

Finite-Sample Analysis of Prediction-Powered Linear Regression

1 Introduction

This document presents a finite-sample analysis of the Prediction-Powered Inference (PPI) estimator in the context of linear regression. The PPI framework leverages both labeled and unlabeled data, along with a pre-trained predictor, to improve estimation accuracy. However, the standard PPI approach only provides a asymptotic analysis, but in reality, the reason why we use PPI is that we have a limited amount of labeled data, and we hope to utilize the unlabeled data for variance reduction. In another perspective, we can see PPI as Therefore, a finite-sample analysis is crucial for understanding its practical performance.

The basic intuition for stratified PPI is that, the bias-correction term treats all data points equally, but in reality, the prediction residuals may vary significantly for different X . By stratifying the data and applying separate bias corrections within each stratum, we can better account for this heterogeneity and gain statistical efficiency.

2 Related Work

3 Preliminaries

We consider a linear regression model:

$$Y = X\theta^* + \varepsilon,$$

where $\theta^* \in \mathbb{R}^d$ is the target parameter and ε is a noise term with $\mathbb{E}[\varepsilon | X] = 0$.

We observe independent labeled and unlabeled samples:

$$\{(X_i, Y_i)\}_{i=1}^n, \quad \{\tilde{X}_j\}_{j=1}^N, \quad \text{and} \quad f(X_i), f(\tilde{X}_j),$$

where $f : \mathbb{R}^d \rightarrow \mathbb{R}$ is a black-box predictor which is fixed or independent of both labeled and unlabeled samples (e.g. pre-trained).

Let $X \in \mathbb{R}^{n \times d}$ and $\tilde{X} \in \mathbb{R}^{N \times d}$ denote the design matrices with rows X_i^\top and \tilde{X}_j^\top , and let $Y \in \mathbb{R}^n$ be the vector of labeled responses. We adopt the following empirical PPI loss with tuning parameter $\lambda \in \mathbb{R}$:

$$L_{PP,n,N}^\lambda(\theta) := \frac{1}{2n} \|Y - X\theta\|_2^2 + \lambda \left(\frac{1}{2N} \|f(\tilde{X}) - \tilde{X}\theta\|_2^2 - \frac{1}{2n} \|f(X) - X\theta\|_2^2 \right). \quad (1)$$

Here $f(X) \in \mathbb{R}^n$ and $f(\tilde{X}) \in \mathbb{R}^N$ are understood componentwise.

PPI estimator. The PPI estimator is any minimizer of the empirical loss:

$$\hat{\theta} := \hat{\theta}_{PP}^\lambda \in \arg \min_{\theta \in \mathbb{R}^d} L_{PP,n,N}^\lambda(\theta).$$

By definition of the argmin, for every $\theta \in \mathbb{R}^d$ and in particular for θ^* ,

$$L_{PP,n,N}^\lambda(\hat{\theta}) \leq L_{PP,n,N}^\lambda(\theta^*).$$

This is the starting point of our basic inequality.

4 Finite Sample Bound for PPI Linear Regression

Let

$$\Delta := \hat{\theta} - \theta^*$$

denote the parameter error. Define the labeled and unlabeled prediction residual vectors

$$r(X) := f(X) - X\theta^* \quad b(X) := Y - f(X) = \varepsilon - r(X)$$

Theorem 1 (Finite-sample error bound for PPI linear regression). *Suppose Assumptions A1–A4 hold:*

- A1. **Linear model:** $\{(X_i, Y_i)\}_{i=1}^n \stackrel{i.i.d.}{\sim} P_{X,Y}$ and $\{\tilde{X}_j\}_{j=1}^N \stackrel{i.i.d.}{\sim} P_X$, where P_X is the marginal distribution of X under $P_{X,Y}$, and the two samples are independent. $f : \mathbb{R}^d \rightarrow \mathbb{R}$ is a fixed predictor (e.g., pre-trained).
- A2. **Design conditions:** The rows of X and \tilde{X} are sub-Gaussian random vectors with covariance Σ_X satisfying $\lambda_{\min}(\Sigma_X) \geq \kappa_0 > 0$. Moreover, there exists $\sigma_X > 0$ such that $\sup_{\|u\|_2=1} \|u^\top X\|_{\psi_2} \leq \sigma_X$.
- A3. **Noise:** The noise ε_i is sub-Gaussian with parameter σ_ε and independent of X_i , with $\mathbb{E}[\varepsilon_i | X_i] = 0$.
- A4. **Prediction residual:** Define the prediction residual $r(X_i) := f(X_i) - X_i^\top \theta^*$ and the vector $Z_i := r(X_i)X_i \in \mathbb{R}^d$. Assume either:
 - (a) **Boundedness:** $|r(X)| \leq R$ and $\|X\|_2 \leq M$ almost surely, OR
 - (b) **Centered sub-Gaussian:** Let $\mu := \mathbb{E}[Z | f]$. Conditional on f , for any unit vector $u \in \mathbb{R}^d$, $u^\top (Z - \mu)$ is sub-Gaussian with parameter σ_Z , i.e.,

$$\sup_{\|u\|_2=1} \|u^\top (Z - \mu)\|_{\psi_2} \leq \sigma_Z.$$

The labeled and unlabeled copies $\{Z_i\}_{i=1}^n$ and $\{\tilde{Z}_j\}_{j=1}^N$ are i.i.d.

Let $\hat{\theta}$ be the (unregularized) PPI estimator

$$\hat{\theta} \in \arg \min_{\theta \in \mathbb{R}^d} L_{PP,n,N}^\lambda(\theta),$$

where

$$L_{PP,n,N}^\lambda(\theta) := \frac{1}{2n} \|Y - X\theta\|_2^2 + \lambda \left(\frac{1}{2N} \|f(\tilde{X}) - \tilde{X}\theta\|_2^2 - \frac{1}{2n} \|f(X) - X\theta\|_2^2 \right).$$

Then for any fixed $\lambda \in (0, 1)$, there exist constants $c, C > 0$ (depending only on the sub-Gaussian parameters and κ_0) such that, with probability at least $1 - \delta$,

$$\|\hat{\theta} - \theta^*\|_2 \leq \frac{C}{\kappa_0} \left(\sigma_X \sigma_\varepsilon \sqrt{\frac{d + \log(1/\delta)}{n}} + \lambda \sigma_Z \sqrt{\left(\frac{1}{n} + \frac{1}{N}\right) (d + \log(1/\delta))} \right).$$

In particular, when $N \gg n$, the second term scales as $\lambda \sigma_Z \sqrt{d/n}$, so the PPI estimator attains the usual parametric rate $\sqrt{d/n}$ while trading off the noise scales σ_ε and σ_Z .

Proof. Since $\hat{\theta}$ minimizes $L_{PP,n,N}^\lambda$, we have $L_{PP,n,N}^\lambda(\hat{\theta}) \leq L_{PP,n,N}^\lambda(\theta^*)$.

Step 1: Algebraic expansion. Write $\Delta := \hat{\theta} - \theta^*$. Using $Y = X\theta^* + \varepsilon$ and $r(X_i) = f(X_i) - X_i^\top \theta^*$ (so that $f(X) = X\theta^* + r(X)$ componentwise), we expand:

$$\begin{aligned} \|Y - X\hat{\theta}\|_2^2 - \|Y - X\theta^*\|_2^2 &= \|X\theta^* + \varepsilon - X\hat{\theta}\|_2^2 - \|\varepsilon\|_2^2 \\ &= \|X\Delta\|_2^2 - 2\varepsilon^\top X\Delta, \end{aligned}$$

$$\begin{aligned} \|f(\tilde{X}) - \tilde{X}\hat{\theta}\|_2^2 - \|f(\tilde{X}) - \tilde{X}\theta^*\|_2^2 &= \|\tilde{X}\theta^* + r(\tilde{X}) - \tilde{X}\hat{\theta}\|_2^2 - \|r(\tilde{X})\|_2^2 \\ &= \|\tilde{X}\Delta\|_2^2 - 2r(\tilde{X})^\top \tilde{X}\Delta, \end{aligned}$$

$$\|f(X) - X\hat{\theta}\|_2^2 - \|f(X) - X\theta^*\|_2^2 = \|X\Delta\|_2^2 - 2r(X)^\top X\Delta.$$

Substituting into $L_{PP,n,N}^\lambda(\hat{\theta}) - L_{PP,n,N}^\lambda(\theta^*) \leq 0$ and rearranging yields the basic inequality:

$$\frac{1-\lambda}{2n} \|X\Delta\|_2^2 + \frac{\lambda}{2N} \|\tilde{X}\Delta\|_2^2 \leq \underbrace{\frac{1}{n} \varepsilon^\top X\Delta}_{=:T_1} + \underbrace{\lambda \left(\frac{1}{N} r(\tilde{X})^\top \tilde{X}\Delta - \frac{1}{n} r(X)^\top X\Delta \right)}_{=:T_2}. \quad (2)$$

Step 2: Refined lower-bounding revealing the λ -tradeoff under random design. The left-hand side of (2) can be expressed as:

$$\frac{1-\lambda}{2n} \|X\Delta\|_2^2 + \frac{\lambda}{2N} \|\tilde{X}\Delta\|_2^2 = \frac{1-\lambda}{2} \cdot \Delta^\top \hat{\Sigma}_n \Delta + \frac{\lambda}{2} \cdot \Delta^\top \hat{\Sigma}_N \Delta,$$

where $\hat{\Sigma}_n := \frac{1}{n} X^\top X$ and $\hat{\Sigma}_N := \frac{1}{N} \tilde{X}^\top \tilde{X}$ are random matrices.

Let $\Sigma_X := \mathbb{E}[XX^\top]$ with $\lambda_{\min}(\Sigma_X) \geq \kappa_0 > 0$. By the sub-Gaussian random matrix concentration (Vershynin, Theorem 4.6.1), there exist constants $C_1, C_2 > 0$ such that with probability at least $1 - \delta/4$:

$$\|\hat{\Sigma}_n - \Sigma_X\|_{\text{op}} \leq \epsilon_n, \quad \|\hat{\Sigma}_N - \Sigma_X\|_{\text{op}} \leq \epsilon_N,$$

where

$$\epsilon_n = C_1 \sigma_X^2 \left(\sqrt{\frac{d}{n}} + \sqrt{\frac{\log(4/\delta)}{n}} + \frac{\log(4/\delta)}{n} \right),$$

$$\epsilon_N = C_2 \sigma_X^2 \left(\sqrt{\frac{d}{N}} + \sqrt{\frac{\log(4/\delta)}{N}} + \frac{\log(4/\delta)}{N} \right).$$

Using Weyl's inequality, we have:

$$|\lambda_{\min}(\hat{\Sigma}_n) - \lambda_{\min}(\Sigma_X)| \leq \|\hat{\Sigma}_n - \Sigma_X\|_{\text{op}} \leq \epsilon_n,$$

$$|\lambda_{\min}(\hat{\Sigma}_N) - \lambda_{\min}(\Sigma_X)| \leq \|\hat{\Sigma}_N - \Sigma_X\|_{\text{op}} \leq \epsilon_N.$$

Thus, with the same probability:

$$\lambda_{\min}(\hat{\Sigma}_n) \geq \kappa_0 - \epsilon_n, \quad \lambda_{\min}(\hat{\Sigma}_N) \geq \kappa_0 - \epsilon_N.$$

Since $\lambda_{\min}(\cdot)$ is concave on the cone of positive semidefinite matrices:

$$\begin{aligned} \lambda_{\min}\left((1-\lambda)\hat{\Sigma}_n + \lambda\hat{\Sigma}_N\right) &\geq (1-\lambda)\lambda_{\min}(\hat{\Sigma}_n) + \lambda\lambda_{\min}(\hat{\Sigma}_N) \\ &\geq (1-\lambda)(\kappa_0 - \epsilon_n) + \lambda(\kappa_0 - \epsilon_N) \\ &= \kappa_0 - [(1-\lambda)\epsilon_n + \lambda\epsilon_N]. \end{aligned}$$

Therefore, for all $\Delta \in \mathbb{R}^d$:

$$\frac{1-\lambda}{2n}\|X\Delta\|_2^2 + \frac{\lambda}{2N}\|\tilde{X}\Delta\|_2^2 \geq \frac{1}{2}[\kappa_0 - (1-\lambda)\epsilon_n - \lambda\epsilon_N]\|\Delta\|_2^2. \quad (3)$$

Key probabilistic insight: The error terms ϵ_n and ϵ_N are random but concentrate around zero at rates $\sqrt{d/n}$ and $\sqrt{d/N}$ respectively. When N is sufficiently larger than n , we typically have $\epsilon_N \ll \epsilon_n$ with high probability.

Step 3: Controlling T_1 .

$$T_1 = \frac{1}{n}\varepsilon^\top X\Delta = \left(\frac{1}{n}X^\top \varepsilon\right)^\top \Delta,$$

so by Cauchy–Schwarz,

$$|T_1| \leq \left\| \frac{1}{n}X^\top \varepsilon \right\|_2 \|\Delta\|_2. \quad (4)$$

It remains to control $\left\| \frac{1}{n}X^\top \varepsilon \right\|_2$. Write

$$\frac{1}{n}X^\top \varepsilon = \frac{1}{n} \sum_{i=1}^n \varepsilon_i X_i.$$

Under Assumption A3, ε_i is independent of X_i and satisfies $\mathbb{E}[\varepsilon_i | X_i] = 0$, hence $\mathbb{E}[\varepsilon_i X_i] = 0$. Moreover, by Assumption A2 and the product property of sub-Gaussian random variables, for any fixed $u \in \mathbb{S}^{d-1}$ the scalar random variable

$$u^\top (\varepsilon_i X_i) = \varepsilon_i (u^\top X_i)$$

is mean-zero sub-exponential with

$$\|\varepsilon_i(u^\top X_i)\|_{\psi_1} \lesssim \|\varepsilon_i\|_{\psi_2} \|u^\top X_i\|_{\psi_2} \leq C \sigma_\varepsilon \sigma_X.$$

Therefore, for each fixed $u \in \mathbb{S}^{d-1}$, Bernstein's inequality implies that

$$\mathbb{P}\left(\left|u^\top \left(\frac{1}{n}X^\top \varepsilon\right)\right| > t\right) \leq 2 \exp\left(-cn \min\left\{\frac{t^2}{\sigma_X^2 \sigma_\varepsilon^2}, \frac{t}{\sigma_X \sigma_\varepsilon}\right\}\right). \quad (5)$$

Next, note the dual characterization of the Euclidean norm:

$$\left\| \frac{1}{n} X^\top \varepsilon \right\|_2 = \sup_{\|u\|_2=1} u^\top \left(\frac{1}{n} X^\top \varepsilon \right).$$

Applying (5) on a $1/2$ -net of \mathbb{S}^{d-1} and taking a union bound (or equivalently invoking a standard vector Bernstein inequality; see, e.g., Wainwright), we obtain that there exists a constant $C_1 > 0$ such that with probability at least $1 - \delta/4$, (using that a $1/2$ -net can be chosen with cardinality at most 5^d).

$$\left\| \frac{1}{n} X^\top \varepsilon \right\|_2 \leq C_1 \sigma_X \sigma_\varepsilon \sqrt{\frac{d + \log(4/\delta)}{n}}. \quad (6)$$

Combining (4) and (6) yields, with probability at least $1 - \delta/4$,

$$|T_1| \leq C_1 \sigma_X \sigma_\varepsilon \sqrt{\frac{d + \log(4/\delta)}{n}} \|\Delta\|_2. \quad (7)$$

Step 4: Controlling T_2 via the difference structure. Recall that the predictor f is externally pre-trained and treated as fixed. Define

$$Z_i := r(X_i)X_i, \quad \tilde{Z}_j := r(\tilde{X}_j)\tilde{X}_j \in \mathbb{R}^d,$$

and let

$$\mu := \mathbb{E}[Z \mid f] = \mathbb{E}[r(X)X \mid f]$$

denote the common population mean. Writing

$$\hat{\mu}_n := \frac{1}{n} \sum_{i=1}^n Z_i, \quad \hat{\mu}_N := \frac{1}{N} \sum_{j=1}^N \tilde{Z}_j,$$

we may express

$$T_2 = \lambda (\hat{\mu}_N - \hat{\mu}_n)^\top \Delta.$$

Since $\{X_i\}_{i=1}^n$ and $\{\tilde{X}_j\}_{j=1}^N$ are independent samples drawn from the same distribution and f is fixed, the random vectors $\{Z_i\}_{i=1}^n$ and $\{\tilde{Z}_j\}_{j=1}^N$ are independent i.i.d. copies of Z . Consequently,

$$\mathbb{E}[\hat{\mu}_N - \hat{\mu}_n \mid f] = \mu - \mu = 0.$$

Thus, although the predictor f may be systematically biased (so that $\mathbb{E}[r(X)] \neq 0$), the difference $\hat{\mu}_N - \hat{\mu}_n$ is centered and reflects only sampling variability.

To control its fluctuations, fix an arbitrary unit vector $u \in \mathbb{S}^{d-1}$. We write

$$u^\top (\hat{\mu}_N - \hat{\mu}_n) = \frac{1}{N} \sum_{j=1}^N u^\top (\tilde{Z}_j - \mu) - \frac{1}{n} \sum_{i=1}^n u^\top (Z_i - \mu).$$

Under Assumption A4(b), the centered projections $u^\top (Z - \mu)$ and $u^\top (\tilde{Z} - \mu)$ are mean-zero sub-Gaussian random variables with ψ_2 -norm bounded by σ_Z (conditional on f). Since the two sums are independent, standard properties of sub-Gaussian random variables imply that $u^\top (\hat{\mu}_N - \hat{\mu}_n)$ is itself sub-Gaussian with

$$\|u^\top (\hat{\mu}_N - \hat{\mu}_n)\|_{\psi_2} \leq C \sigma_Z \sqrt{\frac{1}{n} + \frac{1}{N}},$$

for a universal constant $C > 0$. In particular, there exists $c > 0$ such that for all $t > 0$,

$$\mathbb{P}\left(|u^\top(\hat{\mu}_N - \hat{\mu}_n)| > t\right) \leq 2 \exp\left(-\frac{ct^2}{\sigma_Z^2(\frac{1}{n} + \frac{1}{N})}\right). \quad (8)$$

To obtain a bound uniform over all directions, we use the identity

$$\|\hat{\mu}_N - \hat{\mu}_n\|_2 = \sup_{\|u\|_2=1} u^\top(\hat{\mu}_N - \hat{\mu}_n),$$

together with a standard ε -net argument on \mathbb{S}^{d-1} . Applying (8) on a $1/2$ -net and taking a union bound yields that, with probability at least $1 - \delta/4$,

$$\|\hat{\mu}_N - \hat{\mu}_n\|_2 \leq C_2 \sigma_Z \sqrt{\left(\frac{1}{n} + \frac{1}{N}\right)(d + \log(1/\delta))}, \quad (9)$$

for a constant $C_2 > 0$ depending only on the sub-Gaussian parameters.

Finally, applying the Cauchy–Schwarz inequality gives

$$|T_2| = \lambda |(\hat{\mu}_N - \hat{\mu}_n)^\top \Delta| \leq \lambda \|\hat{\mu}_N - \hat{\mu}_n\|_2 \|\Delta\|_2 \leq C_2 \lambda \sigma_Z \sqrt{\left(\frac{1}{n} + \frac{1}{N}\right)(d + \log(1/\delta))} \|\Delta\|_2. \quad (10)$$

Step 5: Combining all bounds. On the intersection of all high-probability events (overall probability $\geq 1 - \delta$), combining (??), (10), and (7) with the basic inequality (2), we obtain:

$$\frac{\kappa_0}{4} \|\Delta\|_2^2 \leq |T_1| + |T_2| \leq \left(C_1 \sigma_X \sigma_\varepsilon \sqrt{\frac{d + \log(1/\delta)}{n}} + C_2 \lambda \sigma_Z \sqrt{\left(\frac{1}{n} + \frac{1}{N}\right)(d + \log(1/\delta))}\right) \|\Delta\|_2.$$

If $\Delta \neq 0$, dividing both sides by $(\kappa_0/4)\|\Delta\|_2$ yields:

$$\|\Delta\|_2 = \|\hat{\theta} - \theta^*\|_2 \leq \frac{4}{\kappa_0} \left(C_1 \sigma_X \sigma_\varepsilon \sqrt{\frac{d + \log(1/\delta)}{n}} + C_2 \lambda \sigma_Z \sqrt{\left(\frac{1}{n} + \frac{1}{N}\right)(d + \log(1/\delta))}\right).$$

Absorbing all constants into $C := 4 \max\{C_1, C_2\}$ gives the claimed bound. \square

5 Another proof

Recall the basic inequality: for $\Delta := \hat{\theta} - \theta^*$,

$$\frac{1-\lambda}{2n} \|X\Delta\|_2^2 + \frac{\lambda}{2N} \|\tilde{X}\Delta\|_2^2 \leq \underbrace{\frac{1}{n} \varepsilon^\top X \Delta}_{=: T_1} + \underbrace{\lambda \left(\frac{1}{N} r(\tilde{X})^\top \tilde{X} \Delta - \frac{1}{n} r(X)^\top X \Delta\right)}_{=: T_2}. \quad (11)$$

Which is equivalent to:

$$\frac{1-\lambda}{2n} \|X\Delta\|_2^2 + \frac{\lambda}{2N} \|\tilde{X}\Delta\|_2^2 \leq \frac{1}{n} (\varepsilon - \lambda r(X))^\top X \Delta + \lambda \frac{1}{N} r(\tilde{X})^\top \tilde{X} \Delta.$$

If we take $r(X) := f(X) - X^T \theta^*$, $b(X, Y) := Y - f(X) = \varepsilon - r(X)$

$$\frac{1-\lambda}{2n} \|X\Delta\|_2^2 + \frac{\lambda}{2N} \|\tilde{X}\Delta\|_2^2 \leq \frac{1}{n} ((1-\lambda)\varepsilon - \lambda b)^\top X \Delta + \lambda \frac{1}{N} r(\tilde{X})^\top \tilde{X} \Delta.$$

Assumptions. (i) $X \in \mathbb{R}^d$ is mean-zero sub-Gaussian with covariance $\Sigma = \mathbb{E}[XX^\top]$ and $\lambda_{\min}(\Sigma) \geq k_0 > 0$. The labeled covariates $\{X_i\}_{i=1}^n$ and unlabeled $\{\tilde{X}_j\}_{j=1}^N$ are i.i.d. draws from P_X and independent across the two samples. (ii) ε_i satisfies $\mathbb{E}[\varepsilon_i | X_i] = 0$ and is sub-Gaussian (conditional on X_i) with proxy σ_ε . (iii) $r(\cdot)$ is deterministic (given the externally trained predictor f) and $\mathbb{E}[r(X)^2] \leq \sigma_r^2$. Define the approximation bias

$$\delta := \|\mathbb{E}[Xr(X)]\|_2.$$

(iv) (Vector mean concentration) The random vector $Z := Xr(X)$ is sub-exponential in the sense that for all unit vectors $u \in \mathbb{S}^{d-1}$, $u^\top(Z - \mathbb{E}Z)$ has ψ_1 -norm bounded by K_Z .

Step 1: Rewriting the RHS to expose the residual structure. Define the labeled residual-like vector

$$u := \varepsilon - \lambda r(X) \in \mathbb{R}^n, \quad u_i = \varepsilon_i - \lambda r(X_i).$$

Then

$$T_1 + T_2 = \frac{1}{n}(\varepsilon - \lambda r(X))^\top X\Delta + \lambda \frac{1}{N}r(\tilde{X})^\top \tilde{X}\Delta = \frac{1}{n}u^\top X\Delta + \lambda \frac{1}{N}r(\tilde{X})^\top \tilde{X}\Delta.$$

In particular, when $\lambda = 1$,

$$u_i = \varepsilon_i - r(X_i) = (Y_i - X_i^\top \theta^*) - (f(X_i) - X_i^\top \theta^*) = Y_i - f(X_i),$$

so the labeled term depends on the residual $Y - f(X)$.

Step 2: Lower bounding the LHS. Let $\hat{\Sigma}_n := \frac{1}{n}X^\top X$ and $\hat{\Sigma}_N := \frac{1}{N}\tilde{X}^\top \tilde{X}$. Then

$$\frac{1-\lambda}{2n}\|X\Delta\|_2^2 + \frac{\lambda}{2N}\|\tilde{X}\Delta\|_2^2 = \frac{1}{2}\Delta^\top \left((1-\lambda)\hat{\Sigma}_n + \lambda\hat{\Sigma}_N\right)\Delta \geq \frac{1}{2}\left((1-\lambda)\lambda_{\min}(\hat{\Sigma}_n) + \lambda\lambda_{\min}(\hat{\Sigma}_N)\right)\|\Delta\|_2^2.$$

By Weyl's inequality, $\lambda_{\min}(\hat{\Sigma}_m) \geq k_0 - \|\hat{\Sigma}_m - \Sigma\|_{\text{op}}$ for $m \in \{n, N\}$. Moreover, for sub-Gaussian X , there exists $C > 0$ such that for any $t > 0$, with probability at least $1 - 4e^{-t}$,

$$\|\hat{\Sigma}_m - \Sigma\|_{\text{op}} \leq C\|\Sigma\|_{\text{op}} \left(\sqrt{\frac{d+t}{m}} + \frac{d+t}{m} \right), \quad m \in \{n, N\}.$$

Define

$$\delta_m := C\|\Sigma\|_{\text{op}} \left(\sqrt{\frac{d+t}{m}} + \frac{d+t}{m} \right).$$

On the same event,

$$\frac{1-\lambda}{2n}\|X\Delta\|_2^2 + \frac{\lambda}{2N}\|\tilde{X}\Delta\|_2^2 \geq \frac{1}{2}\left((1-\lambda)(k_0 - \delta_n) + \lambda(k_0 - \delta_N)\right)\|\Delta\|_2^2 \geq \frac{1}{2}\left(k_0 - (1-\lambda)\delta_n - \lambda\delta_N\right)\|\Delta\|_2^2. \quad (12)$$

Step 3: Upper bounding the RHS. By Cauchy–Schwarz,

$$|T_1 + T_2| \leq \left\| \frac{1}{n}X^\top u \right\|_2 \|\Delta\|_2 + \lambda \left\| \frac{1}{N}\tilde{X}^\top r(\tilde{X}) \right\|_2 \|\Delta\|_2. \quad (13)$$

(a) *Labeled residual term.* Under Assumption (ii) and sub-Gaussian X , standard vector Bernstein bounds imply that there exists $C_1 > 0$ such that with probability at least $1 - 2e^{-t}$,

$$\left\| \frac{1}{n} X^\top u \right\|_2 \leq C_1 \sqrt{\|\Sigma\|_{\text{op}}} \sigma_u \sqrt{\frac{d+t}{n}}, \quad \sigma_u^2 := \mathbb{E}[u^2]. \quad (14)$$

Specifically,

$$\sigma_u^2 := \mathbb{E}[u^2] = \mathbb{E}[(\varepsilon - \lambda r(X))^2] = \mathbb{E}[\varepsilon^2] - 2\lambda\mathbb{E}[\varepsilon r(X)] + \lambda^2\mathbb{E}[r(X)^2]$$

When $\lambda = 1$, this simplifies to $\sigma_u^2 = \mathbb{E}[(Y - f(X))^2]$ which is small if the predictor is accurate.

(b) *Unlabeled prediction term.* Write

$$\frac{1}{N} \tilde{X}^\top r(\tilde{X}) = \mathbb{E}[Xr(X)] + \left(\frac{1}{N} \sum_{j=1}^N \tilde{Z}_j - \mathbb{E}Z \right), \quad \tilde{Z}_j := \tilde{X}_j r(\tilde{X}_j).$$

Hence,

$$\left\| \frac{1}{N} \tilde{X}^\top r(\tilde{X}) \right\|_2 \leq \delta + \left\| \frac{1}{N} \sum_{j=1}^N (\tilde{Z}_j - \mathbb{E}Z) \right\|_2.$$

By Assumption (iv) (vector sub-exponential mean concentration), there exists $C_2 > 0$ such that with probability at least $1 - 2e^{-t}$,

$$\left\| \frac{1}{N} \sum_{j=1}^N (\tilde{Z}_j - \mathbb{E}Z) \right\|_2 \leq C_2 K_Z \left(\sqrt{\frac{d+t}{N}} + \frac{d+t}{N} \right). \quad (15)$$

Combining,

$$\left\| \frac{1}{N} \tilde{X}^\top r(\tilde{X}) \right\|_2 \leq \delta + C_2 K_Z \left(\sqrt{\frac{d+t}{N}} + \frac{d+t}{N} \right).$$

Plugging (14) and (15) into (13), on an event of probability at least $1 - 4e^{-t}$,

$$|T_1 + T_2| \leq \left[C_1 \sqrt{\|\Sigma\|_{\text{op}}} \sigma_u \sqrt{\frac{d+t}{n}} + \lambda\delta + C_2 \lambda K_Z \left(\sqrt{\frac{d+t}{N}} + \frac{d+t}{N} \right) \right] \|\Delta\|_2. \quad (16)$$

Step 4: Combine and interpret the improvement. Combining (11), (12), and (16) yields: with probability at least $1 - 4e^{-t}$,

$$\frac{1}{2} \left(k_0 - (1-\lambda)\delta_n - \lambda\delta_N \right) \|\Delta\|_2^2 \leq \left[C_1 \sqrt{\|\Sigma\|_{\text{op}}} \sigma_u \sqrt{\frac{d+t}{n}} + \lambda\delta + C_2 \lambda K_Z \left(\sqrt{\frac{d+t}{N}} + \frac{d+t}{N} \right) \right] \|\Delta\|_2.$$

Whenever the curvature term on the left is bounded below by a positive constant (e.g., $\delta_n, \delta_N \leq k_0/2$), we obtain

$$\|\Delta\|_2 \lesssim \sigma_u \sqrt{\frac{d+t}{n}} + \lambda\delta + \lambda \left(\sqrt{\frac{d+t}{N}} + \frac{d+t}{N} \right),$$

up to multiplicative constants depending on k_0 and $\|\Sigma\|_{\text{op}}$. In particular, for $\lambda = 1$, $\sigma_u^2 = \mathbb{E}[(Y - f(X))^2]$ and the leading labeled stochastic term scales with the residual size. Thus, when the predictor is accurate (so $Y - f(X)$ is small) and N is large (so $\sqrt{(d+t)/N}$ is small), the resulting bound can be strictly smaller than the baseline $\lambda = 0$ bound, whose labeled stochastic term is governed by $\sigma_\varepsilon^2 = \mathbb{E}[\varepsilon^2]$.

6 Finite-Sample Analysis of Stratified Prediction-Powered Linear Regression

We study the finite-sample behavior of the stratified prediction-powered estimator introduced in Section ???. Throughout, let θ^* denote the true parameter in the linear model

$$Y = X^\top \theta^* + \varepsilon, \quad \mathbb{E}[\varepsilon | X] = 0.$$

We observe n labeled samples (X_i, Y_i) and N unlabeled samples \tilde{X}_j , together with predictions from a fixed, externally trained predictor $f(\cdot)$. Define the prediction residual

$$r(X) := f(X) - X^\top \theta^*.$$

The data are partitioned into K strata indexed by a random variable $S \in \{1, \dots, K\}$ with $\mathbb{P}(S = k) = w_k$. Let n_k and N_k denote the numbers of labeled and unlabeled samples in stratum k , and define the stratified PPI estimator

$$\hat{\theta} = \arg \min_{\theta} \sum_{k=1}^K w_k L_{k,\lambda_k}^{\text{PP}}(\theta),$$

where

$$L_{k,\lambda_k}^{\text{PP}}(\theta) = \frac{1}{2n_k} \|Y_{(k)} - X_{(k)}\theta\|_2^2 + \lambda_k \left(\frac{1}{2N_k} \|f(\tilde{X}_{(k)}) - \tilde{X}_{(k)}\theta\|_2^2 - \frac{1}{2n_k} \|f(X_{(k)}) - X_{(k)}\theta\|_2^2 \right).$$

6.1 Assumptions

We impose the following standard conditions.

- (A1) $X \in \mathbb{R}^d$ is mean-zero sub-Gaussian with covariance $\Sigma = \mathbb{E}[XX^\top]$ and $\lambda_{\min}(\Sigma) \geq \kappa_0 > 0$.
- (A2) The noise ε is conditionally mean-zero and sub-Gaussian, with $\mathbb{E}[\varepsilon^2 | S = k] \leq \sigma_{\varepsilon,k}^2$.
- (A3) The residual $r(X)$ is deterministic given f , and $\mathbb{E}[r(X)^2 | S = k] \leq \sigma_{r,k}^2$.

6.2 Main finite-sample bound

For each stratum, define the residual

$$u_k := \varepsilon - \lambda_k r(X), \quad \sigma_{u,k}^2 := \mathbb{E}[u_k^2 | S = k],$$

and let $\Delta := \hat{\theta} - \theta^*$.

Theorem 2 (Finite-sample bound for stratified PPI). *Under Assumptions (A1)–(A3), there exists a universal constant $C > 0$ such that, with probability at least $1 - \delta$,*

$$\|\Delta\|_2 \leq \frac{C}{\kappa_0} \left(\sqrt{\frac{d + \log(1/\delta)}{n}} \left(\sum_{k=1}^K w_k \sigma_{u,k}^2 \right)^{1/2} + \sqrt{\frac{d + \log(1/\delta)}{N}} \left(\sum_{k=1}^K w_k \lambda_k^2 \sigma_{Z,k}^2 \right)^{1/2} \right), \quad (17)$$

where $\sigma_{Z,k}^2 := \mathbb{E}[\|Xr(X)\|_2^2 | S = k]$.

Proof. Let $\Delta = \hat{\theta} - \theta^*$. By optimality of $\hat{\theta}$,

$$\sum_{k=1}^K w_k (L_{k,\lambda_k}^{\text{PP}}(\hat{\theta}) - L_{k,\lambda_k}^{\text{PP}}(\theta^*)) \leq 0.$$

Substituting $Y = X\theta^* + \varepsilon$ and $f(X) = X\theta^* + r(X)$ yields

$$\begin{aligned} \sum_{k=1}^K w_k \left(\frac{1-\lambda_k}{2n_k} \|X_{(k)}\Delta\|_2^2 + \frac{\lambda_k}{2N_k} \|\tilde{X}_{(k)}\Delta\|_2^2 \right) &\leq \sum_{k=1}^K w_k \left[\frac{1}{n_k} (\varepsilon_{(k)} - \lambda_k r(X_{(k)}))^\top X_{(k)} \Delta \right. \\ &\quad \left. + \lambda_k \frac{1}{N_k} r(\tilde{X}_{(k)})^\top \tilde{X}_{(k)} \Delta \right]. \end{aligned} \quad (18)$$

Define the weighted empirical covariance

$$\hat{\Sigma} = \sum_{k=1}^K w_k \left(\frac{1-\lambda_k}{n_k} X_{(k)}^\top X_{(k)} + \frac{\lambda_k}{N_k} \tilde{X}_{(k)}^\top \tilde{X}_{(k)} \right).$$

By sub-Gaussian covariance concentration and $\lambda_{\min}(\Sigma) \geq \kappa_0$, with probability at least $1 - \delta/2$, $\lambda_{\min}(\hat{\Sigma}) \geq \kappa_0/2$, so the left-hand side of (18) is bounded below by $\frac{\kappa_0}{4} \|\Delta\|_2^2$.

For the labeled term, using $w_k = n_k/n$,

$$\sum_{k=1}^K \frac{w_k}{n_k} X_{(k)}^\top u_{(k)} = \frac{1}{n} \sum_{i=1}^n X_i u_i,$$

where $u_i = u_{S_i}$. A vector Bernstein inequality yields

$$\left\| \frac{1}{n} \sum_{i=1}^n X_i u_i \right\|_2 \leq C \sqrt{\frac{d + \log(1/\delta)}{n}} \left(\sum_{k=1}^K w_k \sigma_{u,k}^2 \right)^{1/2}.$$

An analogous argument for the unlabeled samples gives

$$\left\| \sum_{k=1}^K w_k \lambda_k \frac{1}{N_k} \tilde{X}_{(k)}^\top r(\tilde{X}_{(k)}) \right\|_2 \leq C \sqrt{\frac{d + \log(1/\delta)}{N}} \left(\sum_{k=1}^K w_k \lambda_k^2 \sigma_{Z,k}^2 \right)^{1/2}.$$

Combining these bounds completes the proof. \square

6.3 Comparison with unstratified PPI

The benefit of stratification can be made explicit by comparing the effective variance terms in (17). Define, for each stratum,

$$V_k(\lambda) := \mathbb{E}[(\varepsilon - \lambda r(X))^2 \mid S = k].$$

If a single global λ is used (unstratified PPI), the smallest achievable variance proxy is

$$V_{\text{glob}}^* := \min_{\lambda} \sum_{k=1}^K w_k V_k(\lambda).$$

By contrast, stratification allows stratum-specific choices λ_k , yielding

$$V_{\text{str}}^* := \sum_{k=1}^K w_k \min_{\lambda} V_k(\lambda).$$

Since $\sum_k w_k \min_{\lambda} V_k(\lambda) \leq \min_{\lambda} \sum_k w_k V_k(\lambda)$, we always have $V_{\text{str}}^* \leq V_{\text{glob}}^*$, with strict inequality whenever the optimal λ_k differ across strata. Consequently, for the same sample sizes (n, N) and curvature constant κ_0 , the stratified estimator admits a finite-sample error bound no larger than that of the optimal unstratified PPI estimator, and is strictly tighter under heterogeneity across strata.

In particular, when $\lambda_k = 1$ for strata in which the predictor is accurate, the labeled variance term involves $\mathbb{E}[(Y - f(X))^2 | S = k]$ rather than $\mathbb{E}[\varepsilon^2 | S = k]$, leading to a potentially substantial reduction in estimation error.