PREDICTION-POWERED INFERENCE

DOWOO KIM
MENTOR: YANEES DOBBERSTEIN
DIRECTED READING PROGRAM WINTER 2025
MCGILL UNIVERSITY MATHEMATICS & STATISTICS

Contents

1 Introduction

Prediction-powered inference (PPI) bridges the gap between modern predictive modeling and classical statistical inference. While machine learning models—especially neural networks—excel at capturing

complex patterns in high-dimensional data, they often lack mechanisms for valid uncertainty quantification. Classical inference, by contrast, guarantees coverage but requires restrictive assumptions and sufficient labeled data.

PPI addresses this tension by introducing the *rectifier*, an empirical correction term that calibrates predictions using a small labeled dataset. This allows complex prediction models to be integrated into valid inferential procedures without sacrificing statistical rigor.

In this project, we explore the theoretical foundations and algorithmic formulations of PPI, focusing on mean estimation, quantile inference, and linear and logistic regression. We present efficient confidence interval constructions that leverage both labeled and unlabeled data and show how PPI generalizes to convex estimation problems.

Our goal is to summarize this paper: `https://arxiv.org/abs/2301.09633` and demonstrate how prediction-powered inference enables valid and efficient estimation even under model complexity, limited supervision, or distributional shift—making it broadly applicable to fields such as genomics, ecology, and population health. All of codes of generated graphs are available here: `https://github.com/dk1028/DRP/tree/main/codes` and the poster for 2025 Undergraduate Science Showcase is available here: `https://www.mcgill.ca/ose/files/ose/prediction-powered_inference_evaluating_efficiency_in_genomic_data_analysis.pdf`.

2 Fundamentals and analysis

2.1 Prediction-Powered Estimation: From Classical to Corrected Inference

Let $(X, Y) = \{(X_1, Y_1), \ldots, (X_n, Y_n)\}$ be labeled data and $(\tilde{X}) = \{\tilde{X}_1, \ldots, \tilde{X}_N\}$ unlabeled features, with predictions $f(X_i)$ and $f(\tilde{X}_i)$ from a model f. We will use this notation for the setup of our procedure for the following.

First, we introduce how PPI works with a simple example where the investigator is interested in the mean of some population. The *classical estimator* for the mean in a sample of size n is

$$\hat{\theta}_{\text{class}} = \frac{1}{n} \sum_{i=1}^{n} Y_i. \tag{2.1}$$

This estimator is unbiased but may have high variance for small n.

The *imputation estimator*, which uses predictions only, is

$$\hat{\theta}_f = \frac{1}{N}\sum_{i=1}^{N} f(\tilde{X}_i), \tag{2.2}$$

efficient but potentially biased if the prediction function f is not well-calibrated or systematically deviates from true labels. See, for example, [LR02; BM02] for discussions on predictive bias in statistical learning.

The classical estimator ignores data for which we do not have labels but have predicted labels—potentially leading to unnecessary variance. On the other hand, if we use imputation, we implicitly rely on the quality of our predictions.

To correct this bias, prediction-powered inference introduces the *rectifier*:

$$\Delta = \frac{1}{n}\sum_{i=1}^{n} \left(f(X_i) - Y_i\right). \tag{2.3}$$

The *prediction-powered estimator* is then defined by

$$\hat{\theta}_{\text{PP}} = \hat{\theta}_f - \Delta = \frac{1}{N}\sum_{i=1}^{N} f(\tilde{X}_i) - \frac{1}{n}\sum_{i=1}^{n}(f(X_i) - Y_i). \tag{2.4}$$

which is both unbiased and efficient when $N \gg n$ and prediction accuracy is moderate or high [GL21].

This framework generalizes to convex problems where $\mathbb{E}[g_\theta(X, Y)] = 0$. The corresponding rectifier is

$$\Delta_\theta = \mathbb{E}[g_\theta(X, Y) - g_\theta(X, f(X))]. \tag{2.5}$$

A corresponding 95% confidence interval around (2.4) is

$$\hat{\theta}_{\text{PP}} \pm 1.96\sqrt{\frac{\hat{\sigma}^2_{f-Y}}{n} + \frac{\hat{\sigma}^2_f}{N}}, \tag{2.6}$$

where $\hat{\sigma}^2_{f-Y}$ is the variance of the rectifier and $\hat{\sigma}^2_f$ is the variance of predicted values. This interval retains correct coverage by the Central Limit Theorem, as shown in [GL21].

2.2 Comparative Analysis of Estimators

We summarize three estimators:

- **Classical:** Equation (2.1) – unbiased, but high variance for small n.

- **Imputation:** Equation (2.2) – efficient, but potentially biased.
- **Prediction-powered:** Equation (2.4) – combines efficiency and unbiasedness.

As illustrated in Figure 1, the prediction-powered estimator consistently yields narrower confidence intervals while maintaining valid coverage. It outperforms classical methods and corrects the overconfidence issues seen in naïve imputation [CSZ06].

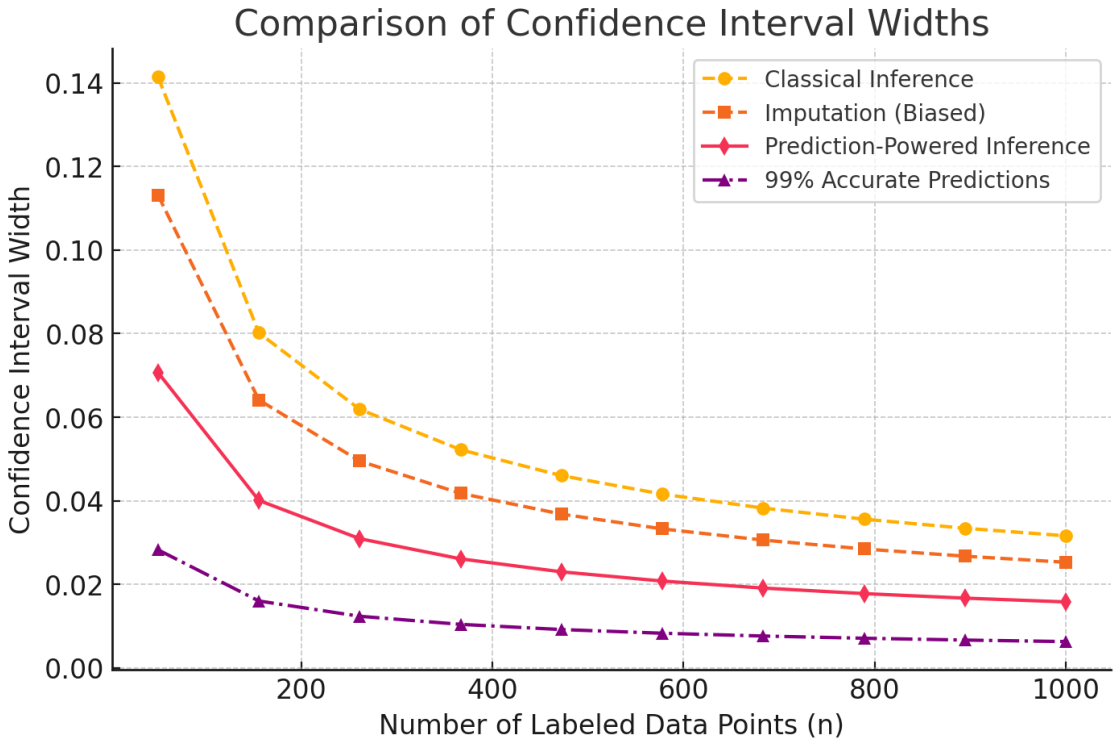

FIGURE 1. Comparison of Confidence Interval Widths. Random genotype data (10 SNPs per individual) are simulated and a continuous phenotype is generated via a linear model with added noise.

2.3 Effect of Prediction Accuracy on Prediction-Powered Confidence Intervals

Building on the mean estimator introduced earlier on (2.4), we examine how its accuracy and variance behave under varying prediction quality.

The accuracy of the predictive model f directly affects the variance of the PPI estimator:

$$\mathrm{Var}(\hat{\theta}_{\mathrm{PP}}) = \frac{\mathrm{Var}(f(X))}{N} + \frac{\mathrm{Var}(f(X) - Y)}{n}. \tag{2.7}$$

As prediction accuracy improves, the second term decreases, leading to narrower confidence intervals.

The associated confidence interval is:

$$\hat{\theta}_{\text{PP}} \pm z_{1-\alpha/2}\sqrt{\frac{\hat{\sigma}_f^2}{N} + \frac{\hat{\sigma}_{f-Y}^2}{n}}, \tag{2.8}$$

where $\hat{\sigma}_f^2$ and $\hat{\sigma}_{f-Y}^2$ are the empirical variances of predictions and prediction errors, respectively.

If the predictor is unbiased in the sense that $\mathbb{E}[f(X)] = \mathbb{E}[Y]$, the rectifier ensures unbiasedness of the estimator, and the confidence interval remains valid:

$$\mathbb{E}[f(X) - Y] = 0 \quad \Rightarrow \quad \mathbb{E}[\hat{\Delta}] = 0. \tag{2.9}$$

However, if the predictor is biased (i.e., $\mathbb{E}[f(X)] \neq \mathbb{E}[Y]$), then $\hat{\Delta}$ may not fully correct the bias. Thus, the PPI estimator is only guaranteed to be unbiased in the zero-bias case. As long as the variance of the prediction error is small, the estimator remains efficient and valid under the Central Limit Theorem assumptions [GL21].

Illustration: Consider a 1D example with $\theta^* = 0$. As model accuracy decreases from 100% to 50%, confidence intervals widen but remain valid. The table below summarizes this effect:

Scenario	Model Accuracy	PP Confidence Interval Width
Perfect prediction	100%	Narrow (smallest possible)
Good prediction	80%	Moderately narrow
Poor prediction	50% or less	Wide

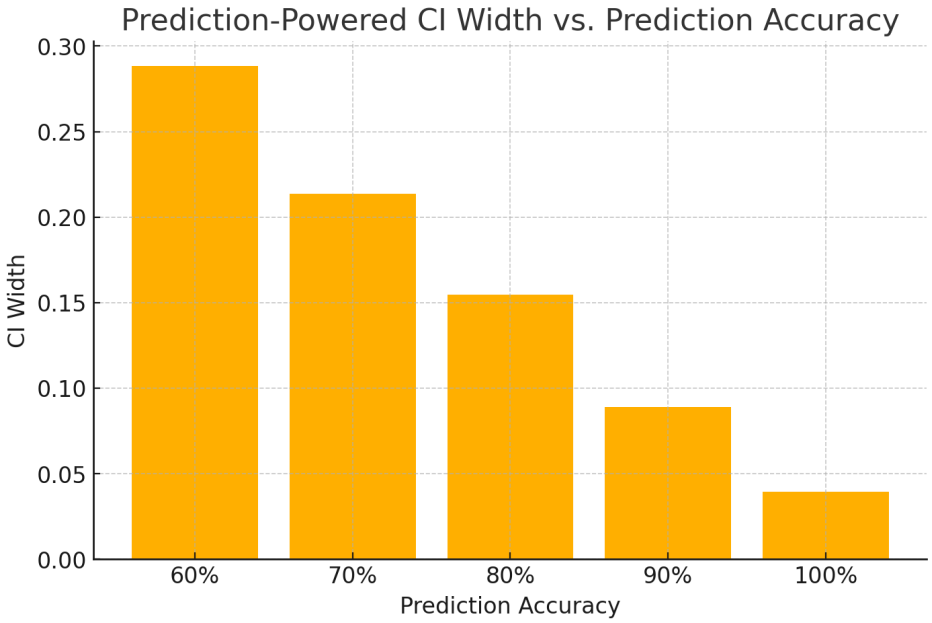

FIGURE 2. Prediction-powered estimates with 95% confidence intervals under varying prediction accuracy.

In summary, PPI adapts to prediction quality: better models yield tighter intervals, while validity is preserved even when predictions are noisy.

3 Algorithms and inferences

3.1 Mean Estimation

Problem. Estimate the population mean $\theta^* = \mathbb{E}[Y]$, which minimizes expected squared loss:

$$\theta^* = \arg\min_{\theta \in \mathbb{R}} \mathbb{E}\left[\tfrac{1}{2}(Y-\theta)^2\right]. \tag{3.1}$$

Estimators.

- *Classical:* $\hat{\theta}_{\text{class}} = \frac{1}{n}\sum Y_i$ — unbiased but high variance.
- *Imputation:* $\tilde{\theta}_f = \frac{1}{N}\sum f(\tilde{X}_i)$ — efficient but biased.
- *Prediction-powered:*

$$\hat{\theta}_{PP} = \tilde{\theta}_f - \hat{\Delta} = \frac{1}{N}\sum f(\tilde{X}_i) - \frac{1}{n}\sum (f(X_i) - Y_i). \tag{3.2}$$

This correction ensures unbiasedness: $\mathbb{E}[\hat{\theta}_{PP}] = \mathbb{E}[Y]$.

Confidence Interval. Since $\hat{\theta}_{PP}$ is a sum of two independent terms, a valid $(1-\alpha)$ interval is:

$$C_\alpha^{PP} = \hat{\theta}_{PP} \pm z_{1-\alpha/2}\sqrt{\frac{\hat{\sigma}_{f-Y}^2}{n} + \frac{\hat{\sigma}_f^2}{N}}, \tag{3.3}$$

where $\hat{\sigma}_{f-Y}^2$ and $\hat{\sigma}_f^2$ are sample variances of the rectifier and prediction terms, respectively.

Algorithm 1 Prediction-Powered Mean Estimation

Require: Labeled data (X_i, Y_i), unlabeled $X_j^{\sim}$, predictor f, level $1-\alpha$

1: Compute $\tilde{\theta}_f = \frac{1}{N}\sum f(X_j^{\sim})$
2: Compute $\hat{\Delta} = \frac{1}{n}\sum(f(X_i) - Y_i)$
3: Form $\hat{\theta}_{PP} = \tilde{\theta}_f - \hat{\Delta}$
4: Estimate variances $\hat{\sigma}_f^2$, $\hat{\sigma}_{f-Y}^2$
5: Compute margin $w = z_{1-\alpha/2}\sqrt{\hat{\sigma}_{f-Y}^2/n + \hat{\sigma}_f^2/N}$
6: **return** $[\hat{\theta}_{PP} - w, \hat{\theta}_{PP} + w]$

3.2 Linear Regression

Problem. Estimate $\theta^* \in \mathbb{R}^d$ minimizing squared error:

$$\theta^* = \arg\min_\theta \mathbb{E}[(Y - X^\top\theta)^2]. \tag{3.4}$$

Estimators.

- *Classical:* $\hat{\theta}_{\text{OLS}} = (X^\top X)^{-1}X^\top Y$
- *Imputed:* $\tilde{\theta}_f = (X^{\sim\top}X^{\sim})^{-1}X^{\sim\top}f(X^{\sim})$
- *Prediction-powered:*

$$\hat{\theta}_{PP} = \tilde{\theta}_f - \hat{\Delta}, \quad \hat{\Delta} = (X^\top X)^{-1}X^\top(f(X) - Y) \tag{3.5}$$

Inference. Let residuals $\varepsilon_j^{\sim} = f(X_j^{\sim}) - X_j^{\sim\top}\tilde{\theta}_f$ and $\varepsilon_i = f(X_i) - Y_i - X_i^\top\hat{\Delta}$. The covariance estimate is:

$$\hat{V} = A_f + A_r, \tag{3.6}$$

where each term represents the contribution from imputation and rectification:

$$A_f = (X^{\sim\top} X^{\sim})^{-1} \left(\frac{1}{N} \sum \varepsilon_j^{\sim 2} X_j^{\sim} X_j^{\sim\top} \right) (X^{\sim\top} X^{\sim})^{-1},$$

$$A_r = (X^\top X)^{-1} \left(\frac{1}{n} \sum \varepsilon_i^2 X_i X_i^\top \right) (X^\top X)^{-1}$$

A coordinate-wise $(1-\alpha)$ interval for θ_{PP,j^*} is:

$$C^{PP}_{\alpha,j^*} = \left[\hat{\theta}_{PP,j^*} \pm z_{1-\alpha/2} \sqrt{\hat{V}_{j^*,j^*}} \right].$$

Algorithm 2 Prediction-Powered Linear Regression

Require: Labeled (X, Y), unlabeled $X^{\sim}$, predictor f, target coordinate j^*, level $1-\alpha$

1: Compute $\tilde{\theta}_f$, $\hat{\Delta}$, and $\hat{\theta}_{PP}$

2: Compute residuals and covariance matrix $\hat{V}$

3: **return** Confidence interval for $\hat{\theta}_{PP,j^*}$

3.3 Convex Estimation and Theoretical Guarantees

Let θ^* minimize a convex risk:

$$\theta^* = \arg\min_\theta \mathbb{E}[\ell_\theta(X, Y)], \tag{3.7}$$

with subgradient $g_\theta(X, Y)$ satisfying $\mathbb{E}[g_{\theta^*}(X, Y)] = 0$.

The prediction-powered estimating equation augments with a rectifier:

$$\Delta_\theta = \mathbb{E}[g_\theta(X, Y) - g_\theta(X, f(X))].$$

Empirical Procedure.

- Compute empirical rectifier: $\hat{\Delta}(\theta) = \frac{1}{n} \sum [g_\theta(X_i, Y_i) - g_\theta(X_i, f(X_i))]$
- Compute predicted gradient: $\hat{g}^f_\theta = \frac{1}{N} \sum g_\theta(\tilde{X}_i, f(\tilde{X}_i))$

Confidence Set. Under CLT assumptions, the confidence set is:

$$C^{PP}_\alpha = \left\{ \theta : \|\hat{g}^f_\theta + \hat{\Delta}(\theta)\| \le w_\alpha(\theta) \right\}, \tag{3.8}$$

where $\|\cdot\|$ denotes the Euclidean norm. This means we include all parameter values θ such that the magnitude of the corrected gradient (combining labeled and unlabeled data) is sufficiently small to be consistent with the true minimizer.

with

$$w_\alpha(\theta) = z_{1-\alpha/2}\sqrt{\frac{\hat{\sigma}^2_\Delta(\theta)}{n} + \frac{\hat{\sigma}^2_g(\theta)}{N}}.$$

Coverage Guarantee. As shown in [GL21], if:

$$P(\Delta_\theta \in R_\delta(\theta)) \geq 1-\delta, \quad P(g^f_\theta \in T_{\alpha-\delta}(\theta)) \geq 1-(\alpha-\delta),$$

then the confidence region

$$C^{PP}_\alpha = \{\theta : 0 \in R_\delta(\theta) + T_{\alpha-\delta}(\theta)\}$$

achieves asymptotic coverage at level $1-\alpha$ via the union bound.

Algorithm 3 Prediction-Powered Convex Estimation

Require: Data $(X, Y), (\tilde{X})$, predictor f, grid Θ, level $1-\alpha$

1: **for** $\theta \in \Theta$ **do**
2: Compute $\hat{g}^f_\theta$, $\hat{\Delta}(\theta)$, variances
3: Form $w_\alpha(\theta)$ and include θ if condition (3.8) holds
4: **end for**
5: **return** Confidence set C^{PP}_α

3.4 Cases Where Prediction-Powered Inference is Underpowered

Prediction-powered inference (PPI) for mean estimation performs well when the prediction model is accurate and the unlabeled dataset is large. However, it can become underpowered when either condition fails.

Mean Inference.

Based on the Section 3.1, the variance of the mean inference becomes:

$$\mathrm{Var}(\hat{\theta}_{\mathrm{PPI}}) = \frac{\mathrm{Var}(f(X))}{N} + \frac{\mathrm{Var}(f(X)-Y)}{n}. \tag{3.9}$$

and PPI becomes underpowered when:

$$\mathrm{Var}(\hat{\theta}_{\mathrm{PPI}}) > \mathrm{Var}(\hat{\theta}_{\mathrm{class}}). \tag{3.10}$$

Key Conditions for Underpowering:

- **Poor predictions:** High $\mathrm{Var}(f(X) - Y)$ [BO17].
- **Limited unlabeled data:** Small N, so the $\frac{1}{N}$ term does not vanish.

Bernoulli Example. For $Y \sim \text{Bernoulli}(p)$, the threshold prediction error rate $\eta = \mathbb{P}(f(X) \neq Y)$ must satisfy:

$$\eta < 0.25 \quad (\text{when } p = 0.5),$$

for PPI to outperform classical estimation.

Figure 3 illustrates: The comparison of confidence interval widths for classical, imputation, and PPI estimators.

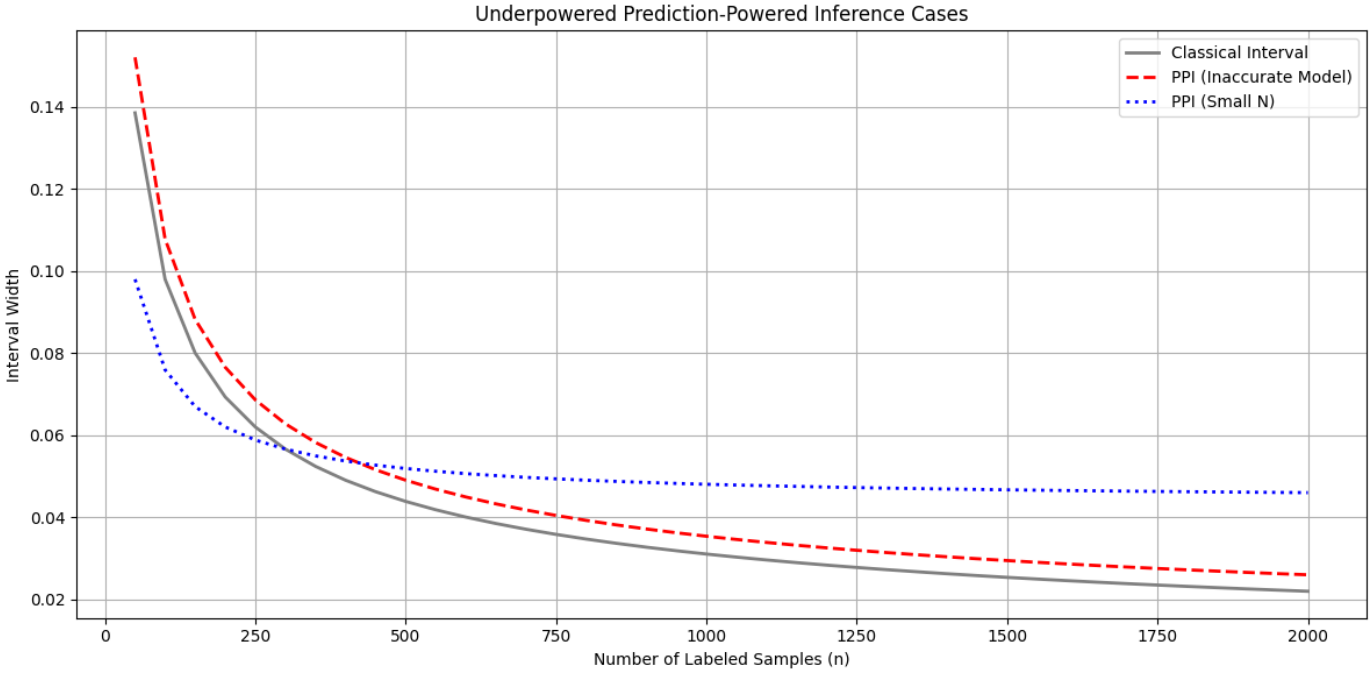

FIGURE 3. Confidence interval widths under different estimators and prediction qualities.

3.5 Prediction-Powered Inference for Structured Parameters

Prediction-powered inference generalizes to various estimands, including quantiles, logistic regression coefficients, and linear regression. In each case, PPI adjusts imputed estimates by a rectifier term to debias and construct valid confidence intervals. Below, we summarize key formulations and conditions for underperformance.

Quantile Estimation (e.g., Median)

The target parameter is the q-quantile:

$$\theta^* = \inf\{\theta \in \mathbb{R} : P(Y \leq \theta) \geq q\}, \quad q \in (0,1). \tag{3.11}$$

PPI replaces the empirical CDF $P(Y \leq \theta)$ with $P(f(X) \leq \theta)$ and corrects the discrepancy. The variance of the PPI estimator becomes:

$$\frac{1}{n}\operatorname{Var}\left(\mathbf{1}\{Y \leq \theta\} - \mathbf{1}\{f(X) \leq \theta\}\right) + \frac{1}{N}\operatorname{Var}\left(\mathbf{1}\{f(X) \leq \theta\}\right). \tag{3.12}$$

Underpowered Cases. PPI may underperform if:

- $f(X)$ poorly approximates the distribution near θ^*, especially when q is close to 0 or 1, i.e., when θ^* lies in the tails of the distribution. In such cases, there may be too few training samples near θ^* to ensure accurate approximation.
- or the unlabeled sample size N is not large enough to reduce the second variance term.

Logistic Regression Coefficients (Binary Response)

The target parameter minimizes the logistic loss:

$$\theta^* = \arg\min_{\theta \in \mathbb{R}^d} \mathbb{E}\left[-YX^\top\theta + \log\left(1 + e^{X^\top\theta}\right)\right]. \tag{3.13}$$

PPI substitutes predictions $f(X)$ for Y, then debiases the estimating equation. The variance for each coordinate j is approximated by:

$$\frac{1}{n}\operatorname{Var}\left(X_j(f(X) - Y)\right) + \frac{1}{N}\operatorname{Var}\left(X_j(h_\theta(X) - f(X))\right), \tag{3.14}$$

where $h_\theta(X) = \mathbb{P}(Y = 1 \mid X)$ is the true conditional probability. For example, if $Y = 1$ indicates a rare event (e.g., disease presence), $f(X)$ must closely approximate this probability even for infrequent outcomes.

Underpowered Cases. Performance degrades when:

- $f(X)$ is a poor approximation of $h_\theta(X)$,
- or N is small relative to d.

Linear Regression Coefficients (Continuous Outcomes)

Based on the estimand (3.4) of the Section 3.2, the variance of the PPI estimator decomposes as:

$$\frac{1}{n}\operatorname{Var}\left(X^\top(f(X)-Y)\right)+\frac{1}{N}\operatorname{Var}\left(X^\top(X\theta-f(X))\right). \tag{3.15}$$

Underpowered Cases. PPI performs poorly when:

- $f(X)$ lacks linear structure aligned with $X^\top\theta$; this may reflect model bias, and PPI is not guaranteed to correct for biased predictors.
- or N is too small to offset the imputation error.

Figure 4 summarizes: The comparison of confidence interval widths for classical, imputation, and PPI estimators.

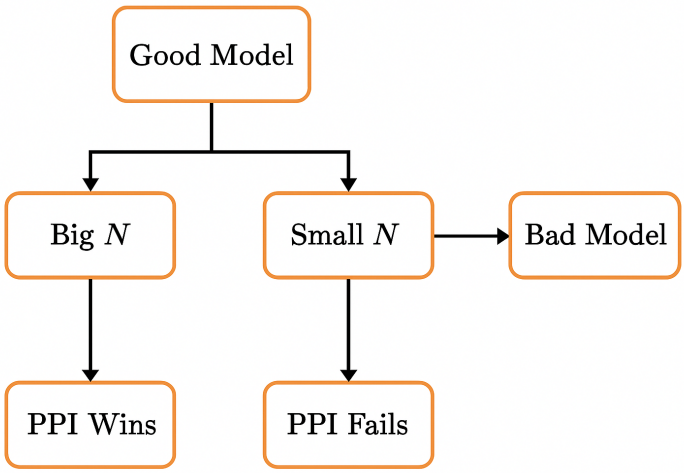

FIGURE 4. Conditions for Success or Failure of Prediction-Powered Inference across different estimators.

4 Conclusion

Prediction-powered inference (PPI) provides a powerful framework for integrating machine learning predictions into classical statistical inference while preserving valid coverage guarantees. By combining a small labeled dataset with a large unlabeled one, PPI leverages model predictions efficiently, correcting for systematic prediction error via an empirical rectifier. This enables the construction of confidence intervals that are both valid and often narrower than classical alternatives.

Across multiple settings—mean, quantile, and regression estimation—PPI offers a general, convex-optimization-based strategy for achieving unbiased inference, even under complex or misspecified models. Empirical evaluations confirm that PPI performs optimally when predictions are accurate and unlabeled data is abundant, while still maintaining robustness in less ideal conditions.

However, PPI is designed for scenarios where the predictor is sufficiently accurate—typically meaning low variance and ideally low bias—to improve power using predictions on unlabeled data. For a fixed-quality predictor, the inferential power increases with the size of the unlabeled dataset N. When the predictor has high variance or systematic bias, PPI may no longer yield valid or efficient inference. In summary, the success of prediction-powered inference critically depends on the quality of the predictive model.

Ultimately, PPI bridges the gap between predictive modeling and inferential rigor, offering a flexible and scalable methodology for modern data analysis. Future extensions may further expand its applicability to dependent data, model selection, and causal inference contexts.

Acknowledgements

I would like to thank my mentor, Yanees Dobberstein, for his guidance throughout the Directed Reading Program. I also appreciate the support of the DRP committee at McGill University Mathematics & Statistics.

References

[BM02] Peter L. Bartlett and Shahar Mendelson. "Rademacher and Gaussian Complexities: Risk Bounds and Structural Results". In: *Journal of Machine Learning Research* 3 (2002), pp. 463–482. URL: https://www.jmlr.org/papers/volume3/bartlett02a/bartlett02a.pdf.

[BO17] F. Jay Breidt and Jean D. Opsomer. "Model-assisted survey estimation with applications". In: *Statistical Science* 32.2 (2017), pp. 195–210.

[CSZ06] Olivier Chapelle, Bernhard Schölkopf, and Alexander Zien. *Semi-Supervised Learning*. Cambridge, MA: MIT Press, 2006. ISBN: 9780262033589.

[GL21] Yue Guo and Lihua Lei. "Confidence sets from prediction-powered inference". In: *Journal of the American Statistical Association* (2021). DOI: 10.1080/01621459.2021.1996374.

[LR02] Roderick J. A. Little and Donald B. Rubin. *Statistical Analysis with Missing Data*. 2nd ed. New York: Wiley, 2002. ISBN: 9780471183860.