Finite-Sample Analysis of Prediction-Powered Linear Regression

1 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.

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 $f : \mathbb{R}^d \to \mathbb{R}$ 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 $\lambda \in \mathbb{R}$:

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

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:

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

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

$$L^\lambda_{PP,n,N}(\hat{\theta}) \;\le\; L^\lambda_{PP,n,N}(\theta^*).$$

This is the starting point of our basic inequality.

2 Finite-Sample Error Bound for the PPI Estimator

Let $\theta^\star := \arg\min_\theta \mathbb{E}\big[\frac{1}{2}(X^\top\theta - Y)^2\big]$ denote the population risk minimizer under squared loss, and define the residual and prediction bias

$$\varepsilon := Y - X^\top\theta^\star, \qquad r(X) := f(X) - X^\top\theta^\star.$$

Write $\Delta := \hat{\theta}_\lambda - \theta^\star$.

Assumptions. Throughout the finite-sample analysis, we assume:

(A1) The design $X \in \mathbb{R}^d$ is mean-zero sub-Gaussian with $\|X\|_{\psi_2} \le K$ and covariance matrix Σ_X satisfying $\lambda_{\min}(\Sigma_X) \ge \kappa_0 > 0$.

(A2) The noise ε satisfies $\mathbb{E}[\varepsilon \mid X] = 0$ and $\|\varepsilon\|_{\psi_2} \le K_\varepsilon$.

(A3) The prediction residual $r(X) = f(X) - X^\top \theta^\star$ satisfies $\|r(X)\|_{\psi_2} \le K_r$.

Step 1: Algebraic expansion. Recalling the optimality condition $L^\lambda_{PP,n,N}(\hat{\theta}) \le L^\lambda_{PP,n,N}(\theta^*)$ and the definition:

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

We can derive the basic inequality by algebraically expanding the differences in the squared norms:

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

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

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

Substituting into $L^\lambda_{PP,n,N}(\hat{\theta}) - L^\lambda_{PP,n,N}(\theta^*) \le 0$, we obtain:

$$\frac{1-\lambda}{2n}\|X\Delta\|_2^2 + \frac{\lambda}{2N}\|\tilde{X}\Delta\|_2^2 \ \le\ \frac{1}{n}\varepsilon^\top X\Delta + \frac{\lambda}{N} r(\tilde{X})^\top \tilde{X}\Delta - \frac{\lambda}{n} r(X)^\top X\Delta \tag{3}$$

Combining the labeled and unlabeled terms on the right-hand side, we can rewrite the basic inequality as:

$$\frac{1-\lambda}{2n}\|X\Delta\|_2^2 + \frac{\lambda}{2N}\|\tilde{X}\Delta\|_2^2 \ \le\ \underbrace{\frac{1}{n}(\varepsilon - \lambda\, r(X))^\top X\Delta}_{=: T_1(\lambda)} \ +\ \underbrace{\lambda\,\frac{1}{N}\, r(\tilde{X})^\top \tilde{X}\Delta}_{=: T_2(\lambda)}. \tag{4}$$

Step 2: lower bound on the left-hand side. Let

$$\hat{\Sigma}_n := \frac{1}{n}X^\top X, \qquad \hat{\Sigma}_N := \frac{1}{N}\tilde{X}^\top \tilde{X}, \qquad \Sigma := \mathbb{E}[XX^\top],$$

and assume $\lambda_{\min}(\Sigma) \ge \kappa_0 > 0$. Under the sub-Gaussian design assumption $\|X\|_{\psi_2} \le K$, Vershynin's covariance concentration theorem (e.g., [**?**, Theorem 4.6.1]) implies that there exist absolute constants $c, C > 0$ such that, whenever

$$\min(n, N)\ \ge\ c\,K^4\,\frac{d + \log(6/\delta)}{\kappa_0^2}, \tag{5}$$

the following event holds with probability at least $1 - \delta/3$:

$$\mathcal{E}_{\rm op} := \Big\{\|\hat{\Sigma}_n - \Sigma\|_{\rm op} \le \tfrac{\kappa_0}{2}, \quad \|\hat{\Sigma}_N - \Sigma\|_{\rm op} \le \tfrac{\kappa_0}{2}\Big\}. \tag{6}$$

On $\mathcal{E}_{\mathrm{op}}$, Weyl's inequality yields

$$\lambda_{\min}(\hat{\Sigma}_n) \ge \lambda_{\min}(\Sigma) - \|\hat{\Sigma}_n - \Sigma\|_{\mathrm{op}} \ge \tfrac{\kappa_0}{2}, \qquad \lambda_{\min}(\hat{\Sigma}_N) \ge \tfrac{\kappa_0}{2}.$$

Therefore, for any $\lambda \in [0,1]$ and any $\Delta \in \mathbb{R}^d$,

$$\begin{aligned}
\frac{1-\lambda}{2n}\|X\Delta\|_2^2 + \frac{\lambda}{2N}\|\tilde{X}\Delta\|_2^2 &= \frac{1}{2}\,\Delta^\top\Big((1-\lambda)\hat{\Sigma}_n + \lambda\hat{\Sigma}_N\Big)\Delta \\
&\ge \frac{1}{2}\,\lambda_{\min}\Big((1-\lambda)\hat{\Sigma}_n + \lambda\hat{\Sigma}_N\Big)\,\|\Delta\|_2^2 \\
&\ge \frac{1}{2}\,\Big((1-\lambda)\lambda_{\min}(\hat{\Sigma}_n) + \lambda\lambda_{\min}(\hat{\Sigma}_N)\Big)\,\|\Delta\|_2^2 \\
&\ge \frac{1}{2}\cdot\frac{\kappa_0}{2}\,\|\Delta\|_2^2 = \frac{\kappa_0}{4}\,\|\Delta\|_2^2. \qquad (7)
\end{aligned}$$

(Optional refinement). One may equivalently track the explicit concentration radii $\epsilon_n(\delta), \epsilon_N(\delta)$ defined by

$$\|\hat{\Sigma}_n - \Sigma\|_{\mathrm{op}} \le \epsilon_n(\delta), \qquad \|\hat{\Sigma}_N - \Sigma\|_{\mathrm{op}} \le \epsilon_N(\delta),$$

which implies

$$\frac{1}{2}\,\Delta^\top\Big((1-\lambda)\hat{\Sigma}_n + \lambda\hat{\Sigma}_N\Big)\Delta \ge \frac{1}{2}\Big(\kappa_0 - (1-\lambda)\epsilon_n(\delta) - \lambda\epsilon_N(\delta)\Big)\|\Delta\|_2^2.$$

For readability, we use the uniform bound (7) in the main proof.

Step 3: Decomposition of $T_1(\lambda)$ and $T_2(\lambda)$. We expose the bias–fluctuation structure of each term and then combine them to exploit the cancellation built into the PPI objective.

Labeled term. Define

$$A_n(\lambda) := \frac{1}{n}\sum_{i=1}^{n} X_i\big(\varepsilon_i - \lambda r(X_i)\big) = m_1(\lambda) + Z_1(\lambda),$$

where

$$m_1(\lambda) := \mathbb{E}[X\varepsilon] - \lambda\,\mathbb{E}[Xr(X)], \qquad Z_1(\lambda) := A_n(\lambda) - m_1(\lambda).$$

Since $\theta^\star$ minimizes the population squared loss, $\mathbb{E}[X\varepsilon] = 0$, hence

$$m_1(\lambda) = -\lambda\,\mathbb{E}[Xr(X)]. \qquad (8)$$

Unlabeled term. Define

$$B_N := \frac{1}{N}\sum_{j=1}^{N}\tilde{X}_j\,r(\tilde{X}_j) = m_2 + Z_2, \qquad m_2 := \mathbb{E}[Xr(X)], \qquad Z_2 := B_N - m_2.$$

Step 4: Combined bound via bias cancellation. Recall

$$T_1(\lambda) + T_2(\lambda) = \langle A_n(\lambda), \Delta \rangle + \lambda \langle B_N, \Delta \rangle.$$

Substituting the decompositions,

$$T_1(\lambda) + T_2(\lambda) = \langle m_1(\lambda) + \lambda m_2, \Delta \rangle + \langle Z_1(\lambda) + \lambda Z_2, \Delta \rangle.$$

The bias cancels exactly:

$$m_1(\lambda) + \lambda m_2 = -\lambda \mathbb{E}[X r(X)] + \lambda \mathbb{E}[X r(X)] = 0.$$

Therefore, by Cauchy–Schwarz and the triangle inequality,

$$T_1(\lambda) + T_2(\lambda) \le \big(\|Z_1(\lambda)\|_2 + \lambda \|Z_2\|_2 \big) \, \|\Delta\|_2. \tag{9}$$

It remains to control $\|Z_1(\lambda)\|_2$ and $\|Z_2\|_2$.

Bounding $\|Z_1(\lambda)\|_2$. Assume X is mean-zero sub-Gaussian with

$$K_X := \sup_{\|u\|_2 = 1} \|u^\top X\|_{\psi_2} < \infty,$$

and assume ε and $r(X)$ are sub-Gaussian with $\|\varepsilon\|_{\psi_2} \le K_\varepsilon$ and $\|r(X)\|_{\psi_2} \le K_r$. Then each coordinate of $X(\varepsilon - \lambda r(X))$ is sub-exponential. By a vector Bernstein inequality, for any $\delta \in (0,1)$, with probability at least $1 - \delta/2$,

$$\|Z_1(\lambda)\|_2 \;\le\; C\, K_X \sqrt{\frac{d + \log(2/\delta)}{n}} \, \sqrt{\mathrm{Var}\big(\varepsilon - \lambda r(X)\big)}, \tag{10}$$

where

$$\mathrm{Var}\big(\varepsilon - \lambda r(X)\big) = \mathrm{Var}(\varepsilon) + \lambda^2 \,\mathrm{Var}(r(X)) - 2\lambda \,\mathrm{Cov}\big(\varepsilon, r(X)\big). \tag{11}$$

Bounding $\lambda \|Z_2\|_2$. For $u \in \mathbb{S}^{d-1}$,

$$\|u^\top (X r(X))\|_{\psi_1} = \|(u^\top X)\, r(X)\|_{\psi_1} \;\le\; C\, \|u^\top X\|_{\psi_2} \, \|r(X)\|_{\psi_2} \;\le\; C\, K_X\, K_r,$$

so $X r(X)$ is a sub-exponential random vector with parameter $C K_X K_r$. By vector Bernstein, for any $\delta \in (0,1)$, with probability at least $1 - \delta/2$,

$$\lambda \|Z_2\|_2 \;\le\; \lambda\, C\, K_X\, K_r \sqrt{\frac{d + \log(2/\delta)}{N}}. \tag{12}$$

Combined upper bound. Applying a union bound to (10) and (12), we obtain that, with probability at least $1 - \delta$,

$$T_1(\lambda) + T_2(\lambda) \;\le\; \left[C\, K_X \sqrt{\frac{d + \log(2/\delta)}{n}} \, \sqrt{\mathrm{Var}\big(\varepsilon - \lambda r(X)\big)} \;+\; \lambda\, C\, K_X\, K_r \sqrt{\frac{d + \log(2/\delta)}{N}} \right] \|\Delta\|_2. \tag{13}$$

Step 5: Final bound and optimization over λ. Let

$$A := C\,K_X \sqrt{\frac{d+\log(3/\delta)}{n}}, \qquad B := C\,K_X\,K_r \sqrt{\frac{d+\log(3/\delta)}{N}}.$$

On the event $\mathcal{E}_{\mathrm{op}}$ in (6), we have the lower bound (7):

$$\frac{1-\lambda}{2n}\|X\Delta\|_2^2 + \frac{\lambda}{2N}\|\tilde{X}\Delta\|_2^2 \;\geq\; \frac{\kappa_0}{4}\|\Delta\|_2^2, \qquad \forall\lambda\in[0,1].$$

Moreover, by (9)–(13) (with $\delta/3$ in each Bernstein bound), with probability at least $1-2\delta/3$,

$$T_1(\lambda)+T_2(\lambda) \;\leq\; \Big[A\sqrt{\mathrm{Var}(\varepsilon-\lambda r(X))}+B\,\lambda\Big]\|\Delta\|_2.$$

By a union bound with $\mathbb{P}(\mathcal{E}_{\mathrm{op}})\geq 1-\delta/3$, the following holds with probability at least $1-\delta$:

$$\frac{\kappa_0}{4}\|\Delta\|_2^2 \;\leq\; \Big[A\sqrt{\mathrm{Var}(\varepsilon-\lambda r(X))}+B\,\lambda\Big]\|\Delta\|_2.$$

Hence, for any $\lambda\in[0,1]$,

$$\|\Delta\|_2 \;\leq\; \frac{4}{\kappa_0}\Big[A\sqrt{\mathrm{Var}(\varepsilon-\lambda r(X))}+B\,\lambda\Big]. \tag{14}$$

Using the variance expansion

$$\mathrm{Var}(\varepsilon-\lambda r(X)) = \sigma_\varepsilon^2+\lambda^2\sigma_r^2-2\lambda\rho, \quad \sigma_\varepsilon^2 := \mathrm{Var}(\varepsilon),\ \sigma_r^2 := \mathrm{Var}(r(X)),\ \rho := \mathrm{Cov}(\varepsilon, r(X)),$$

we may minimize the right-hand side of (14) over $\lambda\in[0,1]$.

Combining these two displays with the basic inequality (4) yields

$$\|\Delta\|_2 \;\leq\; \frac{4}{\kappa_0}\left[C\,K_X\sqrt{\frac{d+\log(1/\delta)}{n}}\;\sqrt{\sigma_\varepsilon^2+\sigma_r^2\lambda^2-2\rho\lambda} \;+\; \lambda\,C\,K_X\,K_r\sqrt{\frac{d+\log(1/\delta)}{N}}\right]. \tag{15}$$

Applying a union bound over the three events and adjusting constant.

Closed-form minimizer. Consider

$$g(\lambda) = A\sqrt{\sigma_\varepsilon^2+\sigma_r^2\lambda^2-2\rho\lambda}+B\lambda, \qquad \lambda\in[0,1].$$

If $A^2\sigma_r^2\neq B^2$, any interior stationary point satisfies

$$A\frac{\sigma_r^2\lambda-\rho}{\sqrt{\sigma_\varepsilon^2+\sigma_r^2\lambda^2-2\rho\lambda}} = -B,$$

which yields

$$\lambda_{\mathrm{unc}} = \frac{\rho\,(A^2\sigma_r^2-B^2) \;+\; B\sqrt{(\sigma_\varepsilon^2\sigma_r^2-\rho^2)\,(A^2\sigma_r^2-B^2)}}{\sigma_r^2\,(A^2\sigma_r^2-B^2)}. \tag{16}$$

The bound-optimal choice is the projection onto $[0,1]$:

$$\lambda^\star = \min\{1, \max\{0, \lambda_{\mathrm{unc}}\}\}. \tag{17}$$

3 Finite-Sample Bound for Stratified PPI

The data are partitioned into K strata indexed by a random variable $S \in \{1, \ldots, K\}$ 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

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

where

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

3.1 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 $\mathbb{E}[\varepsilon^2 \mid S = k] \leq \sigma^2_{\varepsilon,k}$.

(A3) The residual $r(X)$ is deterministic given f, and $\mathbb{E}[r(X)^2 \mid S = k] \leq \sigma^2_{r,k}$.

3.2 Main Proof

Step 1: Basic inequality. Let $\Delta = \hat{\theta} - \theta^*$. By optimality of $\hat{\theta}$,

$$\sum_{k=1}^{K} w_k\big(L^{\mathrm{PP}}_{k,\lambda_k}(\hat{\theta}) - L^{\mathrm{PP}}_{k,\lambda_k}(\theta^*)\big) \leq 0.$$

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

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

Under the basic inequality, we have

$$\frac{1}{2}\Delta^\top\hat{\Sigma}\Delta \;\leq\; T_1 + T_2,$$

where

$$T_1 := \sum_{k=1}^{K}\frac{w_k}{n_k}\left(\varepsilon_{(k)} - \lambda_k r(X_{(k)})\right)^\top X_{(k)}\Delta, \qquad T_2 := \sum_{k=1}^{K} w_k\frac{\lambda_k}{N_k}\tilde{X}_{(k)} r(\tilde{X}_{(k)})^\top\Delta.$$

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

With these definitions, (18) can be written compactly as

$$\frac{1}{2}\,\Delta^\top\hat{\Sigma}\Delta \;\le\; T_1 + T_2, \tag{19}$$

where

$$T_1 := \sum_{k=1}^{K}\frac{w_k}{n_k}\big(\varepsilon_{(k)} - \lambda_k r(X_{(k)})\big)^\top X_{(k)}\Delta, \qquad T_2 := \sum_{k=1}^{K} w_k\frac{\lambda_k}{N_k} r(\tilde{X}_{(k)})^\top\tilde{X}_{(k)}\Delta.$$

Step 2: Lower bound on the left-hand side. For each stratum, define the sample Gram matrices

$$\hat{\Sigma}_{n,k} := \frac{1}{n_k}X_{(k)}^\top X_{(k)}, \qquad \hat{\Sigma}_{N,k} := \frac{1}{N_k}\tilde{X}_{(k)}^\top\tilde{X}_{(k)}, \qquad \Sigma_k := \mathbb{E}[XX^\top \mid S = k].$$

Under (A1), we use the same sub-Gaussian parameter K for all strata and assume $\lambda_{\min}(\Sigma_k) \ge \kappa_0$ for all k.[1]

By covariance concentration (e.g., Vershynin), there exist absolute constants $c, C > 0$ such that if

$$\min_k(n_k, N_k) \;\ge\; c\,K^4\,\frac{d + \log(6K/\delta)}{\kappa_0^2}, \tag{20}$$

then with probability at least $1 - \delta/3$ the event

$$\mathcal{E}_{\text{op}} := \Big\{\max_{k\in[K]}\|\hat{\Sigma}_{n,k} - \Sigma_k\|_{\text{op}} \le \tfrac{\kappa_0}{2},\;\; \max_{k\in[K]}\|\hat{\Sigma}_{N,k} - \Sigma_k\|_{\text{op}} \le \tfrac{\kappa_0}{2}\Big\}$$

holds. On $\mathcal{E}_{\text{op}}$, Weyl's inequality gives

$$\lambda_{\min}(\hat{\Sigma}_{n,k}) \ge \kappa_0/2, \qquad \lambda_{\min}(\hat{\Sigma}_{N,k}) \ge \kappa_0/2, \qquad \forall k \in [K].$$

Therefore,

$$\begin{aligned}
\frac{1}{2}\,\Delta^\top\hat{\Sigma}\Delta &= \frac{1}{2}\,\Delta^\top\sum_{k=1}^{K} w_k\Big((1-\lambda_k)\hat{\Sigma}_{n,k} + \lambda_k\hat{\Sigma}_{N,k}\Big)\Delta \\
&\ge \frac{1}{2}\sum_{k=1}^{K} w_k\Big((1-\lambda_k)\lambda_{\min}(\hat{\Sigma}_{n,k}) + \lambda_k\lambda_{\min}(\hat{\Sigma}_{N,k})\Big)\|\Delta\|_2^2 \\
&\ge \frac{1}{2}\sum_{k=1}^{K} w_k\cdot\frac{\kappa_0}{2}\,\|\Delta\|_2^2 = \frac{\kappa_0}{4}\,\|\Delta\|_2^2.
\end{aligned} \tag{21}$$

Step 3: Decomposition of T_1 and T_2 (bias cancellation). Define, for each stratum k,

$$A_{n,k}(\lambda_k) := \frac{w_k}{n_k}\,X_{(k)}^\top\big(\varepsilon_{(k)} - \lambda_k r(X_{(k)})\big), \qquad B_{N,k} := \frac{w_k}{N_k}\,\tilde{X}_{(k)}^\top r(\tilde{X}_{(k)}).$$

Then

$$T_1 + T_2 = \Big\langle \sum_{k=1}^{K} A_{n,k}(\lambda_k) + \sum_{k=1}^{K}\lambda_k B_{N,k},\;\Delta\Big\rangle. \tag{22}$$

[1] Equivalently, one may strengthen (A1) to hold conditionally within each stratum.

Let

$$m_k := \mathbb{E}[Xr(X) \mid S = k] \in \mathbb{R}^d.$$

Under (A2) and $\mathbb{E}[X\varepsilon \mid S = k] = 0$, we have

$$\mathbb{E}[A_{n,k}(\lambda_k)] = -w_k \lambda_k \, m_k, \qquad \mathbb{E}[B_{N,k}] = w_k \, m_k.$$

Hence the bias cancels exactly at the level of the weighted sum:

$$\mathbb{E}\left[\sum_{k=1}^K A_{n,k}(\lambda_k) + \sum_{k=1}^K \lambda_k B_{N,k}\right] = 0. \tag{23}$$

Define the centered fluctuations

$$Z_{1,k}(\lambda_k) := A_{n,k}(\lambda_k) - \mathbb{E}[A_{n,k}(\lambda_k)], \qquad Z_{2,k} := B_{N,k} - \mathbb{E}[B_{N,k}].$$

Then by (22)–(23),

$$T_1 + T_2 = \left\langle \sum_{k=1}^K Z_{1,k}(\lambda_k) + \sum_{k=1}^K \lambda_k Z_{2,k}, \; \Delta \right\rangle \le \left(\left\| \sum_{k=1}^K Z_{1,k}(\lambda_k) \right\|_2 + \left\| \sum_{k=1}^K \lambda_k Z_{2,k} \right\|_2 \right) \|\Delta\|_2.$$

Step 4: Fluctuation bounds (vector Bernstein). Fix $\delta \in (0, 1)$. Under (A1)–(A3), each coordinate of $X(\varepsilon - \lambda_k r(X)) \mid (S = k)$ and $Xr(X) \mid (S = k)$ is sub-exponential. A vector Bernstein inequality (applied within each stratum and then combined via Cauchy–Schwarz) yields that with probability at least $1 - \delta/3$,

$$\left\| \sum_{k=1}^K Z_{1,k}(\lambda_k) \right\|_2 \le CK\sqrt{d + \log(6/\delta)} \left(\sum_{k=1}^K \frac{w_k^2}{n_k} \operatorname{Var}\big(\varepsilon - \lambda_k r(X) \mid S = k\big) \right)^{1/2}, \tag{24}$$

and similarly,

$$\left\| \sum_{k=1}^K \lambda_k Z_{2,k} \right\|_2 \le CK\sqrt{d + \log(6/\delta)} \left(\sum_{k=1}^K \frac{w_k^2 \lambda_k^2}{N_k} \mathbb{E}\big[r(X)^2 \mid S = k\big] \right)^{1/2}. \tag{25}$$

Using (A3), $\mathbb{E}[r(X)^2 \mid S = k] \le \sigma_{r,k}^2$.

Combining (3.2)–(25), we obtain that with probability at least $1 - \delta/3$,

$$T_1 + T_2 \le CK\sqrt{d + \log(6/\delta)} \left[\left(\sum_{k=1}^K \frac{w_k^2}{n_k} \operatorname{Var}\big(\varepsilon - \lambda_k r(X) \mid S = k\big) \right)^{1/2} + \left(\sum_{k=1}^K \frac{w_k^2 \lambda_k^2}{N_k} \sigma_{r,k}^2 \right)^{1/2} \right] \|\Delta\|_2. \tag{26}$$

Step 5: Final bound. On the intersection of $\mathcal{E}_{\mathrm{op}}$ and the event in (26), which holds with probability at least $1 - \delta$ by a union bound, the basic inequality (19), the curvature bound (21), and (26) imply

$$\frac{\kappa_0}{4} \|\Delta\|_2^2 \le T_1 + T_2 \le \Gamma(\lambda_{1:K}) \, \|\Delta\|_2,$$

where

$$\Gamma(\lambda_{1:K}) := CK\sqrt{d + \log(6/\delta)} \left[\left(\sum_{k=1}^K \frac{w_k^2}{n_k} \operatorname{Var}\big(\varepsilon - \lambda_k r(X) \mid S = k\big) \right)^{1/2} + \left(\sum_{k=1}^K \frac{w_k^2 \lambda_k^2}{N_k} \sigma_{r,k}^2 \right)^{1/2} \right].$$

Therefore,

$$\|\hat{\theta} - \theta^*\|_2 = \|\Delta\|_2 \leq \frac{4}{\kappa_0}\,\Gamma(\lambda_{1:K}). \tag{27}$$

(Explicit variance expansion). For each stratum,

$$\mathrm{Var}\big(\varepsilon - \lambda_k r(X) \mid S = k\big) = \sigma^2_{\varepsilon,k} + \lambda_k^2 \sigma^2_{r,k} - 2\lambda_k \rho_k, \qquad \rho_k := \mathrm{Cov}(\varepsilon, r(X) \mid S = k),$$

so (27) can be optimized over $\lambda_k \in [0, 1]$ separately for each k.

3.3 Why stratification can improve over unstratified PPI

The bound (27) depends on stratum-wise variance proxies

$$V_k(\lambda) := \mathrm{Var}\big(\varepsilon - \lambda r(X) \mid S = k\big),$$

and on the residual scales $\sigma^2_{r,k} = \mathbb{E}[r(X)^2 \mid S = k]$.

Variance reduction from stratum-specific tuning. An unstratified PPI estimator constrained to use a single global λ is governed by

$$V^*_{\mathrm{glob}} := \min_{\lambda \in [0,1]} \sum_{k=1}^{K} \frac{w_k^2}{n_k}\, V_k(\lambda),$$

whereas stratification allows λ_k to be tuned separately, yielding

$$V^*_{\mathrm{strat}} := \sum_{k=1}^{K} \frac{w_k^2}{n_k} \min_{\lambda \in [0,1]} V_k(\lambda).$$

Since pointwise minimization dominates joint minimization, $V^*_{\mathrm{strat}} \leq V^*_{\mathrm{glob}}$, with strict improvement whenever the stratum-wise optimal λ_k^* differ.

Better use of unlabeled data under heterogeneity. The unlabeled term in (27) is

$$U(\lambda_{1:K}, N_{1:K}) = \left(\sum_{k=1}^{K} \frac{w_k^2 \lambda_k^2}{N_k}\, \sigma^2_{r,k} \right)^{1/2}.$$

Fix $\lambda_{1:K}$ and impose $\sum_{k=1}^{K} N_k = N$. Minimizing

$$\sum_{k=1}^{K} \frac{w_k^2 \lambda_k^2}{N_k}\, \sigma^2_{r,k}$$

over $N_k > 0$ yields (by Lagrange multipliers)

$$N_k^\star = \frac{N\, w_k \lambda_k \sigma_{r,k}}{\sum_{j=1}^{K} w_j \lambda_j \sigma_{r,j}},$$

and therefore

$$U(\lambda_{1:K}, N^\star_{1:K}) = \frac{1}{\sqrt{N}} \sum_{k=1}^{K} w_k \lambda_k \sigma_{r,k}.$$

Hence the optimal unlabeled allocation scales as

$$N_k^\star \propto w_k \lambda_k \sigma_{r,k}.$$

4 Asymptotic Results: PPI for Linear regression

Assume squared loss

$$\ell_\theta(x,y) = \tfrac{1}{2}(x^\top \theta - y)^2,$$

and define

$$\varepsilon := Y - X^\top \theta^\star, \qquad e_f := f(X) - X^\top \theta^\star, \qquad \Sigma_X := \mathbb{E}[XX^\top].$$

Then

$$\nabla \ell_{\theta^\star}(X,Y) = -X\varepsilon, \qquad \nabla \ell^f_{\theta^\star}(X) = -Xe_f, \qquad H_{\theta^\star} = \Sigma_X.$$

Asymptotic covariance. The asymptotic covariance reduces to

$$\Sigma^\lambda = \Sigma_X^{-1}\Big(r\lambda^2 \operatorname{Cov}(Xe_f) + \operatorname{Cov}\big(X(\lambda e_f - \varepsilon)\big)\Big)\Sigma_X^{-1}$$

If the linear model is correctly specified, $\operatorname{Cov}(e_f, \varepsilon) = 0$, then

$$\operatorname{Cov}(Xe_f, X\varepsilon) = 0,$$

and therefore

$$\Sigma^\lambda = \Sigma_X^{-1}\Big((1+r)\lambda^2 \operatorname{Cov}(Xe_f) + \operatorname{Cov}(X\varepsilon)\Big)\Sigma_X^{-1}.$$

Optimal $\lambda^\star$. The trace-minimizing $\lambda^\star$ becomes

$$\lambda^\star = \frac{\operatorname{Tr}\Big(\Sigma_X^{-1}\operatorname{Cov}(X\varepsilon, Xe_f)\Sigma_X^{-1}\Big)}{(1+r)\operatorname{Tr}\Big(\Sigma_X^{-1}\operatorname{Cov}(Xe_f)\Sigma_X^{-1}\Big)}.$$

Under correct linear specification (i.e., $\operatorname{Cov}(X\varepsilon, Xe_f) = 0$),

$$\lambda^\star = 0.$$