BEYOND POINT PREDICTIONS: DISTRIBUTION-INFORMED PREDICTION-POWERED INFERENCE

Anonymous authors
Paper under double-blind review

ABSTRACT

Prediction-powered inference (PPI) typically relies on point predictions on unlabeled data. When predictive distributions are available as in a wide range of applications, including predictions from LLMs, we introduce *distribution-informed prediction-powered inference* (DiPPI), a general framework for further improving statistical efficiency by using predictive distributions as auxiliary information in the spirit of PPI. We characterize the optimal use of this information through score calibration, derive the oracle efficiency for a finite-dimensional representation of the predictive distribution, and provide theoretical guarantees for positive learning with the cross-fitted DiPPI estimator. Through simulations and three real-data applications, we show that DiPPI achieves better efficiency than PPI methods based on point predictions. These gains arise when the predictive distribution contains score-relevant information that is partially lost in point predictions.

1 INTRODUCTION

In many scientific and social-economic studies, reliable outcome labeling or measurements often require costly experiments or human annotation. As a result, responses and covariates are jointly observed in a small labeled sample, whereas covariates alone are available for a much larger unlabeled sample. Pretrained machine learning models, including large language models (LLMs), can predict or impute responses from the available covariates and additional information, but treating these predictions as observed responses can lead to biased estimates and invalid confidence intervals. Prediction-powered inference (PPI; Angelopoulos et al., 2023a) addresses this problem by combining predictions on the unlabeled sample with a correction estimated from the labeled sample, yielding valid inference on population parameters even when the predictions are biased. Using predictions in this way can improve efficiency relative to the estimator based on the labeled sample alone. A rapidly growing literature has since developed adaptive weighting, recalibration and other strategies to further improve the efficiency of prediction-powered inference (Angelopoulos et al., 2023b; Miao et al., 2025; Gan et al., 2024; Ji et al., 2026), among others.

Existing PPI methods typically use a point prediction of the response to construct the correction and to improve statistical efficiency. However, prediction systems such as LLMs are designed to provide a predictive distribution for each case, that is, a probability distribution over the response space. For example, an LLM used to predict a product rating from review text may provide a probability distribution over the possible ratings. Reducing such a predictive distribution to a point prediction, such as its mean or mode, may discard information relevant to estimation of the population parameter of interest. This suggests shifting the focus of prediction-powered inference from point predictions to the richer information contained in predictive distributions.

To illustrate, suppose we wish to estimate a school district coefficient in a linear regression of home sale price (per square foot) on school district and other property characteristics, where the coefficient captures the school district premium in property values. An LLM may use a property's address to retrieve listing descriptions and recent sales records for comparable properties in the area, and then produce a predictive distribution for the property price. Two properties with the same observed covariates in the regression may have different predictive distributions because the retrievable information differs. If these distributional differences contain information about the expected sale price

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

beyond the regression covariates and the point prediction, they can improve statistical efficiency within the PPI framework. More broadly, this motivates a principled shift in PPI methodology from relying on point predictions for unlabeled data toward exploiting predictive distributions, or their low-dimensional summaries, as richer auxiliary information.

In this paper, we propose distribution-informed prediction-powered inference (DiPPI), a conceptually simple approach, in the spirit of PPI, that uses predictive distributions as auxiliary information for inference on parameters defined by estimating equations. Although the idea behind DiPPI is conceptually natural, we are not aware of any existing PPI variant that adopts it. One reason may be the lack of a straightforward way to exploit the predictive distribution as a whole. To address this challenge, DiPPI builds on recalibrated prediction-powered inference (RePPI; Ji et al., 2026) and uses the labeled sample to estimate the conditional expectation of the estimating function given the covariates and predictive distribution. The fitted conditional expectation is then evaluated for both labeled and unlabeled observations and used to construct the DiPPI estimating equation.

The shift from point predictions to predictive distributions requires rigorous justification. We provide asymptotic results and numerical experiments to address when and why the proposed DiPPI yields meaningful efficiency gains over existing methods. For the DiPPI estimator, we establish asymptotic normality and provide a consistent covariance estimator for confidence intervals. Under appropriate regularity conditions, DiPPI is asymptotically no less efficient than labeled-only estimation, even when the conditional expectation is not estimated consistently. When the conditional expectation is estimated consistently, DiPPI achieves the smallest asymptotic covariance among estimators within our estimating-equation framework. In other words, DiPPI shares two desirable asymptotic properties with RePPI: (i) it is never less efficient than the analysis based on the labeled sample alone; and (ii) if the conditional expectation is estimated consistently, it achieves the minimum asymptotic variance among all prediction-powered estimators in the class under consideration. Moreover, when the predictive distributions capture (some level of) heterogeneity in the uncertainty of the point predictions, DiPPI delivers additional efficiency gains over RePPI.

We evaluate DiPPI through simulations and three real-data applications: estimating the log-price coefficient for wine ratings (Ji et al., 2026), slope coefficients in multinomial logistic regression of forest cover types (Blackard, 1998), and the hedging-indicator coefficient for politeness ratings (Danescu-Niculescu-Mizil et al., 2013). Wine and Politeness use predictive rating distributions from GPT-5.4 nano, while Forest uses class-probability predictions from a held-fixed XGBoost classifier (Chen & Guestrin, 2016). Across applications and labeled sample sizes, DiPPI produces shorter confidence intervals than PPI++ (Angelopoulos et al., 2023b), RePPI (Ji et al., 2026) and labeled-only estimation, with coverage close to the nominal level.

1.1 Related Work

PPI (Angelopoulos et al., 2023a) combines model predictions on a large unlabeled sample with a correction estimated from a small labeled sample, following earlier work on post-prediction inference (Wang et al., 2020). The literature on PPI is expanding rapidly, and we focus here on developments most closely related to our setting. Subsequent work improves efficiency through power tuning (Angelopoulos et al., 2023b), prediction de-correlation (Gan et al., 2024), and adaptive weighting (Miao et al., 2025), post-hoc regression and calibration (Eyre & Madras, 2025; van der Laan & van der Laan, 2026), empirical likelihood with auxiliary moment conditions (Wang et al., 2025), among others. Closest to our work, RePPI (Ji et al., 2026) estimates the conditional expectation of the estimating function given covariates and a point prediction. These methods typically use point predictions as auxiliary information and are closely related to classical control-variate methods and estimation with surrogate outcomes or partially observed responses (Robins et al., 1994; Chen & Chen, 2000; Chen et al., 2008; Gronsbell et al., 2024).

2 Distribution-Informed Prediction-Powered Inference (DiPPI)

Let $X \in \mathcal{X} \subseteq \mathbb{R}^p$ denote a covariate vector and $Y \in \mathcal{Y} \subseteq \mathbb{R}$ a response. For a known estimating function $U_\theta : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}^d$, the parameter of interest $\theta^\star \in \Theta \subseteq \mathbb{R}^d$ satisfies

$$\mathbb{E}[U_{\theta^\star}(X, Y)] = 0,$$

where the expectation is taken under the joint distribution of (X, Y). For example, $U_{\theta^\star}(X, Y)$ may be a score function $\nabla_\theta \log p_\theta(Y|X)$ for a conditional probability mass function or density p_θ.

A fixed prediction system $\mathcal{A}$ uses (X, Z) to produce a predictive distribution $Q = \mathcal{A}(X, Z) \in \mathcal{P}(\mathcal{Y})$, where $\mathcal{P}(\mathcal{Y})$ denotes the space of probability distributions on $\mathcal{Y}$ and Z denotes additional supplied or retrieved inputs. For example, Q may specify probabilities over possible 1-10 ratings or a predictive distribution for a continuous response. A point prediction, such as the mean or mode of Q, summarizes this distribution but may discard information relevant to the target estimating equation. Our framework uses Q as auxiliary information and does not require it to equal the conditional distribution of Y given (X, Z).

We observe a labeled sample of size n and an unlabeled sample of size N,

$$\mathcal{D}_L = \{(X_i, Y_i, Q_i) : i = 1, \ldots, n\}, \qquad \mathcal{D}_U = \{(X_j, Q_j) : j = n+1, \ldots, n+N\},$$

where $Q_i = \mathcal{A}(X_i, Z_i)$. We assume that the labeled observations are i.i.d. from the joint distribution of (X, Y, Q), that the unlabeled observations are i.i.d. from the joint distribution of (X, Q), and that the two samples are independent. Thus, both samples share the same marginal distribution of (X, Q), while responses are observed only in $\mathcal{D}_L$.

Remark 1. *The inputs Z may be available to us but difficult to model directly, or accessible only to the prediction system. For example, in a regression of wine ratings, Z may contain a textual wine description supplied to an LLM. In a regression of housing price, Z may contain listing descriptions and historical property records retrieved using an address not included in X. Our procedure uses Z only through Q, without requiring direct access to or modeling of Z.*

2.1 Distribution-informed estimation

Estimating $\theta^\star$ requires estimating $\mathbb{E}\{U_\theta(X, Y)\}$ as a function of θ. To use the unlabeled sample for this purpose, we use the predictive distribution Q and X to predict $U_\theta(X, Y)$. Let $h_\theta : \mathcal{X} \times \mathcal{P}(\mathcal{Y}) \to \mathbb{R}^d$ be a fixed prediction function. We write

$$\mathbb{E}[U_\theta(X, Y)] = \mathbb{E}[U_\theta(X, Y) - h_\theta(X, Q)] + \mathbb{E}[h_\theta(X, Q)].$$

Since (X, Q) is available in both samples, the second term can be estimated from all $n + N$ observations, whereas the first requires Y and can be estimated only from the n labeled observations. We therefore estimate $\theta^\star$ by solving

$$\Psi_{n,N}(\theta; h) = \frac{1}{n}\sum_{i=1}^{n}\{U_\theta(X_i, Y_i) - h_\theta(X_i, Q_i)\} + \frac{1}{n+N}\sum_{k=1}^{n+N} h_\theta(X_k, Q_k) = 0. \tag{1}$$

Under our sampling assumptions, the left-hand side of (1) has expectation $\mathbb{E}\{U_\theta(X, Y)\}$ for every h_θ, without requiring Q to be the conditional distribution of Y given (X, Z). Taking $h_\theta = 0$ recovers the labeled-only estimating equation.

The choice of h_θ affects the efficiency of the resulting estimator, denoted as $\widehat{\theta}_h$. The following result gives the asymptotic covariance of $\widehat{\theta}_h$ and identifies the optimal h_θ that minimizes this covariance under two regularity conditions.

Assumption 1. *There exists $\rho > 0$ such that $\mathcal{B}_\rho := \{\theta \in \mathbb{R}^d : \|\theta - \theta^\star\| \le \rho\} \subset \mathrm{int}(\Theta)$. The function $U_\theta(X, Y)$ is continuously differentiable in θ on $\mathcal{B}_\rho$. Moreover, $H_{\theta^\star} = \mathbb{E}\{\nabla_\theta U_{\theta^\star}(X, Y)\}$ is nonsingular, and $\mathbb{E}\Big[\|U_{\theta^\star}(X, Y)\|^2 + \sup_{\theta\in\mathcal{B}_\rho}\|\nabla_\theta U_\theta(X, Y)\|\Big] < \infty$.*

Assumption 2. *The function $h_\theta(X, Q)$ is continuously differentiable in θ on $\mathcal{B}_\rho$, and $\mathbb{E}\Big[\|h_{\theta^\star}(X, Q)\|^2 + \sup_{\theta\in\mathcal{B}_\rho}\|\nabla_\theta h_\theta(X, Q)\|\Big] < \infty$.*

Theorem 1. *Suppose Assumptions 1-2 hold and $n/N \to r \in (0, \infty)$. Then there exists $a \in (0, \rho)$ such that, with probability tending to one, $\Psi_{n,N}(\theta; h) = 0$ has a unique root $\widehat{\theta}_h$ in $\mathcal{B}_a = \{\theta : \|\theta - \theta^\star\| \le a\}$, and this root belongs to $\mathrm{int}(\mathcal{B}_a)$. Moreover,*

$$\sqrt{n}(\widehat{\theta}_h - \theta^\star) \xrightarrow{d} N(0, \Sigma_h),$$
$$\Sigma_h = H_{\theta^\star}^{-1}\left[\frac{1}{1+r}\mathrm{Cov}\{U_{\theta^\star}(X, Y) - h_{\theta^\star}(X, Q)\} + \frac{r}{1+r}\mathrm{Cov}\{U_{\theta^\star}(X, Y)\}\right]H_{\theta^\star}^{-\top}. \tag{2}$$

Among all h_θ satisfying Assumption 2, Σ_h is minimized by any $h^\star$ satisfying

$$h^\star_{\theta^\star}(X,Q) = \mathbb{E}\{U_{\theta^\star}(X,Y) \mid X,Q\}. \tag{3}$$

The corresponding minimum covariance is

$$\Sigma_{h^\star} = H^{-1}_{\theta^\star}\left[\mathrm{Cov}\{U_{\theta^\star}(X,Y)\} - \frac{1}{1+r}\mathrm{Cov}\{\mathbb{E}[U_{\theta^\star}(X,Y) \mid X,Q]\}\right] H^{-\top}_{\theta^\star}. \tag{4}$$

Theorem 1 shows that using Q can improve efficiency when $\mathbb{E}\{U_{\theta^\star}(X,Y) \mid X,Q\}$ differs from $\mathbb{E}\{U_{\theta^\star}(X,Y) \mid X\}$. This conditional expectation may depend on features of Q beyond its mean, such as shape, variance or tail probabilities. Using these features can therefore improve prediction of the estimating function and reduce the asymptotic covariance. We discuss the estimation of $\mathbb{E}\{U_{\theta^\star}(X,Y) \mid X,Q\}$ in Section 2.3. Taking $h_\theta = 0$ recovers labeled-only estimation, while $h_\theta(X,Q) = (1+n/N)U_\theta(X,f(X))$ recovers PPI with point prediction $f(X)$. Thus, the optimal $h^\star_{\theta^\star}(X,Q)$ in Theorem 1 yields asymptotic covariance no greater than that of either estimator.

2.2 Beyond point predictions

Let $\widehat{Y} = T(Q)$ denote a point prediction obtained from Q through a summary function T. We study when using Q improves asymptotic efficiency. A direct way to predict $U_\theta(X,y)$ using Q is

$$s_\theta(X,Q) = \int U_\theta(X,y)\, Q(dy).$$

When s_θ satisfies Assumption 2, it can be used as h_θ in (1). However, $s_{\theta^\star}(X,Q)$ need not equal the optimal predictor $h^\star_{\theta^\star}(X,Q) = \mathbb{E}[U_{\theta^\star}(X,Y) \mid X,Q]$.

For mean estimation and the usual linear and logistic regression estimating functions,

$$s_{\theta^\star}(X,Q) = U_{\theta^\star}(X,m(Q)), \qquad h^\star_{\theta^\star}(X,Q) = U_{\theta^\star}(X,\mathbb{E}[Y \mid X,Q]),$$

where $m(Q) = \int y\, Q(dy)$. The two response predictions differ by

$$\mathbb{E}[Y \mid X,Q] - m(Q) = \mathbb{E}[Y - m(Q) \mid X,Q],$$

the conditional mean prediction error of $m(Q)$. Features such as the variance, skewness, or tail probabilities of Q may help predict this error conditional on $(X,m(Q))$. If $\mathbb{E}[Y|X,Q] = \mathbb{E}[Y|X,m(Q)]$, then the optimal prediction of $U_\theta(X,Y)$ based on (X,Q) is also a function of $(X,m(Q))$. In this case, having access to the full predictive distribution gives the same minimum asymptotic covariance as having access to only its mean.

For other estimating functions, s_θ may depend on features of Q beyond its mean. In multinomial logistic regression with class K as the reference category, let $p_{\theta,k}(X)$ and q_k denote the model and predicted probabilities for class k. For $k = 1,\ldots,K-1$,

$$U_{\theta,k}(X,Y) = X\{\mathbf{1}(Y=k) - p_{\theta,k}(X)\}, \qquad s_{\theta,k}(X,Q) = X\{q_k - p_{\theta,k}(X)\}.$$

Hence, s_θ depends on the class probabilities, which are generally not determined by the predictive mean or modal class; reducing Q to either summary can therefore lose relevant information.

We now compare the minimum asymptotic covariances achievable with Q and $\widehat{Y} = T(Q)$. Define

$$h^{\star,T}_{\theta^\star}(X,\widehat{Y}) = \mathbb{E}[U_{\theta^\star}(X,Y) \mid X,\widehat{Y}].$$

By Theorem 1, this predictor minimizes the asymptotic covariance among prediction functions based on $(X,\widehat{Y})$. Denote this minimum by $\Sigma^\star_T$. The corresponding minimum based on (X,Q) is $\Sigma_{h^\star}$. The law of total covariance gives

$$\Sigma^\star_T - \Sigma_{h^\star} = \frac{1}{1+r} H^{-1}_{\theta^\star}\mathrm{Cov}\left\{h^\star_{\theta^\star}(X,Q) - h^{\star,T}_{\theta^\star}(X,\widehat{Y})\right\} H^{-\top}_{\theta^\star} \succeq 0. \tag{5}$$

Thus, replacing Q by $\widehat{Y}$ gives no loss of efficiency if and only if

$$\mathbb{E}[U_{\theta^\star}(X,Y) \mid X,Q] = \mathbb{E}[U_{\theta^\star}(X,Y) \mid X,\widehat{Y}].$$

The predictor $h^{\star,T}_{\theta^\star}$ is the conditional expectation targeted by RePPI (Ji et al., 2026). Thus, (5) shows that retaining Q can improve efficiency even relative to a point-based method that consistently estimates this conditional expectation. This comparison assumes that $\widehat{Y} = T(Q)$ and does not establish an efficiency ordering for separately generated predictions that are not determined by (X,Q).

2.3 CROSS-FITTED ESTIMATION AND MATRIX CALIBRATION

To use Q as a regression input, we choose a finite-dimensional representation $\Phi(Q) \in \mathbb{R}^q$. For a discrete response with K possible values, $\Phi(Q)$ can consist of $K-1$ class probabilities. For a continuous response, the choice depends on how the prediction system provides Q. An available density or CDF can be represented by basis coefficients or functional principal component scores. If only generated responses are available, these functions must be approximated from repeated samples. In this case, selected moments or quantiles may offer a simpler representation; a practical starting point for $\Phi(Q)$ is the predictive mean, standard deviation, and a few quantiles.

We then estimate $\mathbb{E}\{U_{\theta^\star}(X,Y) \mid X, \Phi(Q)\}$ from the labeled data using three-fold cross-fitting. Because $\theta^\star$ is unknown, one fold provides an initial estimator $\widehat{\theta}_0$. A second fold is used to construct $\widehat{\mu}(X,Q) \approx \mathbb{E}\{U_{\widehat{\theta}_0}(X,Y) \mid X, \Phi(Q)\}$. Since an inaccurate estimate of the conditional score can increase variance, we calibrate $\widehat{\mu}(X,Q)$ by a matrix M chosen to minimize the asymptotic covariance. This step is related to optimal control-variate and power-tuning constructions (Chen & Chen, 2000; Angelopoulos et al., 2023b; Gan et al., 2024; Gronsbell et al., 2024). Replacing $\theta^\star$ by $\widehat{\theta}_0$ and the population covariances in this minimizer by empirical covariances on a third labeled fold gives

$$\widehat{M} = \widehat{\mathrm{Cov}}\Big\{U_{\widehat{\theta}_0}(X,Y), \widehat{\mu}(X,Q)\Big\}\, \widehat{\mathrm{Cov}}\{\widehat{\mu}(X,Q)\}^{-1}. \tag{6}$$

We refer to this step as *matrix calibration*. If the empirical covariance is singular, its generalized inverse may be used. The final estimator is $\widehat{h}(X,Q) = \widehat{M}\widehat{\mu}(X,Q)$.

The procedure is summarized in Algorithm 1. We split the labeled sample into three folds $\mathcal{D}_1, \mathcal{D}_2, \mathcal{D}_3$. For each $k \in \{1,2,3\}$, $\mathcal{D}_k$ is used for calibration and inference, while the other two folds separately provide $\widehat{\theta}_0^k$ and $\widehat{\mu}^k$. Hence, $(\widehat{\theta}_0^k, \widehat{\mu}^k)$ is independent of $\mathcal{D}_k$ and the unlabeled sample. We combine the resulting estimates with weights $\omega_k = |\mathcal{D}_k|/n$.

Algorithm 1: Distribution-Informed Prediction-Powered Inference (DiPPI)

Step 1: Split the labeled sample into three folds $\mathcal{D}_1$, $\mathcal{D}_2$, and $\mathcal{D}_3$.
Step 2: On $\mathcal{D}_2$, compute $\widehat{\theta}_0^1$ as the labeled-only local root of $|\mathcal{D}_2|^{-1}\sum_{i\in\mathcal{D}_2} U_\theta(X_i,Y_i) = 0$.
Step 3: On $\mathcal{D}_3$, fit $\widehat{\mu}^1(X,Q) \approx \mathbb{E}\{U_{\widehat{\theta}_0^1}(X,Y) \mid X, \Phi(Q)\}$.
Step 4: On $\mathcal{D}_1$, compute $\widehat{M}^1$ using (6).
Step 5: Using $\mathcal{D}_1$ and the full unlabeled sample, solve

$$0 = \frac{1}{|\mathcal{D}_1|}\sum_{i\in\mathcal{D}_1}\Big[U_\theta(X_i,Y_i) - \frac{1}{1+n/N}\widehat{M}^1\widehat{\mu}^1(X_i,Q_i)\Big] + \frac{1}{N}\sum_{j=n+1}^{n+N}\frac{1}{1+n/N}\widehat{M}^1\widehat{\mu}^1(X_j,Q_j),$$

and denote the solution by $\widehat{\theta}^1$.
Step 6: Repeat Steps 2–5 with $(\mathcal{D}_3,\mathcal{D}_1,\mathcal{D}_2)$ and $(\mathcal{D}_1,\mathcal{D}_2,\mathcal{D}_3)$; obtain $\widehat{\theta}^2$ and $\widehat{\theta}^3$.
Step 7: Return $\widehat{\theta}^{\mathrm{DiPPI}} = \sum_{k=1}^3 \omega_k \widehat{\theta}^k$.

In Step 3, $\widehat{\mu}(X,Q)$ may be estimated using any suitable regression method with input $(X, \Phi(Q))$, including linear, regularized, additive, tree-based, or neural-network models. The choice should reflect the labeled sample size and the complexity and dimension of the regression problem. Model selection and tuning are performed within the fold used to fit $\widehat{\mu}$.

2.4 THEORETICAL GUARANTEES AND INFERENCE

We establish asymptotic guarantees for the cross-fitted estimator $\widehat{\theta}^{\mathrm{DiPPI}}$ in Algorithm 1, with estimation of the prediction functions and calibration matrices. We also establish consistency of the covariance estimator and asymptotic validity of the resulting Wald confidence intervals.

Assumption 3. *For each $k = 1,2,3$, the fold size satisfies $n_k/n \to \pi_k \in (0,1)$, and $\mathbb{E}\Big[\sup_{\theta\in\mathcal{B}_\rho} \|\nabla_\theta U_\theta(X,Y)\|^2\Big] < \infty$.*

Theorem 2. *Suppose Assumptions 1 and 3 hold and $n/N \to r \in (0,\infty)$. If, for some square-integrable $\mu(X,Q)$ with $\mathrm{Cov}\{\mu(X,Q)\} \succ 0$, $\max_{1\le k\le 3}\mathbb{E}\{\|\widehat{\mu}^k(X,Q)-\mu(X,Q)\|^2\} = o_p(1)$, then $\widehat{\theta}^{\mathrm{DiPPI}} \xrightarrow{p} \theta^\star$ and $\sqrt{n}(\widehat{\theta}^{\mathrm{DiPPI}}-\theta^\star) \xrightarrow{d} N\left(0,\Sigma_\mu^{\mathrm{DiPPI}}\right)$, where*

$$\Sigma_\mu^{\mathrm{DiPPI}} = H_{\theta^\star}^{-1}\Big[\mathrm{Cov}\{U_{\theta^\star}(X,Y)\} - \frac{1}{1+r}\mathrm{Cov}\{U_{\theta^\star}(X,Y),\mu(X,Q)\}\mathrm{Cov}\{\mu(X,Q)\}^{-1} \times \mathrm{Cov}\{\mu(X,Q),U_{\theta^\star}(X,Y)\}\Big]H_{\theta^\star}^{-\top}. \tag{7}$$

Corollary 3. *Under the conditions of Theorem 2, let $\Sigma^{\mathrm{C}} = H_{\theta^\star}^{-1}\mathrm{Cov}\{U_{\theta^\star}(X,Y)\}H_{\theta^\star}^{-\top}$ denote the labeled-only asymptotic covariance. Then*

$$\Sigma^{\mathrm{C}} - \Sigma_\mu^{\mathrm{DiPPI}} = \frac{1}{1+r}H_{\theta^\star}^{-1}\mathrm{Cov}\{U_{\theta^\star}(X,Y),\mu(X,Q)\}\mathrm{Cov}\{\mu(X,Q)\}^{-1} \times \mathrm{Cov}\{\mu(X,Q),U_{\theta^\star}(X,Y)\}H_{\theta^\star}^{-\top} \succeq 0. \tag{8}$$

Thus, DiPPI is asymptotically no less efficient than labeled-only inference, even when $\mu(X,Q)$ differs from $\mathbb{E}[U_{\theta^\star}(X,Y) \mid X,\Phi(Q)]$. If, in addition, $\mu(X,Q) = E\{U_{\theta^\star}(X,Y) \mid X,\Phi(Q)\}$, then

$$\Sigma_\mu^{\mathrm{DiPPI}} = H_{\theta^\star}^{-1}\Big[\mathrm{Cov}\{U_{\theta^\star}(X,Y)\} - \frac{1}{1+r}\mathrm{Cov}\{\mathbb{E}[U_{\theta^\star}(X,Y) \mid X,\Phi(Q)]\}\Big]H_{\theta^\star}^{-\top}. \tag{9}$$

Corollary 3 shows that $\Sigma_\mu^{\mathrm{DiPPI}} \preceq \Sigma^{\mathrm{C}}$ even if $\mu(X,Q) \neq \mathbb{E}\{U_{\theta^\star}(X,Y) \mid X,\Phi(Q)\}$. When $\mu(X,Q) = \mathbb{E}\{U_{\theta^\star}(X,Y) \mid X,\Phi(Q)\}$, DiPPI attains the minimum asymptotic covariance among prediction functions based on $(X,\Phi(Q))$ satisfying Assumption 2.

For variance estimation, let $\widehat{\mathrm{Cov}}_{\mathcal{D}_k}$ and $\widehat{\mathrm{Cov}}_{\mathrm{U}}$ denote empirical covariances over $\mathcal{D}_k$ and the unlabeled sample, respectively. Because $\omega_k/|\mathcal{D}_k| = 1/n$, each labeled observation receives weight $1/n$ after the fold estimates are averaged. All folds use the same unlabeled observations, so their unlabeled corrections are averaged before computing the covariance. We estimate the asymptotic covariance by

$$\widehat{\Sigma}^{\mathrm{DiPPI}} = \widehat{H}^{-1}\left[\sum_{k=1}^{3}\omega_k\widehat{\mathrm{Cov}}_{\mathcal{D}_k}\left\{U_{\widehat{\theta}^{\mathrm{DiPPI}}}(X_i,Y_i) - \frac{\widehat{M}^k\widehat{\mu}^k(X_i,Q_i)}{1+n/N}\right\} + \frac{n}{N}\widehat{\mathrm{Cov}}_{\mathrm{U}}\left\{\frac{1}{1+n/N}\sum_{k=1}^{3}\omega_k\widehat{M}^k\widehat{\mu}^k(X_j,Q_j)\right\}\right]\widehat{H}^{-\top}, \tag{10}$$

For smooth scores, one may take $\widehat{H} = n^{-1}\sum_{i=1}^{n}\nabla_\theta U_{\widehat{\theta}^{\mathrm{DiPPI}}}(X_i,Y_i)$. More generally, we only require $\widehat{H} \xrightarrow{p} H_{\theta^\star}$. The following theorem gives covariance consistency.

Theorem 4. *Suppose the conditions of Theorem 2 hold and $\widehat{H} \xrightarrow{p} H_{\theta^\star}$. Then*

$$\widehat{\Sigma}^{\mathrm{DiPPI}} \xrightarrow{p} \Sigma_\mu^{\mathrm{DiPPI}}. \tag{11}$$

For $a \in \mathbb{R}^d$ satisfying $a^\top\Sigma_\mu^{\mathrm{DiPPI}}a > 0$,

$$\frac{\sqrt{n}\,a^\top(\widehat{\theta}^{\mathrm{DiPPI}}-\theta^\star)}{\{a^\top\widehat{\Sigma}^{\mathrm{DiPPI}}a\}^{1/2}} \xrightarrow{d} N(0,1), \tag{12}$$

and the Wald interval $a^\top\widehat{\theta}^{\mathrm{DiPPI}} \pm z_{1-\alpha/2}(\frac{a^\top\widehat{\Sigma}^{\mathrm{DiPPI}}a}{n})^{1/2}$ has asymptotic coverage $1-\alpha$ for $a^\top\theta^\star$.

3 Simulation Studies

We consider two simulation settings in which predictive variance provides information beyond the predictive mean. In the first setting, observations with the same covariates and predictive mean can have different expected responses, and the predictive variance provides information about these differences. In the second setting, predictive variance indicates which point predictions are more reliable, while the overall prediction bias and mean squared error remain fixed. Additional simulations are reported in Appendix A.

We compare labeled-only estimation (Classical), PPI++, RePPI, and DiPPI. We specify Q directly, with PPI++ and RePPI using its mean $\widehat{Y} = \mathbb{E}_Q(Y)$ and DiPPI using the distributional features $\Phi(Q)$ defined below:

$$\Phi(Q) = \big(\mathrm{mean}(Q), \mathrm{sd}(Q), q_{0.1}(Q), q_{0.25}(Q), q_{0.5}(Q), q_{0.75}(Q), q_{0.9}(Q)\big).$$

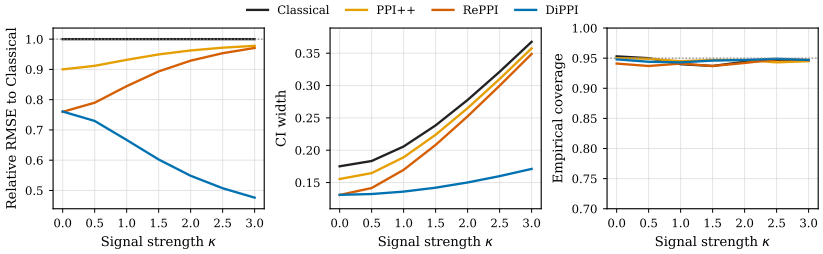
We estimate the conditional mean of Y using linear regression in Experiment 1 and histogram gradient boosting in Experiment 2. Both experiments use $N = 10{,}000$ and 200 replications. We report relative RMSE to Classical, mean 95% confidence-interval width, and empirical coverage. For method m, relative $\mathrm{RMSE}_m = \left\{\sum_{b=1}^{R} \|\widehat{\theta}_m^{(b)} - \theta\|_2^2 / \sum_{b=1}^{R} \|\widehat{\theta}_{\mathrm{C}}^{(b)} - \theta\|_2^2\right\}^{1/2}$, where $\widehat{\theta}_{\mathrm{C}}^{(b)}$ denotes the Classical estimator. Classical therefore has relative RMSE one, with smaller values indicating greater accuracy. We report confidence-interval width and coverage averaged across the slope coefficients.

Experiment 1: Outcome-relevant predictive variance. We examine whether predictive variance can improve inference beyond the predictive mean. Let $X, Z \in \mathbb{R}^5$ have independent standard-normal coordinates, let $V \sim \mathrm{Unif}(1, 4)$, and let $\varepsilon \sim N(0, 1)$, independently. We generate

$$Y = X^\top\theta + Z^\top\beta + \kappa(V - 2.5) + \varepsilon, \qquad Q = N(Z^\top\beta, V), \qquad \widehat{Y} = Z^\top\beta.$$

Here, V affects the conditional mean of Y but appears only in the variance of Q. Thus, Q provides information about Y beyond $(X, \widehat{Y})$ when $\kappa \neq 0$. We estimate the slope coefficients θ with $n = 1,000$ and $N = 10,000$, varying $\kappa \in \{0, 0.5, 1, 1.5, 2, 2.5, 3\}$. Figure 1(A) shows similar performance for DiPPI and RePPI when $\kappa = 0$. As κ increases, DiPPI achieves larger efficiency gains over RePPI, with coverage remaining close to nominal.

(A) Experiment 1: Outcome-relevant predictive variance



(B) Experiment 2: Heterogeneous prediction-error variances

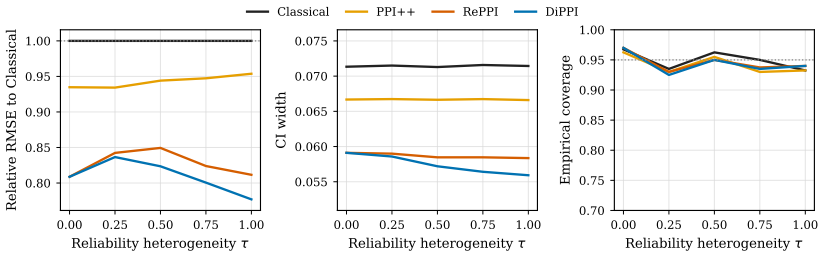


Figure 1: Relative RMSE to Classical, mean 95% CI width, and empirical coverage.

Experiment 2: Heterogeneous prediction-error variances.

We next consider point predictions whose accuracy differs across observations. Generate

$$X_i \sim N(0, I_2), \qquad Y_i = X_i^\top\theta^\star + \varepsilon_i, \qquad \theta^\star = (1, -1)^\top/\sqrt{2}, \qquad \varepsilon_i \sim N(0, 1).$$

Independently, draw A_i uniformly from $\{-1, 1\}$ and set

$$S_i = \frac{\exp(\tau A_i)}{\sqrt{\cosh(2\tau)}}, \qquad \xi_i \mid S_i \sim N(-2, S_i^2), \qquad \widehat{Y}_i = Y_i + \xi_i, \qquad Q_i = N(\widehat{Y}_i, S_i^2).$$

The parameter $\tau \in \{0, 0.25, 0.5, 0.75, 1\}$ controls the variation in prediction-error variance across observations. The normalization gives $\mathbb{E}(S_i^2) = 1$ and $\mathbb{E}\{(\widehat{Y}_i - Y_i)^2\} = 5$, so the prediction bias and mean squared error remain fixed as τ varies. At $\tau = 0$, $S_i = 1$, and Q_i provides no information beyond $\widehat{Y}_i$. For $\tau > 0$, its variance indicates how accurately $\widehat{Y}_i$ measures Y_i. This information helps estimate the conditional mean of Y_i given X_i and $\widehat{Y}_i$, since predictions with smaller error variances receive greater weight. We estimate the regression slopes $\theta^\star$ with $n = 3{,}000$ and $N = 10{,}000$. Figure 1(B) shows similar performance for DiPPI and RePPI at $\tau = 0$, with increasing efficiency gains for DiPPI as τ increases.

4 Real-Data Experiments

We evaluate DiPPI using Wine, Forest, and Politeness data. Wine and Politeness examine whether predictive distributions improve inference beyond their means, while Forest examines whether class probabilities improve inference beyond point predictions in multinomial regression.

We compare Classical, PPI++, RePPI, and DiPPI using point predictions derived from the same predictive distribution Q used by DiPPI. DiPPI and RePPI use identical data folds and histogram gradient boosting to estimate conditional scores. For each labeled ratio r, we independently draw $n = rN$ labeled and N unlabeled observations with replacement from the analysis population and repeat the experiment 200 times. Using the full-population estimate as the ground-truth target, we report relative RMSE, mean 95% CI width, and empirical coverage, as in Section 3.

4.1 Wine ratings

The Wine data [1] contain $10,000$ U.S. reviews with ratings from 80 to 99. We regress rating on log price and indicators for California, Washington, and Oregon, and target the log-price coefficient. We set $r \in \{0.05, 0.10, 0.15, 0.20, 0.25\}$. We use GPT-5.4 nano to construct a predictive distribution over all 20 possible ratings from the written review. For each of PPI++ and RePPI, we evaluate two variants for point predictions, one using the predictive mean $\widehat{Y} = \mathbb{E}_Q(Y)$ and the other using the modal rating $\widehat{Y} = \arg\max_{y \in \{80,\ldots,99\}} Q(y)$, whereas DiPPI uses the full probability vector. This setting tests whether features of the rating distribution beyond its mean or mode, such as dispersion or asymmetry, contain additional information about the OLS score. Figure 2(A) shows that DiPPI has the lowest relative RMSE and shortest intervals across all labeled ratios, with coverage near the nominal level.

4.2 Forest cover types

The Forest CoverType data[2] contain forest cover labels and cartographic covariates from four wilderness areas in northern Colorado (Blackard, 1998). We retain spruce/fir, lodgepole pine, ponderosa pine, and krummholz as separate classes, and combine cottonwood/willow, aspen, and Douglas-fir into a fifth class, Rare. The inferential target is the four class-specific coefficients of standardized slope in a five-class multinomial-logit regression with an intercept and standardized noon hillshade, using Rare as the reference class.

We train an XGBoost classifier (Chen & Guestrin, 2016) on 15,000 observations sampled from Rawah, Neota, and Cache la Poudre, using all 54 covariates. The analysis population consists of 40,000 observations sampled from Comanche Peak, which contains all five classes. The fitted classifier is held fixed and returns a class-probability vector Q for each observation. We set $r \in \{0.10, 0.15, 0.20, 0.25, 0.30\}$. PPI++ uses Q directly in the averaged multinomial loss. RePPI uses $\widehat{Y} = \mathbb{E}_Q(Y)$, whereas DiPPI uses the full vector Q. We also evaluate PPI++ and RePPI using the modal class $\widehat{Y} = \arg\max_{k \in \{1,\ldots,5\}} Q_k$ as the point prediction. Figure 2(B) shows that DiPPI has the lowest relative RMSE and shortest intervals across all labeled ratios, with coverage near the nominal level.

[1] https://www.kaggle.com/datasets/mysarahmadbhat/wine-tasting

[2] https://archive.ics.uci.edu/dataset/31/covertype

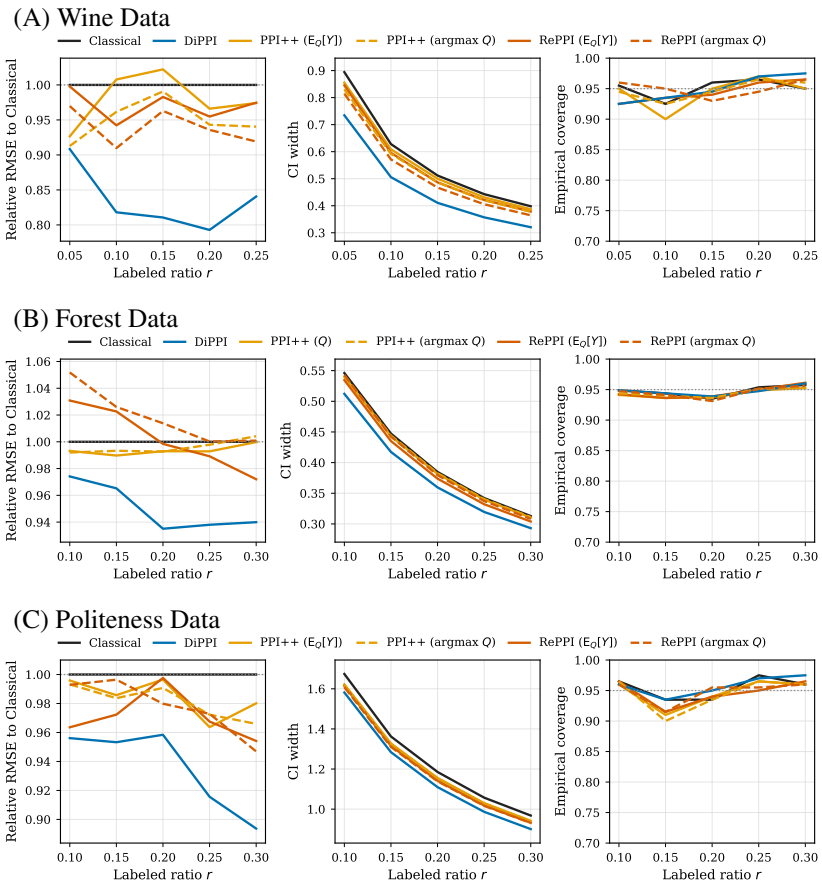


Figure 2: Relative RMSE to Classical, mean 95% CI width, and empirical coverage.

4.3 Politeness requests

The Politeness data contain Stack Exchange and Wikipedia requests with human ratings on a 1–25 politeness scale (Danescu-Niculescu-Mizil et al., 2013). We regress the politeness rating on an indicator for a hedging expression and target its coefficient.

We use GPT-5.4 nano to construct a predictive distribution over the 25 possible ratings. For each of PPI++ and RePPI, we also evaluate two point predictions, one using the predictive mean $\widehat{Y} = \mathbb{E}_Q(Y)$ and the other using the modal rating $\widehat{Y} = \arg\max_{y \in \{1,\ldots,25\}} Q(y)$, whereas DiPPI uses the full probability vector. This setting therefore examines whether features of the rating distribution beyond its mean or mode contain additional information about the OLS score. The analysis population contains 4386 requests. We set $r \in \{0.10, 0.15, 0.20, 0.25, 0.30\}$. Figure 2(C) shows that DiPPI gives the shortest intervals throughout and the lowest or nearly lowest relative RMSE, while maintaining near-nominal coverage.

5 Discussion

We introduce distribution-informed prediction-powered inference (DiPPI), which uses predictive distributions as auxiliary information for inference based on general estimating equations. We characterize when a predictive distribution provides information about the estimating score beyond a point prediction and quantify the resulting oracle efficiency gain. We establish efficiency results, asymptotic guarantees for the cross-fitted estimator, and valid inference under matrix calibration. Simulations and real-data experiments show that DiPPI improves efficiency over PPI methods based on point predictions when the predictive distribution contains additional score-relevant information.

AI USE STATEMENT

Generative AI tools were used to assist with code implementation, provide feedback on proof strategies, identify relevant literature, and improve the clarity of the writing. The research ideas and methodological development originated from the authors. All mathematical claims, proofs, code, and other AI-assisted outputs were independently reviewed and verified by the authors, who take full responsibility for the final content of this work.

REPRODUCIBILITY STATEMENT

To facilitate reproducibility, we provide the code used to reproduce all experiments and figures in the Supplementary Mabsaterial.

REFERENCES

Anastasios N. Angelopoulos, Stephen Bates, Clara Fannjiang, Michael I. Jordan, and Tijana Zrnic. Prediction-powered inference. *Science*, 382(6671):669–674, 2023a. doi: 10.1126/science.adi6000.

Anastasios N. Angelopoulos, John C. Duchi, and Tijana Zrnic. PPI++: Efficient prediction-powered inference, 2023b. arXiv:2311.01453.

Jock Blackard. Covertype [dataset]. UCI Machine Learning Repository, 1998. URL https://doi.org/10.24432/C50K5N.

Song Xi Chen, Denis H. Y. Leung, and Jing Qin. Improving semiparametric estimation by using surrogate data. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 70(4):803–823, 2008. doi: 10.1111/j.1467-9868.2008.00662.x.

Tianqi Chen and Carlos Guestrin. XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 785–794, 2016. doi: 10.1145/2939672.2939785.

Yi-Hau Chen and Hung Chen. A unified approach to regression analysis under double-sampling designs. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 62(3):449–460, 2000. doi: 10.1111/1467-9868.00243.

Cristian Danescu-Niculescu-Mizil, Moritz Sudhof, Dan Jurafsky, Jure Leskovec, and Christopher Potts. A computational approach to politeness with application to social factors. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 250–259, Sofia, Bulgaria, 2013. Association for Computational Linguistics. URL https://aclanthology.org/P13-1025/.

Benjamin Eyre and David Madras. Regression for the mean: Auto-evaluation and inference with few labels through post-hoc regression. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 15596–15612. PMLR, 2025. URL https://proceedings.mlr.press/v267/eyre25a.html.

Feng Gan, Wanfeng Liang, and Changliang Zou. Prediction de-correlated inference: A safe approach for post-prediction inference. *Australian & New Zealand Journal of Statistics*, 66(4):417–440, 2024. doi: 10.1111/anzs.12429.

Jessica Gronsbell, Jianhui Gao, Zachary R. McCaw, Yaqi Shi, and David Cheng. Another look at statistical inference with machine learning-imputed data, 2024. arXiv:2411.19908.

Wenlong Ji, Lihua Lei, and Tijana Zrnic. Predictions as surrogates: Revisiting surrogate outcomes in the age of AI. *Biometrika*, art. asag053, 2026. doi: 10.1093/biomet/asag053.

Jiacheng Miao, Xinran Miao, Yixuan Wu, Jiwei Zhao, and Qiongshi Lu. Assumption-lean and data-adaptive post-prediction inference. *Journal of Machine Learning Research*, 26(179):1–31, 2025. URL https://jmlr.org/papers/v26/24-0056.html.

James M. Robins, Andrea Rotnitzky, and Lue Ping Zhao. Estimation of regression coefficients when some regressors are not always observed. *Journal of the American Statistical Association*, 89(427):846–866, 1994. doi: 10.1080/01621459.1994.10476818.

Lars van der Laan and Mark van der Laan. Calibeating prediction-powered inference, 2026. arXiv:2604.21260.

Guanghui Wang, Mengtao Wen, and Changliang Zou. Empirical likelihood meets prediction-powered inference, 2025. arXiv:2512.16363.

Siruo Wang, Tyler H. McCormick, and Jeffrey T. Leek. Methods for correcting inference based on outcomes predicted by machine learning. *Proceedings of the National Academy of Sciences*, 117 (48):30266–30275, 2020. doi: 10.1073/pnas.2001238117.

Appendix

A Additional Simulation Studies

A.1 Predictive variance across labeled ratios

We further examine the predictive-variance setting in Section 3 as the amount of labeled data varies. We fix $\kappa = 1.5$, set $N = 10{,}000$, and vary $r = n/N \in \{0.05, 0.10, 0.15, 0.20, 0.25, 0.30\}$. Each configuration uses 200 replications.

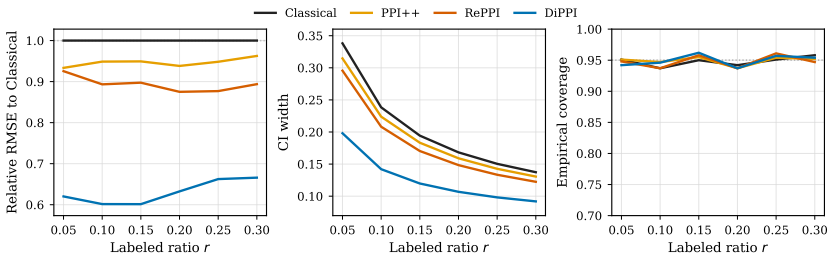


Figure 3: Outcome-relevant predictive variance as the labeled ratio varies. Panels show relative RMSE to Classical, mean 95% CI width, and empirical coverage over 200 replications.

Figure 3 shows that DiPPI has the lowest relative RMSE and shortest intervals across the labeled ratios, while maintaining near-nominal coverage.

A.2 Predictive skewness

We also consider information carried by predictive skewness rather than predictive variance. We use the methods, representation, and evaluation metrics from Section 3. Let $X, Z \in \mathbb{R}^5$ have independent standard-normal coordinates, let S be uniform on $\{-1, 1\}$, $G \sim \mathrm{Gamma}(4, 1)$, and $\varepsilon \sim N(0, 1)$, all independent. Define $W_+ = (G - 4)/2$, and $W_- = -W_+$. Given (Z, S), the prediction system returns the distribution Q of $Z^\top \beta + W_S$, and we generate

$$Y = X^\top \theta + Z^\top \beta + \kappa S + \varepsilon.$$

Every Q has the same mean $Z^\top \beta$ and variance one, but its skewness reveals S. PPI++ and RePPI therefore use the same point prediction $\widehat{Y} = Z^\top \beta$ regardless of S, whereas the asymmetric quantiles in $\Phi(Q)$ retain the tail direction. We estimate the five OLS slopes in θ. We use $N = 10{,}000$ and 200 replications. The top row of Figure 4 varies $r \in \{0.05, 0.10, 0.15, 0.20, 0.25, 0.30\}$ at $\kappa = 1$, and the bottom row varies $\kappa \in \{0, 0.25, 0.5, 0.75, 1\}$ at $n = 1{,}000$.

Figure 4 shows that DiPPI and RePPI perform similarly when $\kappa = 0$. As the skewness becomes more informative about the outcome, DiPPI achieves lower relative RMSE and shorter intervals while maintaining near-nominal coverage.

B Regularity Conditions and Proofs

This section states sufficient conditions and proves the results in Section 2. For each cross-fitted estimator, we condition on the data used to compute the initial estimator and estimate the conditional score. The remaining labeled fold and the unlabeled sample are then independent of the resulting estimates. Lemma 5 bounds the difference between their correction averages.

We use $\|\cdot\|$ for the Euclidean norm and its induced operator norm.

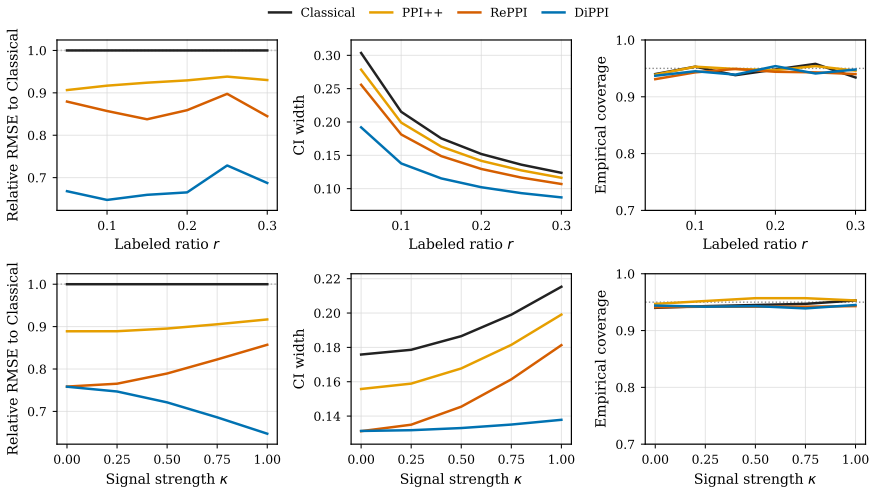


Figure 4: Outcome-relevant predictive skewness. The top row varies the labeled ratio at $\kappa = 1$, and the bottom row varies κ at $n = 1{,}000$. Columns show relative RMSE to Classical, mean 95% CI width, and empirical coverage over 200 replications.

B.1 Proof of Theorem 1

Proof. We first establish existence and uniqueness of a local root, together with its $n^{-1/2}$ rate. Define

$$\Psi_{n,N}(\theta; h) := \frac{1}{n}\sum_{i=1}^{n}\{U_\theta(X_i, Y_i) - h_\theta(X_i, Q_i)\} + \frac{1}{n+N}\sum_{k=1}^{n+N} h_\theta(X_k, Q_k).$$

$$T_{n,N}(\theta) = \theta - H_{\theta^\star}^{-1}\Psi_{n,N}(\theta; h).$$

Since $H_{\theta^\star}$ is nonsingular,

$$\Psi_{n,N}(\theta; h) = 0 \quad \Longleftrightarrow \quad T_{n,N}(\theta) = \theta.$$

Thus, proving that the estimating equation has a unique root in $\mathcal{B}_a$ is equivalent to proving that $T_{n,N}$ has a unique fixed point in $\mathcal{B}_a$. Since $\mathcal{B}_a$ is a closed subset of $\mathbb{R}^d$, it is complete. By the Banach fixed-point theorem, $T_{n,N}$ has a unique fixed point in $\mathcal{B}_a$ whenever, for some $\kappa \in (0, 1)$,

$$T_{n,N}(\mathcal{B}_a) \subseteq \mathcal{B}_a, \text{ and } \|T_{n,N}(\theta_1) - T_{n,N}(\theta_2)\| \quad \leq \kappa\|\theta_1 - \theta_2\|, \qquad \forall\, \theta_1, \theta_2 \in \mathcal{B}_a.$$

Because $\mathcal{B}_a$ is convex, $\theta_2 + t(\theta_1 - \theta_2) \in \mathcal{B}_a$ for every $\theta_1, \theta_2 \in \mathcal{B}_a$ and $t \in [0, 1]$. The fundamental theorem of calculus along this line segment yields

$$T_{n,N}(\theta_1) - T_{n,N}(\theta_2) = \left[\int_0^1 \nabla_\theta T_{n,N}\{\theta_2 + t(\theta_1 - \theta_2)\}\, dt\right](\theta_1 - \theta_2).$$

Hence,

$$\sup_{\theta\in\mathcal{B}_a} \|\nabla_\theta T_{n,N}(\theta)\| \leq \kappa \quad \Longrightarrow \quad \|T_{n,N}(\theta_1) - T_{n,N}(\theta_2)\| \leq \kappa\|\theta_1 - \theta_2\|. \tag{13}$$

We now verify the derivative bound. Since

$$\nabla_\theta T_{n,N}(\theta) = I_d - H_{\theta^\star}^{-1}\nabla_\theta \Psi_{n,N}(\theta; h),$$

it is enough to control $\nabla_\theta \Psi_{n,N}(\theta; h)$ uniformly. We proceed in two steps. We first prove the uniform law

$$\sup_{\theta\in\mathcal{B}_\rho} \|\nabla_\theta \Psi_{n,N}(\theta; h) - \mathbb{E}\{\nabla_\theta U_\theta(X, Y)\}\| = o_p(1). \tag{14}$$

We then combine (14) with the continuity of $\theta \mapsto \mathbb{E}\{\nabla_\theta U_\theta(X,Y)\}$ at $\theta^\star$ and the nonsingularity of $H_{\theta^\star}$ to choose $a \in (0,\rho)$ such that

$$\Pr\Big\{ \sup_{\theta\in\mathcal{B}_a} \|\nabla_\theta T_{n,N}(\theta)\| \le \kappa \Big\} \longrightarrow 1.$$

To prove (14), write $c_n = N/(n+N) = 1/(1+n/N)$. Differentiating the sample estimating equation gives

$$\nabla_\theta \Psi_{n,N}(\theta;h) = \frac{1}{n}\sum_{i=1}^{n} \nabla_\theta U_\theta(X_i,Y_i) - c_n \frac{1}{n}\sum_{i=1}^{n} \nabla_\theta h_\theta(X_i,Q_i) + c_n \frac{1}{N}\sum_{j=n+1}^{n+N} \nabla_\theta h_\theta(X_j,Q_j).$$

Since the labeled and unlabeled samples have the same (X,Q) marginal distribution, the two h terms have the same population expectation. Define

$$\begin{aligned}
A_{1n}(\theta) &= \frac{1}{n}\sum_{i=1}^{n} \nabla_\theta U_\theta(X_i,Y_i) - \mathbb{E}\{\nabla_\theta U_\theta(X,Y)\},\\
A_{2n}(\theta) &= \frac{1}{n}\sum_{i=1}^{n} \nabla_\theta h_\theta(X_i,Q_i) - \mathbb{E}\{\nabla_\theta h_\theta(X,Q)\},\\
A_{3N}(\theta) &= \frac{1}{N}\sum_{j=n+1}^{n+N} \nabla_\theta h_\theta(X_j,Q_j) - \mathbb{E}\{\nabla_\theta h_\theta(X,Q)\}.
\end{aligned}$$

Then we have

$$\nabla_\theta \Psi_{n,N}(\theta;h) - \mathbb{E}\{\nabla_\theta U_\theta(X,Y)\} = A_{1n}(\theta) - c_n A_{2n}(\theta) + c_n A_{3N}(\theta),$$

and, since $0 < c_n \le 1$, the triangle inequality gives

$$\sup_{\theta\in\mathcal{B}_\rho} \|\nabla_\theta \Psi_{n,N}(\theta;h) - \mathbb{E}\{\nabla_\theta U_\theta(X,Y)\}\| \le \sup_{\theta\in\mathcal{B}_\rho} \|A_{1n}(\theta)\| + \sup_{\theta\in\mathcal{B}_\rho} \|A_{2n}(\theta)\| + \sup_{\theta\in\mathcal{B}_\rho} \|A_{3N}(\theta)\|. \tag{15}$$

We first prove that $\sup_{\theta\in\mathcal{B}_\rho} \|A_{1n}(\theta)\| = o_p(1)$. For any $\delta > 0$, let $\widetilde{\theta}_1, \ldots, \widetilde{\theta}_{K_\delta}$ be a finite δ-net of $\mathcal{B}_\rho$. The triangle inequality gives

$$\begin{aligned}
\sup_{\theta\in\mathcal{B}_\rho} \|A_{1n}(\theta)\| \le \max_{1\le k\le K_\delta} \|A_{1n}(\widetilde{\theta}_k)\| &+ \frac{1}{n}\sum_{i=1}^{n} \sup_{\substack{\theta_1,\theta_2\in\mathcal{B}_\rho \\ \|\theta_1-\theta_2\|\le\delta}} \|\nabla_{\theta_1} U_{\theta_1}(X_i,Y_i) - \nabla_{\theta_2} U_{\theta_2}(X_i,Y_i)\| \\
&+ \mathbb{E}\left[\sup_{\substack{\theta_1,\theta_2\in\mathcal{B}_\rho \\ \|\theta_1-\theta_2\|\le\delta}} \|\nabla_{\theta_1} U_{\theta_1}(X,Y) - \nabla_{\theta_2} U_{\theta_2}(X,Y)\| \right].
\end{aligned} \tag{16}$$

For fixed δ and each $k = 1, \ldots, K_\delta$, Assumption 1 gives

$$\mathbb{E}\Big[\|\nabla_{\widetilde{\theta}_k} U_{\widetilde{\theta}_k}(X,Y)\|\Big] \le \mathbb{E}\left[\sup_{\theta\in\mathcal{B}_\rho} \|\nabla_\theta U_\theta(X,Y)\|\right] < \infty.$$

Hence, the weak law of large numbers, applied componentwise, gives

$$\|A_{1n}(\widetilde{\theta}_k)\| = \left\| \frac{1}{n}\sum_{i=1}^{n} \nabla_{\widetilde{\theta}_k} U_{\widetilde{\theta}_k}(X_i,Y_i) - \mathbb{E}\{\nabla_{\widetilde{\theta}_k} U_{\widetilde{\theta}_k}(X,Y)\} \right\| \xrightarrow{p} 0.$$

Since $K_\delta < \infty$, the union bound gives, for every $\varepsilon > 0$,

$$\Pr\Big\{ \max_{1\le k\le K_\delta} \|A_{1n}(\widetilde{\theta}_k)\| > \varepsilon \Big\} \le \sum_{k=1}^{K_\delta} \Pr\Big\{ \|A_{1n}(\widetilde{\theta}_k)\| > \varepsilon \Big\} \longrightarrow 0.$$

Therefore, the maximum over the finite net is $o_p(1)$. For $\delta > 0$, define

$$W_\delta(x, y) = \sup_{\substack{\theta_1, \theta_2 \in \mathcal{B}_\rho \\ \|\theta_1 - \theta_2\| \le \delta}} \|\nabla_{\theta_1} U_{\theta_1}(x, y) - \nabla_{\theta_2} U_{\theta_2}(x, y)\|.$$

In (16), the last two terms can be written as $\frac{1}{n}\sum_{i=1}^n W_\delta(X_i, Y_i)$ and $\mathbb{E}\{W_\delta(X, Y)\}$. Since

$$\mathbb{E}\{W_\delta(X, Y)\} \le 2\mathbb{E}\left[\sup_{\theta \in \mathcal{B}_\rho} \|\nabla_\theta U_\theta(X, Y)\|\right] < \infty,$$

the weak law of large numbers gives, for every fixed $\delta > 0$,

$$\frac{1}{n}\sum_{i=1}^n W_\delta(X_i, Y_i) - \mathbb{E}\{W_\delta(X, Y)\} \xrightarrow{p} 0.$$

Equivalently,

$$\frac{1}{n}\sum_{i=1}^n W_\delta(X_i, Y_i) = \mathbb{E}\{W_\delta(X, Y)\} + o_p(1).$$

Substituting this result into (16) gives

$$\sup_{\theta \in \mathcal{B}_\rho} \|A_{1n}(\theta)\| \le 2\mathbb{E}\{W_\delta(X, Y)\} + o_p(1). \tag{17}$$

Almost-sure continuity on the compact set $\mathcal{B}_\rho$ implies $W_\delta(X, Y) \to 0$ almost surely as $\delta \to 0$. Moreover,

$$0 \le W_\delta(X, Y) \le 2 \sup_{\theta \in \mathcal{B}_\rho} \|\nabla_\theta U_\theta(X, Y)\|,$$

and the dominating random variable is integrable by Assumption 1. The dominated convergence theorem therefore implies

$$\lim_{\delta \to 0} \mathbb{E}\{W_\delta(X, Y)\} = 0. \tag{18}$$

For any $\varepsilon > 0$, (18) permits a fixed $\delta_\varepsilon > 0$ such that $2\mathbb{E}\{W_{\delta_\varepsilon}(X, Y)\} < \varepsilon/2$. By (17),

$$\sup_{\theta \in \mathcal{B}_\rho} \|A_{1n}(\theta)\| \le \frac{\varepsilon}{2} + o_p(1),$$

and hence

$$\Pr\left\{\sup_{\theta \in \mathcal{B}_\rho} \|A_{1n}(\theta)\| > \varepsilon\right\} \longrightarrow 0.$$

Thus,

$$\sup_{\theta \in \mathcal{B}_\rho} \|A_{1n}(\theta)\| = o_p(1). \tag{19}$$

Since $n/N \to r \in (0, \infty)$, both n and N diverge. By the same finite-net argument and Assumption 2, with $\nabla_\theta U_\theta(X, Y)$ replaced by $\nabla_\theta h_\theta(X, Q)$, we obtain

$$\sup_{\theta \in \mathcal{B}_\rho} \|A_{2n}(\theta)\| + \sup_{\theta \in \mathcal{B}_\rho} \|A_{3N}(\theta)\| = o_p(1). \tag{20}$$

Substituting (19) and (20) into (15) proves (14).

We next prove that $\theta \mapsto \mathbb{E}\{\nabla_\theta U_\theta(X, Y)\}$ is continuous at $\theta^\star$. To this end, let $\{\theta_m\}_{m \ge 1} \subset \mathcal{B}_\rho$ satisfy $\theta_m \to \theta^\star$. Almost-sure continuity gives

$$\nabla_{\theta_m} U_{\theta_m}(X, Y) \longrightarrow \nabla_{\theta^\star} U_{\theta^\star}(X, Y) \quad \text{a.s.}$$

Moreover, we have

$$\|\nabla_{\theta_m} U_{\theta_m}(X, Y) - \nabla_{\theta^\star} U_{\theta^\star}(X, Y)\| \le 2 \sup_{\theta \in \mathcal{B}_\rho} \|\nabla_\theta U_\theta(X, Y)\|,$$

whose expectation is finite by Assumption 1. Hence, the dominated convergence theorem gives

$$\mathbb{E}[\|\nabla_{\theta_m} U_{\theta_m}(X, Y) - \nabla_{\theta^\star} U_{\theta^\star}(X, Y)\|] \longrightarrow 0.$$

Therefore,

$$\|\mathbb{E}\{\nabla_{\theta_m} U_{\theta_m}(X,Y)\} - \mathbb{E}\{\nabla_{\theta^\star} U_{\theta^\star}(X,Y)\}\| \le \mathbb{E}[\|\nabla_{\theta_m} U_{\theta_m}(X,Y) - \nabla_{\theta^\star} U_{\theta^\star}(X,Y)\|] \longrightarrow 0.$$

Since the sequence $\{\theta_m\}_{m\ge 1}$ was arbitrary, the sequential characterization of continuity implies that $\theta \mapsto \mathbb{E}\{\nabla_\theta U_\theta(X,Y)\}$ is continuous at $\theta^\star$. Moreover,

$$I_d - H_{\theta^\star}^{-1}\mathbb{E}\{\nabla_\theta U_{\theta^\star}(X,Y)\} = I_d - H_{\theta^\star}^{-1} H_{\theta^\star} = 0.$$

Fix $\kappa \in (0,1)$. By continuity, there exists $\delta > 0$ such that

$$\|\mathbb{E}\{\nabla_\theta U_\theta(X,Y)\} - \mathbb{E}\{\nabla_\theta U_{\theta^\star}(X,Y)\}\| < \frac{\kappa}{2\|H_{\theta^\star}^{-1}\|}$$

whenever $\|\theta - \theta^\star\| < \delta$. Choose $a \in (0, \min\{\rho, \delta\})$. Then, for every $\theta \in \mathcal{B}_a$,

$$\begin{aligned}\left\|I_d - H_{\theta^\star}^{-1}\mathbb{E}\{\nabla_\theta U_\theta(X,Y)\}\right\| &= \left\|H_{\theta^\star}^{-1}\left[\mathbb{E}\{\nabla_\theta U_{\theta^\star}(X,Y)\} - \mathbb{E}\{\nabla_\theta U_\theta(X,Y)\}\right]\right\| \\ &\le \|H_{\theta^\star}^{-1}\|\,\|\mathbb{E}\{\nabla_\theta U_{\theta^\star}(X,Y)\} - \mathbb{E}\{\nabla_\theta U_\theta(X,Y)\}\| \le \frac{\kappa}{2}.\end{aligned}$$

Hence,

$$\sup_{\theta\in\mathcal{B}_a}\left\|I_d - H_{\theta^\star}^{-1}\mathbb{E}\{\nabla_\theta U_\theta(X,Y)\}\right\| \le \frac{\kappa}{2}. \tag{21}$$

For every $\theta \in \mathcal{B}_a$, adding and subtracting $H_{\theta^\star}^{-1}\mathbb{E}\{\nabla_\theta U_\theta(X,Y)\}$ gives

$$\begin{aligned}I_d - H_{\theta^\star}^{-1}\nabla_\theta \Psi_{n,N}(\theta;h) &= I_d - H_{\theta^\star}^{-1}\mathbb{E}\{\nabla_\theta U_\theta(X,Y)\} \\ &\quad + H_{\theta^\star}^{-1}\left[\mathbb{E}\{\nabla_\theta U_\theta(X,Y)\} - \nabla_\theta \Psi_{n,N}(\theta;h)\right].\end{aligned}$$

Therefore, the triangle inequality and $\|AB\| \le \|A\|\|B\|$ give

$$\begin{aligned}\left\|I_d - H_{\theta^\star}^{-1}\nabla_\theta \Psi_{n,N}(\theta;h)\right\| &\le \left\|I_d - H_{\theta^\star}^{-1}\mathbb{E}\{\nabla_\theta U_\theta(X,Y)\}\right\| \\ &\quad + \|H_{\theta^\star}^{-1}\|\,\|\nabla_\theta \Psi_{n,N}(\theta;h) - \mathbb{E}\{\nabla_\theta U_\theta(X,Y)\}\|.\end{aligned}$$

Taking the supremum over $\theta \in \mathcal{B}_a$, and using (21) and $\mathcal{B}_a \subset \mathcal{B}_\rho$, gives

$$\begin{aligned}\sup_{\theta\in\mathcal{B}_a}\left\|I_d - H_{\theta^\star}^{-1}\nabla_\theta \Psi_{n,N}(\theta;h)\right\| &\le \frac{\kappa}{2} + \|H_{\theta^\star}^{-1}\| \sup_{\theta\in\mathcal{B}_\rho}\|\nabla_\theta \Psi_{n,N}(\theta;h) - \mathbb{E}\{\nabla_\theta U_\theta(X,Y)\}\| \\ &= \frac{\kappa}{2} + o_p(1),\end{aligned}$$

where the last equality follows from (14) and $\|H_{\theta^\star}^{-1}\| < \infty$. Consequently, since $\nabla_\theta T_{n,N}(\theta) = I_d - H_{\theta^\star}^{-1}\nabla_\theta \Psi_{n,N}(\theta;h)$, we have

$$\Pr\left\{\sup_{\theta\in\mathcal{B}_a}\|\nabla_\theta T_{n,N}(\theta)\| \le \kappa\right\} = \Pr\left\{\sup_{\theta\in\mathcal{B}_a}\left\|I_d - H_{\theta^\star}^{-1}\nabla_\theta \Psi_{n,N}(\theta;h)\right\| \le \kappa\right\} \longrightarrow 1. \tag{22}$$

It remains to verify that $T_{n,N}$ maps $\mathcal{B}_a$ into itself. We first control its displacement at the center:

$$T_{n,N}(\theta^\star) - \theta^\star = -H_{\theta^\star}^{-1}\Psi_{n,N}(\theta^\star;h).$$

Because $\mathbb{E}\{U_{\theta^\star}(X,Y)\} = 0$ and the two samples have the same (X,Q) marginal distribution,

$$\begin{aligned}\Psi_{n,N}(\theta^\star;h) &= \frac{1}{n}\sum_{i=1}^{n}\Big[U_{\theta^\star}(X_i,Y_i) - c_n h_{\theta^\star}(X_i,Q_i) - \mathbb{E}\{U_{\theta^\star}(X,Y) - c_n h_{\theta^\star}(X,Q)\}\Big] \\ &\quad + c_n\frac{1}{N}\sum_{j=n+1}^{n+N}\Big[h_{\theta^\star}(X_j,Q_j) - \mathbb{E}\{h_{\theta^\star}(X,Q)\}\Big].\end{aligned}$$

By sample independence, Assumptions 1 and 2, and $n/N \to r$, we have

$$\begin{aligned}\mathbb{E}\left\|\sqrt{n}\,\Psi_{n,N}(\theta^\star;h)\right\|^2 &= \mathbb{E}\left\|U_{\theta^\star}(X,Y) - c_n h_{\theta^\star}(X,Q) - \mathbb{E}\{U_{\theta^\star}(X,Y) - c_n h_{\theta^\star}(X,Q)\}\right\|^2 \\ &\quad + \frac{n}{N}c_n^2\mathbb{E}\left\|h_{\theta^\star}(X,Q) - \mathbb{E}\{h_{\theta^\star}(X,Q)\}\right\|^2 \\ &\le 2\mathbb{E}\|U_{\theta^\star}(X,Y)\|^2 + \left(2 + \frac{n}{N}\right)\mathbb{E}\|h_{\theta^\star}(X,Q)\|^2 = O(1).\end{aligned}$$

Therefore, Markov's inequality gives

$$\sqrt{n}\,\Psi_{n,N}(\theta^\star; h) = O_p(1), \qquad H_{\theta^\star}^{-1}\Psi_{n,N}(\theta^\star; h) = O_p(n^{-1/2}) = o_p(1). \tag{23}$$

Since $(1-\kappa)a > 0$, (23) implies

$$\Pr\big\{\|H_{\theta^\star}^{-1}\Psi_{n,N}(\theta^\star; h)\| \le (1-\kappa)a\big\} \longrightarrow 1. \tag{24}$$

Equations (22) and (24), together with the union bound, give

$$\Pr\left(\left\{\sup_{\theta\in\mathcal{B}_a}\|\nabla_\theta T_{n,N}(\theta)\| \le \kappa\right\} \cap \big\{\|H_{\theta^\star}^{-1}\Psi_{n,N}(\theta^\star; h)\| \le (1-\kappa)a\big\}\right) \longrightarrow 1. \tag{25}$$

On the event inside (25), (13) shows that $T_{n,N}$ is a κ-contraction on $\mathcal{B}_a$. For every $\theta \in \mathcal{B}_a$, the two conditions inside (25) give

$$\|T_{n,N}(\theta) - \theta^\star\| \le \kappa\|\theta - \theta^\star\| + \|H_{\theta^\star}^{-1}\Psi_{n,N}(\theta^\star; h)\| \le \kappa a + (1-\kappa)a = a.$$

Thus, on the event inside (25), $T_{n,N}$ is a contraction from $\mathcal{B}_a$ into itself. The Banach fixed-point theorem and (25) therefore yield

$$\Pr\{\exists!\,\theta \in \mathcal{B}_a : \Psi_{n,N}(\theta; h) = 0\} \longrightarrow 1. \tag{26}$$

On the event inside (25), the contraction inequality and the fixed-point identity $\widehat{\theta}_h = T_{n,N}(\widehat{\theta}_h)$ give

$$\begin{aligned}\|\widehat{\theta}_h - \theta^\star\| &= \|T_{n,N}(\widehat{\theta}_h) - \theta^\star\| \\ &\le \|T_{n,N}(\widehat{\theta}_h) - T_{n,N}(\theta^\star)\| + \|T_{n,N}(\theta^\star) - \theta^\star\| \\ &\le \kappa\|\widehat{\theta}_h - \theta^\star\| + \|H_{\theta^\star}^{-1}\Psi_{n,N}(\theta^\star; h)\|,\end{aligned}$$

where the last step uses $T_{n,N}(\theta^\star) - \theta^\star = -H_{\theta^\star}^{-1}\Psi_{n,N}(\theta^\star; h)$. Therefore,

$$(1-\kappa)\|\widehat{\theta}_h - \theta^\star\| \le \|H_{\theta^\star}^{-1}\Psi_{n,N}(\theta^\star; h)\|,$$

and hence

$$\|\widehat{\theta}_h - \theta^\star\| \le \frac{1}{1-\kappa}\|H_{\theta^\star}^{-1}\Psi_{n,N}(\theta^\star; h)\| \le \frac{\|H_{\theta^\star}^{-1}\|}{1-\kappa}\|\Psi_{n,N}(\theta^\star; h)\|. \tag{27}$$

Equations (25), (23), and (27) imply

$$\|\widehat{\theta}_h - \theta^\star\| = O_p(n^{-1/2}). \tag{28}$$

Equation (28) also implies

$$\Pr\{\widehat{\theta}_h \in \mathrm{int}(\mathcal{B}_a)\} = \Pr\{\|\widehat{\theta}_h - \theta^\star\| < a\} \longrightarrow 1.$$

We next derive the asymptotic distribution. Define

$$\overline{H}_{n,N} = \int_0^1 \nabla_\theta \Psi_{n,N}\{\theta^\star + t(\widehat{\theta}_h - \theta^\star); h\}\,dt.$$

On the event $\{\widehat{\theta}_h \in \mathcal{B}_a\}$,

$$\begin{aligned}\|\overline{H}_{n,N} - H_{\theta^\star}\| \le{}& \sup_{\theta\in\mathcal{B}_\rho}\|\nabla_\theta\Psi_{n,N}(\theta; h) - \mathbb{E}\{\nabla_\theta U_\theta(X,Y)\}\| \\ &+ \sup_{0\le t\le 1}\Big\|\mathbb{E}\Big[\nabla_{\theta^\star + t(\widehat{\theta}_h - \theta^\star)}U_{\theta^\star + t(\widehat{\theta}_h - \theta^\star)}(X,Y)\Big] - H_{\theta^\star}\Big\| = o_p(1),\end{aligned}$$

The $o_p(1)$ comes from Equations (14) and (28). Specifically, (14) gives

$$\sup_{\theta\in\mathcal{B}_\rho}\|\nabla_\theta\Psi_{n,N}(\theta; h) - \mathbb{E}\{\nabla_\theta U_\theta(X,Y)\}\| = o_p(1).$$

(28) and continuity of $\theta \mapsto \mathbb{E}\{\nabla_\theta U_\theta(X,Y)\}$ give

$$\sup_{0\le t\le 1}\Big\|\mathbb{E}\Big[\nabla_{\theta^\star + t(\widehat{\theta}_h - \theta^\star)}U_{\theta^\star + t(\widehat{\theta}_h - \theta^\star)}(X,Y)\Big] - H_{\theta^\star}\Big\| = o_p(1).$$

Hence, $\overline{H}_{n,N} \xrightarrow{p} H_{\theta^\star}$. Since $H_{\theta^\star}$ is nonsingular, the continuous mapping theorem gives

$$\Pr\{\overline{H}_{n,N} \text{ is nonsingular}\} \longrightarrow 1, \qquad \overline{H}_{n,N}^{-1} \xrightarrow{p} H_{\theta^\star}^{-1}. \tag{29}$$

Equations (26) and (29) imply

$$\Pr\Big(\{\Psi_{n,N}(\widehat{\theta}_h; h) = 0\} \cap \{\overline{H}_{n,N} \text{ is nonsingular}\}\Big) \longrightarrow 1. \tag{30}$$

On the event inside (30), the fundamental theorem of calculus along the line segment from $\theta^\star$ to $\widehat{\theta}_h$ gives

$$0 = \Psi_{n,N}(\theta^\star; h) + \overline{H}_{n,N}(\widehat{\theta}_h - \theta^\star).$$

Therefore,

$$\sqrt{n}(\widehat{\theta}_h - \theta^\star) = -\overline{H}_{n,N}^{-1}\sqrt{n}\,\Psi_{n,N}(\theta^\star; h). \tag{31}$$

Equations (29) and (23) give

$$\Big\|(\overline{H}_{n,N}^{-1} - H_{\theta^\star}^{-1})\sqrt{n}\,\Psi_{n,N}(\theta^\star; h)\Big\| = o_p(1). \tag{32}$$

Combining (30), (31), and (32) yields

$$\sqrt{n}(\widehat{\theta}_h - \theta^\star) = -H_{\theta^\star}^{-1}\sqrt{n}\,\Psi_{n,N}(\theta^\star; h) + o_p(1). \tag{33}$$

Moreover,

$$\begin{aligned}\sqrt{n}\,\Psi_{n,N}(\theta^\star; h) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\Big[U_{\theta^\star}(X_i, Y_i) - c_n h_{\theta^\star}(X_i, Q_i) - \mathbb{E}\{U_{\theta^\star}(X, Y) - c_n h_{\theta^\star}(X, Q)\}\Big] \\ + c_n\frac{\sqrt{n}}{N}\sum_{j=n+1}^{n+N}\Big[h_{\theta^\star}(X_j, Q_j) - \mathbb{E}\{h_{\theta^\star}(X, Q)\}\Big].\end{aligned} \tag{34}$$

Set $c = (1+r)^{-1}$. Since $n/N \to r$, we have $c_n \to c$. By sample independence and the multivariate central limit theorem, (34) gives

$$\sqrt{n}\,\Psi_{n,N}(\theta^\star; h) \xrightarrow{d} N\left(0, \frac{1}{1+r}\mathrm{Cov}\{U_{\theta^\star}(X, Y) - h_{\theta^\star}(X, Q)\} + \frac{r}{1+r}\mathrm{Cov}\{U_{\theta^\star}(X, Y)\}\right). \tag{35}$$

The covariance in (35) follows from

$$\begin{aligned}&\mathrm{Cov}\{U_{\theta^\star}(X, Y) - c h_{\theta^\star}(X, Q)\} + r c^2 \mathrm{Cov}\{h_{\theta^\star}(X, Q)\} \\ &\qquad = c\mathrm{Cov}\{U_{\theta^\star}(X, Y) - h_{\theta^\star}(X, Q)\} + (1-c)\mathrm{Cov}\{U_{\theta^\star}(X, Y)\}.\end{aligned}$$

Equations (33) and (35), together with Slutsky's theorem, yield

$$\sqrt{n}(\widehat{\theta}_h - \theta^\star) \xrightarrow{d} N\left(0, H_{\theta^\star}^{-1}\left[\frac{1}{1+r}\mathrm{Cov}\{U_{\theta^\star}(X, Y) - h_{\theta^\star}(X, Q)\} + \frac{r}{1+r}\mathrm{Cov}\{U_{\theta^\star}(X, Y)\}\right] H_{\theta^\star}^{-\top}\right),$$

which proves (2). It remains to prove the oracle optimality statement. For any h satisfying Assumption 2,

$$\begin{aligned}&\frac{1}{1+r}\mathrm{Cov}\{U_{\theta^\star}(X, Y) - h_{\theta^\star}(X, Q)\} + \frac{r}{1+r}\mathrm{Cov}\{U_{\theta^\star}(X, Y)\} \\ &\quad = \frac{1}{1+r}\mathbb{E}[\mathrm{Cov}\{U_{\theta^\star}(X, Y) \mid X, Q\}] + \frac{r}{1+r}\mathrm{Cov}\{U_{\theta^\star}(X, Y)\} \\ &\qquad + \frac{1}{1+r}\mathrm{Cov}\left[\mathbb{E}\{U_{\theta^\star}(X, Y) \mid X, Q\} - h_{\theta^\star}(X, Q)\right],\end{aligned} \tag{36}$$

The last term on the right-hand side of (36) is positive semidefinite and equals zero for (3). Substituting (3) into (36) and using the law of total covariance for $U_{\theta^\star}(X, Y)$, yields (4). More generally,

$$\Sigma_h - \Sigma_{h^\star} = \frac{1}{1+r}H_{\theta^\star}^{-1}\mathrm{Cov}[\mathbb{E}\{U_{\theta^\star}(X, Y) \mid X, Q\} - h_{\theta^\star}(X, Q)]\, H_{\theta^\star}^{-\top} \succeq 0.$$

This proves $\Sigma_{h^\star} \preceq \Sigma_h$. □

B.2 Proof of Theorem 2

The following lemma isolates the effect of nuisance estimation. It shows that each learned fold correction is first-order equivalent to the correction formed with the fixed population quantities M_μ and μ. The theorem proof then treats this population correction as the leading term.

Lemma 5. *Suppose the conditions of Theorem 2 hold. Define*

$$M_\mu = \mathrm{Cov}\{U_{\theta^\star}(X,Y), \mu(X,Q)\}\mathrm{Cov}\{\mu(X,Q)\}^{-1}.$$

Then, for every $k = 1, 2, 3$, $\widehat{M}^k \xrightarrow{p} M_\mu$, and

$$R_{n,k} := \sqrt{n}\Bigg[\frac{\widehat{M}^k}{1+n/N}\left\{\frac{1}{N}\sum_{j=n+1}^{n+N}\widehat{\mu}^k(X_j,Q_j) - \frac{1}{n_k}\sum_{i\in\mathcal{D}_k}\widehat{\mu}^k(X_i,Q_i)\right\} - \frac{M_\mu}{1+r}\left\{\frac{1}{N}\sum_{j=n+1}^{n+N}\mu(X_j,Q_j) - \frac{1}{n_k}\sum_{i\in\mathcal{D}_k}\mu(X_i,Q_i)\right\}\Bigg] = o_p(1).$$

The proof of Lemma 5 is given in Appendix B.3.

Proof. The proof has three steps. We first establish existence, uniqueness, and the $n^{-1/2}$ rate of each fold-specific root. We then derive an asymptotic linear representation of the weighted estimator. Finally, we apply the multivariate central limit theorem and calculate the limiting covariance.

For $k = 1, 2, 3$, define the fold estimating function

$$\Psi_k(\theta) = \frac{1}{n_k}\sum_{i\in\mathcal{D}_k} U_\theta(X_i,Y_i) + \frac{\widehat{M}^k}{1+n/N}\left\{\frac{1}{N}\sum_{j=n+1}^{n+N}\widehat{\mu}^k(X_j,Q_j) - \frac{1}{n_k}\sum_{i\in\mathcal{D}_k}\widehat{\mu}^k(X_i,Q_i)\right\}.$$

Following the center-control step in the proof of Theorem 1, with n replaced by n_k, we first control $\Psi_k(\theta^\star)$. Since $\mathbb{E}\{U_{\theta^\star}(X,Y)\} = 0$, Assumption 1, and $n_k/n \to \pi_k \in (0,1)$ give

$$\mathbb{E}\left\|\frac{\sqrt{n}}{n_k}\sum_{i\in\mathcal{D}_k} U_{\theta^\star}(X_i,Y_i)\right\|^2 = \frac{n}{n_k}\mathbb{E}\|U_{\theta^\star}(X,Y)\|^2 = O(1).$$

Markov's inequality therefore gives

$$\frac{1}{n_k}\sum_{i\in\mathcal{D}_k} U_{\theta^\star}(X_i,Y_i) = O_p(n^{-1/2}). \tag{37}$$

Since $\mu(X,Q)$ is square integrable, sample independence, $n/N \to r$, $n_k/n \to \pi_k$, and the multivariate central limit theorem give

$$\sqrt{n}\left\{\frac{1}{N}\sum_{j=n+1}^{n+N}\mu(X_j,Q_j) - \frac{1}{n_k}\sum_{i\in\mathcal{D}_k}\mu(X_i,Q_i)\right\} = O_p(1).$$

Lemma 5 then gives

$$\frac{\widehat{M}^k}{1+n/N}\left\{\frac{1}{N}\sum_{j=n+1}^{n+N}\widehat{\mu}^k(X_j,Q_j) - \frac{1}{n_k}\sum_{i\in\mathcal{D}_k}\widehat{\mu}^k(X_i,Q_i)\right\} = O_p(n^{-1/2}). \tag{38}$$

By the definition of Ψ_k, (37) and (38) imply

$$\max_{1\le k\le 3}\|\Psi_k(\theta^\star)\| = O_p(n^{-1/2}). \tag{39}$$

We next verify the contraction conditions. The correction in Ψ_k does not depend on θ, so

$$\nabla_\theta\Psi_k(\theta) = \frac{1}{n_k}\sum_{i\in\mathcal{D}_k}\nabla_\theta U_\theta(X_i,Y_i).$$

Since $n_k \to \infty$ and there are only three folds, the finite-net argument used to prove (14), with n replaced by n_k, gives

$$\max_{1\le k\le 3} \sup_{\theta\in\mathcal{B}_\rho} \|\nabla_\theta \Psi_k(\theta) - \mathbb{E}\{\nabla_\theta U_\theta(X,Y)\}\| = o_p(1). \tag{40}$$

Define

$$T_k(\theta) = \theta - H_{\theta^\star}^{-1}\Psi_k(\theta).$$

Then choose $a \in (0,\rho)$ and $\kappa \in (0,1)$ so that the population derivative bound in (21) holds. Equation (40) then gives

$$\Pr\left\{\max_{1\le k\le 3} \sup_{\theta\in\mathcal{B}_a} \|\nabla_\theta T_k(\theta)\| \le \kappa\right\} \longrightarrow 1.$$

Whenever this derivative bound holds, for every $\theta \in \mathcal{B}_a$ and $k = 1,2,3$,

$$\|T_k(\theta) - \theta^\star\| \le \kappa\|\theta - \theta^\star\| + \|H_{\theta^\star}^{-1}\Psi_k(\theta^\star)\|.$$

Equation (39) therefore gives

$$\Pr\left\{\max_{1\le k\le 3} \|H_{\theta^\star}^{-1}\Psi_k(\theta^\star)\| \le (1-\kappa)a\right\} \longrightarrow 1.$$

Consequently,

$$\Pr\left(\left[\bigcap_{k=1}^{3}\{T_k(\mathcal{B}_a)\subseteq\mathcal{B}_a\}\right]\cap\left\{\max_{1\le k\le 3}\sup_{\theta\in\mathcal{B}_a}\|\nabla_\theta T_k(\theta)\|\le\kappa\right\}\right)\longrightarrow 1.$$

On this event, the Banach fixed-point theorem gives a unique root $\widehat{\theta}^k \in \mathcal{B}_a$ of $\Psi_k(\theta) = 0$ for every k. Moreover, the contraction inequality and $\widehat{\theta}^k = T_k(\widehat{\theta}^k)$ give, for each k,

$$(1-\kappa)\|\widehat{\theta}^k - \theta^\star\| \le \|H_{\theta^\star}^{-1}\Psi_k(\theta^\star)\|.$$

Taking the maximum over the three folds gives

$$(1-\kappa)\max_{1\le k\le 3}\|\widehat{\theta}^k - \theta^\star\| \le \max_{1\le k\le 3}\|H_{\theta^\star}^{-1}\Psi_k(\theta^\star)\| \le \|H_{\theta^\star}^{-1}\|\max_{1\le k\le 3}\|\Psi_k(\theta^\star)\|.$$

Therefore, (39) yields

$$\max_{1\le k\le 3}\|\widehat{\theta}^k - \theta^\star\| = O_p(n^{-1/2}). \tag{41}$$

Since $\widehat{\theta}^{\mathrm{DiPPI}} = \sum_{k=1}^3 \omega_k\widehat{\theta}^k$ and $\sum_{k=1}^3\omega_k = 1$,

$$\|\widehat{\theta}^{\mathrm{DiPPI}} - \theta^\star\| \le \sum_{k=1}^{3}\omega_k\|\widehat{\theta}^k - \theta^\star\| = O_p(n^{-1/2}).$$

This proves consistency.

We next derive the asymptotic linear representation. For each k, define

$$\overline{H}_k = \frac{1}{n_k}\sum_{i\in\mathcal{D}_k}\int_0^1 \nabla_\theta U_{\theta^\star + t(\widehat{\theta}^k - \theta^\star)}(X_i,Y_i)\,dt.$$

By (40), (41), and the continuity of $\theta \mapsto \mathbb{E}\{\nabla_\theta U_\theta(X,Y)\}$ at $\theta^\star$,

$$\max_{1\le k\le 3}\|\overline{H}_k - H_{\theta^\star}\| = o_p(1).$$

Since $H_{\theta^\star}$ is nonsingular,

$$\Pr\{\overline{H}_k \text{ is nonsingular for all } k\} \longrightarrow 1, \qquad \max_{1\le k\le 3}\|\overline{H}_k^{-1} - H_{\theta^\star}^{-1}\| = o_p(1).$$

The fundamental theorem of calculus along the segment from $\theta^\star$ to $\widehat{\theta}^k$ gives

$$0 = \Psi_k(\widehat{\theta}^k) = \Psi_k(\theta^\star) + \overline{H}_k(\widehat{\theta}^k - \theta^\star).$$

Therefore, using (39),

$$\begin{aligned}\widehat{\theta}^k - \theta^\star &= -\overline{H}_k^{-1}\Psi_k(\theta^\star) \\ &= -H_{\theta^\star}^{-1}\Psi_k(\theta^\star) - \{\overline{H}_k^{-1} - H_{\theta^\star}^{-1}\}\Psi_k(\theta^\star) \\ &= -H_{\theta^\star}^{-1}\Psi_k(\theta^\star) + o_p(n^{-1/2}).\end{aligned}$$

Substituting the definition of $\Psi_k(\theta^\star)$ and applying Lemma 5 give

$$\widehat{\theta}^k - \theta^\star = -H_{\theta^\star}^{-1}\left[\frac{1}{n_k}\sum_{i\in\mathcal{D}_k} U_{\theta^\star}(X_i, Y_i) + \frac{M_\mu}{1+r}\left\{\frac{1}{N}\sum_{j=n+1}^{n+N}\mu(X_j, Q_j) - \frac{1}{n_k}\sum_{i\in\mathcal{D}_k}\mu(X_i, Q_i)\right\}\right] + o_p(n^{-1/2}). \quad (42)$$

The remainder is $o_p(n^{-1/2})$ simultaneously for all three folds.

Since the folds partition the labeled sample,

$$\sum_{k=1}^{3}\omega_k\frac{1}{n_k}\sum_{i\in\mathcal{D}_k} f_i = \frac{1}{n}\sum_{i=1}^{n} f_i, \qquad \sum_{k=1}^{3}\omega_k = 1.$$

Multiplying (42) by ω_k, summing over k, and using $\mathbb{E}\{U_{\theta^\star}(X, Y)\} = 0$ yield

$$\begin{aligned}\sqrt{n}(\widehat{\theta}^{\mathrm{DiPPI}} - \theta^\star) = -H_{\theta^\star}^{-1}\Bigg[&\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left[U_{\theta^\star}(X_i, Y_i) - \frac{1}{1+r}M_\mu\{\mu(X_i, Q_i) - \mathbb{E}\mu(X, Q)\}\right] \\ &+ \frac{\sqrt{n}}{N}\sum_{j=n+1}^{n+N}\frac{1}{1+r}M_\mu\{\mu(X_j, Q_j) - \mathbb{E}\mu(X, Q)\}\Bigg] + o_p(1).\end{aligned} \quad (43)$$

It remains to derive the limiting distribution. The labeled-sample sum and the unlabeled-sample sum in (43) are independent and have mean zero. By square integrability, the multivariate central limit theorem, and $n/N \to r$, their limiting covariance is

$$H_{\theta^\star}^{-1}\left[\mathrm{Cov}\left\{U_{\theta^\star}(X, Y) - \frac{1}{1+r}M_\mu\mu(X, Q)\right\} + r\mathrm{Cov}\left\{\frac{1}{1+r}M_\mu\mu(X, Q)\right\}\right]H_{\theta^\star}^{-\top}.$$

By the definition of M_μ, we have

$$\begin{aligned}&\mathrm{Cov}\{U_{\theta^\star}(X, Y), M_\mu\mu(X, Q)\} \\ &\quad= \mathrm{Cov}\{M_\mu\mu(X, Q), U_{\theta^\star}(X, Y)\} \\ &\quad= \mathrm{Cov}\{M_\mu\mu(X, Q)\} \\ &\quad= \mathrm{Cov}\{U_{\theta^\star}(X, Y), \mu(X, Q)\}\mathrm{Cov}\{\mu(X, Q)\}^{-1}\mathrm{Cov}\{\mu(X, Q), U_{\theta^\star}(X, Y)\}.\end{aligned} \quad (44)$$

Expanding $\mathrm{Cov}\{U_{\theta^\star}(X, Y) - (1+r)^{-1}M_\mu\mu(X, Q)\} + r\mathrm{Cov}\{(1+r)^{-1}M_\mu\mu(X, Q)\}$ and using (44) gives

$$\begin{aligned}&\mathrm{Cov}\left\{U_{\theta^\star}(X, Y) - \frac{1}{1+r}M_\mu\mu(X, Q)\right\} + r\mathrm{Cov}\left\{\frac{1}{1+r}M_\mu\mu(X, Q)\right\} \\ &\quad= \mathrm{Cov}\{U_{\theta^\star}(X, Y)\} - \frac{2}{1+r}\mathrm{Cov}\{M_\mu\mu(X, Q)\} + \frac{1+r}{(1+r)^2}\mathrm{Cov}\{M_\mu\mu(X, Q)\} \\ &\quad= \mathrm{Cov}\{U_{\theta^\star}(X, Y)\} - \frac{1}{1+r}\mathrm{Cov}\{M_\mu\mu(X, Q)\}.\end{aligned} \quad (45)$$

Substituting (44) into (45), the limiting covariance is

$$\begin{aligned}H_{\theta^\star}^{-1}\Big[&\mathrm{Cov}\{U_{\theta^\star}(X, Y)\} - \frac{1}{1+r}\mathrm{Cov}\{U_{\theta^\star}(X, Y), \mu(X, Q)\}\mathrm{Cov}\{\mu(X, Q)\}^{-1} \\ &\times \mathrm{Cov}\{\mu(X, Q), U_{\theta^\star}(X, Y)\}\Big]H_{\theta^\star}^{-\top} = \Sigma_\mu^{\mathrm{DiPPI}}.\end{aligned}$$

Slutsky's theorem therefore yields

$$\sqrt{n}(\widehat{\theta}^{\mathrm{DiPPI}} - \theta^\star) \xrightarrow{d} N\left(0, \Sigma_\mu^{\mathrm{DiPPI}}\right),$$

where $\Sigma_\mu^{\mathrm{DiPPI}}$ is given in (7). □

B.3 Proof of Lemma 5

Proof. Fix k and condition on the two folds used to construct $\widehat{\theta}_0^k$ and $\widehat{\mu}^k$. The fitted quantities are then fixed, while $\mathcal{D}_k$ and the unlabeled sample remain independent. Expectations involving fitted quantities below are taken over an independent observation, conditional on the training folds.

We first prove $\widehat{M}^k \xrightarrow{p} M_\mu$. By (6),

$$\widehat{M}^k = \widehat{\mathrm{Cov}}_{\mathcal{D}_k}\{U_{\widehat{\theta}_0^k}(X,Y), \widehat{\mu}^k(X,Q)\} \left[\widehat{\mathrm{Cov}}_{\mathcal{D}_k}\{\widehat{\mu}^k(X,Q)\}\right]^{-1}.$$

Since $\mathrm{Cov}\{\mu(X,Q)\} \succ 0$, continuity of matrix inversion shows that it suffices to prove

$$\begin{aligned}\widehat{\mathrm{Cov}}_{\mathcal{D}_k}\{U_{\widehat{\theta}_0^k}(X,Y), \widehat{\mu}^k(X,Q)\} &= \mathrm{Cov}\{U_{\theta^\star}(X,Y), \mu(X,Q)\} + o_p(1),\\ \widehat{\mathrm{Cov}}_{\mathcal{D}_k}\{\widehat{\mu}^k(X,Q)\} &= \mathrm{Cov}\{\mu(X,Q)\} + o_p(1).\end{aligned} \tag{46}$$

To prove (46), write the two empirical covariances as

$$\begin{aligned}\widehat{\mathrm{Cov}}_{\mathcal{D}_k}\{U_{\widehat{\theta}_0^k}(X,Y), \widehat{\mu}^k(X,Q)\} &= \frac{1}{n_k}\sum_{i\in\mathcal{D}_k} U_{\widehat{\theta}_0^k}(X_i,Y_i)\widehat{\mu}^k(X_i,Q_i)^\top\\ &\quad - \left\{\frac{1}{n_k}\sum_{i\in\mathcal{D}_k} U_{\widehat{\theta}_0^k}(X_i,Y_i)\right\}\left\{\frac{1}{n_k}\sum_{i\in\mathcal{D}_k}\widehat{\mu}^k(X_i,Q_i)\right\}^\top,\\ \widehat{\mathrm{Cov}}_{\mathcal{D}_k}\{\widehat{\mu}^k(X,Q)\} &= \frac{1}{n_k}\sum_{i\in\mathcal{D}_k}\widehat{\mu}^k(X_i,Q_i)\widehat{\mu}^k(X_i,Q_i)^\top\\ &\quad - \left\{\frac{1}{n_k}\sum_{i\in\mathcal{D}_k}\widehat{\mu}^k(X_i,Q_i)\right\}\left\{\frac{1}{n_k}\sum_{i\in\mathcal{D}_k}\widehat{\mu}^k(X_i,Q_i)\right\}^\top.\end{aligned} \tag{47}$$

Thus, to prove (46), it remains to establish convergence of the four empirical moments displayed in (47). We first derive the empirical L_2 bounds used for those four limits. Since every fold size is of order n, applying Theorem 1 with $h_\theta \equiv 0$ to the three pilot estimating equations gives

$$\max_{1\le k\le 3}\|\widehat{\theta}_0^k - \theta^\star\| = O_p(n^{-1/2}).$$

On the event $\widehat{\theta}_0^k \in \mathcal{B}_\rho$, the integral mean-value theorem and Assumption 3 give

$$\mathbb{E}\|U_{\widehat{\theta}_0^k}(X,Y) - U_{\theta^\star}(X,Y)\|^2 \le \|\widehat{\theta}_0^k - \theta^\star\|^2 \mathbb{E}\left\{\sup_{\theta\in\mathcal{B}_\rho}\|\nabla_\theta U_\theta(X,Y)\|^2\right\} = o_p(1).$$

Cross-fitting makes $\mathcal{D}_k$ independent of $(\widehat{\theta}_0^k, \widehat{\mu}^k)$. Thus, for every $\varepsilon > 0$, conditional Markov's inequality gives

$$\begin{aligned}\Pr\left\{\frac{1}{n_k}\sum_{i\in\mathcal{D}_k}\|\widehat{\mu}^k(X_i,Q_i) - \mu(X_i,Q_i)\|^2 > \varepsilon \,\middle|\, \widehat{\mu}^k\right\} &\le \frac{1}{\varepsilon}\mathbb{E}\|\widehat{\mu}^k(X,Q) - \mu(X,Q)\|^2\\ &= o_p(1),\\ \Pr\left\{\frac{1}{n_k}\sum_{i\in\mathcal{D}_k}\|U_{\widehat{\theta}_0^k}(X_i,Y_i) - U_{\theta^\star}(X_i,Y_i)\|^2 > \varepsilon \,\middle|\, \widehat{\theta}_0^k\right\} &\le \frac{1}{\varepsilon}\mathbb{E}\|U_{\widehat{\theta}_0^k}(X,Y) - U_{\theta^\star}(X,Y)\|^2\\ &= o_p(1).\end{aligned}$$

Therefore, we can obtain

$$\begin{aligned}\frac{1}{n_k}\sum_{i\in\mathcal{D}_k}\|\widehat{\mu}^k(X_i,Q_i) - \mu(X_i,Q_i)\|^2 &= o_p(1),\\ \frac{1}{n_k}\sum_{i\in\mathcal{D}_k}\|U_{\widehat{\theta}_0^k}(X_i,Y_i) - U_{\theta^\star}(X_i,Y_i)\|^2 &= o_p(1).\end{aligned} \tag{48}$$

Assumption 1 gives $\mathbb{E}\|U_{\theta^\star}(X,Y)\|^2 < \infty$, while Theorem 2 assumes $\mathbb{E}\|\mu(X,Q)\|^2 < \infty$. Hence, the ordinary law of large numbers gives

$$\frac{1}{n_k}\sum_{i\in\mathcal{D}_k}\|\mu(X_i,Q_i)\|^2 = O_p(1), \qquad \frac{1}{n_k}\sum_{i\in\mathcal{D}_k}\|U_{\theta^\star}(X_i,Y_i)\|^2 = O_p(1). \tag{49}$$

Moreover,

$$\begin{aligned}\frac{1}{n_k}\sum_{i\in\mathcal{D}_k}\|\widehat{\mu}^k(X_i,Q_i)\|^2 &\le \frac{2}{n_k}\sum_{i\in\mathcal{D}_k}\|\widehat{\mu}^k(X_i,Q_i)-\mu(X_i,Q_i)\|^2 \\ &\quad + \frac{2}{n_k}\sum_{i\in\mathcal{D}_k}\|\mu(X_i,Q_i)\|^2 \\ &= O_p(1).\end{aligned} \tag{50}$$

The triangle inequality and the empirical Cauchy–Schwarz inequality give

$$\begin{aligned}&\left\|\frac{1}{n_k}\sum_{i\in\mathcal{D}_k}U_{\widehat{\theta}_0^k}(X_i,Y_i)\widehat{\mu}^k(X_i,Q_i)^\top - \frac{1}{n_k}\sum_{i\in\mathcal{D}_k}U_{\theta^\star}(X_i,Y_i)\mu(X_i,Q_i)^\top\right\|_F \\ &\le \left\{\frac{1}{n_k}\sum_{i\in\mathcal{D}_k}\|U_{\widehat{\theta}_0^k}(X_i,Y_i)-U_{\theta^\star}(X_i,Y_i)\|^2\right\}^{1/2}\left\{\frac{1}{n_k}\sum_{i\in\mathcal{D}_k}\|\widehat{\mu}^k(X_i,Q_i)\|^2\right\}^{1/2} \\ &\quad + \left\{\frac{1}{n_k}\sum_{i\in\mathcal{D}_k}\|U_{\theta^\star}(X_i,Y_i)\|^2\right\}^{1/2}\left\{\frac{1}{n_k}\sum_{i\in\mathcal{D}_k}\|\widehat{\mu}^k(X_i,Q_i)-\mu(X_i,Q_i)\|^2\right\}^{1/2} \\ &= o_p(1)O_p(1) + O_p(1)o_p(1) = o_p(1).\end{aligned} \tag{51}$$

Similarly,

$$\begin{aligned}&\left\|\frac{1}{n_k}\sum_{i\in\mathcal{D}_k}\widehat{\mu}^k(X_i,Q_i)\widehat{\mu}^k(X_i,Q_i)^\top - \frac{1}{n_k}\sum_{i\in\mathcal{D}_k}\mu(X_i,Q_i)\mu(X_i,Q_i)^\top\right\|_F \\ &\le \left\{\frac{1}{n_k}\sum_{i\in\mathcal{D}_k}\|\widehat{\mu}^k(X_i,Q_i)-\mu(X_i,Q_i)\|^2\right\}^{1/2} \\ &\quad \times\left[\left\{\frac{1}{n_k}\sum_{i\in\mathcal{D}_k}\|\widehat{\mu}^k(X_i,Q_i)\|^2\right\}^{1/2} + \left\{\frac{1}{n_k}\sum_{i\in\mathcal{D}_k}\|\mu(X_i,Q_i)\|^2\right\}^{1/2}\right] \\ &= o_p(1)\{O_p(1)+O_p(1)\} = o_p(1).\end{aligned} \tag{52}$$

For the empirical means,

$$\begin{aligned}\left\|\frac{1}{n_k}\sum_{i\in\mathcal{D}_k}\{U_{\widehat{\theta}_0^k}(X_i,Y_i)-U_{\theta^\star}(X_i,Y_i)\}\right\| &\le \frac{1}{n_k}\sum_{i\in\mathcal{D}_k}\|U_{\widehat{\theta}_0^k}(X_i,Y_i)-U_{\theta^\star}(X_i,Y_i)\| \\ &\le \left\{\frac{1}{n_k}\sum_{i\in\mathcal{D}_k}\|U_{\widehat{\theta}_0^k}(X_i,Y_i)-U_{\theta^\star}(X_i,Y_i)\|^2\right\}^{1/2} \\ &= o_p(1),\end{aligned} \tag{53}$$

$$\begin{aligned}\left\|\frac{1}{n_k}\sum_{i\in\mathcal{D}_k}\{\widehat{\mu}^k(X_i,Q_i)-\mu(X_i,Q_i)\}\right\| &\le \frac{1}{n_k}\sum_{i\in\mathcal{D}_k}\|\widehat{\mu}^k(X_i,Q_i)-\mu(X_i,Q_i)\| \\ &\le \left\{\frac{1}{n_k}\sum_{i\in\mathcal{D}_k}\|\widehat{\mu}^k(X_i,Q_i)-\mu(X_i,Q_i)\|^2\right\}^{1/2} \\ &= o_p(1).\end{aligned} \tag{54}$$

Equations (51)–(54) follow from (48), (49), and (50). We next show that the corresponding oracle empirical moments converge to their population values. The same moment conditions and Cauchy–Schwarz give

$$\mathbb{E}\|U_{\theta^\star}(X,Y)\mu(X,Q)^\top\| \le \{\mathbb{E}\|U_{\theta^\star}(X,Y)\|^2\}^{1/2}\{\mathbb{E}\|\mu(X,Q)\|^2\}^{1/2} < \infty,$$
$$\mathbb{E}\|\mu(X,Q)\mu(X,Q)^\top\| = \mathbb{E}\|\mu(X,Q)\|^2 < \infty.$$

Combining the ordinary law of large numbers with (51)–(54) gives

$$\begin{aligned}
\frac{1}{n_k}\sum_{i\in\mathcal{D}_k} U_{\widehat{\theta}_0^k}(X_i,Y_i)\widehat{\mu}^k(X_i,Q_i)^\top &= \mathbb{E}\{U_{\theta^\star}(X,Y)\mu(X,Q)^\top\} + o_p(1),\\
\frac{1}{n_k}\sum_{i\in\mathcal{D}_k} \widehat{\mu}^k(X_i,Q_i)\widehat{\mu}^k(X_i,Q_i)^\top &= \mathbb{E}\{\mu(X,Q)\mu(X,Q)^\top\} + o_p(1),\\
\frac{1}{n_k}\sum_{i\in\mathcal{D}_k} U_{\widehat{\theta}_0^k}(X_i,Y_i) &= \mathbb{E}\{U_{\theta^\star}(X,Y)\} + o_p(1),\\
\frac{1}{n_k}\sum_{i\in\mathcal{D}_k} \widehat{\mu}^k(X_i,Q_i) &= \mathbb{E}\{\mu(X,Q)\} + o_p(1).
\end{aligned} \tag{55}$$

Substituting (55) into (47) proves (46). Since $\mathrm{Cov}\{\mu(X,Q)\} \succ 0$, continuity of matrix inversion gives

$$\left[\widehat{\mathrm{Cov}}_{\mathcal{D}_k}\{\widehat{\mu}^k(X,Q)\}\right]^{-1} = \mathrm{Cov}\{\mu(X,Q)\}^{-1} + o_p(1).$$

Hence,

$$\widehat{M}^k = M_\mu + o_p(1) \xrightarrow{p} M_\mu, \qquad \|\widehat{M}^k\| = O_p(1). \tag{56}$$

It remains to replace the fitted correction by its population limit. Conditional on the training folds, the labeled and unlabeled averages of $\widehat{\mu}^k(X,Q)-\mu(X,Q)$ have the same mean. Their independence gives

$$\begin{aligned}
&\mathbb{E}\left\|\sqrt{n}\left[\frac{1}{N}\sum_{j=n+1}^{n+N}\{\widehat{\mu}^k(X_j,Q_j)-\mu(X_j,Q_j)\} - \frac{1}{n_k}\sum_{i\in\mathcal{D}_k}\{\widehat{\mu}^k(X_i,Q_i)-\mu(X_i,Q_i)\}\right]\right\|^2\\
&= \left(\frac{n}{N}+\frac{n}{n_k}\right)\mathbb{E}\left\|\widehat{\mu}^k(X,Q)-\mu(X,Q)-\mathbb{E}\{\widehat{\mu}^k(X,Q)-\mu(X,Q)\}\right\|^2\\
&\le \left(\frac{n}{N}+\frac{n}{n_k}\right)\mathbb{E}\|\widehat{\mu}^k(X,Q)-\mu(X,Q)\|^2 = o_p(1).
\end{aligned}$$

Conditional Markov's inequality and boundedness of conditional probabilities therefore imply

$$\sqrt{n}\left[\frac{1}{N}\sum_{j=n+1}^{n+N}\{\widehat{\mu}^k(X_j,Q_j)-\mu(X_j,Q_j)\} - \frac{1}{n_k}\sum_{i\in\mathcal{D}_k}\{\widehat{\mu}^k(X_i,Q_i)-\mu(X_i,Q_i)\}\right] = o_p(1). \tag{57}$$

Since $\mathbb{E}\|\mu(X,Q)\|^2 < \infty$, the multivariate central limit theorem gives

$$\begin{aligned}
&\sqrt{n}\left\{\frac{1}{N}\sum_{j=n+1}^{n+N}\mu(X_j,Q_j) - \frac{1}{n_k}\sum_{i\in\mathcal{D}_k}\mu(X_i,Q_i)\right\}\\
&= \sqrt{\frac{n}{N}}\frac{1}{\sqrt{N}}\sum_{j=n+1}^{n+N}\{\mu(X_j,Q_j)-\mathbb{E}\mu(X,Q)\}\\
&\quad - \sqrt{\frac{n}{n_k}}\frac{1}{\sqrt{n_k}}\sum_{i\in\mathcal{D}_k}\{\mu(X_i,Q_i)-\mathbb{E}\mu(X,Q)\} = O_p(1).
\end{aligned} \tag{58}$$

Finally, by the definition of $R_{n,k}$ in Lemma 5,

$$R_{n,k} = \frac{\widehat{M}^k}{1+n/N}\sqrt{n}\left[\frac{1}{N}\sum_{j=n+1}^{n+N}\{\widehat{\mu}^k(X_j,Q_j)-\mu(X_j,Q_j)\} - \frac{1}{n_k}\sum_{i\in\mathcal{D}_k}\{\widehat{\mu}^k(X_i,Q_i)-\mu(X_i,Q_i)\}\right]$$
$$+ \frac{\widehat{M}^k - M_\mu}{1+n/N}\sqrt{n}\left\{\frac{1}{N}\sum_{j=n+1}^{n+N}\mu(X_j,Q_j) - \frac{1}{n_k}\sum_{i\in\mathcal{D}_k}\mu(X_i,Q_i)\right\}$$
$$+ \left\{\frac{1}{1+n/N} - \frac{1}{1+r}\right\}M_\mu\sqrt{n}\left\{\frac{1}{N}\sum_{j=n+1}^{n+N}\mu(X_j,Q_j) - \frac{1}{n_k}\sum_{i\in\mathcal{D}_k}\mu(X_i,Q_i)\right\}.$$

By (56), (57), and (58), the three terms on the right are $o_p(1)$. Therefore, $R_{n,k} = o_p(1)$, which proves the lemma. □

B.4 Proof of Corollary 3

Proof. Because $\mathrm{Cov}\{\mu(X,Q)\}$ is positive definite,

$$a^\top \mathrm{Cov}\{U_{\theta^\star}(X,Y),\mu(X,Q)\}\mathrm{Cov}\{\mu(X,Q)\}^{-1}\mathrm{Cov}\{\mu(X,Q),U_{\theta^\star}(X,Y)\}a$$
$$= \left\|\mathrm{Cov}\{\mu(X,Q)\}^{-1/2}\mathrm{Cov}\{\mu(X,Q),U_{\theta^\star}(X,Y)\}a\right\|^2 \geq 0$$

for every $a\in\mathbb{R}^d$. Subtracting (7) from the labeled-only covariance gives (8). Premultiplication by $H_{\theta^\star}^{-1}$ and postmultiplication by $H_{\theta^\star}^{-\top}$ preserve positive semidefiniteness.

For the oracle statement, suppose $\mu(X,Q) = \mathbb{E}\{U_{\theta^\star}(X,Y)\mid X,\Phi(Q)\}$. By iterated expectations and $\mathbb{E}\{U_{\theta^\star}(X,Y)\}=0$,

$$\mathbb{E}\{\mu(X,Q)\} = \mathbb{E}\{U_{\theta^\star}(X,Y)\} = 0.$$

Since $\mu(X,Q)$ is measurable with respect to $(X,\Phi(Q))$, another application of iterated expectations gives

$$\begin{aligned}\mathrm{Cov}\{U_{\theta^\star}(X,Y),\mu(X,Q)\} &= \mathbb{E}\{U_{\theta^\star}(X,Y)\mu(X,Q)^\top\}\\ &= \mathbb{E}\big[\mathbb{E}\{U_{\theta^\star}(X,Y)\mid X,\Phi(Q)\}\mu(X,Q)^\top\big]\\ &= \mathbb{E}\{\mu(X,Q)\mu(X,Q)^\top\}\\ &= \mathrm{Cov}\{\mu(X,Q)\}.\end{aligned}$$

Thus, $M_\mu = I_d$. Substituting this into (7) gives (9). Theorem 1, applied with Q replaced by $\Phi(Q)$, shows that (9) is the minimum asymptotic covariance among corrections measurable with respect to $(X,\Phi(Q))$ □

B.5 Proof of Theorem 4

Proof. We first prove covariance consistency in (11). Since $\widehat{H}\xrightarrow{p}H_{\theta^\star}$ and $H_{\theta^\star}$ is nonsingular, the continuous mapping theorem gives

$$\widehat{H}^{-1}\xrightarrow{p}H_{\theta^\star}^{-1}.$$

Thus, in view of (10) and $n/N\to r$, it suffices to prove

$$\sum_{k=1}^{3}\omega_k\widehat{\mathrm{Cov}}_{\mathcal{D}_k}\left\{U_{\widehat{\theta}^{\mathrm{DiPPI}}}(X_i,Y_i) - \frac{\widehat{M}^k\widehat{\mu}^k(X_i,Q_i)}{1+n/N}\right\}\xrightarrow{p}\mathrm{Cov}\left\{U_{\theta^\star}(X,Y) - \frac{M_\mu\mu(X,Q)}{1+r}\right\} \tag{59}$$

and

$$\widehat{\mathrm{Cov}}_{\mathrm{U}}\left\{\frac{1}{1+n/N}\sum_{k=1}^{3}\omega_k\widehat{M}^k\widehat{\mu}^k(X_j,Q_j)\right\}\xrightarrow{p}\mathrm{Cov}\left\{\frac{1}{1+r}M_\mu\mu(X,Q)\right\}. \tag{60}$$

After proving (59) and (60), we combine the resulting covariance consistency with Theorem 2 and Slutsky's theorem to prove (12) and the stated Wald coverage.

We first prove (59). After multiplication by $\omega_k = |\mathcal{D}_k|/n$, every observation in $\mathcal{D}_k$ has weight $1/n$ in (43). The labeled folds are disjoint, so their covariance contributions add with weights ω_k. The same unlabeled observations are used for all three fold estimates, so the three unlabeled corrections are averaged before their covariance is computed, as in (10).

For each fixed k, write the difference between the fitted and limiting corrections as

$$\frac{1}{1+n/N}\widehat{M}^k\{\widehat{\mu}^k(X_i,Q_i)-\mu(X_i,Q_i)\}+\left\{\frac{1}{1+n/N}\widehat{M}^k-\frac{1}{1+r}M_\mu\right\}\mu(X_i,Q_i).$$

Their empirical squared norms satisfy

$$\begin{aligned}
&\frac{1}{n_k}\sum_{i\in\mathcal{D}_k}\left\|\frac{\widehat{M}^k}{1+n/N}\{\widehat{\mu}^k(X_i,Q_i)-\mu(X_i,Q_i)\}\right\|^2 \\
&\quad\le \frac{\|\widehat{M}^k\|^2}{(1+n/N)^2}\times\frac{1}{n_k}\sum_{i\in\mathcal{D}_k}\|\widehat{\mu}^k(X_i,Q_i)-\mu(X_i,Q_i)\|^2=o_p(1), \\
&\frac{1}{n_k}\sum_{i\in\mathcal{D}_k}\left\|\left\{\frac{\widehat{M}^k}{1+n/N}-\frac{M_\mu}{1+r}\right\}\mu(X_i,Q_i)\right\|^2 \\
&\quad\le\left\|\frac{\widehat{M}^k}{1+n/N}-\frac{M_\mu}{1+r}\right\|^2\times\frac{1}{n_k}\sum_{i\in\mathcal{D}_k}\|\mu(X_i,Q_i)\|^2=o_p(1).
\end{aligned}$$

Equation (48) and $\|\widehat{M}^k\|=O_p(1)$ prove the first bound. Equation (56), $n/N\to r$, and

$$\frac{1}{n_k}\sum_{i\in\mathcal{D}_k}\|\mu(X_i,Q_i)\|^2=O_p(1)$$

prove the second bound. Combining them and using the fixed number of folds gives

$$\sum_{k=1}^{3}\omega_k\frac{1}{n_k}\sum_{i\in\mathcal{D}_k}\left\|\frac{\widehat{M}^k\widehat{\mu}^k(X_i,Q_i)}{1+n/N}-\frac{M_\mu\mu(X_i,Q_i)}{1+r}\right\|^2=o_p(1). \tag{61}$$

Theorem 2 gives $\widehat{\theta}^{\mathrm{DiPPI}}\xrightarrow{p}\theta^\star$; hence,

$$\Pr\{\widehat{\theta}^{\mathrm{DiPPI}}\in\mathrm{int}(\mathcal{B}_\rho)\}=\Pr\{\|\widehat{\theta}^{\mathrm{DiPPI}}-\theta^\star\|<\rho\}\longrightarrow 1.$$

On the event $\widehat{\theta}^{\mathrm{DiPPI}}\in\mathcal{B}_\rho$, a first-order Taylor expansion in integral form and Assumptions 1 and 3 give

$$\begin{aligned}
&\frac{1}{n}\sum_{i=1}^{n}\|U_{\widehat{\theta}^{\mathrm{DiPPI}}}(X_i,Y_i)-U_{\theta^\star}(X_i,Y_i)\|^2 \\
&\quad\le\|\widehat{\theta}^{\mathrm{DiPPI}}-\theta^\star\|^2\frac{1}{n}\sum_{i=1}^{n}\sup_{\xi\in\mathcal{B}_\rho}\|\nabla_\theta U_\xi(X_i,Y_i)\|^2 \\
&\quad=o_p(1).
\end{aligned}\tag{62}$$

Since $\omega_k/n_k=1/n$, (62) and (61) imply

$$\begin{aligned}
&\sum_{k=1}^{3}\omega_k\frac{1}{n_k}\sum_{i\in\mathcal{D}_k}\left\|U_{\widehat{\theta}^{\mathrm{DiPPI}}}(X_i,Y_i)-\frac{\widehat{M}^k\widehat{\mu}^k(X_i,Q_i)}{1+n/N}-U_{\theta^\star}(X_i,Y_i)+\frac{M_\mu\mu(X_i,Q_i)}{1+r}\right\|^2 \\
&\le\frac{2}{n}\sum_{i=1}^{n}\|U_{\widehat{\theta}^{\mathrm{DiPPI}}}(X_i,Y_i)-U_{\theta^\star}(X_i,Y_i)\|^2 \\
&\quad+2\sum_{k=1}^{3}\omega_k\frac{1}{n_k}\sum_{i\in\mathcal{D}_k}\left\|\frac{\widehat{M}^k\widehat{\mu}^k(X_i,Q_i)}{1+n/N}-\frac{M_\mu\mu(X_i,Q_i)}{1+r}\right\|^2=o_p(1).
\end{aligned}\tag{63}$$

Since every term on the left-hand side of (63) is nonnegative and $\omega_k^{-1} \to \pi_k^{-1} < \infty$, for each $k = 1, 2, 3$,

$$\frac{1}{n_k} \sum_{i \in \mathcal{D}_k} \left\| U_{\widehat{\theta}^{\mathrm{DiPPI}}}(X_i, Y_i) - \frac{\widehat{M}^k \widehat{\mu}^k(X_i, Q_i)}{1 + n/N} - U_{\theta^\star}(X_i, Y_i) + \frac{M_\mu \mu(X_i, Q_i)}{1 + r} \right\|^2 = o_p(1). \quad (64)$$

Moreover,

$$\mathbb{E} \left\| U_{\theta^\star}(X, Y) - \frac{M_\mu \mu(X, Q)}{1 + r} \right\|^2 \leq 2\mathbb{E}\|U_{\theta^\star}(X, Y)\|^2 + \frac{2\|M_\mu\|^2}{(1 + r)^2} \mathbb{E}\|\mu(X, Q)\|^2 < \infty.$$

Since $n_k/n \to \pi_k > 0$ implies $n_k \to \infty$, the weak law of large numbers therefore gives

$$\frac{1}{n_k} \sum_{i \in \mathcal{D}_k} \left\| U_{\theta^\star}(X_i, Y_i) - \frac{M_\mu \mu(X_i, Q_i)}{1 + r} \right\|^2 \xrightarrow{p} \mathbb{E} \left\| U_{\theta^\star}(X, Y) - \frac{M_\mu \mu(X, Q)}{1 + r} \right\|^2 < \infty, \quad (65)$$

Thus, the empirical second moment in (65) is $O_p(1)$. The triangle inequality and (64) give

$$\frac{1}{n_k} \sum_{i \in \mathcal{D}_k} \left\| U_{\widehat{\theta}^{\mathrm{DiPPI}}}(X_i, Y_i) - \frac{\widehat{M}^k \widehat{\mu}^k(X_i, Q_i)}{1 + n/N} \right\|^2 \quad (66)$$

$$\leq \frac{2}{n_k} \sum_{i \in \mathcal{D}_k} \left\| U_{\widehat{\theta}^{\mathrm{DiPPI}}}(X_i, Y_i) - \frac{\widehat{M}^k \widehat{\mu}^k(X_i, Q_i)}{1 + n/N} - U_{\theta^\star}(X_i, Y_i) + \frac{M_\mu \mu(X_i, Q_i)}{1 + r} \right\|^2 \quad (67)$$

$$+ \frac{2}{n_k} \sum_{i \in \mathcal{D}_k} \left\| U_{\theta^\star}(X_i, Y_i) - \frac{M_\mu \mu(X_i, Q_i)}{1 + r} \right\|^2 = O_p(1). \quad (68)$$

Applying the empirical Cauchy–Schwarz inequality separately to the empirical means and second moments, using (64), (65), and (68), gives

$$\left\| \widehat{\mathrm{Cov}}_{\mathcal{D}_k} \left\{ U_{\widehat{\theta}^{\mathrm{DiPPI}}}(X_i, Y_i) - \frac{\widehat{M}^k \widehat{\mu}^k(X_i, Q_i)}{1 + n/N} \right\} \right. \quad (69)$$

$$\left. - \widehat{\mathrm{Cov}}_{\mathcal{D}_k} \left\{ U_{\theta^\star}(X_i, Y_i) - \frac{M_\mu \mu(X_i, Q_i)}{1 + r} \right\} \right\|_F = o_p(1). \quad (70)$$

The ordinary law of large numbers gives

$$\widehat{\mathrm{Cov}}_{\mathcal{D}_k} \left\{ U_{\theta^\star}(X_i, Y_i) - \frac{M_\mu \mu(X_i, Q_i)}{1 + r} \right\} \xrightarrow{p} \mathrm{Cov} \left\{ U_{\theta^\star}(X, Y) - \frac{M_\mu \mu(X, Q)}{1 + r} \right\}. \quad (71)$$

Combining (70) and (71), and using $\sum_{k=1}^3 \omega_k = 1$ and the fixed number of folds, proves (59).

We next prove (60). For the unlabeled contribution, condition on the labeled data. For each fixed k and every $\varepsilon > 0$, conditional Markov's inequality gives

$$\Pr\left\{ \frac{1}{N} \sum_{j=n+1}^{n+N} \|\widehat{\mu}^k(X_j, Q_j) - \mu(X_j, Q_j)\|^2 > \varepsilon \,\middle|\, \widehat{\mu}^k \right\} \leq \frac{1}{\varepsilon} \mathbb{E}\|\widehat{\mu}^k(X, Q) - \mu(X, Q)\|^2 = o_p(1).$$

Since the conditional probability is bounded by one,

$$\Pr\left\{ \frac{1}{N} \sum_{j=n+1}^{n+N} \|\widehat{\mu}^k(X_j, Q_j) - \mu(X_j, Q_j)\|^2 > \varepsilon \right\}$$

$$= \mathbb{E}\left[\Pr\left\{ \frac{1}{N} \sum_{j=n+1}^{n+N} \|\widehat{\mu}^k(X_j, Q_j) - \mu(X_j, Q_j)\|^2 > \varepsilon \,\middle|\, \widehat{\mu}^k \right\} \right] \longrightarrow 0.$$

Combining this convergence with the ordinary law of large numbers and $\mathbb{E}\|\mu(X,Q)\|^2 < \infty$ gives

$$\frac{1}{N}\sum_{j=n+1}^{n+N} \|\widehat{\mu}^k(X_j,Q_j) - \mu(X_j,Q_j)\|^2 = o_p(1), \qquad \frac{1}{N}\sum_{j=n+1}^{n+N} \|\mu(X_j,Q_j)\|^2 = O_p(1). \tag{72}$$

Together with $\widehat{M}^k = M_\mu + o_p(1)$ and $n/N \to r$, (72) gives

$$\frac{1}{N}\sum_{j=n+1}^{n+N} \left\| \frac{\widehat{M}^k\widehat{\mu}^k(X_j,Q_j)}{1+n/N} - \frac{M_\mu\mu(X_j,Q_j)}{1+r} \right\|^2 = o_p(1). \tag{73}$$

Since $\sum_{k=1}^3 \omega_k = 1$, convexity gives

$$\begin{aligned}
&\frac{1}{N}\sum_{j=n+1}^{n+N} \left\| \frac{1}{1+n/N}\sum_{k=1}^3 \omega_k \widehat{M}^k\widehat{\mu}^k(X_j,Q_j) - \frac{1}{1+r}M_\mu\mu(X_j,Q_j) \right\|^2 \\
&\quad \le \sum_{k=1}^3 \omega_k \frac{1}{N}\sum_{j=n+1}^{n+N} \left\| \frac{\widehat{M}^k\widehat{\mu}^k(X_j,Q_j)}{1+n/N} - \frac{M_\mu\mu(X_j,Q_j)}{1+r} \right\|^2.
\end{aligned}$$

By (73) and the fixed number of folds, the right-hand side is $o_p(1)$. Hence

$$\frac{1}{N}\sum_{j=n+1}^{n+N} \left\| \frac{1}{1+n/N}\sum_{k=1}^3 \omega_k \widehat{M}^k\widehat{\mu}^k(X_j,Q_j) - \frac{1}{1+r}M_\mu\mu(X_j,Q_j) \right\|^2 = o_p(1). \tag{74}$$

The ordinary law of large numbers gives

$$\frac{1}{N}\sum_{j=n+1}^{n+N} \left\| \frac{1}{1+r}M_\mu\mu(X_j,Q_j) \right\|^2 = O_p(1). \tag{75}$$

The triangle inequality and (74) give

$$\frac{1}{N}\sum_{j=n+1}^{n+N} \left\| \frac{1}{1+n/N}\sum_{k=1}^3 \omega_k \widehat{M}^k\widehat{\mu}^k(X_j,Q_j) \right\|^2 \tag{76}$$

$$\le \frac{2}{N}\sum_{j=n+1}^{n+N} \left\| \frac{1}{1+n/N}\sum_{k=1}^3 \omega_k \widehat{M}^k\widehat{\mu}^k(X_j,Q_j) - \frac{1}{1+r}M_\mu\mu(X_j,Q_j) \right\|^2 \tag{77}$$

$$+ \frac{2}{N}\sum_{j=n+1}^{n+N} \left\| \frac{1}{1+r}M_\mu\mu(X_j,Q_j) \right\|^2 = O_p(1). \tag{78}$$

Applying the empirical Cauchy–Schwarz inequality to the empirical means and second moments, using (74), (75), and (78), gives

$$\left\| \widehat{\mathrm{Cov}}_{\mathrm{U}} \left\{ \frac{1}{1+n/N}\sum_{k=1}^3 \omega_k \widehat{M}^k\widehat{\mu}^k(X_j,Q_j) \right\} - \widehat{\mathrm{Cov}}_{\mathrm{U}} \left\{ \frac{1}{1+r}M_\mu\mu(X_j,Q_j) \right\} \right\|_F = o_p(1). \tag{79}$$

The ordinary law of large numbers gives

$$\widehat{\mathrm{Cov}}_{\mathrm{U}} \left\{ \frac{1}{1+r}M_\mu\mu(X_j,Q_j) \right\} \xrightarrow{p} \mathrm{Cov} \left\{ \frac{1}{1+r}M_\mu\mu(X,Q) \right\}. \tag{80}$$

Combining (79) and (80) proves (60). The average over k is taken before the covariance because all three estimators use the same unlabeled observations.

By the definition

$$M_\mu = \mathrm{Cov}\{U_{\theta^\star}(X,Y), \mu(X,Q)\}\mathrm{Cov}\{\mu(X,Q)\}^{-1},$$

we have

$$\begin{aligned}\mathrm{Cov}\{U_{\theta^\star}(X,Y), M_\mu\mu(X,Q)\} &= \mathrm{Cov}\{M_\mu\mu(X,Q), U_{\theta^\star}(X,Y)\}\\ &= \mathrm{Cov}\{M_\mu\mu(X,Q)\}\\ &= \mathrm{Cov}\{U_{\theta^\star}(X,Y), \mu(X,Q)\}\mathrm{Cov}\{\mu(X,Q)\}^{-1}\\ &\quad\times \mathrm{Cov}\{\mu(X,Q), U_{\theta^\star}(X,Y)\}.\end{aligned}$$

Consequently,

$$\begin{aligned}&\mathrm{Cov}\left\{U_{\theta^\star}(X,Y) - \frac{1}{1+r}M_\mu\mu(X,Q)\right\} + r\mathrm{Cov}\left\{\frac{1}{1+r}M_\mu\mu(X,Q)\right\}\\ &\quad= \mathrm{Cov}\{U_{\theta^\star}(X,Y)\} - \frac{1}{1+r}\mathrm{Cov}\{U_{\theta^\star}(X,Y), \mu(X,Q)\}\mathrm{Cov}\{\mu(X,Q)\}^{-1}\\ &\qquad\times \mathrm{Cov}\{\mu(X,Q), U_{\theta^\star}(X,Y)\}. \qquad (81)\end{aligned}$$

Combining (59), (60), $n/N \to r$, and (81) gives

$$\begin{aligned}&\sum_{k=1}^{3}\omega_k\widehat{\mathrm{Cov}}_{\mathcal{D}_k}\left\{U_{\widehat{\theta}^{\mathrm{DiPPI}}}(X_i,Y_i) - \frac{\widehat{M}^k\widehat{\mu}^k(X_i,Q_i)}{1+n/N}\right\} + \frac{n}{N}\widehat{\mathrm{Cov}}_{\mathrm{U}}\left\{\frac{1}{1+n/N}\sum_{k=1}^{3}\omega_k\widehat{M}^k\widehat{\mu}^k(X_j,Q_j)\right\}\\ &\xrightarrow{p} \mathrm{Cov}\{U_{\theta^\star}(X,Y)\} - \frac{1}{1+r}\mathrm{Cov}\{U_{\theta^\star}(X,Y), \mu(X,Q)\}\mathrm{Cov}\{\mu(X,Q)\}^{-1}\mathrm{Cov}\{\mu(X,Q), U_{\theta^\star}(X,Y)\}.\end{aligned}$$

Since $H_{\theta^\star}$ is nonsingular,

$$\widehat{H} \xrightarrow{p} H_{\theta^\star} \quad\Longrightarrow\quad \widehat{H}^{-1} \xrightarrow{p} H_{\theta^\star}^{-1}.$$

Therefore,

$$\widehat{\Sigma}^{\mathrm{DiPPI}} \xrightarrow{p} \Sigma_\mu^{\mathrm{DiPPI}},$$

which proves (11).

For any fixed a satisfying $a^\top\Sigma_\mu^{\mathrm{DiPPI}}a > 0$, Theorem 2 and (11) give

$$\frac{\sqrt{n}\,a^\top(\widehat{\theta}^{\mathrm{DiPPI}} - \theta^\star)}{\{a^\top\Sigma_\mu^{\mathrm{DiPPI}}a\}^{1/2}} \xrightarrow{d} N(0,1),$$

$$\frac{a^\top\widehat{\Sigma}^{\mathrm{DiPPI}}a}{a^\top\Sigma_\mu^{\mathrm{DiPPI}}a} \xrightarrow{p} 1.$$

In particular,

$$\Pr\{a^\top\widehat{\Sigma}^{\mathrm{DiPPI}}a > 0\} \longrightarrow 1, \qquad \left\{\frac{a^\top\widehat{\Sigma}^{\mathrm{DiPPI}}a}{a^\top\Sigma_\mu^{\mathrm{DiPPI}}a}\right\}^{1/2} \xrightarrow{p} 1.$$

Slutsky's theorem therefore proves (12). Finally,

$$\begin{aligned}&\Pr\left\{a^\top\theta^\star \in a^\top\widehat{\theta}^{\mathrm{DiPPI}} \pm z_{1-\alpha/2}\left(\frac{a^\top\widehat{\Sigma}^{\mathrm{DiPPI}}a}{n}\right)^{1/2}\right\}\\ &\quad= \Pr\left\{\left|\frac{\sqrt{n}\,a^\top(\widehat{\theta}^{\mathrm{DiPPI}} - \theta^\star)}{\{a^\top\widehat{\Sigma}^{\mathrm{DiPPI}}a\}^{1/2}}\right| \le z_{1-\alpha/2}\right\} \longrightarrow 1-\alpha,\end{aligned}$$

which proves the stated coverage. □