{PT}}\theta_j-\tilde{\sigma}_j^2)\right| \tag{23}$$

The following lemma would help us for handling the RHS of (23), as long as many other similar occurances below.

Lemma A.4. *Let $A_1,\ldots,A_n$ be real numbers. Then*

$$\max_{1\geq c_1\geq\cdots\geq c_n\geq 0}\left|\sum_{i=1}^n c_i\,A_i\right| \;=\; \max_{1\leq k\leq n}\left|\sum_{i=1}^k A_i\right|.$$

Proof. Define $S_k=\sum_{i=1}^k A_i$ for $k=1,\ldots,n$, and let $c_1,\ldots,c_n$ be real numbers satisfying $1\geq c_1\geq\cdots\geq c_n\geq 0$. Set $c_{n+1}=0$. Then we can rewrite

$$\sum_{i=1}^n c_i\,A_i \;=\; \sum_{k=1}^n (c_k-c_{k+1})\Big(\sum_{i=1}^k A_i\Big) \;=\; \sum_{k=1}^n (c_k-c_{k+1})\,S_k.$$

Since $c_k\geq c_{k+1}$, each $\alpha_k:=c_k-c_{k+1}$ is nonnegative, and

$$\sum_{k=1}^n \alpha_k = c_1-c_{n+1}\leq 1.$$

Hence,

$$\left|\sum_{i=1}^n c_iA_i\right| \;=\; \Big|\sum_{k=1}^n \alpha_k\,S_k\Big| \;\leq\; \sum_{k=1}^n \alpha_k\,|S_k| \;\leq\; \Big(\max_{1\leq k\leq n}|S_k|\Big)\Big(\sum_{k=1}^n \alpha_k\Big) \;\leq\; \max_{1\leq k\leq n}|S_k|.$$

This shows

$$\max_{1\geq c_1\geq\cdots\geq c_n\geq 0}\left|\sum_{i=1}^n c_i\,A_i\right| \;\leq\; \max_{1\leq k\leq n}\left|\sum_{i=1}^k A_i\right|.$$

To see that this upper bound can be attained, consider for each k the choice

$$c_1=c_2=\cdots=c_k=1,\quad c_{k+1}=c_{k+2}=\cdots=c_n=0.$$

Clearly $1 \geq c_1 \geq \cdots \geq c_n \geq 0$, and in this case

$$\sum_{i=1}^{n} c_i A_i = \sum_{i=1}^{k} A_i = S_k.$$

Taking the maximum over all such $k \in \{1, \ldots, n\}$ matches $\max_{1 \leq k \leq n} |S_k|$. Thus,

$$\max_{1 \geq c_1 \geq \cdots \geq c_n \geq 0} \left| \sum_{i=1}^{n} c_i A_i \right| \;=\; \max_{1 \leq k \leq n} \left| \sum_{i=1}^{k} A_i \right|,$$

as claimed. □

With Lemma A.4, we easily have

$$\max_{1 \geq c_1 \geq \ldots \geq c_m \geq 0} \left| \frac{2}{m} \sum_{j=1}^{m} c_j \left(\left(\hat{\theta}_j^{\mathrm{PT}}\right)^2 - \hat{\theta}_j^{\mathrm{PT}} \theta_j - \tilde{\sigma}_j^2 \right) \right| = \max_{1 \leq k \leq m} \left| \frac{2}{m} \sum_{j=1}^{k} \left(\left(\hat{\theta}_j^{\mathrm{PT}}\right)^2 - \hat{\theta}_j^{\mathrm{PT}} \theta_j - \tilde{\sigma}_j^2 \right) \right|$$

Let $M_k := \sum_{j=1}^{k} \left(\left(\hat{\theta}_j^{\mathrm{PT}}\right)^2 - \hat{\theta}_j^{\mathrm{PT}} \theta_j - \tilde{\sigma}_j^2 \right)$, it is easy to see that $\{M_k\}_{k=1}^{m}$ forms a martingale conditional on $\boldsymbol{\eta}$. Therefore, by a standard L^2 maximal inequality (e.g. Theorem 4.4.6 in Durrett [2019]), we have

$$\mathbb{E}_{\boldsymbol{\eta}} \left[\max_{1 \leq k \leq m} M_k^2 \right] \leq 4 \mathbb{E}_{\boldsymbol{\eta}} \left[M_m^2 \right] = 4 \sum_{j=1}^{m} \mathrm{Var}_{\boldsymbol{\eta}} \left[\left(\hat{\theta}_j^{\mathrm{PT}}\right)^2 - \hat{\theta}_j^{\mathrm{PT}} \theta_j \right] \tag{24}$$

which then implies

$$\mathbb{E}_{\mathbb{P}_\eta} \left[\left(\sup_{\omega \geq 0} \left| \frac{1}{m} \sum_{j=1}^{m} 2\omega_j \left(\left(\hat{\theta}_j^{\mathrm{PT}}\right)^2 - \hat{\theta}_j^{\mathrm{PT}} \theta_j - \tilde{\sigma}_j^2 \right) \right| \right)^2 \right] \leq \frac{4}{m^2} \mathbb{E}_{\mathbb{P}_\eta} \left[\max_{1 \leq k \leq m} M_k^2 \right] = \frac{16}{m} \sup_j \mathrm{Var}_{\mathbb{P}_\eta} \left[\left(\hat{\theta}_j^{\mathrm{PT}}\right)^2 - \hat{\theta}_j^{\mathrm{PT}} \theta_j \right] \tag{25}$$

$$\Longrightarrow \quad \mathbb{E}_{\mathbb{P}_\eta} \left[\sup_{\omega \geq 0} \left| \frac{1}{m} \sum_{j=1}^{m} 2\omega_j \left(\left(\hat{\theta}_j^{\mathrm{PT}}\right)^2 - \hat{\theta}_j^{\mathrm{PT}} \theta_j - \tilde{\sigma}_j^2 \right) \right| \right] \leq \left(\frac{16}{m} \sup_j \mathrm{Var}_{\mathbb{P}_\eta} \left[\left(\hat{\theta}_j^{\mathrm{PT}}\right)^2 - \hat{\theta}_j^{\mathrm{PT}} \theta_j \right] \right)^{1/2}. \tag{26}$$

Next, we bound the last expression in (22): $\frac{2}{m} \sup_{\omega \geq 0} \left| \sum_{j=1}^{m} (1 - \omega_j) \mu_j (\hat{\theta}_j^{\mathrm{PT}} - \theta_j) \right|$. Note that $(1 - \omega_j)$ is also monotonic in $\tilde{\sigma}_j^2$, and the random sequence $M'_k := \sum_{j=1}^{k} \mu_j (\hat{\theta}_j^{\mathrm{PT}} - \theta_j)$ forms another martingale. Therefore, following the same argument as (23)–(24) gives

$$\begin{aligned} \frac{4}{m^2} \mathbb{E}_{\mathbb{P}_\eta} \left[\sup_{\omega \geq 0} \left| \sum_{j=1}^{m} (1 - \omega_j) \mu_j (\hat{\theta}_j^{\mathrm{PT}} - \theta_j) \right|^2 \right] &\leq \frac{4}{m^2} \mathbb{E}_{\mathbb{P}_\eta} \left[\max_{1 \leq k \leq m} {M'_k}^2 \right] \\ &\leq \frac{16}{m^2} \mathbb{E}_{\mathbb{P}_\eta} \left[{M'_m}^2 \right] = \frac{16}{m} \sup_j \mathbb{E}_{\mathbb{P}_\eta} \left[\mathrm{Var}_{\eta_j} \left[\hat{\theta}_j^{\mathrm{PT}} \right] \mu_j^2 \right] \\ \Longrightarrow \mathbb{E}_{\mathbb{P}_\eta} \left[\frac{2}{m} \sup_{\omega \geq 0} \left| \sum_{j=1}^{m} (1 - \omega_j) \mu_j (\hat{\theta}_j^{\mathrm{PT}} - \theta_j) \right| \right] &\leq \left(\frac{16}{m} \sup_j \mathbb{E}_{\mathbb{P}_\eta} \left[\mathrm{Var}_{\eta_j} \left[\hat{\theta}_j^{\mathrm{PT}} \right] \mu_j^2 \right] \right)^{1/2}. \end{aligned}$$

The above arguments establish control on

$$\mathbb{E}_{\mathbb{P}_\eta} \left[\sup_{\omega \geq 0} |\Delta_1(\omega)| \right].$$

Step 3: Bounding the second difference $\Delta_2(\omega) := \mathbb{III}(\omega) - \mathbb{III}^*(\omega)$.

We next cancel out identical terms in the second difference in (21) and get

$$\Delta_2(\omega) = \frac{2}{m} \sum_{j=1}^{m} (1 - \omega_j) \left[\tilde{\gamma}_j - (\tilde{Z}_j^f - \mu_j)(\hat{\theta}_j^{\mathrm{PT}} - \theta_j) \right] \tag{27}$$

By the same proof logic that has been applied twice above, we now have a function $(1-\omega_j)$ monotonic in $\tilde{\sigma}_j^2$, and a martingale $Q_k := \sum_{j=1}^{k}\left[\tilde{\gamma}_j - (\tilde{Z}_j^f - \mu_j)(\hat{\theta}_j^{\mathrm{PT}} - \theta_j)\right]$ for $k = 1, \ldots, m$. The steps from (23)–(24) follows, and we have

$$\frac{4}{m^2}\mathbb{E}_{\mathbb{P}_\eta}\left[\left(\sup_{\omega\geq 0}\left|\sum_{j=1}^{m}(1-\omega_j)\left[\tilde{\gamma}_j - (\tilde{Z}_j^f - \mu_j)(\hat{\theta}_j^{\mathrm{PT}} - \theta_j)\right]\right|\right)^2\right] \leq \frac{16}{m}\sup_j \mathbb{E}_{\mathbb{P}_\eta}\left[\mathrm{Var}_{\eta_j}\left[(\tilde{Z}_j^f - \mu_j)(\hat{\theta}_j^{\mathrm{PT}} - \theta_j)\right]\right],$$

and so,

$$\mathbb{E}_{\mathbb{P}_\eta}\left[\frac{2}{m}\sup_{\omega\geq 0}\left|\sum_{j=1}^{m}(1-\omega_j)\left[\tilde{\gamma}_j - (\tilde{Z}_j^f - \mu_j)(\hat{\theta}_j^{\mathrm{PT}} - \theta_j)\right]\right|\right] \leq \left(\frac{16}{m}\sup_j \mathbb{E}_{\mathbb{P}_\eta}\left[\mathrm{Var}_{\eta_j}\left[(\tilde{Z}_j^f - \mu_j)(\hat{\theta}_j^{\mathrm{PT}} - \theta_j)\right]\right]\right)^{1/2}.$$

This establishes control on

$$\mathbb{E}_{\mathbb{P}_\eta}\left[\sup_{\omega\geq 0}|\Delta_2(\omega)|\right].$$

Step 4: Wrap everything up .

Finally, based on Steps 1–3, we have that

$$\mathbb{E}_{\mathbb{P}_\eta}\left[\sup_{\omega\geq 0}\left|\mathrm{CURE}(\hat{\boldsymbol{\theta}}_\omega^{\mathrm{PAS}}) - \ell_m(\hat{\boldsymbol{\theta}}_\omega^{\mathrm{PAS}}, \boldsymbol{\theta})\right|\right] \leq \mathbb{E}_{\mathbb{P}_\eta}\left[\sup_{\omega\geq 0}|\Delta_1(\omega)|\right] + \mathbb{E}_{\mathbb{P}_\eta}\left[\sup_{\omega\geq 0}|\Delta_2(\omega)|\right],$$

and both terms on the right hand side converge to zero by our preceding bounds and the moment assumptions in the statement of the theorem. □

A.4 Proof of Theorem 5.2

We apply a standard argument used to prove consistency of M-estimators.

Let ω_* be the oracle choice of $\omega \geq 0$ that minimizes the Bayes risk $\mathcal{B}_m^{\mathbb{P}_\eta}(\hat{\boldsymbol{\theta}}_\omega^{\mathrm{PAS}})$. Notice that by definition of $\hat{\omega}$ as the minimizer of CURE,

$$\mathrm{CURE}(\hat{\boldsymbol{\theta}}_{\hat{\omega}}^{\mathrm{PAS}}) \leq \mathrm{CURE}(\hat{\boldsymbol{\theta}}_{\omega_*}^{\mathrm{PAS}}).$$

Then:

$$\ell_m(\hat{\boldsymbol{\theta}}_{\hat{\omega}}^{\mathrm{PAS}}, \boldsymbol{\theta}) - \ell_m(\hat{\boldsymbol{\theta}}_{\omega_*}^{\mathrm{PAS}}, \boldsymbol{\theta}) \leq 2\sup_{\omega\geq 0}\left|\mathrm{CURE}(\hat{\boldsymbol{\theta}}_\omega^{\mathrm{PAS}}) - \ell_m(\hat{\boldsymbol{\theta}}_\omega^{\mathrm{PAS}}, \boldsymbol{\theta})\right|.$$

Taking expectations,

$$\mathcal{B}_m^{\mathbb{P}_\eta}(\hat{\boldsymbol{\theta}}_{\hat{\omega}}^{\mathrm{PAS}}) - \mathcal{B}_m^{\mathbb{P}_\eta}(\hat{\boldsymbol{\theta}}_{\omega_*}^{\mathrm{PAS}}) \leq 2\,\mathbb{E}_{\mathbb{P}_\eta}\left[\sup_{\omega\geq 0}\left|\mathrm{CURE}(\hat{\boldsymbol{\theta}}_\omega^{\mathrm{PAS}}) - \ell_m(\hat{\boldsymbol{\theta}}_\omega^{\mathrm{PAS}}, \boldsymbol{\theta})\right|\right].$$

Noting that the right hand side converges to 0 as $m \to \infty$, and recalling the definition of ω_*, we prove the desired result

$$\mathcal{B}_m^{\mathbb{P}_\eta}(\hat{\boldsymbol{\theta}}_{\hat{\omega}}^{\mathrm{PAS}}) \leq \inf_{\omega\geq 0}\mathcal{B}_m^{\mathbb{P}_\eta}(\hat{\boldsymbol{\theta}}_\omega^{\mathrm{PAS}}) + o(1).$$

□

A.5 Proof of Proposition 5.3

We start with the result in Theorem 5.2

$$\mathcal{B}_m^{\mathbb{P}_\eta}(\hat{\boldsymbol{\theta}}_{\hat{\omega}}^{\mathrm{PAS}}, \boldsymbol{\theta}) \leq \inf_{\omega}\ \mathcal{B}_m^{\mathbb{P}_\eta}(\hat{\boldsymbol{\theta}}_\omega^{\mathrm{PAS}}, \boldsymbol{\theta}) + o(1).$$

Now, since we are integrating over $\eta \sim \mathbb{P}_\eta$ for all problems

$$\begin{aligned}\mathcal{B}_m^{\mathbb{P}_\eta}(\hat{\boldsymbol{\theta}}_\omega^{\mathrm{PAS}}, \boldsymbol{\theta}) &= \frac{1}{m}\sum_{j=1}^{m} \mathbb{E}_{\mathbb{P}_\eta}\left[\left(\hat{\theta}_{j,\omega}^{\mathrm{PAS}} - \theta_j\right)^2\right] \\ &= \mathbb{E}_{\mathbb{P}_\eta}\left[\left(\hat{\theta}_{j,\omega}^{\mathrm{PAS}} - \theta_j\right)^2\right],\end{aligned}$$

by definition $\hat{\theta}_{j,\omega}^{\mathrm{PAS}} = \omega_j \hat{\theta}_j^{\mathrm{PT}} + (1-\omega_j)\tilde{Z}_j^f$, where $\omega_j = \omega/(\omega + \tilde{\sigma}^2)$. Therefore

$$\begin{aligned}\mathbb{E}_{\mathbb{P}_\eta}\left[\left(\hat{\theta}_{j,\omega}^{\mathrm{PAS}} - \theta_j\right)^2\right] &= \mathbb{E}_{\mathbb{P}_\eta}\left[\left(\omega_j\left(\hat{\theta}_j^{\mathrm{PT}} - \theta_j\right) + (1-\omega_j)\left(\tilde{Z}_j^f - \theta_j\right)\right)^2\right] \\ &= \frac{\omega^2}{(\omega+\tilde{\sigma}^2)^2}\mathbb{E}_{\mathbb{P}_\eta}\left[\left(\hat{\theta}_j^{\mathrm{PT}} - \theta_j\right)^2\right] + \frac{\tilde{\sigma}^4}{(\omega+\tilde{\sigma}^2)^2}\mathbb{E}_{\mathbb{P}_\eta}\left[\left(\tilde{Z}_j^f - \theta_j\right)^2\right] \\ &\quad + 2\frac{\tilde{\sigma}^2\omega}{(\omega+\tilde{\sigma}^2)^2}\mathbb{E}_{\mathbb{P}_\eta}\left[\left(\hat{\theta}_j^{\mathrm{PT}} - \theta_j\right)\left(\tilde{Z}_j^f - \theta_j\right)\right].\end{aligned}$$

By our assumption, second moment terms like $\tilde{\sigma}^2$ and $\tilde{\gamma}$ are now fixed, so we have (by iterated expectation)

$$\mathbb{E}_{\mathbb{P}_\eta}\left[\left(\hat{\theta}_j^{\mathrm{PT}} - \theta_j\right)^2\right] = \tilde{\sigma}^2.$$

Noting that $\tilde{\gamma}_j = 0$ since $N_j = \infty$, we have

$$\mathbb{E}_{\mathbb{P}_\eta}\left[\left(\hat{\theta}_{j,\omega}^{\mathrm{PAS}} - \theta_j\right)^2\right] = \frac{\omega^2\tilde{\sigma}^2}{(\omega+\tilde{\sigma}^2)^2} + \frac{\tilde{\sigma}^4}{(\omega+\tilde{\sigma}^2)^2}\mathbb{E}_{\mathbb{P}_\eta}\left[\left(\tilde{Z}_j^f - \theta_j\right)^2\right].$$

For the first two terms, plugging in $\omega = \mathbb{E}_{\mathbb{P}_\eta}\left[\left(\tilde{Z}_j^f - \theta_j\right)^2\right]$ gives

$$\frac{\omega^2\tilde{\sigma}^2}{(\omega+\tilde{\sigma}^2)^2} + \frac{\tilde{\sigma}^4}{(\omega+\tilde{\sigma}^2)^2}\mathbb{E}_{\mathbb{P}_\eta}\left[\left(\tilde{Z}_j^f - \theta_j\right)^2\right] = \frac{\tilde{\sigma}^2\mathbb{E}_{\mathbb{P}_\eta}\left[(\tilde{Z}_j^f - \theta_j)^2\right]}{\tilde{\sigma}^2 + \mathbb{E}_{\mathbb{P}_\eta}\left[(\tilde{Z}_j^f - \theta_j)^2\right]}.$$

We finally have

$$\mathcal{B}_m^{\mathbb{P}_\eta}(\hat{\boldsymbol{\theta}}^{\mathrm{PAS}}) \le \frac{\tilde{\sigma}^2\mathbb{E}_{\mathbb{P}_\eta}\left[(\tilde{Z}_j^f - \theta_j)^2\right]}{\tilde{\sigma}^2 + \mathbb{E}_{\mathbb{P}_\eta}\left[(\tilde{Z}_j^f - \theta_j)^2\right]} + o(1).$$

□

B Baseline Shrinkage Estimator

This appendix details the "shrinkage-only" baseline referenced in Section 6 of the main text. This estimator applies shrinkage directly to the classical estimator $\bar{Y}_j$, using the prediction mean $\tilde{Z}_j^f$ as a shrinkage target *without* first applying power-tuned PPI. We include this baseline to isolate the benefits of power tuning from the `PAS` estimator as an ablation study.

Formulation. The shrinkage-only estimator for problem j takes the form:

$$\begin{aligned}\hat{\theta}_{j,\omega}^{\mathrm{Shrink}} &:= \omega_j \bar{Y}_j + (1-\omega_j)\tilde{Z}_j^f, \\ \text{where} \quad \omega_j &:= \omega/(\omega + \tilde{\sigma}_j^2), \quad \tilde{\sigma}_j^2 := \mathrm{Var}[\bar{Y}_j] = \sigma_j^2/n_j.\end{aligned}$$

Here $\omega \ge 0$ is a global shrinkage parameter analogous to Section 4.2. The key difference from `PAS` is that we shrink the *classical estimator* $\bar{Y}_j$ (which is independent of $\tilde{Z}_j^f$) rather than the power-tuned estimator $\hat{\theta}_j^{\mathrm{PT}}$ (which is correlated).

Optimizing ω via CURE Since $\bar{Y}_j$ and $\tilde{Z}_j^f$ are independent, Theorem 4.1 simplifies considerably. Let $\tilde{\gamma}_j := \mathrm{Cov}\left[\bar{Y}_j, \tilde{Z}_j^f\right] = 0$ and $\tilde{\sigma}_j^2 = \sigma_j^2/n_j$. The correlation-aware unbiased risk estimate (CURE) becomes:

$$\mathrm{CURE}(\hat{\theta}_{j,\omega}^{\mathrm{Shrink}}) = (2\omega_j - 1)\tilde{\sigma}_j^2 + \left[(1-\omega_j)(\bar{Y}_j - \tilde{Z}_j^f)\right]^2.$$

This follows from Theorem 4.1 by setting $\tilde{\gamma}_j = 0$. The global shrinkage parameter ω is selected by minimizing the average CURE across all m problems.

$$\hat{\theta}_j^{\mathrm{Shrink}} := \hat{\omega}_j \bar{Y}_j + (1-\hat{\omega}_j)\tilde{Z}_j^f, \quad \hat{\omega}_j = \hat{\omega}/(\hat{\omega} + \tilde{\sigma}_j^2),$$
$$\text{where} \quad \hat{\omega} = \arg\min_{\omega \geq 0} \frac{1}{m}\sum_{j=1}^{m} \mathrm{CURE}(\hat{\theta}_{j,\omega}^{\mathrm{Shrink}}) \tag{28}$$

The optimal $\hat{\omega}$ does not admit a closed-form expression, but we can easily compute it numerically by grid search. Below we detail the pseudo-code for implementing the shrinkage-only estimator.

Algorithm 2 Baseline Shrinkage Estimator

Require: $\{(X_{ij}, Y_{ij})_{i=1}^{n_j}\}, \{\tilde{X}_{ij}\}_{i=1}^{N_j}$ for $j \in [m]$, variance parameters $\{\sigma_j\}_{j=1}^m$, predictive model f
1: **for** $j = 1$ to m **do**
2: $\quad \bar{Y}_j, \tilde{Z}_j^f = \texttt{get\_means}((X_{ij}, Y_{ij})_{i=1}^{n_j}, (\tilde{X}_{ij})_{i=1}^{N_j}, f)$
3: $\quad \tilde{\sigma}_j^2 \leftarrow \sigma_j^2/n_j$ $\quad$ {variance of $\bar{Y}_j$}
4: **end for**
5: $\hat{\omega} = \texttt{get\_shrink\_param}(\{\bar{Y}_j\}_{j=1}^m, \{\tilde{Z}_j^f\}_{j=1}^m, \{\tilde{\sigma}_j^2\}_{j=1}^m)$ {use Eq. (28)}
6: **for** $j = 1$ to m **do**
7: $\quad \hat{\omega}_j \;=\; \hat{\omega} \Big/ (\hat{\omega} + \tilde{\sigma}_j^2)$
8: $\quad \hat{\theta}_j^{\mathrm{Shrink}} = \hat{\omega}_j\, \bar{Y}_j \;+\; (1-\hat{\omega}_j)\, \tilde{Z}_j^f$
9: **end for**
10: **return** $\{\hat{\theta}_j^{\mathrm{Shrink}}\}_{j=1}^m$

C Experiment Details

C.1 Synthetic Model

Motivation. In Example 2.2, we described the following data generation process (copied from Eq. (8))

$$\eta_j \sim \mathcal{U}[-1,1], \quad j = 1, \ldots, m,$$
$$X_{ij} \sim \mathcal{N}(\eta_j, \psi^2), \quad Y_{ij}|X_{ij} \sim \mathcal{N}(2\eta_j X_{ij} - \eta_j^2, c), \quad i = 1, \ldots, n_j,$$

and the same for $(\tilde{X}_{ij}, \tilde{Y}_{ij})$. ψ and c are two hyperparameters that we chose to be 0.1 and 0.05, respectively. The (marginal) mean and variance of Y_{ij} are

$$\theta_j := \mathbb{E}_{\eta_j}[Y_{ij}] = \eta_j^2, \quad \sigma_j^2 := \mathrm{Var}_{\eta_j}[Y_{ij}] = 4\eta_j^2\psi^2 + c.$$

To understand the motivation behind this setup, we can further inspect the covariance between X_{ij} and Y_{ij}, which can easily be verified to be $\mathrm{Cov}_{\eta_j}[X_{ij}, Y_{ij}] = 2\eta_j\psi^2$. Therefore, if we consider the *ratio between the absolute covariance and the variance* (of Y_{ij}) as a characterization of the "inherent predictability" of a problem (which is also very relevant to the power-tuning parameter λ_j^* when we use a "perfect" predictor that returns $f^*(X_{ij}) = Y_{ij}$), we see that

$$\frac{|\mathrm{Cov}_{\eta_j}[X_{ij}, Y_{ij}]|}{\mathrm{Var}_{\eta_j}[Y_{ij}]} = \frac{2|\eta_j|\psi^2}{4\eta_j^2\psi^2 + c} = \left(2\eta_j + \frac{c}{2\eta_j\psi^2}\right)^{-1},$$

which has its minimum when $\eta_j = 0$ and increases as $|\eta_j| \to 1$ with our chosen ψ and c. In other words, problems with covariates near the origin has a lower predictability, while as we move away from zero the problems appear "easier to solve". This quantitatively reflects the pattern we see in Figure 3, where we display the power-tuning parameters as a function of η_j.

Identifying $\mu_j, \tau_j^2, \gamma_j$ with $f(x) = |x|$. When we work with the synthetic model using the "flawed" predictor $f(x) = |x|$, we can match the form of our dataset with the general setting in Assumption 2.1 by identifying closed-form expressions for the model parameters $\mu_j, \tau_j^2, \gamma_j$.

$$\gamma_j = 2\eta_j\psi^2\sqrt{\frac{2}{\pi}}e^{-\eta_j^2/(2\psi^2)}, \quad \mu_j = \sqrt{\frac{2}{\pi}}\psi\exp\left(-\frac{\eta_j}{2\psi^2}\right) + \eta_j\left[\Phi\left(\frac{\eta_j}{\psi}\right) - \frac{1}{2}\right],$$
$$\tau_j^2 = \eta_j^2 + \psi^2 - \left[\sqrt{\frac{2\psi^2}{\pi}}\exp\left(-\frac{\eta_j^2}{2\psi^2}\right) + \eta_j\left(2\Phi\left(\frac{\eta_j}{\psi}\right) - 1\right)\right]^2$$

where $\Phi(\cdot)$ denotes the standard normal CDF. In experiments involving the synthetic model, we supplement these ground truth second-moment parameters together with the datasets.

(Note: similar and even simpler analytical expressions can be carried out with $f(x) = x^2$. For more complicated predictors, since the data generation process is controlled in the synthetic setting, we can take on a Monte Carlo approach: sample much more data points than N_j or n_j and estimate these parameters from the predictions on them.)

Interpretation of MSE. In the synthetic experiments, since we have accessed to the true prior for η_j (therefore for θ_j) and resample them for each problem across K trials, the MSE we obtained in Table 2 is an unbiased estimate of the *Bayes Risk* defined in Eq. (11).

C.2 Amazon Review Ratings Dataset

Dataset & Preprocessing. The *Amazon Fine Food Reviews* dataset, provided by the Stanford Network Analysis Project (SNAP; SNAP [2014]) on Kaggle,[9] comes in a clean format. We group reviews by their `ProductID`. For each review, we concatenate the title and body text to form the covariate, while the response is the reviewer's score/rating (1 to 5 stars). Here's a sample review:

> **Score:** 4 **Product:** BBQ Pop Chips
> **Title:** Delicious!
> **Text:** BBQ Pop Chips are a delicious tasting healthier chip than many on the market. They are light and full of flavor. The 3 oz bags are a great size to have. I would recommend them to anyone.

We focus on the top $m = 200$ products with the most reviews for our compound mean estimation of average ratings. This approach mitigates extreme heteroscedasticity across estimators for different problems, which could unduly favor shrinkage-based methods when considering unweighted compound risk. There are a total of 74,913 reviews for all 200 products.

Fine-tuning `BERT`. The Bidirectional Encoder Representations from Transformers (`BERT`) model is a state-of-the-art language model for many NLP tasks including text classification [Devlin, 2018]. However, pretraining `BERT` from scratch is time-consuming and require large amounts of data (than only product reviews). We thus decide to use the `bert-base-multilingual-uncased-sentiment` model[10] from NLP Town [2023] as the base model, denoted as `BERT-base`. `BERT-base` is pre-trained on general product reviews (not exclusive to Amazon) in six languages. It achieves 67.5% prediction accuracy on a validation set of 100 products (∼46k reviews).

Then, we further fine-tune it on the held-out review data, i.e. reviews outside the top 200 products, for 2 full epochs. The fine-tuning is done using Hugging Face's `transformers` library [Wolf, 2019]. After fine-tuning, the `BERT-tuned` model achieves 78.8% accuracy on the same validation set.

C.3 Spiral Galaxy Fractions (Galaxy Zoo 2)

Dataset & Preprocessing. The *Galaxy Zoo 2* (GZ2) project[11] contains a huge collection of human-annotated classification results for galaxy images from SDSS. However, instead of having a single dataframe, GZ2 has many different tables—each for some subsets of the SDSS raw data. We begin

[9] https://www.kaggle.com/datasets/snap/amazon-fine-food-reviews
[10] https://huggingface.co/nlptown/bert-base-multilingual-uncased-sentiment
[11] https://data.galaxyzoo.org/

with a particular subset of 239,696 images with metadata drawn from Hart et al. [2016]. Our data cleaning pipeline is inspired by Lin et al. [2021], which removes missing data and relabels the class name of each galaxy image to a more readable format:

Class Names: Round Elliptical, In-between Elliptical, Cigar-shaped Elliptical, Edge-on Spiral, Barred Spiral, Unbarred Spiral, Irregular, Merger

In the downstream estimation problems, we consider a galaxy "spiral" if it is classified as one of the three classes ending with "Spiral", otherwise "non-spiral". Below we display a few examples of galaxy images. Each of the image has dimension $424 \times 424 \times 3$, where the last dimension corresponds to the g, r, i filter channels. The cleaned dataset has 155,951 images.

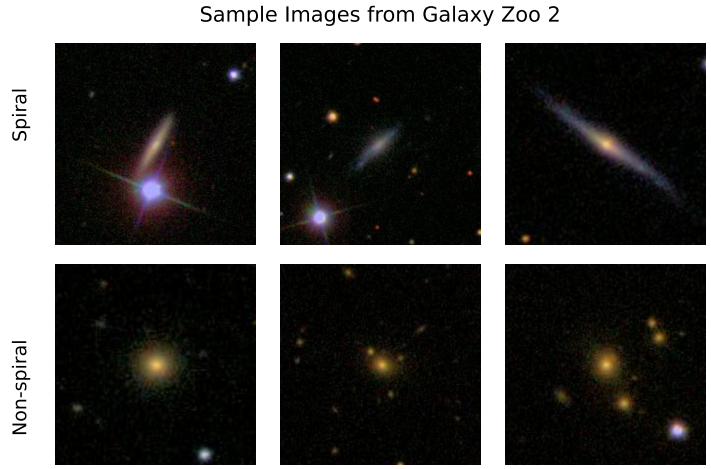

Figure 4: Example of spiral & non-spiral galaxy images from Galaxy Zoo 2.

The additional SDSS metadata for GZ2[12] contains valuable information that directly partitions the galaxies based on a certain attribute, e.g., `REDSHIFT_SIMPLE_BIN` for binning with redshift info, and `WVT_BIN` is the bin number for weighted Voronoi tessellation. These partitions naturally motivate fine-grained compound mean estimation on this dataset.

After partitioning the images based on `REDSHIFT_SIMPLE_BIN`, we consider only the top $m = 122$ partitions based on the cutoff that each problem should have ≥ 150 images (in fact, many partitions have very few galaxy images in them), for the same reason as in the *Amazon Review* dataset. Finally, we have a total of ∼100k images as covariates (either X_{ij} or $\tilde{X}_{ij}$) for our problem.

Training the Predictor. We employ the `ResNet50` architecture [He et al., 2016], utilizing the pre-trained model from `torchvision` initially trained on ImageNet [Deng et al., 2009]. To tailor the model to our task, we fine-tune it on ∼50k images excluded from the top m problems. The model is trained to classify galaxies into eight categories, later condensed into a binary spiral/non-spiral classification for prediction. We use a batch size of 256 and Adam optimizer [Kingma, 2014] with a learning rate of 1e-3. After 20 epochs, the model achieves 87% training accuracy and 83% test accuracy. Despite these promising results, Table 3 indicates that the predictions still require de-biasing for accurate estimation.

C.4 Benchmarking in real-world datasets

We further detail the steps to obtain the MSEs and their standard errors for real-world datasets shown in Table 3.

Let K be the number of experiment trials, T_j be the total number of data points for problem j, i.e. $\{\dot{X}_{ij}, \dot{Y}_{ij}\}_{j=1}^{T_j}$ represents the "raw data" we have, and n_j, N_j be the desired number of labeled/unlabeled data to simulate, usually calculated through a hyperparameter splitting ratio (e.g. $N_j = \lfloor r \cdot T_j \rfloor$, $n_j = T_j - N_j$ for $r = 0.8$ in our case).

[12] column names and their meanings is available at `https://data.galaxyzoo.org/data/gz2/gz2sample.txt`.

1. Following evaluation methodology in existing PPI literature, e.g., [Angelopoulos et al., 2023], we first calculate the mean of all responses for each problem and treat it as the pseudo ground-truth, i.e., $\dot{\theta}_j := \frac{1}{T_j} \sum_i \dot{Y}_{ij}$.

2. For each trial $k \in [K]$, we create a random permutation for the raw data, with indices permuted by $\kappa : \mathbb{N} \to \mathbb{N}$, and obtain the labeled and unlabeled datasets for problem j as

$$\{X_{ij}, Y_{ij}\}_{i=1}^{n_j} = \{\dot{X}_{\kappa(i)j}, \dot{Y}_{\kappa(i)j}\}_{i=1}^{n_j}, \quad \{\tilde{X}_{ij}\}_{i=1}^{N_j} = \{\dot{X}_{\kappa(i)j}\}_{i=n_j+1}^{T_j}$$

3. We proceed with using these datasets to obtain the baseline and `PAS` estimators. Let $\hat{\theta}_j^k$ be an estimator for the j-th problem at trial k, then our final reported MSE and standard error is calculated as

$$\widehat{\mathrm{MSE}}_K(\hat{\boldsymbol{\theta}}) := \frac{1}{K} \sum_{k=1}^{K} \left(\frac{1}{m} \sum_{j=1}^{m} (\hat{\theta}_j^k - \dot{\theta}_j)^2 \right),$$

$$\mathrm{SE}_K(\hat{\boldsymbol{\theta}}) := \frac{1}{\sqrt{K}} \sqrt{\frac{1}{K-1} \sum_{k=1}^{K} \left(\frac{1}{m} \sum_{j=1}^{m} (\hat{\theta}_j^k - \dot{\theta}_j)^2 - \widehat{\mathrm{MSE}}_K(\hat{\theta}) \right)}$$

It should be noted that the standard error only accounts for uncertainty due to the random splits into labeled and unlabeled datasets.

C.5 Computational Resources

All the experiments were conducted on a compute cluster with Intel Xeon Silver 4514Y (16 cores) CPU, Nvidia A100 (80GB) GPU, and 64GB of memory. Fine-tuning the `BERT-tuned` model took 2 hours, and training the `ResNet50` model took 1 hour. All the inferences (predictions) can be done within 10 minutes. The nature of our research problem only requires running the prediction once (per dataset), so benchmarking all the estimators with existing predictions for $K = 200$ trials is extremely fast with our provided datasets.

C.6 Code Availability

The code for reproducing the experiments is anonymized and available at https://anonymous.4open.science/r/prediction-powered-adaptive-shrinkage.