Prediction-Augmented Trees for Reliable Statistical Inference

Vikram Kher[*] [1], Argyris Oikonomou[†], and Manolis Zampetakis [‡] [1]

[1]Department of Computer Science, Yale University

Abstract

Machine learning (ML) is increasingly central to scientific discovery, but translating ML predictions into rigorous scientific claims still requires statistical inference on gold-standard labeled data. A setting that has emerged across the sciences provides the analyst with a small labeled sample, a much larger unlabeled sample, and a pre-trained predictor, with the goal of constructing a confidence interval for a quantity of interest, as was formalized in Angelopoulos et al. [4]. In this ML-driven scientific discovery literature, labeled data is typically the binding constraint on precision while computation is comparatively unconstrained, motivating the design question of how additional computation can extract more inferential value from a fixed labeled set, a fixed unlabeled set, and a fixed predictor.

To address this question, we propose the *Prediction-Augmented Residual Tree* (PART) estimator. PART takes the same inputs as the original PPI estimator and replaces its single global rectifier with locally computed rectifiers obtained from an adaptive partition of the feature space. We show that PART outperforms existing methods by producing tighter confidence intervals across real-world datasets from ecology, astronomy, and census reports, among other domains. We then describe and analyze PAQ, an estimator based on numerical quadrature techniques. Under appropriate assumptions on the input data, we show that the variance of PAQ shrinks at a rate of $O(N^{-1} + n^{-4})$, showing that there are settings where the rate of $O(N^{-1} + n^{-1})$ of existing methods can be provably broken.

1 Introduction

The integration of machine learning (ML) into scientific discovery has fundamentally transformed how researchers across disciplines generate knowledge. From predicting protein structures [31] to discovering new materials [39] to mapping poverty from satellite imagery [29], ML models now serve as the primary engines of scientific inference across the biological [20, 36, 42, 51, 53], physical [15, 34, 50], and social sciences [16, 32, 48]. The 2024 Nobel Prize in Chemistry, awarded in part for AlphaFold [31, 44], cemented what the scientific community already knew: ML is not a peripheral tool but a central pillar of modern scientific discovery.

Translating ML predictions into rigorous scientific claims, however, still requires statistical inference, and statistical inference still requires gold-standard labeled data. One fundamental setting that has emerged across these disciplines is the following: a researcher has access to a small labeled set $\mathcal{L} = \{(X_i, Y_i)\}_{i=1}^{n}$ in which X_i is a feature vector that is inexpensive to obtain and Y_i is a quantity of interest that is expensive to measure; a much larger unlabeled set $\mathcal{U} = \{\widetilde{X}_j\}_{j=1}^{N}$ with $N \gg n$; and a pre-trained black-box predictor f, such as AlphaFold. The objective is to combine

[*]vikram[dot]kher[at]yale.edu

[†]ar[dot]economou[at]outlook.com

[‡]manolis[dot]zampetakis[at]yale.edu

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

these inputs into a confidence interval for a target functional of the joint distribution $P_{X,Y}$. This function can be the population mean, a regression coefficient, or an odds ratio. The goal is to obtain confidence intervals that are meaningfully tighter than the intervals obtained from the labeled data $\mathcal{L}$ alone. Angelopoulos et al. [4] formalized this protocol as *Prediction-Powered Inference* (PPI), and Angelopoulos et al. [5] introduced the refinement PPI++.

The resources needed to perform such statistical inference have very different costs. A single labeled point (X_i, Y_i) is very expensive and typically requires physical experimentation, dedicated data-collection effort, or expert annotation; an unlabeled covariate $\widetilde{X}_j$ is generally obtainable at a small fraction of this cost; and the predictor f is taken as exogenous, with its training cost amortized across the broader research community that consumes its outputs; the computation time needed to compute the confidence interval is negligible compared to the rest. Throughout ML-driven scientific discovery, the labeled-sample size n is therefore the binding constraint on inferential precision, while computation, measured on the same scale, is comparatively unconstrained. Tightening confidence intervals by enlarging the labeled set $\mathcal{L}$ requires new experimental effort and is rarely under the analyst's control; tightening them by spending additional computation on a fixed $\mathcal{L}$ is cheap by comparison. The natural design question is therefore the following, which is the main motivation of this paper.

Question: *How can we use additional computation to extract more inferential value from the same labeled data, the same unlabeled data, and the same predictor?*

On its face, the existing PPI/PPI++ pipeline does not admit a natural place to invest additional computation. Both estimators take the form "predictor plug-in plus residual rectifier," where the rectifier is a single global empirical mean of the form $\frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{\lambda} f(X_i))$, computed in $O(n)$ time. The global rectifier is, however, also the source of the largest unexploited statistical efficiency in these estimators. In essentially every scientific setting in which they are deployed, the predictor f is non-uniformly accurate across the feature space: medical imaging models systematically underperform on rare subgroups while preserving high aggregate accuracy [43]; covariate shift between training and deployment concentrates errors in feature-dependent regions [33]; structural biology predictors are known to vary in reliability across structural classes [31]. A single global rectifier averages over this heterogeneity and discards the associated signal. Recovering it requires a finer-grained rectifier, the construction of which is the natural site for additional computational investment.

1.1 Our Contributions

We propose the *Prediction-Augmented Residual Tree* (PART) estimator, which leverages additional computation to reduce the variance of PPI/PPI++ without altering their input requirements. PART constructs a CART-style binary tree over the feature space whose splits are chosen greedily to minimize the variance contribution of the rectifier, and replaces PPI++'s single global rectifier with a leaf-weighted mixture of locally computed rectifiers. The construction takes the same inputs $(\mathcal{L}, \mathcal{U}, f)$ as PPI/PPI++ and introduces no auxiliary trained model. For mean estimation, setting the tree depth to zero recovers PPI++ exactly. Under our regularity conditions, the PART mean estimator has asymptotic variance no larger than PPI++, with strict improvement when the limiting local rectifiers differ across leaves (Corollary 1).

Empirical Evaluation. We evaluate PART on six datasets that have become standard benchmarks in the prediction-powered-inference literature, drawn from ecology (deforestation rates from satellite imagery; 14), astronomy (spiral-galaxy fractions from Galaxy Zoo 2; 57), socioeconomic statistics (median house values from California census data 46, and an income-by-sex regression on ACS data 22), food science (Vinho Verde wine quality; 19), and structural biology (phosphorylation odds ratios from AlphaFold predictions on UniProt; 10, 55). Across all six benchmarks, PART produces strictly tighter confidence intervals than PPI++ while maintaining nominal coverage, with width reductions in the range of 4–20% depending on the task and the labeled-sample size. These reductions are most informatively expressed in labeled-sample-equivalent terms: because the asymptotic variance scales as n^{-1}, a width reduction of 10–20% corresponds to an effective increase in labeled-sample size of 23–56%. In application domains in which labeled samples are obtained through experimental procedures whose marginal cost is high, a gain of this magnitude obtained solely from additional computation is very valuable.

Theoretical Results. We establish asymptotic normality and asymptotically valid confidence intervals for the PART mean and linear regression estimators (Theorem 1 and Theorem 3) under regularity assumptions (Assumptions 1 and 3, respectively). Our results allow the same labeled observations to be used both to learn the tree and to construct confidence intervals, without sample splitting. We account for this adaptive selection by showing that, under the stated regularity conditions, the fitted leaf regions converge to those of a deterministic population tree.

We complement PART with a companion estimator, *Prediction-Augmented Quadrature* (PAQ), which exploits smooth, deterministic residuals to achieve faster convergence. The estimator approximates the debiasing term by interpolating labeled residuals and averaging the interpolant over the unlabeled sample, giving a numerical quadrature interpretation. Under Assumption 2, Theorem 2 establishes a variance bound of $O(N^{-1} + n^{-4})$ for linear PAQ. When $N \gg n$, this dramatically improves upon the usual $O(N^{-1} + n^{-1})$ variance bound of PPI and PPI++.

2 Preliminaries

In this section, we introduce the statistical framework and model that we use to present our theoretical results on PART and PAQ estimators.

Empirical Expectation and Variance. Let $\{Z_i\}_{i=1}^n$ be a sequence of independently and identically distributed random variables (i.i.d.) with mean $\mu = \mathbb{E}[Z]$ and variance $\sigma^2 = \mathrm{Var}[Z] = \mathbb{E}[(Z-\mu)^2]$. The *empirical mean* is defined as $\widehat{\mu}_n = \frac{1}{n}\sum_{i=1}^n Z_i$, and the *empirical variance* is defined as $\widehat{\sigma}_n^2 = \frac{1}{n-1}\sum_{i=1}^n (Z_i - \widehat{\mu}_n)^2$. It is easy to see that $\mathbb{E}[\widehat{\mu}_n] = \mu$ and $\mathbb{E}[\widehat{\sigma}_n^2] = \sigma^2$, i.e., $\widehat{\mu}_n$ and $\widehat{\sigma}_n^2$ are *unbiased* estimators of the mean and variance respectively. Lastly, for an integer $k \geq 0$, we let $C^k([a,b])$ denote the set of functions with continuous derivatives through order k on the interval $[a,b]$.

Convergence in Distribution. A sequence of random variables $\{Z_n\}_{n=1}^\infty$ is said to converge in distribution to a random variable Y (denoted by $Z_n \xrightarrow{d} Y$) if, for every real number x at which the *cumulative distribution function* (CDF) $F_Y(x)$ is continuous, the CDF $F_{Z_n}(x)$ satisfies $\lim_{n\to\infty} F_{Z_n}(x) = F_Y(x)$. For a cumulative distribution function F_Y, we define the quantile at level $u \in (0,1)$ by $F_Y^{-1}(u) := \inf\{t \in \mathbb{R} : F_Y(t) \geq u\}$. We write $Z_n = O_p(1)$ if Z_n is bounded in probability, and $Z_n = o_p(1)$ if Z_n converges to zero in probability.

Central Limit Theorem (CLT) and Confidence Intervals. The Central Limit Theorem states that for i.i.d. random variables $\{Z_i\}_{i=1}^n$ with finite variance σ^2, the normalized empirical mean converges in distribution to a normal distribution, meaning that $\sqrt{n}\,(\widehat{\mu}_n - \mu) \xrightarrow{d} \mathcal{N}(0, \sigma^2)$.

2.1 Model

Our datasets include two sets of data observations: *labeled* and *unlabeled* observations.

- The *labeled dataset* $\mathcal{L}$ consists of n independent samples $\{(x_i, y_i)\}_{i=1}^n$ drawn from the joint distribution $P_{X,Y}$. Here, $x_i \in \mathcal{X} \subseteq \mathbb{R}^d$ is a d-dimensional feature vector, and $y_i \in \mathbb{R}$ is the corresponding outcome.
- The *unlabeled dataset* $\mathcal{U}$ contains N feature vectors $\{\widetilde{x}_j\}_{j=1}^N$ drawn from P_X which is the marginal distribution on X of $P_{X,Y}$. We assume that $N \gg n$, reflecting the practical scenario where unlabeled feature vectors are considerably easier to obtain than labels.

We further assume access to a machine learning predictor $f : \mathcal{X} \to \mathbb{R}$ that produces predicted outcomes for given feature vectors. We treat f as fixed and independent of the labeled and unlabeled sample. Finally, for any $x \in \mathbb{R}^d$, we denote the k-th coordinate by $x^{(k)}$ for $k = 1, \ldots, d$.

Mean Estimation and Linear Regression. In this work, we address two fundamental estimation problems: mean estimation and linear regression. For mean estimation, our objective is to accurately recover the expected value of the response, $\mu_Y = \mathbb{E}[Y]$, from the observed data. In the linear regression setting, we seek a linear model that minimizes the expected squared prediction error. Formally, we aim to find a parameter vector $\theta^* \in \mathbb{R}^d$ such that

$$\theta^* \in \arg\min_{\theta \in \mathbb{R}^d} \mathbb{E}\left[\left(Y - \theta^\top X\right)^2\right],$$

where $X \in \mathbb{R}^d$ denotes the feature vector. These problems underpin our development of semi-supervised estimators that leverage both labeled and abundant unlabeled data to enhance estimation accuracy.

2.2 PPI & PPI++: The Case of the Mean

We review several competing approaches for prediction-powered inference in mean estimation, highlighting the key insights and underlying intuition behind each method.

Empirical Estimator. A common estimation problem is to estimate the mean outcome $\mu = \mathbb{E}[Y]$ based on the labeled data. The empirical estimator is given by $\widehat{\mu}_E = \frac{1}{n}\sum_{i=1}^{n} y_i$, which satisfies

$$\sqrt{n}(\widehat{\mu}_E - \mu) \xrightarrow{d} \mathcal{N}\left(0, \mathrm{Var}(Y)\right).$$

PPI Estimator. Building on this approach, the PPI estimator of Angelopoulos et al. [4] integrates the predictive model $f(x)$ with the unlabeled data. It is defined as

$$\widehat{\mu}_{\text{PPI}} = \frac{1}{N}\sum_{j=1}^{N} f(\widetilde{x}_j) + \frac{1}{n}\sum_{i=1}^{n}\left[y_i - f(x_i)\right].$$

The first term leverages predictions on the large unlabeled dataset, while the second term corrects for residual errors observed in the labeled data. Applying the CLT, the PPI estimator satisfies

$$\sqrt{n}(\widehat{\mu}_{\text{PPI}} - \mu) \xrightarrow{d} \mathcal{N}\left(0, \mathrm{Var}\left(Y - f(X)\right)\right).$$

Assuming that $N \gg n$ and that the predictor $f(x)$ is sufficiently accurate (i.e., $\mathrm{Var}\left(Y - f(X)\right) < \mathrm{Var}(Y)$), the PPI estimator yields a narrower confidence interval for μ than the empirical estimator while preserving the same coverage probability.

PPI++ Estimator. Angelopoulos et al. [5] introduced a tunable version of PPI defined by

$$\widehat{\mu}_{\text{PPI++}} = \frac{1}{N}\sum_{j=1}^{N} \widehat{\lambda}\cdot f(\widetilde{x}_j) + \frac{1}{n}\sum_{i=1}^{n}\left[y_i - \widehat{\lambda}\cdot f(x_i)\right],$$

where $\widehat{\lambda} = \frac{\widehat{\mathrm{Cov}}_n(Y, f(X))}{(1+n/N)\widehat{\mathrm{Var}}_{n+N}(f(X_i))}$. Like PPI, this estimator remains consistent and satisfies

$$\sqrt{n}(\widehat{\mu}_{\text{PPI++}} - \mu) \xrightarrow{d} \mathcal{N}\left(0, \mathrm{Var}\left(Y - \lambda^* \cdot f(X)\right)\right),$$

where $\lambda^* = \frac{\mathrm{Cov}(Y, f(X))}{\mathrm{Var}(f(X))}$. Unlike PPI, it is straightforward to show that this method produces confidence intervals that are asymptotically at least as small as the empirical estimator.

3 Our Estimators

We are now ready to describe the methodology and the coverage of the PART and PAQ estimators.

3.1 The PART Estimator: *An Adaptive Partitioning Estimator*

In this section, we describe the PART estimator and discuss its application to building confidence intervals for the mean. In Appendix F.3, we adapt PART to estimate linear regression coefficients. At a high level, PART uses a CART-style binary tree to partition the feature space, with splits chosen to greedily minimize an estimate of the variance of the debiasing term. Inside each leaf of the tree, we compute a local residual correction using labeled points in the leaf. We then produce a final debiasing term by taking a mixture of these local corrections based on the fraction of unlabeled data assigned to each leaf. In Appendix E, we provide a simple motivating example to illustrate how partitioning the residuals can reduce the variance of the debiasing term. Below, we give a short outline for how to build the PART estimator and construct valid confidence intervals.

PART Estimator: Construction and Inference

1. Data, Coefficient, and Residuals. Given labeled data $\mathcal{L} = \{(x_i, y_i)\}_{i=1}^n$, unlabeled data $\mathcal{U} = \{\widetilde{x}_j\}_{j=1}^N$ (with $N \gg n$), and a predictor $f : \mathcal{X} \to \mathbb{R}$, we first estimate the coefficient

$$\hat{\lambda} \;\coloneqq\; \frac{\widehat{\mathrm{Cov}}_n\big(Y, f(X)\big)}{\big(1 + n/N\big)\,\widehat{\mathrm{Var}}_{n+N}\big(f(X)\big)},$$

and define the $\hat{\lambda}$-*residuals*

$$r_i^{(\hat{\lambda})} \;\coloneqq\; y_i - \hat{\lambda}\, f(x_i), \qquad i = 1, \ldots, n.$$

2. Adaptive Partitioning. We recursively partition the feature space $\mathcal{X}$ into disjoint axis-aligned regions (leaves) $\{\mathcal{R}_\ell\}_{\ell=1}^L$. Fix a maximum depth D, a minimum number of labeled observations per child $m \geq 2$, and quantile levels $0 < u_1 < \cdots < u_K < 1$, with $K \geq 2$, independently of n and N. For each coordinate k, compute the empirical quantiles $q_1^{(k)}, \ldots, q_K^{(k)}$ at these levels from the full unlabeled sample, and define the candidate thresholds by

$$\mathcal{S}^{(k)} \coloneqq \{q_j^{(k)} : j = 1, \ldots, K\}.$$

We remove duplicate thresholds within each coordinate. These sets are computed once and used at every node. At a node with region $\mathcal{R}$, a candidate (k, s), with $s \in \mathcal{S}^{(k)}$, produces the children

$$\mathcal{R}_L(k, s) \coloneqq \{x \in \mathcal{R} : x^{(k)} \leq s\}, \qquad \mathcal{R}_R(k, s) \coloneqq \{x \in \mathcal{R} : x^{(k)} > s\}.$$

Let $n_{(k,s,\mathrm{L})}$ and $n_{(k,s,\mathrm{R})}$ be the labeled counts in these children, and let $\widehat{\sigma}^2_{(k,s,\mathrm{L})}$ and $\widehat{\sigma}^2_{(k,s,\mathrm{R})}$ be their sample variances of the $\hat{\lambda}$-residuals. Let $p_{(k,s,\mathrm{L})}$ and $p_{(k,s,\mathrm{R})}$ be the corresponding proportions of the full unlabeled sample. Among candidates with at least m labeled observations in each child, minimize the *Variance-Minimizing Split* (VMS) criterion

$$\mathrm{VMS}(k, s) \coloneqq p^2_{(k,s,\mathrm{L})} \frac{\widehat{\sigma}^2_{(k,s,\mathrm{L})}}{n_{(k,s,\mathrm{L})}} + p^2_{(k,s,\mathrm{R})} \frac{\widehat{\sigma}^2_{(k,s,\mathrm{R})}}{n_{(k,s,\mathrm{R})}},$$

using a fixed deterministic rule to break ties. Recurse on both children, stopping at depth D or when no admissible candidate remains. Let T denote the resulting data-dependent tree, with leaves $\{\mathcal{R}_\ell\}_{\ell=1}^L$ and $L \leq 2^D$.

3. Estimation and Inference. Let n_ℓ be the labeled count in leaf ℓ, $\overline{r}_\ell^{(\hat{\lambda})}$ its mean $\hat{\lambda}$-residual and p_ℓ the unlabeled mass proportion in leaf ℓ. Let $\widehat{\sigma}^2_{\ell,\hat{\lambda}}$ be the sample variance of the $\hat{\lambda}$-residuals in leaf ℓ. The PART estimator $\hat{\mu}_{T,\hat{\lambda}}$ and its variance estimate $\widehat{\sigma}^2_{T,\hat{\lambda}}$ are defined as:

$$\hat{\mu}_{T,\hat{\lambda}} \coloneqq \underbrace{\frac{1}{N}\sum_{j=1}^N \hat{\lambda}\, f(\widetilde{x}_j)}_{\text{Base Prediction}} + \underbrace{\sum_{\ell=1}^L p_\ell\, \overline{r}_\ell^{(\hat{\lambda})}}_{\text{Bias Correction}}, \qquad \widehat{\sigma}^2_{T,\hat{\lambda}} \coloneqq \sum_{\ell=1}^L p_\ell^2\, \frac{\widehat{\sigma}^2_{\ell,\hat{\lambda}}}{n_\ell}.$$

The $(1 - \alpha)$ Wald confidence interval is given by $\hat{\mu}_{T,\hat{\lambda}} \pm z_{1-\alpha/2}\, \widehat{\sigma}_{T,\hat{\lambda}}$, where $z_{1-\alpha/2}$ is the $1 - \alpha/2$ standard normal quantile. We prove coverage guarantees for this interval in Theorem 1 and formalize the procedure in Algorithm 1.

Remark. When the tree depth is zero, there is a single leaf $\mathcal{R}_1 = \mathcal{X}$, so $p_1 = 1$ and $\overline{r}_1^{(\hat{\lambda})} = \frac{1}{n}\sum_{i=1}^n (y_i - \hat{\lambda} f(x_i))$. In this case, $\hat{\mu}_{T,\hat{\lambda}}$ reduces exactly to the PPI++ estimator.

We now establish the coverage result. Let $p(\mathcal{R}) = P_X(\mathcal{R})$ be the population probability of region $\mathcal{R}$. The population PPI++ coefficient and its residual are

$$\lambda^* := \frac{\mathrm{Cov}(Y, f(X))}{\mathrm{Var}(f(X))}, \qquad r^{(\lambda^*)}(x, y) := y - \lambda^* f(x).$$

Let T_0 be the population greedy tree obtained by using the corresponding population quantiles as candidate thresholds, admitting splits with two positive-probability children, and minimizing the population VMS criterion

$$\sum_{a \in \{L,R\}} p(\mathcal{R}_a(k,s)) \, \mathrm{Var}\Big(r^{(\lambda^*)}(X,Y) \mid X \in \mathcal{R}_a(k,s)\Big)$$

at each node. The population procedure uses the same quantile levels and depth limit as its empirical counterpart and stops when no admissible split remains. Under the regularity conditions below, the fitted leaf regions converge to the leaves of T_0.

Assumption 1 (Regularity for PART)**.** *The dimension d, maximum depth D, minimum child size $m \geq 2$, and finite grid of candidate quantile levels are fixed. Assume:*

(i) The labeled observations are i.i.d. from $P_{X,Y}$, the unlabeled observations are i.i.d. from P_X, and the two samples are independent. As $n, N \to \infty$, $n/N \to 0$. Furthermore, $\mathbb{E}[Y^2 + f(X)^2] < \infty$ and $\mathrm{Var}(f(X)) > 0$.

(ii) Whenever the population procedure evaluates candidate splits at a node, let $\mathcal{R}_{0,L}$ and $\mathcal{R}_{0,R}$ denote the children of a candidate split. Define their empirical counterparts $\mathcal{R}_L$ and $\mathcal{R}_R$ by replacing every population quantile threshold defining these regions with its unlabeled empirical quantile at the same level and coordinate. For every such correspondence, let n_a be the labeled count in $\mathcal{R}_a$, and assume

$$P_X(\mathcal{R}_a \triangle \mathcal{R}_{0,a}) \xrightarrow{p} 0, \qquad a \in \{L, R\},$$

$$\Pr\Big[\min_{a \in \{L,R\}} n_a \geq m\Big] \longrightarrow \mathbf{1}\Big\{\min_{a \in \{L,R\}} p(\mathcal{R}_{0,a}) > 0\Big\},$$

where $\triangle$ denotes the symmetric difference between two sets.

(iii) At every splitting node of T_0, the population splitting criterion has a unique minimizer among the feasible candidates. Moreover,

$$v_0 := \sum_{\mathcal{R} \text{ a leaf of } T_0} p(\mathcal{R}) \; \mathrm{Var}\Big(r^{(\lambda^*)}(X,Y) \mid X \in \mathcal{R}\Big) > 0.$$

Here v_0 is the probability-weighted within-leaf residual variance of the population tree.

The sampling and moment conditions in *(i)* are standard for asymptotic mean inference. The regime $n/N \to 0$ makes unlabeled sampling uncertainty negligible at the $\sqrt{n}$ scale. The first part of *(ii)* requires empirical and population candidate regions to disagree on the membership of a new feature vector with probability tending to zero. The second requires the minimum-child-size rule to agree with population feasibility. For the candidate thresholds defined above, condition *(ii)* holds if the feature distribution has a density positive almost everywhere on a rectangular support. Alternatively, it holds for continuous marginal distributions under the stronger regime $n/\sqrt{N} \to 0$. These sufficient conditions are justified in Appendix F.1. Together with consistency of the split scores, the unique-minimizer requirement in *(iii)* makes the greedy split choices stabilize. Finally, $v_0 > 0$ requires residual variation within at least one population leaf, ensuring a nondegenerate limiting normal distribution.

Theorem 1 (Asymptotic coverage of PART)**.** *Under Assumption 1, the data-dependent tree T, coefficient $\hat{\lambda}$, and estimates constructed above satisfy*

$$\sqrt{n}\big(\hat{\mu}_{T,\hat{\lambda}} - \mu\big) \xrightarrow{d} \mathcal{N}(0, v_0), \qquad n\widehat{\sigma}^2_{T,\hat{\lambda}} \xrightarrow{p} v_0,$$

where $\mu = \mathbb{E}[Y]$. Consequently, for every fixed $\alpha \in (0,1)$,

$$\Pr\Big[\mu \in \Big[\hat{\mu}_{T,\hat{\lambda}} - z_{1-\alpha/2}\widehat{\sigma}_{T,\hat{\lambda}}, \hat{\mu}_{T,\hat{\lambda}} + z_{1-\alpha/2}\widehat{\sigma}_{T,\hat{\lambda}}\Big]\Big] \longrightarrow 1 - \alpha.$$

It is straightforward to characterize when the PART interval is asymptotically smaller than the interval implied by PPI++.

Corollary 1 (Asymptotic variance reduction)**.** *Under Assumption 1, the PART mean estimator has asymptotic variance no larger than PPI++. In particular,*

$$v_0 \le \mathrm{Var}\Big(r^{(\lambda^*)}(X,Y)\Big).$$

The inequality is strict if and only if at least two leaves of T_0 have different mean λ^*-residuals.

Proof. The depth-zero case of Theorem 1 gives the PPI++ asymptotic variance $\mathrm{Var}(r^{(\lambda^*)}(X,Y))$. By the law of total variance,

$$\begin{aligned}&\mathrm{Var}\Big(r^{(\lambda^*)}(X,Y)\Big) - v_0 \\ &\quad= \sum_{\mathcal{R}\text{ a leaf of }T_0} p(\mathcal{R})\left(\mathbb{E}\Big[r^{(\lambda^*)}(X,Y)\mid X\in\mathcal{R}\Big] - \mathbb{E}\Big[r^{(\lambda^*)}(X,Y)\Big]\right)^2 \ge 0.\end{aligned}$$

All population leaves have positive probability, so the difference is positive exactly when their mean residuals are not all equal. □

3.2 The PAQ Estimator: *Residual Interpolation and Numerical Quadrature*

We now introduce PAQ, an estimator that can achieve faster convergence when the response variable is deterministic and the residual function is smooth. The estimator approximates the debiasing term by interpolating the labeled residuals and averaging the interpolant over the unlabeled sample. This approach is motivated by classical work on weighted Monte Carlo integration, which shows that exploiting smoothness can yield faster convergence than ordinary Monte Carlo averaging (Yakowitz et al. [59]).

Below, we consider linear interpolation when the feature distribution is uniform on $[0,1]$. In Appendix G, we extend the estimator to higher-degree polynomial interpolation, obtaining faster interpolation error rates under stronger smoothness assumptions. We also establish the corresponding results for nonuniform feature distributions satisfying additional regularity conditions.

Assumption 2 (Regularity for Linear PAQ)**.** *Assume:*

(i) The labeled observations are i.i.d. from $P_{X,Y}$, the unlabeled observations are i.i.d. from $P_X = \mathrm{Unif}[0,1]$, and the two samples are independent. Furthermore, $0 < \mathrm{Var}(Y) < \infty$.

(ii) There is a deterministic function y such that $Y = y(X)$ almost surely and the residual function $r(x) = y(x) - f(x)$ belongs to $C^2([0,1])$.

The sampling assumptions are standard for mean estimation. The assumption that the residual is a deterministic function of x is restrictive; however, nonzero conditional label noise generally imposes a variance lower bound of order n^{-1} for mean estimators, precluding the faster variance rates we seek.

PAQ Estimator: Linear Residual Interpolation

1. Residual Interpolation. Given $n \ge 3$ labeled observations and $N \ge 2$ unlabeled observations, sort the labeled features as $x_{(1)} < \cdots < x_{(n)}$ and let $r_i = y_{(i)} - f(x_{(i)})$. For $i = 1, \ldots, n-1$, define

$$s_i := \frac{r_{i+1} - r_i}{x_{(i+1)} - x_{(i)}}, \qquad q_1(x) := r_i + s_i(x - x_{(i)}), \quad x \in [x_{(i)}, x_{(i+1)}].$$

For $x < x_{(1)}$, extend the first line; for $x > x_{(n)}$, extend the last line. Thus, q_1 linearly interpolates the residuals between labeled features and extrapolates linearly at the two boundaries.

2. Estimation. Define the adjusted predictions and their average by

$$Z_j \coloneqq f(\widetilde{x}_j) + q_1(\widetilde{x}_j), \qquad \widehat{\mu}_{\mathrm{PAQ}} \coloneqq \frac{1}{N}\sum_{j=1}^{N} Z_j.$$

3. Confidence Interval. Estimate the variance due to unlabeled averaging by

$$\widehat{\sigma}^2_{\mathrm{PAQ}} \coloneqq \frac{1}{N(N-1)}\sum_{j=1}^{N}(Z_j - \widehat{\mu}_{\mathrm{PAQ}})^2.$$

To account for interpolation error, fix $\kappa > 1$, for example $\kappa = 2$, and compute

$$\widehat{L}_2 \coloneqq 2\kappa \max_{1\le i\le n-2}\left|\frac{s_{i+1} - s_i}{x_{(i+2)} - x_{(i)}}\right|$$

as an inflated estimate of $\|r''\|_\infty$. Let $t_0(x), t_1(x)$ be the two labeled feature values whose line is used to evaluate $q_1(x)$, and define

$$w_1(x) \coloneqq \frac{1}{2}\left|(x - t_0(x))(x - t_1(x))\right|, \qquad \widehat{B}_1 \coloneqq \frac{\widehat{L}_2}{N}\sum_{j=1}^{N} w_1(\widetilde{x}_j).$$

Here $w_1(x)$ captures the effect of the labeled-feature spacing on interpolation error, and $\widehat{B}_1$ is the average interpolation-error allowance over the unlabeled sample. For a fixed $\alpha \in (0,1)$, use the interval

$$C_{n,N} \coloneqq \left[\widehat{\mu}_{\mathrm{PAQ}} - z_{1-\alpha/2}\widehat{\sigma}_{\mathrm{PAQ}} - \widehat{B}_1, \quad \widehat{\mu}_{\mathrm{PAQ}} + z_{1-\alpha/2}\widehat{\sigma}_{\mathrm{PAQ}} + \widehat{B}_1\right],$$

where $z_{1-\alpha/2}$ is the $(1-\alpha/2)$ standard normal quantile. Under Assumption 2, Theorem 2 establishes asymptotic coverage of at least $1-\alpha$ for this interval. Pseudocode is provided in Algorithm 3.

Theorem 2 (Linear PAQ under uniform sampling)**.** *Under Assumption 2, as $n, N \to \infty$, with fixed $\kappa > 1$ and $\alpha \in (0,1)$, the estimator satisfies*

$$|\mathbb{E}[\widehat{\mu}_{\mathrm{PAQ}}] - \mu| = O(n^{-2}), \qquad \mathbb{E}\left[(\widehat{\mu}_{\mathrm{PAQ}} - \mu)^2\right] = O(N^{-1} + n^{-4}),$$

where $\mu = \mathbb{E}[Y]$. The same bound holds for its variance. Moreover,

$$\liminf_{n,N\to\infty} \Pr(\mu \in C_{n,N}) \ge 1 - \alpha, \qquad |C_{n,N}| = O_p(N^{-1/2} + n^{-2}),$$

where $|C_{n,N}|$ denotes the interval length.

When $n/N \to 0$, the interval length is $o_p(n^{-1/2})$, thus improving on the convergence rate of PART and PPI++. In Appendix G, we prove the result and extend it to nonuniform feature distributions and higher-degree interpolation.

4 Experimental Results with Real-World Data

In this section, we evaluate the empirical performance of PART across a diverse set of datasets, ranging from ecological monitoring to census data analysis. We apply PART to the task of mean estimation in Sections 4.1 to 4.4, to odds ratio estimation in Section 4.5, and to the estimation of linear regression coefficients in Appendix F.5. These experiments reveal a consistent pattern that PART outperforms existing estimators. We utilize PPI++ as our primary baseline and omit the classical PPI estimator, as PPI++ consistently outperforms PPI. We also compare the experimental performance of PAQ on a synthetic dataset with smooth and deterministic residuals, showing that it also outperforms PPI++. We defer these results to Appendix G.2. For each dataset, we vary the number of labeled samples n and evaluate both confidence interval width and the empirical coverage

probability over repeated sub-sampling. We include error bars representing one standard deviation on each measured n value in the interval width graph. We include detailed testing methodology and dataset descriptions in Appendix B. Tabulated data from the experiments can be found in Table 1.

4.1 Predicting Mean Deforestation Rate from Satellite Imagery

We first consider satellite imagery data capturing forest cover in the Amazon rainforest. We seek to measure the mean deforestation rate in the forest cover data provided by [14] using the PART estimator, where the predictions are provided by a histogram-based gradient boosting prediction model. Figure 1 demonstrates substantially tighter confidence intervals compared to PPI++. For example, given access to $n = 800$ labeled samples, PART produces intervals that are 9.0% smaller than PPI++ (see Table 1). We further note that this example demonstrates that trees with non-zero depth can result in smaller confidence intervals while still remaining valid.

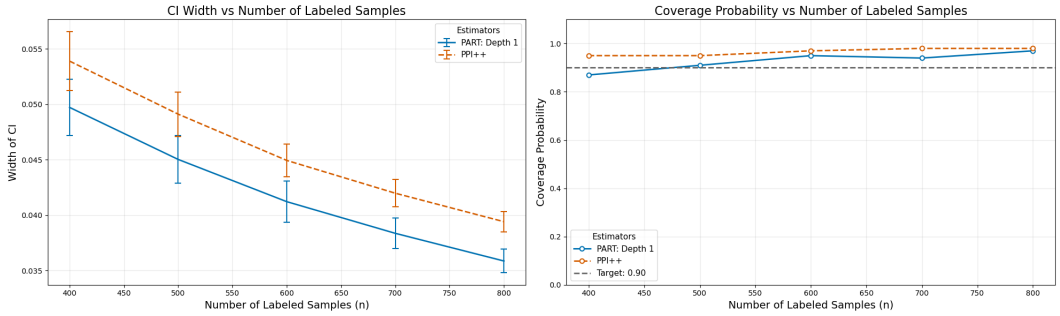


Figure 1: Estimating the Amazon's mean deforestation rate from satellite imagery [14].

4.2 Estimating Mean Fraction of Spiral Galaxies

We next evaluate PART on the Galaxy Zoo 2 dataset, which provides morphological classifications for galaxies in the Sloan Digital Sky Survey. Using auxiliary predictions from a ResNet50 model supplied by [4], PART produces substantially tighter confidence intervals for the mean fraction of spiral galaxies compared to PPI++ (Figure 2). For $n = 500$, PART yields 4.5% smaller intervals compared to PPI++ (see Table 1). As in the previous experiment, we show that increasing the depth of our decision tree beyond zero improves estimation performance.

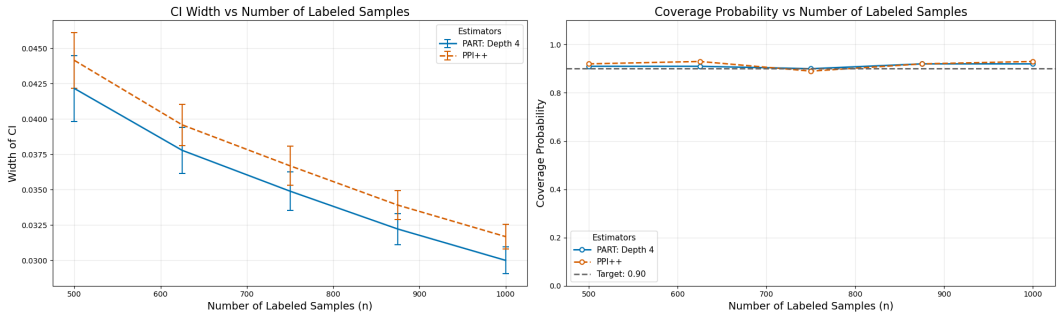


Figure 2: Estimating the mean fraction of spiral galaxies using Galaxy Zoo 2 data [57].

4.3 Measuring Mean Quality of Portuguese Wine

We next turn to the task of estimating the mean quality of Portuguese "Vinho Verde" wine based on physicochemical tests [19]. Using auxiliary predictions from a Random Forest Classifier, PART produces tighter confidence intervals for the mean quality score compared to PPI++ (Figure 3). In particular, for $n = 300$, we report 17.3% smaller intervals compared to PPI++ (see Table 1).

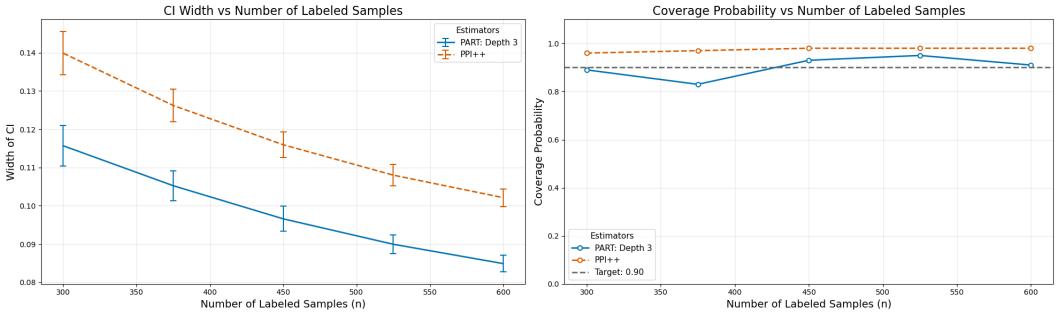


Figure 3: Estimating the mean quality of Portuguese Vinho Verde wine [19].

4.4 Estimating Average Median House Price Using Census Data

We next consider the California Housing dataset [46], a standard benchmark derived from the 1990 US Census. Our objective is to estimate the average median house value across census blocks. Leveraging predictions from a Random Forest Regressor, PART achieves significantly tighter confidence intervals than PPI++ (Figure 4). As shown in Table 1, with $n = 625$ labeled samples, PART reduces the interval width by 13.6% compared to the baseline.

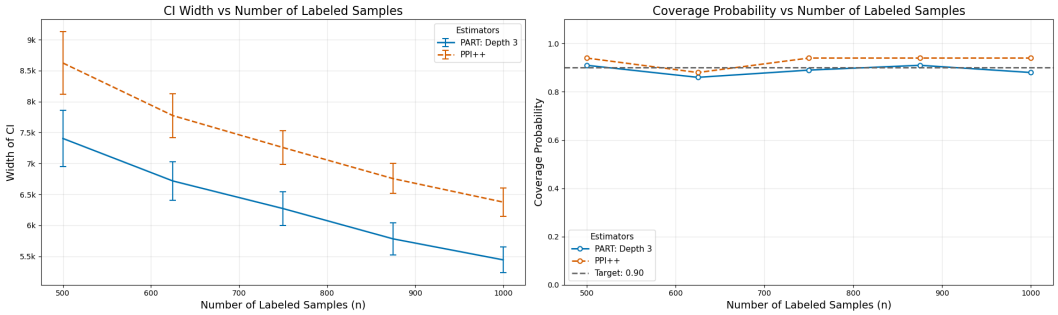


Figure 4: Estimating the average median house price using California census data [26].

4.5 Estimating Odds Ratio of Protein Structures via AlphaFold

We next apply PART to estimate the odds ratio of protein phosphorylation in intrinsically disordered regions (IDRs), a structural property recently analyzed by Bludau et al. [10]. Using AlphaFold-derived structural predictions provided by [4], we estimate this odds ratio on data from the UniProt Consortium [55]. As shown in Figure 5, PART yields tighter confidence intervals compared to the baseline[4].

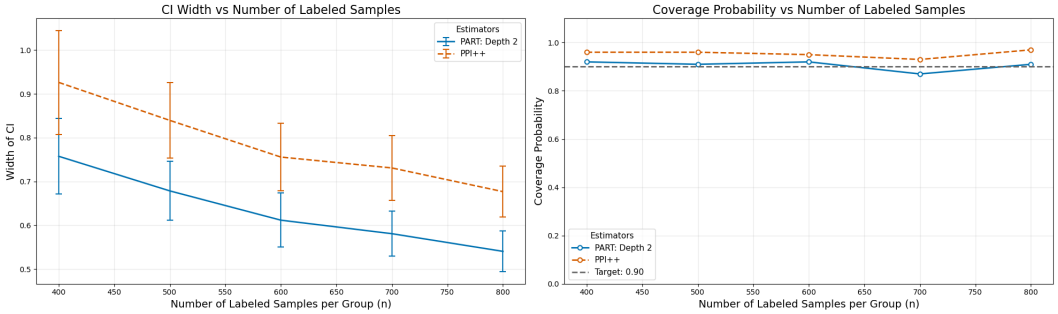


Figure 5: Estimating the odds ratio of a protein being phosphorylated and belonging to an intrinsically disordered region using AlphaFold predictions [55].

[4]For the precise derivation of the confidence interval for the odds ratio via the delta method, see [5].

References

[1] Algorithms with predictions. https://algorithms-with-predictions.github.io/, 2025. Accessed: 2025-09-15.

[2] P. Agrawal, E. Balkanski, V. Gkatzelis, T. Ou, and X. Tan. Learning-augmented mechanism design: Leveraging predictions for facility location. In *Proceedings of the 23rd ACM Conference on Economics and Computation*, EC '22, page 497–528, New York, NY, USA, 2022. Association for Computing Machinery. ISBN 9781450391504. doi: 10.1145/3490486.3538306. URL https://doi.org/10.1145/3490486.3538306.

[3] C. B. Anfinsen. Principles that govern the folding of protein chains. *Science*, 181(4096): 223–230, July 1973.

[4] A. N. Angelopoulos, S. Bates, C. Fannjiang, M. I. Jordan, and T. Zrnic. Prediction-powered inference. *Science*, 382(6671):669–674, 2023. doi: 10.1126/science.adi6000. URL https://www.science.org/doi/abs/10.1126/science.adi6000.

[5] A. N. Angelopoulos, J. C. Duchi, and T. Zrnic. PPI++: Efficient Prediction-Powered Inference. *arXiv e-prints*, art. arXiv:2311.01453, Nov. 2023. doi: 10.48550/arXiv.2311.01453.

[6] E. Balkanski, V. Gkatzelis, and G. Shahkarami. Randomized strategic facility location with predictions. In *Proceedings of the 38th International Conference on Neural Information Processing Systems*, NIPS '24, Red Hook, NY, USA, 2025. Curran Associates Inc. ISBN 9798331314385.

[7] E. Bampis, K. Dogeas, A. Kononov, G. Lucarelli, and F. Pascual. Scheduling with untrusted predictions. In L. D. Raedt, editor, *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI-22*, pages 4581–4587. International Joint Conferences on Artificial Intelligence Organization, 7 2022. doi: 10.24963/ijcai.2022/636. URL https://doi.org/10.24963/ijcai.2022/636. Main Track.

[8] B. Berger, M. Feldman, V. Gkatzelis, and X. Tan. Learning-augmented metric distortion via (p,q)-veto core. In *Proceedings of the 25th ACM Conference on Economics and Computation*, EC '24, page 984, New York, NY, USA, 2024. Association for Computing Machinery. ISBN 9798400707049. doi: 10.1145/3670865.3673456. URL https://doi.org/10.1145/3670865.3673456.

[9] F. D. Blau and L. M. Kahn. The gender wage gap: Extent, trends, and explanations. *Journal of Economic Literature*, 55(3):789–865, September 2017. doi: 10.1257/jel.20160995. URL https://www.aeaweb.org/articles?id=10.1257/jel.20160995.

[10] I. Bludau, S. Willems, W.-F. Zeng, M. T. Strauss, F. M. Hansen, M. C. Tanzer, O. Karayel, B. A. Schulman, and M. Mann. The structural context of posttranslational modifications at a proteome-wide scale. *PLoS Biol.*, 20(5):e3001636, May 2022.

[11] L. Breiman. Bagging predictors. *Mach. Learn.*, 24(2):123–140, Aug. 1996.

[12] L. Breiman. Random forests. *Mach. Learn.*, 45(1):5–32, Oct. 2001.

[13] L. Breiman, J. H. Friedman, R. A. Olshen, and C. J. Stone. *Classification And Regression Trees*. Routledge, Oct. 2017.

[14] E. L. Bullock, C. E. Woodcock, C. Souza Jr., and P. Olofsson. Satellite-based estimates reveal widespread forest degradation in the amazon. *Global Change Biology*, 26(5):2956–2969, 2020. doi: https://doi.org/10.1111/gcb.15029. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/gcb.15029.

[15] G. Carleo and M. Troyer. Solving the quantum many-body problem with artificial neural networks. *Science*, 355(6325):602–606, Feb. 2017.

[16] V. Chernozhukov, D. Chetverikov, M. Demirer, E. Duflo, C. Hansen, W. Newey, and J. Robins. Double/debiased machine learning for treatment and structural parameters. *Econom. J.*, 21(1): C1–C68, Feb. 2018.

[17] E. Cheung, E. Athanassoula, K. L. Masters, R. C. Nichol, A. Bosma, E. F. Bell, S. M. Faber, D. C. Koo, C. Lintott, T. Melvin, K. Schawinski, R. A. Skibba, and K. W. Willett. Galaxy zoo: Observing secular evolution through bars. *Astrophys. J.*, 779(2):162, Dec. 2013.

[18] P. Constantin and C. Foiaş. *Navier-stokes equations*. University of Chicago press, 1988.

[19] P. Cortez, A. Cerdeira, F. Almeida, T. Matos, and J. Reis. Wine Quality. UCI Machine Learning Repository, 2009. DOI: https://doi.org/10.24432/C56S3T.

[20] J. Dauparas, I. Anishchenko, N. Bennett, H. Bai, R. J. Ragotte, L. F. Milles, B. I. M. Wicky, A. Courbet, R. J. de Haas, N. Bethel, P. J. Y. Leung, T. F. Huddy, S. Pellock, D. Tischer, F. Chan, B. Koepnick, H. Nguyen, A. Kang, B. Sankaran, A. K. Bera, N. P. King, and D. Baker. Robust deep learning–based protein sequence design using proteinmpnn. *Science*, 378(6615): 49–56, 2022. doi: 10.1126/science.add2187. URL https://www.science.org/doi/abs/10.1126/science.add2187.

[21] S. Dieleman, K. W. Willett, and J. Dambre. Rotation-invariant convolutional neural networks for galaxy morphology prediction. *Monthly Notices of the Royal Astronomical Society*, 450 (2):1441–1459, 04 2015. ISSN 0035-8711. doi: 10.1093/mnras/stv632. URL https://doi.org/10.1093/mnras/stv632.

[22] F. Ding, M. Hardt, J. Miller, and L. Schmidt. Retiring adult: new datasets for fair machine learning. In *Proceedings of the 35th International Conference on Neural Information Processing Systems*, NIPS '21, Red Hook, NY, USA, 2021. Curran Associates Inc. ISBN 9781713845393.

[23] M. Dinitz, S. Im, T. Lavastida, B. Moseley, and S. Vassilvitskii. Algorithms with prediction portfolios. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NIPS '22, Red Hook, NY, USA, 2022. Curran Associates Inc. ISBN 9781713871088.

[24] A. Dvoretzky, J. Kiefer, and J. Wolfowitz. Asymptotic Minimax Character of the Sample Distribution Function and of the Classical Multinomial Estimator. *The Annals of Mathematical Statistics*, 27(3):642 – 669, 1956. doi: 10.1214/aoms/1177728174. URL https://doi.org/10.1214/aoms/1177728174.

[25] A. Fisch, J. Maynez, R. A. Hofer, B. Dhingra, A. Globerson, and W. W. Cohen. Stratified prediction-powered inference for hybrid language model evaluation. In *Proceedings of the 38th International Conference on Neural Information Processing Systems*, NIPS '24, Red Hook, NY, USA, 2024. Curran Associates Inc. ISBN 9798331314385.

[26] A. Geron. *Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems*. O'Reilly Media, Inc., 2nd edition, 2019. ISBN 1492032646.

[27] V. Gkatzelis, D. Schoepflin, and X. Tan. *Clock Auctions Augmented with Unreliable Advice*, pages 2629–2655. doi: 10.1137/1.9781611978322.86. URL https://epubs.siam.org/doi/abs/10.1137/1.9781611978322.86.

[28] S. Haber. A modified monte-carlo quadrature. ii. *Mathematics of Computation*, 21(99):388–397, 1967. ISSN 00255718, 10886842. URL http://www.jstor.org/stable/2003241.

[29] N. Jean, M. Burke, M. Xie, W. M. Alampay Davis, D. B. Lobell, and S. Ermon. Combining satellite imagery and machine learning to predict poverty. *Science*, 353(6301):790–794, Aug. 2016.

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

[31] J. Jumper, R. Evans, A. Pritzel, T. Green, M. Figurnov, O. Ronneberger, K. Tunyasuvunakool, R. Bates, A. Žídek, A. Potapenko, A. Bridgland, C. Meyer, S. A. A. Kohl, A. J. Ballard, A. Cowie, B. Romera-Paredes, S. Nikolov, R. Jain, J. Adler, T. Back, S. Petersen, D. Reiman, E. Clancy, M. Zielinski, M. Steinegger, M. Pacholska, T. Berghammer, S. Bodenstein, D. Silver, O. Vinyals, A. W. Senior, K. Kavukcuoglu, P. Kohli, and D. Hassabis. Highly accurate protein structure prediction with AlphaFold. *Nature*, 596(7873):583–589, Aug. 2021.

[32] J. Kleinberg, H. Lakkaraju, J. Leskovec, J. Ludwig, and S. Mullainathan. Human decisions and machine predictions. *Q. J. Econ.*, 133(1):237–293, Feb. 2018.

[33] P. W. Koh, S. Sagawa, H. Marklund, S. M. Xie, M. Zhang, A. Balsubramani, W. Hu, M. Yasunaga, R. L. Phillips, I. Gao, T. Lee, E. David, I. Stavness, W. Guo, B. Earnshaw, I. Haque, S. M. Beery, J. Leskovec, A. Kundaje, E. Pierson, S. Levine, C. Finn, and P. Liang. Wilds: A benchmark of in-the-wild distribution shifts. In M. Meila and T. Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 5637–5664. PMLR, 18–24 Jul 2021. URL https://proceedings.mlr.press/v139/koh21a.html.

[34] R. Lam, A. Sanchez-Gonzalez, M. Willson, P. Wirnsberger, M. Fortunato, F. Alet, S. Ravuri, T. Ewalds, Z. Eaton-Rosen, W. Hu, A. Merose, S. Hoyer, G. Holland, O. Vinyals, J. Stott, A. Pritzel, S. Mohamed, and P. Battaglia. Learning skillful medium-range global weather forecasting. *Science*, 382(6677):1416–1421, 2023. doi: 10.1126/science.adi2336. URL https://www.science.org/doi/abs/10.1126/science.adi2336.

[35] D. M. Lapola, P. Pinho, J. Barlow, L. E. O. C. Aragão, E. Berenguer, R. Carmenta, H. M. Liddy, H. Seixas, C. V. J. Silva, C. H. L. Silva-Junior, A. A. C. Alencar, L. O. Anderson, D. Armenteras, V. Brovkin, K. Calders, J. Chambers, L. Chini, M. H. Costa, B. L. Faria, P. M. Fearnside, J. Ferreira, L. Gatti, V. H. Gutierrez-Velez, Z. Han, K. Hibbard, C. Koven, P. Lawrence, J. Pongratz, B. T. T. Portela, M. Rounsevell, A. C. Ruane, R. Schaldach, S. S. da Silva, C. von Randow, and W. S. Walker. The drivers and impacts of amazon forest degradation. *Science*, 379(6630):eabp8622, 2023. doi: 10.1126/science.abp8622. URL https://www.science.org/doi/abs/10.1126/science.abp8622.

[36] Z. Lin, H. Akin, R. Rao, B. Hie, Z. Zhu, W. Lu, N. Smetanin, R. Verkuil, O. Kabeli, Y. Shmueli, A. dos Santos Costa, M. Fazel-Zarandi, T. Sercu, S. Candido, and A. Rives. Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science*, 379(6637):1123–1130, 2023. doi: 10.1126/science.ade2574. URL https://www.science.org/doi/abs/10.1126/science.ade2574.

[37] W.-Y. Loh. Regression tress with unbiased variable selection and interaction detection. *Statistica Sinica*, 12(2):361–386, 2002. ISSN 10170405, 19968507. URL http://www.jstor.org/stable/24306967.

[38] T. Lykouris and S. Vassilvitskii. Competitive caching with machine learned advice. *J. ACM*, 68(4), July 2021. ISSN 0004-5411. doi: 10.1145/3447579. URL https://doi.org/10.1145/3447579.

[39] A. Merchant, S. Batzner, S. S. Schoenholz, M. Aykol, G. Cheon, and E. D. Cubuk. Scaling deep learning for materials discovery. *Nature*, 624(7990):80–85, Dec. 2023.

[40] J. N. Morgan and J. A. Sonquist. Problems in the analysis of survey data, and a proposal. *Journal of the American Statistical Association*, 58(302):415–434, 1963. ISSN 01621459, 1537274X. URL http://www.jstor.org/stable/2283276.

[41] G. Ngo, R. Beard, and R. Chandra. Evolutionary bagging for ensemble learning. *Neurocomputing*, 510:1–14, 2022. ISSN 0925-2312. doi: https://doi.org/10.1016/j.neucom.2022.08.055. URL https://www.sciencedirect.com/science/article/pii/S0925231222010414.

[42] E. Nguyen, M. Poli, M. G. Durrant, B. Kang, D. Katrekar, D. B. Li, L. J. Bartie, A. W. Thomas, S. H. King, G. Brixi, J. Sullivan, M. Y. Ng, A. Lewis, A. Lou, S. Ermon, S. A. Baccus, T. Hernandez-Boussard, C. Ré, P. D. Hsu, and B. L. Hie. Sequence modeling and design from molecular to genome scale with evo. *Science*, 386(6723):eado9336, 2024. doi: 10.1126/science.ado9336. URL https://www.science.org/doi/abs/10.1126/science.ado9336.

[43] L. Oakden-Rayner, J. Dunnmon, G. Carneiro, and C. Re. Hidden stratification causes clinically meaningful failures in machine learning for medical imaging. In *Proceedings of the ACM Conference on Health, Inference, and Learning*, CHIL '20, page 151–159, New York, NY, USA, 2020. Association for Computing Machinery. ISBN 9781450370462. doi: 10.1145/3368555.3384468. URL https://doi.org/10.1145/3368555.3384468.

[44] K. O'Leary. AlphaFold gets an upgrade (and a nobel). *Nat. Med.*, 30(12):3393, Dec. 2024.

[45] A. Omidi, M. H. Møller, N. Malhis, J. M. Bui, and J. Gsponer. Alphafold-multimer accurately captures interactions and dynamics of intrinsically disordered protein regions. *Proceedings of the National Academy of Sciences*, 121(44):e2406407121, 2024. doi: 10.1073/pnas.2406407121. URL https://www.pnas.org/doi/abs/10.1073/pnas.2406407121.

[46] R. K. Pace and R. Barry. Sparse spatial autoregressions. *Statistics & Probability Letters*, 33(3):291–297, 1997. ISSN 0167-7152. doi: https://doi.org/10.1016/S0167-7152(96)00140-X. URL https://www.sciencedirect.com/science/article/pii/S016771529600140X.

[47] A. Philippe. Processing simulation output by riemann sums. *J. Stat. Comput. Simul.*, 59(4):295–314, Dec. 1997.

[48] C. Piech, J. Bassen, J. Huang, S. Ganguli, M. Sahami, L. J. Guibas, and J. Sohl-Dickstein. Deep knowledge tracing. In C. Cortes, N. Lawrence, D. Lee, M. Sugiyama, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 28. Curran Associates, Inc., 2015. URL https://proceedings.neurips.cc/paper_files/paper/2015/file/bac9162b47c56fc8a4d2a519803d51b3-Paper.pdf.

[49] J. O. Sexton, X.-P. Song, M. Feng, P. Noojipady, A. Anand, C. Huang, D.-H. Kim, K. M. Collins, S. Channan, C. DiMiceli, and J. R. Townshend. Global, 30-m resolution continuous fields of tree cover: Landsat-based rescaling of modis vegetation continuous fields with lidar-based estimates of error. *International Journal of Digital Earth*, 6(5):427–448, 2013. doi: 10.1080/17538947.2013.786146. URL https://doi.org/10.1080/17538947.2013.786146.

[50] C. J. Shallue and A. Vanderburg. Identifying Exoplanets with Deep Learning: A Five-planet Resonant Chain around Kepler-80 and an Eighth Planet around Kepler-90. *Astronomical Journal*, 155(2):94, Feb. 2018. doi: 10.3847/1538-3881/aa9e09.

[51] T. Shen, Z. Hu, S. Sun, D. Liu, F. Wong, J. Wang, J. Chen, Y. Wang, L. Hong, J. Xiao, L. Zheng, T. Krishnamoorthi, I. King, S. Wang, P. Yin, J. J. Collins, and Y. Li. Accurate RNA 3D structure prediction using a language model-based deep learning approach. *Nat. Methods*, 21(12):2287–2298, Dec. 2024.

[52] Y. Song, D. M. Kluger, H. Parikh, and T. Gu. Demystifying prediction powered inference, 2026. URL https://arxiv.org/abs/2601.20819.

[53] J. M. Stokes, K. Yang, K. Swanson, W. Jin, A. Cubillos-Ruiz, N. M. Donghia, C. R. MacNair, S. French, L. A. Carfrae, Z. Bloom-Ackermann, V. M. Tran, A. Chiappino-Pepe, A. H. Badran, I. W. Andrews, E. J. Chory, G. M. Church, E. D. Brown, T. S. Jaakkola, R. Barzilay, and J. J. Collins. A deep learning approach to antibiotic discovery. *Cell*, 180(4):688–702.e13, Feb. 2020.

[54] K. Tunyasuvunakool, J. Adler, Z. Wu, T. Green, M. Zielinski, A. Žídek, A. Bridgland, A. Cowie, C. Meyer, A. Laydon, S. Velankar, G. J. Kleywegt, A. Bateman, R. Evans, A. Pritzel, M. Figurnov, O. Ronneberger, R. Bates, S. A. A. Kohl, A. Potapenko, A. J. Ballard, B. Romera-Paredes, S. Nikolov, R. Jain, E. Clancy, D. Reiman, S. Petersen, A. W. Senior, K. Kavukcuoglu, E. Birney, P. Kohli, J. Jumper, and D. Hassabis. Highly accurate protein structure prediction for the human proteome. *Nature*, 596(7873):590–596, Aug. 2021.

[55] UniProt Consortium. UniProt: a hub for protein information. *Nucleic Acids Res.*, 43(Database issue):D204–12, Jan. 2015.

[56] A. W. v. d. Vaart. *Asymptotic Statistics*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 1998.

[57] K. W. Willett, C. J. Lintott, S. P. Bamford, K. L. Masters, B. D. Simmons, K. R. V. Casteels, E. M. Edmondson, L. F. Fortson, S. Kaviraj, W. C. Keel, T. Melvin, R. C. Nichol, M. J. Raddick, K. Schawinski, R. J. Simpson, R. A. Skibba, A. M. Smith, and D. Thomas. 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, 09 2013. ISSN 0035-8711. doi: 10.1093/mnras/stt1458. URL https://doi.org/10.1093/mnras/stt1458.

[58] C. J. Wilson, W.-Y. Choy, and M. Karttunen. AlphaFold2: A role for disordered protein/region prediction? *Int. J. Mol. Sci.*, 23(9):4591, Apr. 2022.

[59] S. Yakowitz, J. E. Krimmel, and F. Szidarovszky. Weighted monte carlo integration. *SIAM Journal on Numerical Analysis*, 15(6):1289–1300, 1978. doi: 10.1137/0715088. URL https://doi.org/10.1137/0715088.

[60] D. G. York, et al. The Sloan Digital Sky Survey: Technical Summary. *Astronomical Journal*, 120(3):1579–1587, Sept. 2000. doi: 10.1086/301513.

[61] B. Zhu, M. Ding, P. Jacobson, M. Wu, W. Zhan, M. Jordan, and J. Jiao. Doubly-robust self-training. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, editors, *Advances in Neural Information Processing Systems*, volume 36, pages 41413–41431. Curran Associates, Inc., 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/819f426947c27eb5067bb6fdbdde93dd-Paper-Conference.pdf.

[62] T. Zrnic. A note on the prediction-powered bootstrap, 2025. URL https://arxiv.org/abs/2405.18379.

[63] T. Zrnic and E. J. Candès. Cross-prediction-powered inference. *Proceedings of the National Academy of Sciences*, 121(15):e2322083121, 2024. doi: 10.1073/pnas.2322083121. URL https://www.pnas.org/doi/abs/10.1073/pnas.2322083121.

A Tabulated Experimental Results for PART

Table 1 summarizes the performance of PART and PPI++ across all datasets. Housing price and census income values are reported in thousands of dollars. We report the mean confidence interval width $\pm$ standard deviation.

Table 1: Comparison of Confidence Interval Widths: PART vs. PPI++

Dataset	n	PART width	PPI++ width	Reduction (%)
Deforestation	400	0.050 ± 0.003	0.054 ± 0.003	7.7
	500	0.045 ± 0.002	0.049 ± 0.002	8.3
	600	0.041 ± 0.002	0.045 ± 0.001	8.3
	700	0.038 ± 0.001	0.042 ± 0.001	8.6
	800	0.036 ± 0.001	0.039 ± 0.001	9.0
Galaxy Zoo	500	0.042 ± 0.002	0.044 ± 0.002	4.5
	625	0.038 ± 0.002	0.040 ± 0.001	4.6
	750	0.035 ± 0.001	0.037 ± 0.001	4.9
	875	0.032 ± 0.001	0.034 ± 0.001	5.0
	1000	0.030 ± 0.001	0.032 ± 0.001	5.3
Wine Quality	300	0.116 ± 0.005	0.140 ± 0.006	17.3
	375	0.105 ± 0.004	0.126 ± 0.004	16.6
	450	0.097 ± 0.003	0.116 ± 0.003	16.7
	525	0.090 ± 0.003	0.108 ± 0.003	16.7
	600	0.085 ± 0.002	0.102 ± 0.002	16.9
Housing Prices	500	7.405 ± 0.454	8.624 ± 0.509	14.1
	625	6.717 ± 0.310	7.774 ± 0.352	13.6
	750	6.272 ± 0.274	7.258 ± 0.269	13.6
	875	5.782 ± 0.258	6.756 ± 0.242	14.4
	1000	5.443 ± 0.208	6.376 ± 0.228	14.6
AlphaFold (Odds Ratio)	400	0.757 ± 0.086	0.926 ± 0.119	18.2
	500	0.678 ± 0.068	0.839 ± 0.086	19.2
	600	0.611 ± 0.062	0.756 ± 0.077	19.1
	700	0.581 ± 0.052	0.731 ± 0.074	20.6
	800	0.540 ± 0.047	0.677 ± 0.058	20.2
Census Income (Linear Regression)	300	9.408 ± 2.696	9.766 ± 2.799	3.7
	400	8.116 ± 1.997	8.508 ± 2.144	4.6
	500	7.078 ± 1.622	7.422 ± 1.732	4.6
	600	6.435 ± 1.301	6.756 ± 1.403	4.8
	700	5.939 ± 1.007	6.258 ± 1.063	5.1

B Experimental Methodology

In this section, we provide an overview of the testing methodology, tree hyperparameters, and datasets used in the experimental results.

Testing Methodology. For our testing methodology, we assume that each dataset has M i.i.d. samples consisting of features, labels, and predicted labels. For various fixed values of n, we randomly obfuscate $N = M - n$ of the labels to produce N pieces of unlabeled data and n pieces of labeled data. We subsequently run the PART estimator using a manually tuned depth limit on the labeled and unlabeled data, recording the width of the confidence interval and whether the interval contained the true mean. We repeat the previous procedure 100 times for each fixed n and compute the average confidence interval width and the coverage probability. We produce two graphs to show how the width of the intervals and coverage probability change as a function of n. For the interval width graph, we include one standard deviation error bars on each measured n value. We show the nominal coverage probability with a constant dashed gray line.

Tree construction and hyperparameters. For the PART experiments, we compute empirical quantiles at the fixed levels $0.05, 0.10, \ldots, 0.95$ for each partitioning feature using the inverse empirical CDF of the full unlabeled sample. We use these quantile values directly as candidate thresholds, remove duplicates within each coordinate, and reuse the same thresholds at every node. For the mean estimation and odds-ratio tasks, we require at least $m = 5$ labeled observations in each child. For the linear regression experiment, we require $m = 100$. We specify the PART depth limits in each experimental plot.

Deforestation Dataset Deforestation in the Amazon rainforest disrupts local ecosystems and releases stored carbon [14, 35]. This motivates remote monitoring of forest cover through satellite imagery [49]. This dataset consists of satellite imagery derived from Landsat sensors. The objective is to measure the proportion of images that contain evidence of deforestation. There are 1,596 observations with binary labels that indicate the presence of forest cover loss. In our experiments, we use predictions generated from a histogram-based gradient-boosted tree trained by [4].

Galaxy Dataset The Galaxy Zoo 2 project [57] is a crowd-sourced initiative focused on providing morphological classifications for 304,122 galaxies captured by the Sloan Digital Sky Survey [60]. These human-annotated labels serve as key aids in helping scientists understand galaxy evolution (e.g., [17]) but require substantial manual effort. To mitigate this, machine learning approaches—such as the convolutional neural network of Dieleman et al. [21]—predict morphology directly from the images. One goal for this dataset is to estimate the mean fraction of spiral galaxies in the local universe. Our dataset consists of 16,743 total observations where each label is the estimated probability that the galaxy is spiral. We use predictions from a ResNet50 model supplied by [4].

Wine Quality Dataset Vinho Verde wine originates from the Minho Province of Portugal, where it is tightly regulated under the Denominação de Origem Controlada (DOC) system. To deter counterfeiting, regulators use physicochemical and sensory tests to assess wine quality. Cortez et al. [19] provide a dataset of 1280 wine samples, each with 11 features (pH, sulphates, alcohol content, etc) gathered via physicochemical tests and a quality score on a 0-10 scale. This dataset has become a standard benchmark for ensemble tree methods (e.g., [41]). The objective is to estimate the average wine quality based on this physicochemical data. For predictions, we train a Random Forest Classifier on 20% of the data and generate predicted scores for the remaining 80% of the data.

Housing Price Dataset California's housing market is shaped by many factors, including its coastal geography and strict zoning laws. One source of information about the housing market comes from census data. Pace and Barry [46] initially used the 1990 California Census to form a dataset consisting of 16,512 instances with 9 characteristics (population, number of households, total rooms, etc.) and 1 outcome variable indicating the median house value in the block (a geographical unit used by the Census Bureau). Many textbooks (e.g., [26]) use this dataset as a standard ML benchmark. The objective for this dataset is to estimate the average median house price in a block group in California. For predictions, we train a Random Forest Regressor on 20% of the data and generate predicted scores for the remaining 80% of the data.

AlphaFold Dataset Since the groundbreaking result of [31], AlphaFold has quickly become an essential tool used by researchers to allow the study of the structure of proteins on a mass scale (e.g., see [45, 54, 58] among many others). Predictions from AlphaFold were recently used by Bludau et al. [10] to understand whether intrinsically disordered regions (IDRs) on proteins experience phosphorylation at a greater rate. This dataset from [55] consists of 10,802 samples. The objective for this dataset is to predict the odds ratio for a protein being phosphorylated and belonging to an IDR. We utilize the AlphaFold-based predictions from [4].

Income Dataset A persistent correlation between sex and earnings is well documented in the economics literature [9]. We utilize census data from Ding et al. [22] for the task of estimating the linear relationship between sex (M/F) and individual income. The dataset consists of 380,091 observations with two features, sex (M/F) and age (0-99), and a label for the person's income. For predictions, we utilize an XGBoost model trained by [5].

Synthetic Datasets for the PAQ Estimator In this section, we overview the synthetic dataset we use to test the PAQ estimator against PPI^{++}. While the PAQ estimator assumes that the residuals

are deterministic and smooth, there are many natural settings where the conditions of Assumption 4 hold exactly or in good approximation. For example, Anfinsen's thermodynamic hypothesis [3] suggests that for many small globular proteins under appropriate conditions the amino-acid sequence uniquely determines its physical structure. Similarly, in physics, there are many equations in which the final state of the system is determined by its initial conditions, such as in the 2-D incompressible Navier-Stokes equation [18].

To construct our synthetic dataset, we let

$$X \sim \mathcal{U}[0, \pi], \quad y(x) = x^2 + \sin(x), \quad f(x) = x^2, \quad r(x) = \sin(x).$$

The residual function is deterministic and twice continuously differentiable, and thus meets Assumption 4. We demonstrate the performance of PAQ in Figure 7.

C Related Work

ML Predictions for Mean Estimation. We briefly survey the relevant landscape of using ML predictions for statistical inference. Angelopoulos et al. [4] formalized the model for prediction powered statistical inference and proposed the PPI estimator for mean estimation, linear regression coefficient estimation, odds ratio estimation, among other targets. In follow-up work, Angelopoulos et al. [5] introduced PPI++, which uses a power-tuning parameter to adaptively down-weight the influence of the predictions based on their accuracy, yielding an estimator that is asymptotically no less efficient than the empirical mean. Zhu et al. [61] arrived at an equivalent estimator to [4] from the self-training perspective within semi-supervised learning. More recently, Ji et al. [30] proposed RePPI, a recalibrated version of PPI, that uses a flexible machine learning model to model $\mathbb{E}[Y|f(X), X]$, rather than fitting a linear relationship as in PPI++. As RePPI requires training an auxiliary model to learn this conditional expectation, we retain PPI and PPI++ as our primary benchmarks. The methodology of PART is most similar to StratPPI [25], an estimator that uses stratified sampling over *predefined* strata to improve the efficiency of PPI/PPI++. In contrast, PART learns stratification adaptively from the labeled residuals. For a detailed survey of existing prediction-enhanced estimators, we refer the reader to Song et al. [52].

Several papers have since followed which improve the practical use of PPI-style estimators. While earlier work assumes access to a pre-trained ML model, Zrnic and Candès [63] use cross-validation techniques to train the ML model from scratch utilizing the small labeled set of data. Zrnic [62] integrates the bootstrap method to prediction powered inference to facilitate applications where the central limit theorem justifications may be tenuous. While we assume access to a pre-trained model in this work, we remark that both cross-validation and bootstrap methods can be applied to our estimator using a formula similar to the previous papers.

Learning-Augmented Algorithms. Learning-augmented (LA) algorithms seek to enhance classical algorithms by incorporating the predictions of ML models trained on historical data. The framework has been successfully applied to many settings such as facility location [2, 6], mechanism design [8, 27], and scheduling problems [7, 23], among many others[5]. To adapt the original framework of Lykouris and Vassilvitskii [38] to our setting, LA algorithms for mean estimation should output a confidence interval that 1) is tighter than the empirical mean when the predictions are accurate (*consistency*) and 2) retains the correct coverage probability when the predictions are arbitrarily poor (*robustness*). Similarly to [4, 5], we characterize the asymptotic distribution of the PART estimator, allowing us to build confidence intervals that satisfy both consistency and robustness.

Decision Tree for Mean Estimation. Decision trees have a long history, beginning with their use in exploratory data analysis by Morgan and Sonquist [40]. Breiman et al. [13] formalized modern regression trees for estimating conditional means with the CART algorithm. Subsequent work has extended and strengthened CART, e.g., unbiased-splitting methods [37] and popular ensemble approaches including "bagging" [11] and random forests [12]. We contribute to this line of literature by adapting the CART node splitting criteria of [13] to build the PART decision tree for mean residual estimation.

[5] See [1] for an up-to-date list of learning-augmented papers.

Monte Carlo Methods for Numerical Integration. A complementary line of work studies replacing empirical averages with carefully weighted sums to accelerate the convergence of integral estimates. It is well-known that the integral of a function on the unit interval can be approximated via an empirical average at a convergence rate of $O(n^{-1/2})$. Yakowitz et al. [59] were the first to show that, for twice-differentiable functions on the unit interval, the canonical $O(n^{-1})$ rate can be improved to $O(n^{-4})$ via quadrature methods. Many results (e.g., [28, 47]) have since followed that observe convergence improvements through weighted sums. We adapt these quadrature methods to estimate the debiasing term in our semi-supervised setting where the data is not uniformly distributed on the unit interval.

D Algorithms

We present the PART algorithms for mean estimation and linear regression from Appendices F.1 and F.3. Both algorithms select splits using labeled residuals and estimate leaf probabilities using the unlabeled sample. We also give the linear PAQ algorithm from Section 3.2, which interpolates the labeled residuals and averages the adjusted predictions over the unlabeled sample. In the following, we recall some notation used in the construction of PART.

Candidate Splits and Regional Statistics. For both PART algorithms, the maximum depth D, minimum child size $m \geq 2$, and quantile levels $0 < u_1 < \cdots < u_K < 1$, with $K \geq 2$, are fixed independently of n and N. For each feature coordinate j, compute the empirical quantiles $q_1^{(j)}, \ldots, q_K^{(j)}$ from the full unlabeled sample, as in Appendix F.1, and set

$$\mathcal{S}^{(j)} := \{q_t^{(j)} : t = 1, \ldots, K\}.$$

Remove duplicate thresholds within each coordinate. These sets are computed once and used at every node. For any region $\mathcal{R}$, write

$$n_{\mathcal{R}} := \sum_{i=1}^{n} \mathbf{1}\{x_i \in \mathcal{R}\}, \qquad p_{\mathcal{R}} := \widehat{p}_N(\mathcal{R}) = \frac{1}{N} \sum_{j=1}^{N} \mathbf{1}\{\widetilde{x}_j \in \mathcal{R}\}.$$

At a node $\mathcal{R}$, a candidate (j, s) with $s \in \mathcal{S}^{(j)}$ produces

$$\mathcal{R}_L(j, s) = \{x \in \mathcal{R} : x^{(j)} \leq s\}, \qquad \mathcal{R}_R(j, s) = \{x \in \mathcal{R} : x^{(j)} > s\}.$$

It is admissible if both children contain at least m labeled observations. Ties between candidate splits are resolved using a fixed deterministic rule.

Algorithm 1 PART Mean Estimation and Inference

Require: Labeled data $\mathcal{L} = \{(x_i, y_i)\}_{i=1}^{n}$, $n \geq 2$; unlabeled data $\mathcal{U} = \{\widetilde{x}_j\}_{j=1}^{N}$, $N \geq 1$; predictor f.
Require: Maximum depth D, minimum child size $m \geq 2$, fixed quantile levels $u_1, \ldots, u_K$, and $\alpha \in (0, 1)$.

Construct $\mathcal{S}^{(j)}$ for each coordinate from the full unlabeled sample using the rule above.

Compute

$$\hat{\lambda} \leftarrow \frac{\widehat{\mathrm{Cov}}_n(Y, f(X))}{(1 + n/N)\widehat{\mathrm{Var}}_{n+N}(f(X))},$$

setting $\hat{\lambda} = 0$ if the denominator is zero.

Set $r_i^{(\hat{\lambda})} \leftarrow y_i - \hat{\lambda} f(x_i)$ for $i = 1, \ldots, n$.

function BUILDMEANTREE($\mathcal{R}, h$)

  if $h = D$ or no admissible candidate remains **then**

    return a leaf with region $\mathcal{R}$.

  end if

  For each admissible (j, s) and $a \in \{L, R\}$, compute $n_{(j,s,a)}$, $p_{(j,s,a)}$, and the sample residual variance $\widehat{\sigma}^2_{(j,s,a)}$ in $\mathcal{R}_a(j, s)$, as in Appendix F.1.

  Choose (j^*, s^*) minimizing

$$\mathrm{VMS}(j, s) = \sum_{a \in \{L,R\}} p^2_{(j,s,a)} \frac{\widehat{\sigma}^2_{(j,s,a)}}{n_{(j,s,a)}}$$

  over admissible candidates, using the fixed tie-breaking rule.

  $T_L \leftarrow$ BUILDMEANTREE($\mathcal{R}_L(j^*, s^*), h + 1$).

  $T_R \leftarrow$ BUILDMEANTREE($\mathcal{R}_R(j^*, s^*), h + 1$).

  return an internal node with split (j^*, s^*) and subtrees T_L, T_R.

end function

$T \leftarrow$BUILDMEANTREE($\mathcal{X}, 0$).

For the leaves $\{\mathcal{R}_\ell\}_{\ell=1}^{L}$ of T, compute each leaf's labeled count n_ℓ, unlabeled proportion p_ℓ, residual mean $\overline{r}_\ell^{(\hat{\lambda})}$, and sample residual variance $\widehat{\sigma}^2_{\ell,\hat{\lambda}}$, as in Appendix F.1.

Compute

$$\hat{\mu}_{T,\hat{\lambda}} \leftarrow \frac{1}{N} \sum_{j=1}^{N} \hat{\lambda} f(\widetilde{x}_j) + \sum_{\ell=1}^{L} p_\ell \overline{r}_\ell^{(\hat{\lambda})},$$

$$\widehat{\sigma}^2_{T,\hat{\lambda}} \leftarrow \sum_{\ell=1}^{L} p_\ell^2 \frac{\widehat{\sigma}^2_{\ell,\hat{\lambda}}}{n_\ell}.$$

return $\hat{\mu}_{T,\hat{\lambda}}$ and the interval

$$\left[\hat{\mu}_{T,\hat{\lambda}} - z_{1-\alpha/2}\widehat{\sigma}_{T,\hat{\lambda}}, \quad \hat{\mu}_{T,\hat{\lambda}} + z_{1-\alpha/2}\widehat{\sigma}_{T,\hat{\lambda}}\right].$$

Algorithm 2 PART Linear Regression Estimation and Inference

Require: Labeled data $\mathcal{L} = \{(x_i, y_i)\}_{i=1}^{n}$, $n \geq 2$; unlabeled data $\mathcal{U} = \{\widetilde{x}_j\}_{j=1}^{N}$, $N \geq 1$; predictor f; target coordinate $k \in \{1, \ldots, d\}$.

Require: Maximum depth D, minimum child size $m \geq 2$, fixed quantile levels $u_1, \ldots, u_K$, and $\alpha \in (0, 1)$.

Compute $\widetilde{\Sigma} \leftarrow N^{-1} \sum_{j=1}^{N} \widetilde{x}_j \widetilde{x}_j^\top$.

if $\widetilde{\Sigma}$ is singular **then**

  return the estimate 0 and the interval $\mathbb{R}$.

end if

Compute $\widetilde{\theta} \leftarrow \widetilde{\Sigma}^{-1} N^{-1} \sum_{j=1}^{N} \widetilde{x}_j f(\widetilde{x}_j)$.

Set $S_i \leftarrow x_i(y_i - f(x_i))$ for $i = 1, \ldots, n$.

Construct $\mathcal{S}^{(j)}$ for each coordinate from the full unlabeled sample using the rule above.

function BUILDREGRESSIONTREE($\mathcal{R}, h$)

  if $h = D$ or no admissible candidate remains **then**

    return a leaf with region $\mathcal{R}$.

  end if

  For each admissible (j, s) and child $\mathcal{R}_a(j, s)$, compute $\widehat{V}^{(k,k)}_{\mathcal{R}_a(j,s)}$, the sample variance of $(\widetilde{\Sigma}^{-1} S_i)^{(k)}$ in that child.

  Choose (j^*, s^*) minimizing

$$\mathrm{VMS}_{\mathrm{LR}}(j, s, \mathcal{R}) = \sum_{a \in \{L,R\}} p^2_{\mathcal{R}_a(j,s)} \frac{\widehat{V}^{(k,k)}_{\mathcal{R}_a(j,s)}}{n_{\mathcal{R}_a(j,s)}}$$

  over admissible candidates, using the fixed tie-breaking rule.

  $T_L \leftarrow$ BUILDREGRESSIONTREE($\mathcal{R}_L(j^*, s^*), h + 1$).

  $T_R \leftarrow$ BUILDREGRESSIONTREE($\mathcal{R}_R(j^*, s^*), h + 1$).

  return an internal node with split (j^*, s^*) and subtrees T_L, T_R.

end function

$T \leftarrow$BUILDREGRESSIONTREE($\mathcal{X}, 0$).

For the leaves $\{\mathcal{R}_\ell\}_{\ell=1}^{L}$ of T, compute $n_{\mathcal{R}_\ell}$, $p_{\mathcal{R}_\ell}$, the mean score $\overline{S}_{\mathcal{R}_\ell}$, the correction $\widehat{R}_{\mathcal{R}_\ell} = \widetilde{\Sigma}^{-1} \overline{S}_{\mathcal{R}_\ell}$, and $\widehat{V}^{(k,k)}_{\mathcal{R}_\ell}$, as in Appendix F.3.

Compute

$$\widehat{\theta}^{(k)}_T \leftarrow \widetilde{\theta}^{(k)} + \sum_{\ell=1}^{L} p_{\mathcal{R}_\ell} \widehat{R}^{(k)}_{\mathcal{R}_\ell},$$

$$\widehat{\sigma}^2 \leftarrow \sum_{\ell=1}^{L} p^2_{\mathcal{R}_\ell} \frac{\widehat{V}^{(k,k)}_{\mathcal{R}_\ell}}{n_{\mathcal{R}_\ell}}.$$

return $\widehat{\theta}^{(k)}_T$ and the interval

$$\left[\widehat{\theta}^{(k)}_T - z_{1-\alpha/2}\widehat{\sigma}, \quad \widehat{\theta}^{(k)}_T + z_{1-\alpha/2}\widehat{\sigma}\right].$$

Algorithm 3 PAQ Mean Estimation and Inference with Linear Interpolation

Require: Labeled data $\mathcal{L} = \{(x_i, y_i)\}_{i=1}^n$, $n \geq 3$, with distinct scalar features; unlabeled data $\mathcal{U} = \{\widetilde{x}_j\}_{j=1}^N$, $N \geq 2$; predictor f.

Require: Inflation factor $\kappa > 1$ and $\alpha \in (0, 1)$, both fixed.

Sort the labeled observations by feature value, obtaining $(x_{(i)}, y_{(i)})_{i=1}^n$ with $x_{(1)} < \cdots < x_{(n)}$.

Set $r_i \leftarrow y_{(i)} - f(x_{(i)})$ for $i = 1, \ldots, n$.

Set $s_i \leftarrow (r_{i+1} - r_i)/(x_{(i+1)} - x_{(i)})$ for $i = 1, \ldots, n-1$.

Compute

$$\widehat{L}_2 \leftarrow 2\kappa \max_{1 \leq i \leq n-2} \left| \frac{s_{i+1} - s_i}{x_{(i+2)} - x_{(i)}} \right|.$$

for $j = 1, \ldots, N$ **do**
  if $\widetilde{x}_j < x_{(1)}$ **then**
    $i \leftarrow 1$.
  else if $\widetilde{x}_j \geq x_{(n)}$ **then**
    $i \leftarrow n - 1$.
  else
    Find i such that $x_{(i)} \leq \widetilde{x}_j < x_{(i+1)}$.
  end if
  $q_1(\widetilde{x}_j) \leftarrow r_i + s_i(\widetilde{x}_j - x_{(i)})$.
  $w_1(\widetilde{x}_j) \leftarrow \frac{1}{2}|(\widetilde{x}_j - x_{(i)})(\widetilde{x}_j - x_{(i+1)})|$.
  $Z_j \leftarrow f(\widetilde{x}_j) + q_1(\widetilde{x}_j)$.
end for

Compute

$$\widehat{\mu}_{\mathrm{PAQ}} \leftarrow \frac{1}{N} \sum_{j=1}^N Z_j,$$

$$\widehat{\sigma}^2_{\mathrm{PAQ}} \leftarrow \frac{1}{N(N-1)} \sum_{j=1}^N (Z_j - \widehat{\mu}_{\mathrm{PAQ}})^2,$$

$$\widehat{B}_1 \leftarrow \frac{\widehat{L}_2}{N} \sum_{j=1}^N w_1(\widetilde{x}_j).$$

Set

$$C_{n,N} \leftarrow \left[\widehat{\mu}_{\mathrm{PAQ}} - z_{1-\alpha/2}\widehat{\sigma}_{\mathrm{PAQ}} - \widehat{B}_1, \quad \widehat{\mu}_{\mathrm{PAQ}} + z_{1-\alpha/2}\widehat{\sigma}_{\mathrm{PAQ}} + \widehat{B}_1 \right].$$

return $\widehat{\mu}_{\mathrm{PAQ}}$ and $C_{n,N}$.

The PART mean and regression coverage results are given in Theorem 1 and Theorem 3, respectively. The PAQ coverage result is given in Theorem 2, with extensions to nonuniform sampling and higher-degree interpolation in Appendix G.

E Motivating Example for PART

We illustrate the variance reduction in Corollary 1 with a simple example. This example shows how partitioning can improve precision beyond the global correction used by PPI++. Afterwards, we also provide a simple example where partitioning provably does not help reduce variance.

Example 1 (Strict asymptotic variance reduction)**.** *Let* $X = (X^{(1)}, X^{(2)})$*, where the two coordinates are independent and each equals* -1 *or* 1 *with probability* $1/2$*. Let* ε *be independent of* X*, with* $\mathbb{E}[\varepsilon] = 0$ *and* $\mathrm{Var}(\varepsilon) = 1$*, and define*

$$Y = X^{(1)} + X^{(2)} + \varepsilon, \qquad f(X) = X^{(2)}.$$

Then $\mu = \mathbb{E}[Y] = 0$*.*

Global Residual Correction. *Since* $\mathrm{Cov}(Y, f(X)) = \mathrm{Var}(f(X)) = 1$*, the population PPI++ coefficient is*

$$\lambda^* = 1, \qquad r^{(\lambda^*)}(X, Y) = X^{(1)} + \varepsilon.$$

Thus the asymptotic variance of PPI++ is

$$\mathrm{Var}\Big(r^{(\lambda^*)}(X, Y)\Big) = 2.$$

Corrections Within Leaves. *Consider the partition according to the first coordinate, with leaves*

$$\mathcal{R}_- = \{x : x^{(1)} \le -1\}, \qquad \mathcal{R}_+ = \{x : x^{(1)} > -1\}.$$

Each leaf has probability $1/2$*. For* $s \in \{-1, 1\}$*,*

$$\mathbb{E}\Big[r^{(\lambda^*)}(X,Y) \mid X^{(1)} = s\Big] = s, \qquad \mathrm{Var}\Big(r^{(\lambda^*)}(X,Y) \mid X^{(1)} = s\Big) = 1.$$

The probability-weighted within-leaf residual variance is therefore

$$v_0 = \frac{1}{2} \cdot 1 + \frac{1}{2} \cdot 1 = 1.$$

The reduction of the variance from two to one is from the variance of the leaf mean residuals, which equal -1 *and* 1 *with equal probability.*

Selection by PART. *Take maximum depth* $D = 1$*, quantile levels* $u_1 = 1/4$ *and* $u_2 = 3/4$*, and any fixed minimum child size* $m \ge 2$*. The population candidate thresholds are* -1 *and* 1 *in each coordinate. Only the threshold* -1 *is feasible, since the threshold* 1 *produces an empty right child. Splitting on the first coordinate has population VMS score one, as calculated above. Splitting on the second coordinate has score two, since the residual* $X^{(1)} + \varepsilon$ *is independent of* $X^{(2)}$*. Thus the population greedy tree* T_0 *uniquely selects the first coordinate and has leaves* $\mathcal{R}_-$ *and* $\mathcal{R}_+$*.*

The empirical quantiles equal their population counterparts with probability tending to one. On this event, the infeasible splits have empty right children. Each child of a feasible population split has probability $1/2$*, so its labeled count eventually exceeds the fixed minimum* m *with probability tending to one. The remaining moment and nondegeneracy conditions also hold. Hence Assumption 1 is satisfied, and Theorem 1 gives*

$$\sqrt{n}\big(\hat{\mu}_{T,\hat{\lambda}} - \mu\big) \xrightarrow{d} \mathcal{N}(0,1), \qquad \sqrt{n}\big(\hat{\mu}_{\mathrm{PPI++}} - \mu\big) \xrightarrow{d} \mathcal{N}(0,2).$$

For a fixed confidence level, the variance consistency in Theorem 1 implies that the ratio of the PART interval width to the PPI++ interval width converges in probability to $1/\sqrt{2}$*, a reduction of approximately* 29%*.*

In this example, the predictor accounts for $X^{(2)}$, while the global residual correction still includes variation from $X^{(1)}$. PART estimates the mean residual separately in the two leaves and weights these corrections using the unlabeled sample. This removes the between-leaf contribution to the asymptotic variance, leaving the variation from ε.

Finally, we conclude by providing an example where PART does not improve over PPI++. Suppose that $Y = f(X) + \epsilon$, where $\epsilon \sim \mathcal{N}(0, \sigma^2)$ is independent of X, and $\mathrm{Var}(f(X)) > 0$. Then $\lambda^* = 1$, so the residual is

$$r^{(\lambda^*)}(X,Y) = Y - f(X) = \epsilon.$$

Every feature region of positive probability has residual mean zero and residual variance σ^2. Therefore, regardless of the feature partition, PART will have a variance of σ^2, which is equal to the global residual variance. Thus, partitioning offers no variance reduction in this example.

F PART Estimator: Implementation and Analysis

F.1 Estimating the Mean

In this section, we describe the construction of PART for estimating $\mu = \mathbb{E}[Y]$. Let $\mathcal{L} = \{(x_i, y_i)\}_{i=1}^n$ be the labeled sample, let $\mathcal{U} = \{\widetilde{x}_j\}_{j=1}^N$ be the unlabeled sample, and let $f : \mathcal{X} \to \mathbb{R}$ be the predictor. As in Section 3.1, compute

$$\hat{\lambda} := \frac{\widehat{\mathrm{Cov}}_n\big(Y, f(X)\big)}{(1 + n/N)\,\widehat{\mathrm{Var}}_{n+N}\big(f(X)\big)}, \qquad r^{(\hat{\lambda})}(x_i, y_i) := y_i - \hat{\lambda} f(x_i).$$

The covariance uses the labeled sample, and the prediction variance uses the pooled labeled and unlabeled features. For any region $\mathcal{R} \subseteq \mathcal{X}$, define its empirical unlabeled mass by

$$\widehat{p}_N(\mathcal{R}) := \frac{1}{N} \sum_{j=1}^{N} \mathbf{1}\{\widetilde{x}_j \in \mathcal{R}\}.$$

We write $p(\mathcal{R}) = P_X(\mathcal{R})$ to denote the population mass of the region $\mathcal{R}$.

Candidate Splitting Points. Fix a maximum depth D, a minimum number of labeled observations per child $m \geq 2$, and quantile levels $0 < u_1 < \cdots < u_K < 1$, with $K \geq 2$. These choices are fixed independently of n and N. For each coordinate $k \in \{1, \ldots, d\}$, compute the empirical quantiles from the full unlabeled sample:

$$q_j^{(k)} := \inf \left\{ t \in \mathbb{R} : \frac{1}{N} \sum_{i=1}^{N} \mathbf{1}\{\widetilde{x}_i^{(k)} \leq t\} \geq u_j \right\}, \qquad j = 1, \ldots, K.$$

We use these quantile values directly as candidate thresholds:

$$\mathcal{S}^{(k)} := \{q_j^{(k)} : j = 1, \ldots, K\}.$$

We remove duplicate thresholds within each coordinate. These sets are computed once and used at every node. Under condition *(ii)* in Assumption 1, corresponding empirical and population candidate regions agree up to vanishing population mass, and their feasibility decisions agree with probability tending to one.

Remark 1 (Sufficient conditions for Assumption 1)**.** *We briefly justify the sufficient conditions stated after Assumption 1 that enable the greedy fitted tree to stabilize as the number of samples increases. Let F_k and $F_{N,k}$ be the population and empirical marginal distribution functions of coordinate k, and let $q_{0,j}^{(k)} = F_k^{-1}(u_j)$. If F_k is continuous, then $F_k(q_{0,j}^{(k)}) = u_j$ and $F_{N,k}(q_j^{(k)}) = \lceil N u_j \rceil / N$ almost surely. Hence*

$$P_X\Big(\{x : x^{(k)} \leq q_j^{(k)}\} \triangle \{x : x^{(k)} \leq q_{0,j}^{(k)}\}\Big) = \Big|F_k(q_j^{(k)}) - u_j\Big| \leq \sup_t |F_{N,k}(t) - F_k(t)| + \frac{1}{N}.$$

By the DKW inequality, the expected supremum is $O(N^{-1/2})$ [24]. Now, let B_N be the set of feature values whose assignment differs under at least one pair of corresponding empirical and population thresholds. Since the grid is fixed, $\mathbb{E}[p(B_N)] = O(N^{-1/2})$. Thus, the disagreement in the candidate regions shrinks, which proves the first part of condition (ii). Next, positive-probability population children eventually have at least m labeled observations in their empirical counterparts by region consistency and the law of large numbers.

Finally, suppose $n/\sqrt{N} \to 0$. By independence of the labeled and unlabeled samples and a union bound,

$$\Pr[\exists i \leq n : x_i \in B_N] \leq n\, \mathbb{E}[p(B_N)] = O(n/\sqrt{N}) \longrightarrow 0.$$

Thus, with probability tending to one, replacing population thresholds with empirical thresholds does not change which labeled observations belong to each candidate region. On this event, the empirical counterpart of a child with zero population probability also contains no labeled observations. Together with the argument above for children with positive probability, this establishes the feasibility part of condition (ii).

Alternatively, suppose the feature distribution has a density positive almost everywhere on a rectangular support. It is then straightforward to verify condition (ii): quantile consistency ensures that corresponding regions differ by vanishing probability mass, while the rectangular support preserves feasibility with probability tending to one. Thus, we can establish condition (ii) without requiring $n/\sqrt{N} \to 0$.

Splitting Function. At a node with region $\mathcal{R}$, a candidate (k, s) with $s \in \mathcal{S}^{(k)}$ produces the children

$$\mathcal{R}_L(k, s) := \{x \in \mathcal{R} : x^{(k)} \leq s\}, \qquad \mathcal{R}_R(k, s) := \{x \in \mathcal{R} : x^{(k)} > s\}.$$

For $a \in \{L, R\}$, let $n_{(k,s,a)}$ be the labeled count in $\mathcal{R}_a(k, s)$, and let

$$p_{(k,s,a)} := \widehat{p}_N(\mathcal{R}_a(k, s))\,.$$

A candidate is admissible if both children contain at least m labeled observations. For each admissible candidate, let $\overline{r}_{(k,s,a)}^{(\hat{\lambda})}$ be the mean residual in child a, and compute its sample residual variance

$$\widehat{\sigma}_{(k,s,a)}^2 := \frac{1}{n_{(k,s,a)} - 1} \sum_{i : x_i \in \mathcal{R}_a(k,s)} \Big(r^{(\hat{\lambda})}(x_i, y_i) - \overline{r}_{(k,s,a)}^{(\hat{\lambda})}\Big)^2.$$

Choose an admissible candidate that minimizes the *Variance-Minimizing Split* (VMS) criterion

$$\mathrm{VMS}(k,s) \coloneqq \sum_{a\in\{L,R\}} p^2_{(k,s,a)} \frac{\widehat{\sigma}^2_{(k,s,a)}}{n_{(k,s,a)}},$$

using a fixed deterministic rule to break ties.

Termination Criterion. Start at the root $\mathcal{X}$, at depth zero, and apply the splitting rule recursively to both children. Stop at depth D or when no admissible candidate remains. Let T be the resulting tree, with leaves $\{\mathcal{R}_\ell\}_{\ell=1}^{L}$ and $L \le 2^D$. For each leaf, let n_ℓ be its labeled count and define

$$p_\ell \coloneqq \widehat{p}_N(\mathcal{R}_\ell), \qquad \overline{r}_\ell^{(\hat{\lambda})} \coloneqq \frac{1}{n_\ell} \sum_{i:x_i\in\mathcal{R}_\ell} r^{(\hat{\lambda})}(x_i,y_i).$$

The PART estimator is given by

$$\hat{\mu}_{T,\hat{\lambda}} \coloneqq \frac{1}{N}\sum_{j=1}^{N} \hat{\lambda} f(\widetilde{x}_j) + \sum_{\ell=1}^{L} p_\ell\, \overline{r}_\ell^{(\hat{\lambda})}.$$

We provide pseudocode of the estimator in Algorithm 1.

F.1.1 Constructing Confidence Intervals

For each leaf, compute the sample residual variance

$$\widehat{\sigma}^2_{\ell,\hat{\lambda}} \coloneqq \frac{1}{n_\ell - 1} \sum_{i:x_i\in\mathcal{R}_\ell} \left(r^{(\hat{\lambda})}(x_i,y_i) - \overline{r}_\ell^{(\hat{\lambda})}\right)^2.$$

Then, we estimate the variance of $\hat{\mu}_{T,\hat{\lambda}}$ by

$$\widehat{\sigma}^2_{T,\hat{\lambda}} \coloneqq \sum_{\ell=1}^{L} p_\ell^2 \frac{\widehat{\sigma}^2_{\ell,\hat{\lambda}}}{n_\ell}.$$

For a fixed $\alpha \in (0,1)$, the nominal $(1-\alpha)$ Wald interval is

$$\left[\hat{\mu}_{T,\hat{\lambda}} - z_{1-\alpha/2}\widehat{\sigma}_{T,\hat{\lambda}},\quad \hat{\mu}_{T,\hat{\lambda}} + z_{1-\alpha/2}\widehat{\sigma}_{T,\hat{\lambda}}\right],$$

where $z_{1-\alpha/2}$ is the $(1-\alpha/2)$ standard normal quantile. Under Assumption 1, Theorem 1 establishes that the coverage probability of this interval converges to $1-\alpha$.

F.2 Establishing Theoretical Coverage Properties

We prove Theorem 1 under Assumption 1. Let T_0 be the population greedy tree defined in Section 3.1, with leaves $\{\mathcal{R}_{0,\ell}\}_{\ell=1}^{L_0}$. Recall that

$$\lambda^* \coloneqq \frac{\mathrm{Cov}(Y,f(X))}{\mathrm{Var}(f(X))}, \qquad r^{(\lambda^*)}(x,y) \coloneqq y - \lambda^* f(x),$$

and

$$v_0 \coloneqq \sum_{\ell=1}^{L_0} p(\mathcal{R}_{0,\ell})\,\mathrm{Var}\Big(r^{(\lambda^*)}(X,Y) \mid X\in\mathcal{R}_{0,\ell}\Big) > 0.$$

Theorem 1 (Asymptotic coverage of PART)**.** *Under Assumption 1, the data-dependent tree T, coefficient $\hat{\lambda}$, and estimates constructed above satisfy*

$$\sqrt{n}\big(\hat{\mu}_{T,\hat{\lambda}} - \mu\big) \xrightarrow{d} \mathcal{N}(0,v_0), \qquad n\widehat{\sigma}^2_{T,\hat{\lambda}} \xrightarrow{p} v_0,$$

where $\mu = \mathbb{E}[Y]$. Consequently, for every fixed $\alpha \in (0,1)$,

$$\Pr\Big[\mu \in \Big[\hat{\mu}_{T,\hat{\lambda}} - z_{1-\alpha/2}\widehat{\sigma}_{T,\hat{\lambda}}, \hat{\mu}_{T,\hat{\lambda}} + z_{1-\alpha/2}\widehat{\sigma}_{T,\hat{\lambda}}\Big]\Big] \longrightarrow 1-\alpha.$$

Proof. We break the proof into three steps.

1. Consistency of the selected tree. The law of large numbers and $n/N \to 0$ give $\hat{\lambda} \xrightarrow{p} \lambda^*$. For each feasible population split, condition *(ii)* and uniform laws of large numbers over rectangles ensure that the labeled and unlabeled child proportions converge to the corresponding population probabilities. The averages of $Y, f(X), Y^2, Yf(X), f(X)^2$ within each empirical child also converge to their corresponding population averages, since these quantities are integrable under condition *(i)* and the population children have positive probability [56, Chapter 19]. Together with $\hat{\lambda} \xrightarrow{p} \lambda^*$, this gives convergence of the child residual means and variances to their population counterparts at λ^*. Consequently, the empirical score $n\,\mathrm{VMS}$ converges to the population splitting criterion for each such candidate.

At each splitting node, the candidate set is finite and the population minimizer is unique by condition *(iii)*. The minimum is therefore separated from the other feasible values by a positive gap, so the chosen children differ from the population children by vanishing population mass. Condition *(ii)* also ensures agreement of the stopping decisions. Repeating this argument through the fixed depth gives $\Pr(L = L_0) \to 1$ and, after indexing corresponding leaves in the same order, gives

$$P_X(\mathcal{R}_\ell \triangle \mathcal{R}_{0,\ell}) \xrightarrow{p} 0, \qquad \ell = 1, \dots, L_0,$$

where $\triangle$ denotes symmetric difference. Define the population mean residual in leaf ℓ by

$$\mu_{\ell,\lambda^*} := \mathbb{E}\Big[r^{(\lambda^*)}(X,Y) \mid X \in \mathcal{R}_{0,\ell}\Big].$$

All population leaves have positive mass. The same moment convergence therefore gives

$$\begin{aligned} p_\ell,\ n_\ell/n &\xrightarrow{p} p(\mathcal{R}_{0,\ell}) > 0, \\ \overline{r}_\ell^{(\hat{\lambda})} &\xrightarrow{p} \mu_{\ell,\lambda^*}, \\ \widehat{\sigma}^2_{\ell,\hat{\lambda}} &\xrightarrow{p} \mathrm{Var}\Big(r^{(\lambda^*)}(X,Y) \mid X \in \mathcal{R}_{0,\ell}\Big). \end{aligned} \tag{1}$$

2. Asymptotic normality. The key step is an exact rewriting of the fitted estimator. Since

$$\sum_{\ell=1}^{L} \frac{n_\ell}{n} \overline{r}_\ell^{(\hat{\lambda})} = \frac{1}{n} \sum_{i=1}^{n} \big(y_i - \hat{\lambda} f(x_i)\big),$$

we have

$$\hat{\mu}_{T,\hat{\lambda}} = \frac{1}{n} \sum_{i=1}^{n} y_i + \hat{\lambda} \left[\frac{1}{N} \sum_{j=1}^{N} f(\widetilde{x}_j) - \frac{1}{n} \sum_{i=1}^{n} f(x_i) \right] + \sum_{\ell=1}^{L} \Big(p_\ell - \frac{n_\ell}{n}\Big) \overline{r}_\ell^{(\hat{\lambda})}. \tag{2}$$

Thus PART equals the PPI++ estimator plus a correction for differences between the labeled and unlabeled leaf proportions. To control these differences, define the centered labeled count

$$Z_n(\mathcal{R}) := \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \left(\mathbf{1}\{x_i \in \mathcal{R}\} - p(\mathcal{R})\right).$$

Rectangle indicators in fixed dimension form a bounded VC class and hence are P_X-Donsker [56, Chapter 19]. In particular,

$$\sup_{\mathcal{R}} |Z_n(\mathcal{R})| = O_p(1),$$

so the scaled sampling errors are controlled uniformly over all rectangles, including the fitted leaves. Applying Vaart [56, Lemma 19.24] to the corresponding indicator functions and using the region convergence from Step 1 yields

$$Z_n(\mathcal{R}_\ell) - Z_n(\mathcal{R}_{0,\ell}) = o_p(1).$$

Thus the fitted-leaf centered counts can be replaced by their population-leaf counterparts, even when the fitted leaves are selected using the labeled responses.

Since $n/N \to 0$ and $f(X)$ has finite variance, we have

$$\sqrt{n}\left[\frac{1}{N}\sum_{j=1}^{N} f(\widetilde{x}_j) - \mathbb{E}[f(X)]\right] = \sqrt{\frac{n}{N}}\, O_p(1) = o_p(1). \tag{3a}$$

For the leaf proportions, the analogous uniform count bound for the unlabeled sample gives $\sqrt{N}[p_\ell - p(\mathcal{R}_\ell)] = O_p(1)$, where $p_\ell = \widehat{p}_N(\mathcal{R}_\ell)$. Adding and subtracting the population leaf probability $p(\mathcal{R}_\ell)$, we obtain

$$\begin{aligned}\sqrt{n}\left(p_\ell - \frac{n_\ell}{n}\right) &= \sqrt{n}\big[p_\ell - p(\mathcal{R}_\ell)\big] - \sqrt{n}\left[\frac{n_\ell}{n} - p(\mathcal{R}_\ell)\right] \\ &= -Z_n(\mathcal{R}_\ell) + \sqrt{\frac{n}{N}}\,\sqrt{N}\big[p_\ell - p(\mathcal{R}_\ell)\big] \\ &= -Z_n(\mathcal{R}_\ell) + o_p(1).\end{aligned} \tag{3b}$$

By Step 1, we may work on the event $L = L_0$, with corresponding leaves indexed as above, since this event has probability tending to one. Now, centering (2) gives

$$\begin{aligned}\sqrt{n}\big(\hat{\mu}_{T,\hat{\lambda}} - \mu\big) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}(y_i - \mu) + \hat{\lambda}\sqrt{n}\left[\frac{1}{N}\sum_{j=1}^{N} f(\widetilde{x}_j) - \mathbb{E}[f(X)]\right] \\ - \hat{\lambda}\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\big(f(x_i) - \mathbb{E}[f(X)]\big) + \sum_{\ell=1}^{L_0}\overline{r}_\ell^{(\hat{\lambda})}\sqrt{n}\left(p_\ell - \frac{n_\ell}{n}\right).\end{aligned}$$

Since $\hat{\lambda} = O_p(1)$, $\overline{r}_\ell^{(\hat{\lambda})} = O_p(1)$, and L_0 is fixed, substituting (3) into this expansion yields

$$\sqrt{n}\big(\hat{\mu}_{T,\hat{\lambda}} - \mu\big) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\Big[y_i - \mu - \hat{\lambda}\big(f(x_i) - \mathbb{E}[f(X)]\big)\Big] - \sum_{\ell=1}^{L_0}\overline{r}_\ell^{(\hat{\lambda})} Z_n(\mathcal{R}_\ell) + o_p(1).$$

The centered labeled prediction sum and the Z_n terms are $O_p(1)$. Thus, by the consistency results in Step 1, replacing $\hat{\lambda}$ by λ^* and $\overline{r}_\ell^{(\hat{\lambda})}$ by μ_{ℓ,λ^*} contributes only $o_p(1)$. Using also $Z_n(\mathcal{R}_\ell) - Z_n(\mathcal{R}_{0,\ell}) = o_p(1)$ and $\mathbb{E}[r^{(\lambda^*)}(X,Y)] = \mu - \lambda^*\mathbb{E}[f(X)]$, we obtain

$$\begin{aligned}\sqrt{n}\big(\hat{\mu}_{T,\hat{\lambda}} - \mu\big) &= \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\big[y_i - \mu - \lambda^*\big(f(x_i) - \mathbb{E}[f(X)]\big)\big] - \sum_{\ell=1}^{L_0}\mu_{\ell,\lambda^*} Z_n(\mathcal{R}_\ell) + o_p(1) \\ &= \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\Big[r^{(\lambda^*)}(x_i,y_i) - \mathbb{E}[r^{(\lambda^*)}(X,Y)]\Big] - \sum_{\ell=1}^{L_0}\mu_{\ell,\lambda^*} Z_n(\mathcal{R}_{0,\ell}) + o_p(1).\end{aligned}$$

Finally, expand the definition of Z_n. Since the population leaves form a partition, $\sum_{\ell=1}^{L_0} p(\mathcal{R}_{0,\ell})\mu_{\ell,\lambda^*} = \mathbb{E}[r^{(\lambda^*)}(X,Y)]$, so the centering terms cancel. In particular,

$$\begin{aligned}\sqrt{n}\big(\hat{\mu}_{T,\hat{\lambda}} - \mu\big) &= \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\Big[r^{(\lambda^*)}(x_i,y_i) - \mathbb{E}[r^{(\lambda^*)}(X,Y)]\Big] \\ &\quad - \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\sum_{\ell=1}^{L_0}\mu_{\ell,\lambda^*}\left[\mathbf{1}\{x_i \in \mathcal{R}_{0,\ell}\} - p(\mathcal{R}_{0,\ell})\right] + o_p(1) \\ &= \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left[r^{(\lambda^*)}(x_i,y_i) - \sum_{\ell=1}^{L_0}\mu_{\ell,\lambda^*}\mathbf{1}\{x_i \in \mathcal{R}_{0,\ell}\}\right] + o_p(1).\end{aligned}$$

The summands in the final expression are i.i.d., have mean zero, and have variance v_0. The central limit theorem therefore proves

$$\sqrt{n}(\hat{\mu}_{T,\hat{\lambda}} - \mu) \xrightarrow{d} \mathcal{N}(0, v_0).$$

3. Variance consistency and coverage. Using (1),

$$n\widehat{\sigma}^2_{T,\hat{\lambda}} = \sum_{\ell=1}^{L} \frac{p_\ell^2}{n_\ell/n} \widehat{\sigma}^2_{\ell,\hat{\lambda}} \xrightarrow{p} \sum_{\ell=1}^{L_0} p(\mathcal{R}_{0,\ell}) \operatorname{Var}\Big(r^{(\lambda^*)}(X,Y) \mid X \in \mathcal{R}_{0,\ell}\Big) = v_0.$$

Since $v_0 > 0$, Slutsky's theorem gives

$$\frac{\hat{\mu}_{T,\hat{\lambda}} - \mu}{\widehat{\sigma}_{T,\hat{\lambda}}} \xrightarrow{d} \mathcal{N}(0,1),$$

which proves the stated coverage limit for every fixed $\alpha \in (0,1)$. □

F.3 Estimating Linear Regression Coefficients

We now adapt PART to estimate a linear regression coefficient. The estimator combines an initial coefficient vector computed from model predictions on the unlabeled sample with a correction estimated from labeled prediction errors. As in mean estimation, we compute the correction within tree leaves and combine the leaf corrections using their unlabeled sample proportions. The partition is chosen using a variance criterion for the coefficient of interest. Let

$$\Sigma := \mathbb{E}[XX^\top], \qquad \theta := \Sigma^{-1}\mathbb{E}[XY],$$

where Σ is positive definite. The vector θ is the population least-squares coefficient. We fix a target coordinate $k \in \{1,\dots,d\}$ of θ, which we would like to estimate.

Let $\widetilde{X}$ be the unlabeled design matrix with rows $\widetilde{x}_j^\top$. The empirical second-moment matrix and prediction-based coefficient are defined by

$$\widetilde{\Sigma} := \frac{1}{N}\sum_{j=1}^{N} \widetilde{x}_j \widetilde{x}_j^\top, \qquad \widetilde{\theta} := \widetilde{\Sigma}^{-1}\frac{1}{N}\sum_{j=1}^{N} \widetilde{x}_j f(\widetilde{x}_j),$$

respectively. Here, $\widetilde{\theta}$ is the ordinary least-squares coefficient obtained by regressing the predicted responses $f(\widetilde{x}_j)$ on the unlabeled covariates $\widetilde{x}_j$. The correction needed to recover the target coefficient follows from the decomposition

$$\theta = \Sigma^{-1}\mathbb{E}[Xf(X)] + \Sigma^{-1}\mathbb{E}[X\{Y - f(X)\}].$$

We estimate the first term by $\widetilde{\theta}$. The second term accounts for prediction errors and is estimated using the labeled sample.

Leaf Corrections and Variances. To estimate the residual term in the above decomposition, define the score correction vector for each labeled observation by

$$S_i := x_i\big(y_i - f(x_i)\big).$$

For a region $\mathcal{R}$, let $n_\mathcal{R}$ be its labeled count, and write

$$p_\mathcal{R} := \widehat{p}_N(\mathcal{R}) = \frac{1}{N}\sum_{j=1}^{N} \mathbf{1}\{\widetilde{x}_j \in \mathcal{R}\}, \qquad p(\mathcal{R}) := P_X(\mathcal{R}).$$

Define the mean score and correction vector in the region by

$$\overline{S}_\mathcal{R} := \frac{1}{n_\mathcal{R}} \sum_{i:x_i \in \mathcal{R}} S_i, \qquad \widehat{R}_\mathcal{R} := \widetilde{\Sigma}^{-1}\overline{S}_\mathcal{R}.$$

For $n_\mathcal{R} \geq 2$, the sample covariance of the score vectors and its transformed version are

$$\widehat{M}_\mathcal{R} := \frac{1}{n_\mathcal{R} - 1} \sum_{i:x_i \in \mathcal{R}} (S_i - \overline{S}_\mathcal{R})(S_i - \overline{S}_\mathcal{R})^\top, \qquad \widehat{V}_\mathcal{R} := \widetilde{\Sigma}^{-1}\widehat{M}_\mathcal{R}\widetilde{\Sigma}^{-1}.$$

Thus $\widehat{V}_\mathcal{R}^{(k,k)}$ is the sample variance of the kth coordinates of the transformed scores $\widetilde{\Sigma}^{-1}S_i$ in the region.

Splitting Function for Linear Regression. We use the empirical quantile values in Appendix F.1 as candidate thresholds, together with the same minimum child size $m \geq 2$, deterministic tie-breaking rule, and termination criterion. For a candidate split (j, s) in region $\mathcal{R}$, let $\mathcal{R}_L(j, s)$ and $\mathcal{R}_R(j, s)$ be its children. Among candidates with at least m labeled observations in each child, choose one minimizing the *Variance-Minimizing Split* criterion

$$\mathrm{VMS}_{\mathrm{LR}}(j, s, \mathcal{R}) \coloneqq \sum_{a \in \{L,R\}} p^2_{\mathcal{R}_a(j,s)} \frac{\widehat{V}^{(k,k)}_{\mathcal{R}_a(j,s)}}{n_{\mathcal{R}_a(j,s)}}.$$

Let T be the resulting tree, with leaves $\{\mathcal{R}_\ell\}_{\ell=1}^L$ and $L \leq 2^D$. The PART regression estimator is

$$\widehat{\theta}_T \coloneqq \widetilde{\theta} + \sum_{\ell=1}^{L} p_{\mathcal{R}_\ell} \widehat{R}_{\mathcal{R}_\ell}.$$

We use its kth coordinate $\widehat{\theta}^{(k)}_T$ to estimate $\theta^{(k)}$.

F.3.1 Constructing Confidence Intervals

Define

$$\widehat{\sigma}^2 \coloneqq \sum_{\ell=1}^{L} p^2_{\mathcal{R}_\ell} \frac{\widehat{V}^{(k,k)}_{\mathcal{R}_\ell}}{n_{\mathcal{R}_\ell}}.$$

For a fixed $\alpha \in (0, 1)$, the nominal $(1 - \alpha)$ Wald interval is

$$\left[\widehat{\theta}^{(k)}_T - z_{1-\alpha/2}\widehat{\sigma}, \quad \widehat{\theta}^{(k)}_T + z_{1-\alpha/2}\widehat{\sigma}\right],$$

where $z_{1-\alpha/2}$ is the $(1 - \alpha/2)$ standard normal quantile. We establish asymptotic coverage in Theorem 3.

F.4 Theoretical Coverage Guarantees

For a region $\mathcal{R}$ with positive population probability, define the population correction and covariance by

$$R_{\mathcal{R}} \coloneqq \Sigma^{-1}\mathbb{E}\big[X\big(Y - f(X)\big) \mid X \in \mathcal{R}\big],$$
$$V_{\mathcal{R}} \coloneqq \Sigma^{-1}\,\mathrm{Cov}\big(X\big(Y - f(X)\big) \mid X \in \mathcal{R}\big)\,\Sigma^{-1}.$$

For the fixed coordinate k, let T_0 be the population regression tree obtained by using the corresponding population quantiles as candidate thresholds, admitting splits with two positive-probability children, and minimizing

$$\sum_{a \in \{L,R\}} p(\mathcal{R}_a(j, s))\, V^{(k,k)}_{\mathcal{R}_a(j,s)}$$

at each node. The population procedure uses the same quantile levels and depth limit as its empirical counterpart and stops when no admissible split remains.

Assumption 3 (Regularity for PART linear regression)**.** *The dimension d, maximum depth D, minimum child size $m \geq 2$, finite grid of candidate quantile levels, and target coordinate k are fixed. Assume:*

(i) The labeled observations are i.i.d. from $P_{X,Y}$, the unlabeled observations are i.i.d. from P_X, and the two samples are independent. As $n, N \to \infty$, $n/N \to 0$. Furthermore, Σ is positive definite and

$$\mathbb{E}\big[\|X\|^4 + (1 + \|X\|^2)\big(Y^2 + f(X)^2\big)\big] < \infty.$$

(ii) The candidate-region and feasibility conditions in part (ii) of Assumption 1 hold for the regression population tree T_0 defined above.

(iii) At every splitting node of T_0, the population regression splitting criterion has a unique minimizer among the feasible candidates. Moreover,

$$v_{0,k} \coloneqq \sum_{\mathcal{R} \text{ a leaf of } T_0} p(\mathcal{R}) V^{(k,k)}_{\mathcal{R}} > 0.$$

The moment condition in *(i)* is a sufficient condition for finite variances of the regression score vectors. For bounded covariates, it reduces to $\mathbb{E}[Y^2 + f(X)^2] < \infty$. Condition *(ii)* requires corresponding empirical and population candidate regions to disagree on a set of vanishing population probability, and their feasibility decisions to agree with probability tending to one. The sufficient conditions discussed after Assumption 1 apply here as well. Together with consistency of the split scores, the unique-minimizer requirement in *(iii)* makes the empirical split choices stabilize. The quantity $v_{0,k}$ is the probability-weighted within-leaf variance of the kth coordinate of $\Sigma^{-1}X(Y - f(X))$; its positivity ensures a nondegenerate limiting variance. As in mean estimation, $n/N \to 0$ makes unlabeled sampling uncertainty negligible at the $\sqrt{n}$ scale under the moment condition.

Theorem 3 (Asymptotic coverage of PART for linear regression)**.** *Under Assumption 3, the data-dependent tree T and estimates constructed above satisfy*

$$\sqrt{n}\big(\widehat{\theta}_T^{(k)} - \theta^{(k)}\big) \xrightarrow{d} \mathcal{N}(0, v_{0,k}), \qquad n\widehat{\sigma}^2 \xrightarrow{p} v_{0,k}.$$

Consequently, for every fixed $\alpha \in (0,1)$,

$$\Pr\Big[\theta^{(k)} \in \Big[\widehat{\theta}_T^{(k)} - z_{1-\alpha/2}\widehat{\sigma}, \quad \widehat{\theta}_T^{(k)} + z_{1-\alpha/2}\widehat{\sigma}\Big]\Big] \longrightarrow 1 - \alpha.$$

Proof. We break the proof into three steps.

1. Consistency of the selected tree. The law of large numbers gives $\widetilde{\Sigma} \xrightarrow{p} \Sigma$. Since Σ is positive definite, $\widetilde{\Sigma}$ is invertible with probability tending to one and

$$\widetilde{\Sigma}^{-1} \xrightarrow{p} \Sigma^{-1}.$$

The moment condition makes the coordinates of $X(Y - f(X))$ square-integrable. Uniform laws of large numbers over rectangles therefore apply to their first and second moments [56, Chapter 19]. Together with condition *(ii)*, these laws imply that, for each feasible population split, the empirical child proportions, score means, and score covariance matrices converge to their population counterparts. Consequently, $n\,\mathrm{VMS}_{\mathrm{LR}}$ converges to the population regression splitting criterion.

The candidate set is finite, so each unique population minimizer is separated from the other feasible values by a positive gap. The same finite-depth argument as in the proof of Theorem 1 then gives $\Pr(L = L_0) \to 1$ and

$$P_X(\mathcal{R}_\ell \triangle \mathcal{R}_{0,\ell}) \xrightarrow{p} 0, \qquad \ell = 1, \dots, L_0,$$

where $\{\mathcal{R}_{0,\ell}\}_{\ell=1}^{L_0}$ are the leaves of T_0, indexed to correspond to the fitted leaves. Condition *(ii)* also ensures that the stopping decisions agree. All population leaves have positive mass, and the same moment convergence gives

$$\begin{aligned} p_{\mathcal{R}_\ell}, \quad n_{\mathcal{R}_\ell}/n &\xrightarrow{p} p(\mathcal{R}_{0,\ell}) > 0, \\ \widehat{R}_{\mathcal{R}_\ell} &\xrightarrow{p} R_{\mathcal{R}_{0,\ell}}, \\ \widehat{V}_{\mathcal{R}_\ell} &\xrightarrow{p} V_{\mathcal{R}_{0,\ell}}. \end{aligned} \tag{4}$$

2. Asymptotic normality. As in mean estimation, we rewrite the estimator using the difference between labeled and unlabeled leaf proportions. To begin, observe that

$$\sum_{\ell=1}^{L} \frac{n_{\mathcal{R}_\ell}}{n} \widehat{R}_{\mathcal{R}_\ell} = \widetilde{\Sigma}^{-1} \frac{1}{n} \sum_{i=1}^{n} S_i.$$

Next, we add and subtract the correction weighted by the labeled leaf proportions. Using the preceding identity gives

$$\begin{aligned} \widehat{\theta}_T - \theta &= \widetilde{\theta} - \theta + \sum_{\ell=1}^{L} \frac{n_{\mathcal{R}_\ell}}{n} \widehat{R}_{\mathcal{R}_\ell} + \sum_{\ell=1}^{L} \left(p_{\mathcal{R}_\ell} - \frac{n_{\mathcal{R}_\ell}}{n}\right) \widehat{R}_{\mathcal{R}_\ell} \\ &= \widetilde{\theta} - \theta + \widetilde{\Sigma}^{-1} \frac{1}{n} \sum_{i=1}^{n} S_i + \sum_{\ell=1}^{L} \left(p_{\mathcal{R}_\ell} - \frac{n_{\mathcal{R}_\ell}}{n}\right) \widehat{R}_{\mathcal{R}_\ell}. \end{aligned}$$

We next express $\widetilde{\theta}-\theta$ in terms of the unlabeled sample. Since $\theta=\widetilde{\Sigma}^{-1}\widetilde{\Sigma}\theta$, the definition of $\widetilde{\theta}$ yields

$$\widetilde{\theta}-\theta=\widetilde{\Sigma}^{-1}\left[\frac{1}{N}\sum_{j=1}^{N}\widetilde{x}_j f(\widetilde{x}_j)-\widetilde{\Sigma}\theta\right]=\widetilde{\Sigma}^{-1}\frac{1}{N}\sum_{j=1}^{N}\widetilde{x}_j\big(f(\widetilde{x}_j)-\widetilde{x}_j^\top\theta\big),$$

where the second equality uses $\widetilde{\Sigma}=N^{-1}\sum_{j=1}^{N}\widetilde{x}_j\widetilde{x}_j^\top$.

By Step 1, we may work on the event $L=L_0$, with corresponding leaves indexed as above, since this event has probability tending to one. Substituting this expression and collecting the terms multiplied by $\widetilde{\Sigma}^{-1}$ gives

$$\widehat{\theta}_T-\theta=\widetilde{\Sigma}^{-1}\left[\frac{1}{n}\sum_{i=1}^{n}S_i+\frac{1}{N}\sum_{j=1}^{N}\widetilde{x}_j\big(f(\widetilde{x}_j)-\widetilde{x}_j^\top\theta\big)\right]+\sum_{\ell=1}^{L}\Big(p_{\mathcal{R}_\ell}-\frac{n_{\mathcal{R}_\ell}}{n}\Big)\,\widehat{R}_{\mathcal{R}_\ell}. \tag{5}$$

The final sum accounts for the difference between the unlabeled and labeled leaf proportions. We next analyze the bracketed term in (5). The definition of θ gives

$$\mathbb{E}[X(Y-f(X))]+\mathbb{E}[X(f(X)-X^\top\theta)]=0.$$

Consequently,

$$\mathbb{E}[X(f(X)-X^\top\theta)]=-\mathbb{E}[X(Y-f(X))]=-\mathbb{E}[S_i].$$

Under condition *(i)*, the summands in the second empirical average in (5) have a finite covariance matrix. Since $n/N\to 0$, the $\sqrt{n}$-scaled sampling error of this average is $\sqrt{n/N}\,O_p(1)=o_p(1)$. Replacing it by its expectation therefore yields

$$\begin{aligned}&\sqrt{n}\left[\frac{1}{n}\sum_{i=1}^{n}S_i+\frac{1}{N}\sum_{j=1}^{N}\widetilde{x}_j\big(f(\widetilde{x}_j)-\widetilde{x}_j^\top\theta\big)\right]\\&\qquad=\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\big(S_i-\mathbb{E}[S_i]\big)+o_p(1).\end{aligned}$$

Since $\widetilde{\Sigma}^{-1}=O_p(1)$ by Step 1, the resulting error in the $\sqrt{n}$-scaled decomposition is also $o_p(1)$. Next, define the centered labeled count

$$Z_n(\mathcal{R}):=\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left(\mathbf{1}\{x_i\in\mathcal{R}\}-p(\mathcal{R})\right).$$

The rectangle-count bounds used in the mean proof give

$$\sup_{\mathcal{R}}|Z_n(\mathcal{R})|=O_p(1),\qquad Z_n(\mathcal{R}_\ell)-Z_n(\mathcal{R}_{0,\ell})=o_p(1),$$

where the supremum is over axis-aligned rectangles. The second relation follows from the region convergence in Step 1 and Vaart [56, Lemma 19.24]. The corresponding uniform bound for the unlabeled counts and $n/N\to 0$ give

$$\sqrt{n}\left(p_{\mathcal{R}_\ell}-\frac{n_{\mathcal{R}_\ell}}{n}\right)=-Z_n(\mathcal{R}_\ell)+o_p(1)$$

as in the PART mean proof. We now expand (5) to yield

$$\begin{aligned}\sqrt{n}(\widehat{\theta}_T-\theta)&=\widetilde{\Sigma}^{-1}\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\big(S_i-\mathbb{E}[S_i]\big)+\sum_{\ell=1}^{L_0}\sqrt{n}\left(p_{\mathcal{R}_\ell}-\frac{n_{\mathcal{R}_\ell}}{n}\right)\widehat{R}_{\mathcal{R}_\ell}+o_p(1)\\&=\widetilde{\Sigma}^{-1}\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\big(S_i-\mathbb{E}[S_i]\big)-\sum_{\ell=1}^{L_0}\widehat{R}_{\mathcal{R}_\ell}Z_n(\mathcal{R}_\ell)+o_p(1),\end{aligned}$$

where the second equality uses that $\sqrt{n}(p_{\mathcal{R}_\ell}-n_{\mathcal{R}_\ell}/n)=-Z_n(\mathcal{R}_\ell)+o_p(1)$, together with $\widehat{R}_{\mathcal{R}_\ell}=O_p(1)$ and the fact that L_0 is fixed.

By the moment condition, it holds that $n^{-1/2}\sum_{i=1}^{n}(S_i - \mathbb{E}[S_i]) = O_p(1)$ and the uniform count bound gives $\sup_{\mathcal{R}} |Z_n(\mathcal{R})| = O_p(1)$. Thus, consistency of $\widetilde{\Sigma}^{-1}$ and the fitted leaf corrections allows us to replace them by their population counterparts. We then use $Z_n(\mathcal{R}_\ell) - Z_n(\mathcal{R}_{0,\ell}) = o_p(1)$ to replace the fitted-leaf counts:

$$\begin{aligned}\sqrt{n}(\widehat{\theta}_T - \theta) &= \Sigma^{-1}\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\big(S_i - \mathbb{E}[S_i]\big) - \sum_{\ell=1}^{L_0} R_{\mathcal{R}_{0,\ell}} Z_n(\mathcal{R}_\ell) + o_p(1)\\ &= \Sigma^{-1}\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\big(S_i - \mathbb{E}[S_i]\big) - \sum_{\ell=1}^{L_0} R_{\mathcal{R}_{0,\ell}} Z_n(\mathcal{R}_{0,\ell}) + o_p(1).\end{aligned}$$

Since the population leaves form a partition, the definition of $R_{\mathcal{R}}$ and the law of total expectation give

$$\sum_{\ell=1}^{L_0} p(\mathcal{R}_{0,\ell}) R_{\mathcal{R}_{0,\ell}} = \Sigma^{-1}\mathbb{E}[S_i].$$

Expanding the definition of Z_n and using this identity to cancel the centering terms yields

$$\begin{aligned}\sqrt{n}(\widehat{\theta}_T - \theta) &= \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\Sigma^{-1}\big(S_i - \mathbb{E}[S_i]\big)\\ &\quad - \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\sum_{\ell=1}^{L_0} R_{\mathcal{R}_{0,\ell}}\left[\mathbf{1}\{x_i \in \mathcal{R}_{0,\ell}\} - p(\mathcal{R}_{0,\ell})\right] + o_p(1)\\ &= \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left[\Sigma^{-1}S_i - \sum_{\ell=1}^{L_0} R_{\mathcal{R}_{0,\ell}}\mathbf{1}\{x_i \in \mathcal{R}_{0,\ell}\}\right] + o_p(1).\end{aligned}$$

Taking the kth coordinate therefore gives

$$\sqrt{n}\big(\widehat{\theta}_T^{(k)} - \theta^{(k)}\big) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left[\big(\Sigma^{-1}S_i\big)^{(k)} - \sum_{\ell=1}^{L_0} R_{\mathcal{R}_{0,\ell}}^{(k)}\mathbf{1}\{x_i \in \mathcal{R}_{0,\ell}\}\right] + o_p(1).$$

These summands are i.i.d., have mean zero, and have variance $v_{0,k}$. The central limit theorem proves asymptotic normality.

3. Variance consistency and coverage. By (4),

$$n\widehat{\sigma}^2 = \sum_{\ell=1}^{L}\frac{p_{\mathcal{R}_\ell}^2}{n_{\mathcal{R}_\ell}/n}\widehat{V}_{\mathcal{R}_\ell}^{(k,k)} \xrightarrow{p} \sum_{\ell=1}^{L_0} p(\mathcal{R}_{0,\ell}) V_{\mathcal{R}_{0,\ell}}^{(k,k)} = v_{0,k}.$$

Since $v_{0,k} > 0$, Slutsky's theorem gives

$$\frac{\widehat{\theta}_T^{(k)} - \theta^{(k)}}{\widehat{\sigma}} \xrightarrow{d} \mathcal{N}(0,1),$$

which proves the coverage statement. □

F.5 Estimating Linear Relationship between Income and Sex on Census Data

In this section, we empirically test the performance of PART in a linear regression task. We follow the same testing methodology defined in Section 4. Our goal is to use census data from Ding et al. [22] to estimate a linear relationship between sex (M/F) and individual income. For predictions, we utilize an XGBoost model trained by [5]. We plot the confidence interval width and coverage probabilities in Figure 6. In Table 1, we show that with $n = 600$ labeled samples, PART reduces the interval width by 4.8% while maintaining valid coverage.

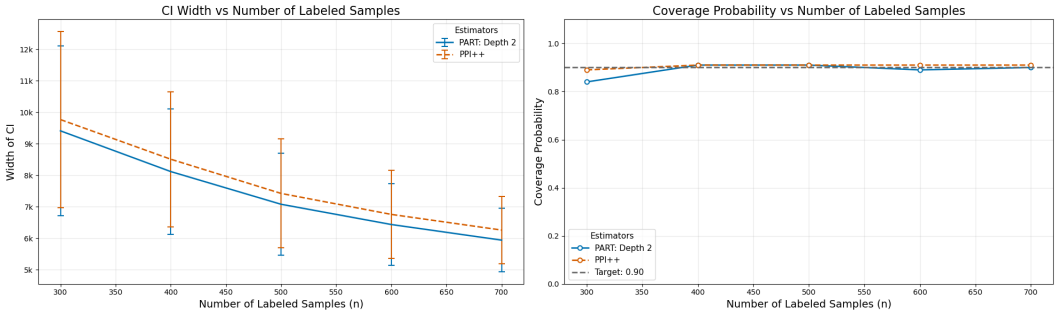


Figure 6: Estimating the linear regression coefficient between income and sex (M/F) on California census data [22].

G PAQ Estimator: Implementation and Analysis

G.1 Estimating the Mean

In this section, we describe the implementation of PAQ, an estimator based on polynomial interpolation that achieves better convergence rates than PPI++ and PART. We derive confidence intervals and establish coverage guarantees using a general polynomial interpolation framework. The result for linear interpolation under uniform features, described in Theorem 2, follows as a special case. We begin with some assumptions on the residual function that help control the interpolation error.

Assumption 4 (Regularity for Polynomial PAQ)**.** *Fix an integer $p \geq 1$. Assume:*

(i) The labeled observations are i.i.d. from $P_{X,Y}$, the unlabeled observations are i.i.d. from P_X, and the two samples are independent.

(ii) The distribution P_X has a density ρ supported on a compact interval $[a, b]$, where $a < b$, with $\rho(x) \geq c > 0$ for almost every $x \in [a, b]$.

(iii) There is a deterministic function y such that $Y = y(X)$ almost surely. The residual function $r(x) = y(x) - f(x)$ belongs to $C^{p+1}([a, b])$, and $0 < \mathrm{Var}(Y) < \infty$.

Condition *(i)* is standard for mean estimation. Condition *(ii)* ensures that the labeled features become dense throughout the interval as the sample size grows. The estimator does not require knowledge of the feature density or its lower bound. Condition *(iii)* assumes deterministic responses and requires $p + 1$ continuous derivatives of the residual, which allows us to bound the error of degree-p interpolation. As expected, higher degree interpolation requires stronger smoothness assumptions.

Local Polynomial Interpolation. We begin by defining our interpolation procedure and the PAQ estimator. Let $n \geq p + 2$ and $N \geq 2$. We sort the labeled features as $x_{(1)} < \cdots < x_{(n)}$ and write $r_i = r(x_{(i)})$. For each $i = 1, \ldots, n - 1$, set

$$j_i := \min\{i, n - p\}.$$

On the interval $[x_{(i)}, x_{(i+1)}]$, define q_p to equal the unique polynomial of degree at most p that interpolates the residuals at the $p + 1$ consecutive labeled features

$$x_{(j_i)}, \ldots, x_{(j_i+p)}.$$

For $x < x_{(1)}$, use the polynomial through the first $p+1$ labeled features; for $x > x_{(n)}$, use the polynomial through the last $p + 1$ labeled features. Thus, q_p is a piecewise polynomial interpolant with polynomial extrapolation at the two boundaries. For $p = 1$, this is exactly the linear construction in Section 3.2. Define

$$Z_j^p := f(\widetilde{x}_j) + q_p(\widetilde{x}_j), \qquad \widehat{\mu}_{\mathrm{PAQ}}^p := \frac{1}{N} \sum_{j=1}^{N} Z_j^p,$$

and

$$(\widehat{\sigma}_{\mathrm{PAQ}}^p)^2 := \frac{1}{N(N-1)} \sum_{j=1}^{N} (Z_j^p - \widehat{\mu}_{\mathrm{PAQ}}^p)^2.$$

The rectified point Z_j^p is formed by taking the initial prediction $f(\widetilde{x}_j)$ and correcting it with an interpolated value given by $q_p(\widetilde{x}_j)$.

Interpolation Error and Confidence Intervals. We next construct a bound for the interpolation error we accrue from using $q_p(\widetilde{x}_j)$ in PAQ. This bound will be used to help form confidence intervals for PAQ. Let $t_0(x), \ldots, t_p(x)$ be the nodes used to compute $q_p(x)$. The Lagrange remainder formula gives

$$r(x) - q_p(x) = \frac{r^{(p+1)}(\xi_x)}{(p+1)!} \prod_{j=0}^{p} \big(x - t_j(x)\big),$$

where ξ_x lies in the smallest interval containing x and its interpolation nodes. Thus, writing

$$L_{p+1} \coloneqq \|r^{(p+1)}\|_\infty, \qquad w_p(x) \coloneqq \frac{1}{(p+1)!} \prod_{j=0}^{p} |x - t_j(x)|,$$

we obtain the pointwise error bound

$$|r(x) - q_p(x)| \le L_{p+1} w_p(x).$$

The factor $w_p(x)$ is directly computable from the selected nodes. The derivative bound L_{p+1} is unknown and hence we must estimate it. To estimate L_{p+1}, we use divided differences, which generalize the slope between two nodes to higher orders and accommodate unequal node spacing. In particular, we may compute the divided difference recursively from the observed residuals:

$$r[t_0] = r(t_0), \qquad r[t_0, \ldots, t_k] = \frac{r[t_1, \ldots, t_k] - r[t_0, \ldots, t_{k-1}]}{t_k - t_0}.$$

By the mean-value theorem, it holds that

$$(p+1)!\, r[t_0, \ldots, t_{p+1}] = r^{(p+1)}(\eta)$$

for some η between the nodes. This motivates taking the largest absolute divided difference over blocks of $p+2$ consecutive labeled features and inflating it by a factor κ. Specifically,

$$\widehat{L}_{p+1} \coloneqq \kappa(p+1)! \max_{1 \le i \le n-p-1} \left| r[x_{(i)}, \ldots, x_{(i+p+1)}] \right|, \qquad \kappa > 1,$$

where κ provides a margin that ensures $\Pr(\widehat{L}_{p+1} \ge L_{p+1}) \to 1$. For $p = 1$, the previous expression reduces to the expression for $\widehat{L}_2$ in the main body. We now average the resulting pointwise error allowances over the unlabeled sample, yielding

$$\widehat{B}_p \coloneqq \frac{\widehat{L}_{p+1}}{N} \sum_{j=1}^{N} w_p(\widetilde{x}_j).$$

On the event $\widehat{L}_{p+1} \ge L_{p+1}$, this quantity bounds the average interpolation error,

$$\left| \frac{1}{N} \sum_{j=1}^{N} \big(q_p(\widetilde{x}_j) - r(\widetilde{x}_j)\big) \right| \le \widehat{B}_p.$$

Consequently, we can add this bound to the standard z-interval to produce the interval

$$C_{n,N}^p \coloneqq \left[\widehat{\mu}_{\mathrm{PAQ}}^p - z_{1-\alpha/2} \widehat{\sigma}_{\mathrm{PAQ}}^p - \widehat{B}_p, \quad \widehat{\mu}_{\mathrm{PAQ}}^p + z_{1-\alpha/2} \widehat{\sigma}_{\mathrm{PAQ}}^p + \widehat{B}_p \right].$$

Below, we show that this interval achieves the desired coverage asymptotically. Furthermore, we also demonstrate improved convergence rates over both PART and PPI.

Theorem 4 (Polynomial PAQ under nonuniform sampling)**.** *Under Assumption 4, as $n, N \to \infty$, with fixed $\kappa > 1$ and $\alpha \in (0, 1)$,*

$$\left| \mathbb{E}[\widehat{\mu}_{\mathrm{PAQ}}^p] - \mu \right| = O(n^{-(p+1)}), \qquad \mathbb{E}\left[(\widehat{\mu}_{\mathrm{PAQ}}^p - \mu)^2 \right] = O(N^{-1} + n^{-2p-2}).$$

The same bound holds for the variance. Moreover,

$$\liminf_{n,N \to \infty} \Pr(\mu \in C_{n,N}^p) \ge 1 - \alpha, \qquad |C_{n,N}^p| = O_p(N^{-1/2} + n^{-(p+1)}).$$

Proof. Let $e_p(x) \coloneqq q_p(x) - r(x)$. Since $f + r = y$, the estimation error decomposes as

$$\widehat{\mu}^p_{\mathrm{PAQ}} - \mu = \left[\frac{1}{N} \sum_{j=1}^{N} y(\widetilde{x}_j) - \mu \right] + \frac{1}{N} \sum_{j=1}^{N} e_p(\widetilde{x}_j).$$

We first control the interpolation term, then use this decomposition to establish the estimation rates and confidence interval coverage. We break the proof into 4 steps.

1. Bounding expected interpolation error.

Recall that the remainder bound gives

$$|e_p(x)| \le L_{p+1} w_p(x), \qquad x \in [a, b].$$

Set $x_{(0)} = a$ and $x_{(n+1)} = b$, and consider the intervals

$$I_i = [x_{(i)}, x_{(i+1)}], \qquad i = 0, \dots, n.$$

For each I_i, let J_i be the smallest interval containing I_i and the interpolation nodes used there. The consecutive-node construction ensures that J_i spans at most $p + 1$ consecutive gaps, including on the two boundary intervals where extrapolation is used. Write

$$H_i \coloneqq |J_i|, \qquad V_i \coloneqq \int_{J_i} \rho(x)\, dx.$$

Because $\rho \ge c > 0$, we have $H_i \le V_i/c$. Moreover, for $x \in I_i$, every factor $|x - t_j(x)|$ in $w_p(x)$ is at most H_i. To control the mean-squared interpolation error, we bound the integrated square of w_p. Indeed, the remainder inequality gives $e_p(x)^2 \le L_{p+1}^2 w_p(x)^2$, so a bound on $\int_a^b w_p(x)^2 \rho(x)\, dx$ also controls the average squared error of the interpolant. In particular, by the previous bounds, it holds that

$$\begin{aligned}
\int_{I_i} w_p(x)^2 \rho(x)\, dx &= \int_{I_i} \left[\frac{1}{(p+1)!} \prod_{j=0}^{p} |x - t_j(x)| \right]^2 \rho(x)\, dx \\
&\le \int_{I_i} \left[\frac{H_i^{p+1}}{(p+1)!} \right]^2 \rho(x)\, dx \\
&= \frac{H_i^{2p+2}}{((p+1)!)^2} \int_{I_i} \rho(x)\, dx \\
&\le \frac{(V_i/c)^{2p+2}}{((p+1)!)^2}\, V_i \\
&= \frac{V_i^{2p+3}}{c^{2p+2}((p+1)!)^2}.
\end{aligned}$$

To bound the expectation of this quantity, we apply the probability integral transform. Let F be the distribution function of X, and set

$$U_j \coloneqq F(x_j), \qquad j = 1, \dots, n.$$

Since F is continuous, the transformed features $U_1, \dots, U_n$ are independent $\mathrm{Unif}[0, 1]$ random variables. Write $U_{(j)} = F(x_{(j)})$ for their order statistics, and set

$$U_{(0)} = 0, \qquad U_{(n+1)} = 1.$$

For our consecutive-node construction, write

$$J_i = [x_{(s_i)}, x_{(s_i + k_i)}], \qquad 1 \le k_i \le p + 1,$$

where the indices are fixed by the interpolation rule. The probability mass of J_i then becomes

$$\begin{aligned} V_i &= \int_{x_{(s_i)}}^{x_{(s_i+k_i)}} \rho(x)\,dx \\ &= F(x_{(s_i+k_i)}) - F(x_{(s_i)}) \\ &= U_{(s_i+k_i)} - U_{(s_i)} \\ &= \sum_{j=s_i}^{s_i+k_i-1} \big(U_{(j+1)} - U_{(j)}\big). \end{aligned}$$

Thus, V_i is the sum of a fixed block of k_i consecutive spacings from a uniform sample. The $n+1$ uniform spacings, including the two boundary spacings, jointly have a Dirichlet distribution with all parameters equal to one. By the aggregation property of the Dirichlet distribution, grouping the k_i spacings in this block and the remaining $n+1-k_i$ spacings gives

$$V_i \sim \mathrm{Beta}(k_i, n+1-k_i).$$

Since p is fixed, the beta moment formula gives

$$\mathbb{E}[V_i^{2p+3}] = \prod_{m=0}^{2p+2} \frac{k_i+m}{n+1+m} = O(n^{-2p-3}),$$

uniformly in i. Summing over the $n+1$ intervals and using the pointwise remainder bound yields

$$\mathbb{E}\left[\int_a^b w_p(x)^2 \rho(x)\,dx\right] = O(n^{-2p-2}),$$

and hence it follows that

$$\mathbb{E}\left[\int_a^b e_p(x)^2 \rho(x)\,dx\right] = O(n^{-2p-2}).$$

2. Bounding bias, mean-squared error, and variance. We now apply the previous bounds to control the error and variance of PAQ. Conditional on the labeled sample, e_p is fixed and the unlabeled features remain independent draws from ρ. Jensen's inequality and the preceding bound therefore give

$$\mathbb{E}\left[\left(\frac{1}{N}\sum_{j=1}^{N} e_p(\widetilde{x}_j)\right)^2\right] \le \mathbb{E}\left[\frac{1}{N}\sum_{j=1}^{N} e_p(\widetilde{x}_j)^2\right] = \mathbb{E}\left[\int_a^b e_p(x)^2\rho(x)\,dx\right] = O(n^{-2p-2}).$$

The first term in the estimation-error decomposition has mean zero and variance $\mathrm{Var}(Y)/N$. Hence, by Cauchy–Schwarz,

$$\begin{aligned} \left|\mathbb{E}[\widehat{\mu}^p_{\mathrm{PAQ}}] - \mu\right| &\le \left[\mathbb{E}\left[\left(\frac{1}{N}\sum_{j=1}^{N} e_p(\widetilde{x}_j)\right)^2\right]\right]^{1/2} \\ &= O(n^{-(p+1)}). \end{aligned}$$

Applying $(u+v)^2 \le 2u^2 + 2v^2$ to the same decomposition gives

$$\begin{aligned} \mathbb{E}\Big[(\widehat{\mu}^p_{\mathrm{PAQ}} - \mu)^2\Big] &\le \frac{2\,\mathrm{Var}(Y)}{N} + 2\mathbb{E}\left[\left(\frac{1}{N}\sum_{j=1}^{N} e_p(\widetilde{x}_j)\right)^2\right] \\ &= O(N^{-1} + n^{-2p-2}). \end{aligned}$$

The variance satisfies the same bound because it is no larger than the mean-squared error.

3. Estimating the derivative bound. We next show that $\widehat{L}_{p+1}$ is at least L_{p+1} with probability tending to one, which justifies its use in the interpolation allowance $\widehat{B}_p$. Write

$$M_{p+1} \coloneqq \frac{\widehat{L}_{p+1}}{\kappa}.$$

The mean-value theorem for divided differences gives

$$0 \le M_{p+1} \le L_{p+1}.$$

For any $\varepsilon > 0$, continuity of $r^{(p+1)}$ implies that there is an interval of positive length on which

$$|r^{(p+1)}(x)| > L_{p+1} - \varepsilon.$$

This interval has positive probability mass because $\rho \ge c > 0$. It therefore contains a block of $p+2$ consecutive labeled features with probability tending to one. Applying the mean-value theorem to that block gives

$$M_{p+1} \ge L_{p+1} - \varepsilon$$

on this event. Thus,

$$M_{p+1} \xrightarrow{p} L_{p+1}.$$

If $L_{p+1} > 0$, the fixed inflation factor $\kappa > 1$ consequently ensures

$$\Pr(\widehat{L}_{p+1} \ge L_{p+1}) \longrightarrow 1.$$

If $L_{p+1} = 0$, the same conclusion holds because $\widehat{L}_{p+1} = 0$ for every sample. Since $M_{p+1} \le L_{p+1}$ and $\widehat{L}_{p+1} = \kappa M_{p+1}$, we also have

$$0 \le \widehat{L}_{p+1} \le \kappa L_{p+1}.$$

We will use this upper bound to control the size of $\widehat{B}_p$ when establishing the rate for the confidence interval length.

4. Confidence interval coverage.

We first verify that the estimated standard error has the correct asymptotic scale. The bound in Step 2 and Markov's inequality imply

$$\frac{1}{N}\sum_{j=1}^{N} e_p(\widetilde{x}_j)^2 = o_p(1).$$

Since

$$Z_j^p = f(\widetilde{x}_j) + q_p(\widetilde{x}_j) = y(\widetilde{x}_j) + e_p(\widetilde{x}_j),$$

Cauchy–Schwarz and the law of large numbers give

$$\begin{aligned}\left|\frac{1}{N}\sum_{j=1}^{N}(Z_j^p)^2 - \frac{1}{N}\sum_{j=1}^{N} y(\widetilde{x}_j)^2\right| &\le 2\left(\frac{1}{N}\sum_{j=1}^{N} y(\widetilde{x}_j)^2\right)^{1/2}\left(\frac{1}{N}\sum_{j=1}^{N} e_p(\widetilde{x}_j)^2\right)^{1/2} + \frac{1}{N}\sum_{j=1}^{N} e_p(\widetilde{x}_j)^2 \\ &= o_p(1).\end{aligned}$$

Together with the consistency of $\widehat{\mu}^p_{\mathrm{PAQ}}$ established in Step 2, this yields

$$N(\widehat{\sigma}^p_{\mathrm{PAQ}})^2 = \frac{N}{N-1}\left[\frac{1}{N}\sum_{j=1}^{N}(Z_j^p)^2 - (\widehat{\mu}^p_{\mathrm{PAQ}})^2\right] \xrightarrow{p} \mathrm{Var}(Y) > 0.$$

The central limit theorem applied to the average of $y(\widetilde{x}_j)$, followed by Slutsky's theorem, now gives

$$\frac{\frac{1}{N}\sum_{j=1}^{N} y(\widetilde{x}_j) - \mu}{\widehat{\sigma}^p_{\mathrm{PAQ}}} \Rightarrow \mathcal{N}(0,1).$$

On the event $\widehat{L}_{p+1} \geq L_{p+1}$, the interpolation allowance bounds the difference between this average and the PAQ estimator as

$$\left| \widehat{\mu}^p_{\mathrm{PAQ}} - \frac{1}{N} \sum_{j=1}^{N} y(\widetilde{x}_j) \right| \leq \widehat{B}_p.$$

The triangle inequality therefore implies

$$\Pr(\mu \in C^p_{n,N}) \geq \Pr\left(\left| \frac{1}{N} \sum_{j=1}^{N} y(\widetilde{x}_j) - \mu \right| \leq z_{1-\alpha/2} \widehat{\sigma}^p_{\mathrm{PAQ}} \right) - \Pr(\widehat{L}_{p+1} < L_{p+1}).$$

The first probability converges to $1-\alpha$, and the second converges to zero, proving the coverage claim. Finally, the deterministic bound $\widehat{L}_{p+1} \leq \kappa L_{p+1}$, Jensen's inequality, and Step 1 give

$$\begin{aligned} \mathbb{E}[\widehat{B}_p^2] &\leq \kappa^2 L_{p+1}^2 \mathbb{E}\left[\frac{1}{N} \sum_{j=1}^{N} w_p(\widetilde{x}_j)^2 \right] \\ &= \kappa^2 L_{p+1}^2 \mathbb{E}\left[\int_a^b w_p(x)^2 \rho(x)\, dx \right] \\ &= O(n^{-2p-2}). \end{aligned}$$

Thus $\widehat{B}_p = O_p(n^{-(p+1)})$. Combining this with $\widehat{\sigma}^p_{\mathrm{PAQ}} = O_p(N^{-1/2})$, we obtain

$$|C^p_{n,N}| = 2 z_{1-\alpha/2} \widehat{\sigma}^p_{\mathrm{PAQ}} + 2\widehat{B}_p = O_p(N^{-1/2} + n^{-(p+1)}).$$

□

It is straightforward to apply the previous result to establish Theorem 2 from the main body.

Theorem 2 (Linear PAQ under uniform sampling)**.** *Under Assumption 2, as $n, N \to \infty$, with fixed $\kappa > 1$ and $\alpha \in (0,1)$, the estimator satisfies*

$$|\mathbb{E}[\widehat{\mu}_{\mathrm{PAQ}}] - \mu| = O(n^{-2}), \qquad \mathbb{E}\big[(\widehat{\mu}_{\mathrm{PAQ}} - \mu)^2\big] = O(N^{-1} + n^{-4}),$$

where $\mu = \mathbb{E}[Y]$. The same bound holds for its variance. Moreover,

$$\liminf_{n,N \to \infty} \Pr(\mu \in C_{n,N}) \geq 1 - \alpha, \qquad |C_{n,N}| = O_p(N^{-1/2} + n^{-2}),$$

where $|C_{n,N}|$ denotes the interval length.

Proof. First, it is clear that Assumption 2 implies Assumption 4. Hence, we may apply Theorem 4 with $p = 1$, $[a,b] = [0,1]$, and $\rho = 1$ to complete the proof. □

G.2 Synthetic Experiment for the PAQ Estimator

In this section, we empirically test the PAQ estimator against PPI++ on a synthetic dataset that meets the assumptions of Theorem 4—specifically, that residuals are deterministic and twice continuously differentiable. In each of 100 trials, we independently generate $M = 10{,}000$ observations with $X \sim \mathrm{Unif}[0, \pi]$, $Y = X^2 + \sin(X)$, and $f(X) = X^2$. For each labeled sample size $n \in \{100, 175, 250, 325, 400\}$, we randomly select n labeled observations and use the remaining $N = M - n$ observations as unlabeled data. We construct nominal 90% intervals, use $\kappa = 2$ for PAQ, and evaluate coverage against $\mu = \pi^2/3 + 2/\pi$. Width error bars show one standard deviation. As illustrated in Figure 7, PAQ yields tighter intervals than PPI++ while maintaining nominal coverage.

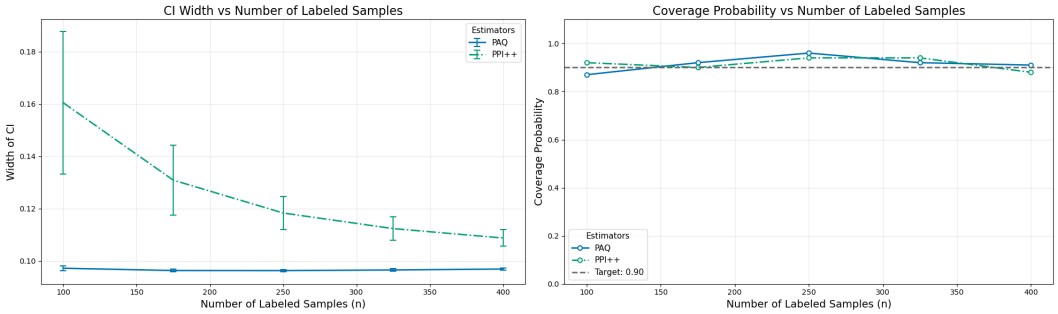


Figure 7: Estimating the mean on a synthetic dataset.