Prediction-Powered Inference via Calibration

Lars van der Laan
Department of Statistics, University of Washington

Mark van der Laan*
Department of Statistics and Division of Biostatistics,
Center for Targeted Machine Learning and Causal Inference,
University of California, Berkeley

April 14, 2026

Abstract

We study semisupervised mean estimation with one labeled sample and one unlabeled sample, in settings where a black-box prediction model is available but may be biased. This problem arises, for example, in modern randomized trials, where machine-learning predictors trained on auxiliary data are used to improve precision, and in recent machine-learning applications with abundant unlabeled data. A standard approach is augmented inverse-probability weighting (AIPW) (Robins et al., 1994), which uses the prediction model to improve efficiency while remaining valid even when the model is misspecified. Our proposal is simple: *"predict, calibrate, and average."* We introduce a class of two-step regression-based estimators that first calibrate the prediction model on the labeled sample and then average the calibrated predictions over the pooled covariate sample. The resulting estimator retains a simple imputation-style plug-in form while also admitting an AIPW representation in which calibration removes the augmentation term; in particular, it can be viewed as a targeted minimum loss estimator (Van Der Laan and Rubin, 2006). We implement calibration using isotonic regression, a lightweight, tuning-free method that learns the best monotone transformation of the prediction score. We show that the resulting estimator is root-n consistent and asymptotically normal under weak conditions. It is efficient among estimators based on the calibrated score, and fully efficient when the calibrated limit equals the true regression function. We also clarify its relationship with recent prediction-powered inference (PPI) methods: PPI is less efficient than AIPW for accurate models, while PPI++ is AIPW with empirical-efficiency maximization (Rubin and van der Laan, 2008) and is first-order equivalent to linear calibration (Mincer and Zarnowitz, 1969) and classical prognostic-score regression adjustment (Hansen, 2008). In synthetic and real-data experiments, calibration-based estimators perform strongly, often outperforming PPI and remaining competitive with or improving on AIPW and PPI++ under meaningful miscalibration. We implement the proposed methods in an accompanying `Python` package, `ppi_aipw`.

1 Introduction

We study semisupervised mean estimation with one labeled dataset $\{(X_i, Y_i)\}_{i=1}^n$ and one unlabeled dataset $\{\widetilde{X}_i\}_{i=1}^N$, where outcomes are observed only for a subset sampled completely at random. This is a classical missing-outcome problem in semiparametric statistics and missing-data theory, for which efficient inference with black-box learners is possible using debiased machine learning methods such as augmented inverse-probability weighting (Robins et al., 1994; van der Laan and Robins, 2003), targeted minimum loss estimation (Van Der Laan and Rubin, 2006; van der Laan and Rose, 2011), and double machine learning (Chernozhukov et al., 2018a). Recent machine-learning work has referred to this setting as prediction-powered inference, especially in applications with black-box predictors and abundant unlabeled data (Angelopoulos et al., 2023a,b; Zrnic and Candès, 2024). Closely related problems also arise in flexible covariate adjustment for randomized trials, where machine-learning predictors trained on rich data sources can improve precision in small-sample settings (Hansen, 2008; Rubin and van der Laan, 2008; Moore and van der Laan, 2009a; Lin, 2013; Schuler et al., 2022; Demirel et al., 2024).

*Targeted ML Solutions, targetedml.com

A standard approach in this setting is to use the labeled sample to debias a plug-in estimator based on a prediction score. If $m(X)$ is a black-box prediction score, the plug-in estimator averages $m(X)$ over the pooled covariate sample, while the labeled sample is used to estimate the residual correction term $\mathbb{E}[Y - m(X)]$. For example, the augmented inverse probability weighted (AIPW) estimator (Robins et al., 1994) is $\frac{1}{N+n}\{\sum_{i=1}^{n} m(X_i) + \sum_{i=1}^{N} m(\widetilde{X}_i)\} + \frac{1}{n}\sum_{i=1}^{n}\{Y_i - m(X_i)\}$, whereas the PPI estimator (Angelopoulos et al., 2023a), given by $\frac{1}{N}\sum_{i=1}^{N} m(\widetilde{X}_i) + \frac{1}{n}\sum_{i=1}^{n}\{Y_i - m(X_i)\}$, is a generally less efficient variant that does not leverage covariate information from the labeled sample. These estimators reduce bias, but they correct only the mean bias of the score.

Calibration provides a stronger and widely studied notion of predictive reliability (Zadrozny and Elkan, 2001, 2002; Niculescu-Mizil and Caruana, 2005). A calibrated score is one whose numerical values accurately reflect the observed outcomes it predicts. Rather than correcting only the average bias $\mathbb{E}[Y - m(X)]$, one can use the labeled sample to construct a calibrated score $m_n^\star(X)$ such that, ideally, $\mathbb{E}[Y \mid m_n^\star(X)] = m_n^\star(X)$. This stronger property immediately implies $\mathbb{E}[m_n^\star(X)] = \mathbb{E}[Y]$, so the resulting plug-in estimator is automatically calibrated for the mean. This is attractive in practice because a useful prediction model may already be available, while a relatively small labeled sample can often recalibrate its output without retraining the underlying model (Roth, 2022). Such recalibration can be implemented using classical post-processing methods such as Platt scaling (Platt et al., 1999), linear calibration (Mincer and Zarnowitz, 1969), histogram binning (Zadrozny and Elkan, 2001), and isotonic regression (Zadrozny and Elkan, 2002). This is especially relevant for modern machine-learning predictors, such as neural networks, which are often poorly calibrated (Bella et al., 2010; Guo et al., 2017; Wang, 2023).

Our proposal is therefore simple: fit a black-box regression model, calibrate it on the labeled sample, and then average the calibrated predictions over the pooled covariate sample. The resulting estimator remains a "predict, calibrate, and average" procedure, yet our main result shows that it is exactly equal to an AIPW estimator built from the calibrated scores, with calibration eliminating the augmentation term. It is therefore a targeted minimum loss estimator (Van Der Laan and Rubin, 2006; van der Laan and Rose, 2011), in which calibration plays the role of the targeting step, and a special case of calibrated debiased machine learning (van der Laan et al., 2024c). Calibration thus gives this simple plug-in rule a debiased interpretation with strong efficiency guarantees.

Contributions. We make three main contributions.

First, in Section 3, we introduce a general calibration-based framework for semisupervised mean estimation under the joint two-sample model. The framework accommodates classical post-hoc calibration methods, including linear calibration, histogram binning, and isotonic regression; see, in particular, Sections 3.2 and 3.3. Among these, isotonic calibration is especially appealing because it is lightweight, tuning-free, and learns the best monotone transformation of a prediction score.

Second, in Section 3, we show that calibration enforces empirical score equations that yield an exact AIPW representation of the resulting estimator, placing the method directly within the semiparametric framework for debiased estimation. In Section 3.3, we further study linear calibration and relate PPI++ to linear regression adjustment using $m(X)$ as a prognostic score, thereby clarifying their connection to classical covariate-adjustment methods in randomized trials and missing-at-random settings.

Third, in Section 4, we establish asymptotic normality, valid inference, and efficiency results for the isotonic-calibrated plug-in estimator under the joint two-sample model. We show that it is efficient in the reduced model defined by the best monotone transformation of the prediction score and therefore more efficient than AIPW based on the uncalibrated score. When this monotone transformation coincides with the true outcome regression, the estimator is nonparametrically efficient.

Section 2 reviews the two-sample semisupervised model, AIPW, PPI, PPI++, and the relevant efficiency theory, and Section 5 presents synthetic and real-data experiments.

Code. The accompanying `Python` package, reproduction code, and documentation are publicly available at github.com/Larsvanderlaan/ppi-aipw. An interactive package website with examples and API documentation

is available at larsvanderlaan.github.io/ppi-aipw/. Self-contained code is provided in Appendix K.

1.1 Related work

Semiparametric statistics and missing data. This paper builds on foundational ideas from the missing-data and semiparametric literatures. Rubin formalized missing at random as a key identification condition for recovering full-data targets from partially observed outcomes (Rubin, 1976). Semiparametric efficiency theory then characterized efficient influence functions, regular asymptotically linear estimators, and one-step debiased procedures (Levit, 1975; Pfanzagl and Wefelmeyer, 1985; Hasminskii and Ibragimov, 1979; Klaassen, 1987; Bickel et al., 1993). In missing-data and causal-inference problems, these ideas led to influence-function-based estimating equations (Robins et al., 1994, 1995; Robins and Rotnitzky, 1995; van der Laan and Robins, 2003; Tsiatis, 2006) that are often doubly robust (Bang and Robins, 2005). From the perspective of asymptotic efficiency, missing-data models under missingness or coarsening at random are now well understood: the class of regular asymptotically linear estimators can be characterized through their influence functions in these models (Robins et al., 1994, 1995; Robins and Rotnitzky, 1995; van der Laan and Robins, 2003; Tsiatis, 2006).

Debiased machine learning and targeted learning. A large body of work combines the semiparametric framework above with machine-learning-based nuisance estimation, often under the label of debiased machine learning. Flexible learning methods within estimating equations and AIPW-style procedures already appear in van der Laan and Robins (2003). Targeted minimum loss estimation (TMLE), or targeted learning, is a machine-learning-based inference framework in which an initial nuisance estimator is updated in a parameter-targeted way to debias the resulting plug-in estimator; in the present setting, the main nuisance is the outcome regression (Van Der Laan and Rubin, 2006; van der Laan and Rose, 2011; Gruber and Van Der Laan, 2009; Hoffman, 2020; Ross et al., 2025). These approaches are often combined with sample splitting and cross-fitting to accommodate generic machine-learning estimators (Schick, 1986; Klaassen, 1987; Zheng and Van Der Laan, 2010; van der Laan and Rose, 2011; Chernozhukov et al., 2018a). Double machine learning emphasizes similar ingredients—orthogonality, cross-fitting, and influence-function expansions—for inference with modern nuisance estimators (Chernozhukov et al., 2018a, 2022), and can be understood within the same semiparametric framework (van der Laan, 2019; Díaz, 2020; Chen et al., 2026). For sufficiently stable estimators, cross-fitting can itself be relaxed (Chen et al., 2022). Closely related work also studies balancing-weight estimators (Hainmueller, 2012; Imai and Ratkovic, 2014; Zubizarreta, 2015; Lendle et al., 2015; Chattopadhyay et al., 2020; Hejazi and van der Laan, 2023), which often admit equivalent regression-adjustment or augmented representations (Chattopadhyay and Zubizarreta, 2023; Bruns-Smith et al., 2025; Rotnitzky et al., 2025). In completely missing-at-random settings, including semisupervised learning and randomized trials, these approaches permit flexible black-box nuisance estimation while still supporting valid inference under appropriate conditions (Moore and van der Laan, 2009a,b; Rosenblum and van der Laan, 2009; Højbjerre-Frandsen et al., 2025; Højbjerre-Frandsen and Schuler, 2026).

Semisupervised inference. Related work studies semisupervised or surrogate-assisted estimation in settings where the primary outcome is observed only on a limited subset, while auxiliary outcomes or surrogate measurements are available more broadly (Pepe, 1992; Pepe et al., 1994; Chen, 2000; Chen et al., 2008; Cheng et al., 2021; Ji et al., 2025; Kallus and Mao, 2025). Two-sample semisupervised inference is a special case of two-stage sampling designs (Scott and Holt, 1982; Rose and van der Laan, 2011; Hejazi et al., 2021; Qiu et al., 2026) and, more generally, of data fusion, for which a substantial semiparametric efficiency literature has been developed (Li and Luedtke, 2023; Li et al., 2025; Graham et al., 2024; Xu et al., 2025). In this context, more recent papers have used the term *prediction-powered inference* (PPI) for the semisupervised mean-estimation setting in which a black-box predictor is combined with a small labeled sample and a large unlabeled sample (Angelopoulos et al., 2023a,b; Zrnic and Candès, 2024; Xu et al., 2025; Song et al., 2026; Poulet et al., 2025; Lee and Kim, 2026). These papers also helped popularize such tools for a broader machine-learning audience and provided accessible software. Closely related problems have also long been studied in randomized trials, where prognostic scores or other regression adjustments learned from larger historical or auxiliary data sources are used to improve precision in smaller trials (Hansen, 2008; Rosenblum and van der Laan, 2009; Moore and van der Laan, 2009a,b; Moore et al., 2011; Schuler et al., 2022; Balzer et al., 2024; Højbjerre-Frandsen et al., 2025). Methodologically, however, PPI estimators are best understood

as instances or variants of existing semiparametric procedures. For example, PPI++ is an AIPW estimator (Robins et al., 1994; van der Laan and Robins, 2003) obtained by empirical efficiency maximization (Rubin and van der Laan, 2008) over a scaling parameter, whereas cross-PPI (Zrnic and Candès, 2024) is an AIPW estimator with cross-fitting, which accommodates machine-learning nuisance estimators and relaxes empirical process conditions, as in Schick (1986); Zheng and Van Der Laan (2010); van der Laan and Rose (2011); Chernozhukov et al. (2018a). More fundamentally, the ability to use black-box prediction models while retaining valid inference is an immediate consequence of the double robustness of AIPW when the missingness probability is known, and is not specific to the semisupervised mean-estimation setting (Bang and Robins, 2005; Seaman and Vansteelandt, 2018; Rotnitzky et al., 2021).

From missing data to causal inference. From the perspective of estimation and efficiency theory, the semisupervised setting with missing-at-random outcomes is statistically equivalent to randomized experiments, controlled trials, and survey-sampling designs once the missingness indicator is identified with the treatment or sampling indicator, so that the corresponding potential outcomes are treated as missing (Rubin, 1978; van der Laan and Robins, 2003; Ding and Li, 2018; Mozer, 2026). Accordingly, tools for covariate adjustment, causal inference, and model-assisted estimation developed in those settings can be imported directly into the semisupervised setting; see, for example, Mozer (2026) for equivalence results relating prediction-powered inference to model-assisted survey-sampling estimators. Moreover, the two-sample setting considered here is statistically equivalent to a one-sample i.i.d. setting (Li and Luedtke, 2023). In particular, the methods we develop also apply to inference on the counterfactual mean $\mathbb{E}[Y(1)]$ in a randomized controlled trial with binary treatment $A_i \in \{0,1\}$, where $\{(X_i, Y_i)\}_{i=1}^n$ are the observed covariates and outcomes for units assigned to treatment $A_i = 1$, so that $Y_i = Y_i(1)$, and the unlabeled dataset $\{\widetilde{X}_i\}_{i=1}^N$ consists of covariate information for units assigned to control $A_i = 0$ (see Appendix B).

Calibration. A final relevant thread concerns calibration. Classical post-processing methods such as Platt scaling (Platt et al., 1999; Cox, 1958), linear calibration (Mincer and Zarnowitz, 1969; Chernozhukov et al., 2018b; Leng and Dimmery, 2021), isotonic regression (Zadrozny and Elkan, 2002; Niculescu-Mizil and Caruana, 2005), and histogram binning (Zadrozny and Elkan, 2001; Gupta and Ramdas, 2021) improve predictive reliability without retraining the underlying model. The idea of calibration has earlier roots in the forecasting literature (Mincer and Zarnowitz, 1969; Lichtenstein et al., 1977; Vovk, 1992; Vovk et al., 2005; Gneiting et al., 2007; Lambert, 2011); see historical discussion of Lee et al. (2023). More recent work studies calibration under broader losses and prediction tasks (Jung et al., 2021; Noarov and Roth, 2023; van der Laan et al., 2023b; Whitehouse et al., 2024; van der Laan and Alaa, 2025). Calibration has also been used in conformal prediction (van der Laan and Alaa, 2024, 2025), importance-weight stabilization (Gutman et al., 2022; Deshpande and Kuleshov, 2023; Ballinari and Bearth, 2024; van der Laan et al., 2024b), treatment effect estimation (Chernozhukov et al., 2018b; van der Laan et al., 2023b; Whitehouse et al., 2024), and value prediction in offline reinforcement learning (van der Laan and Kallus, 2025). Distribution-free guarantees for histogram binning are developed by Gupta and Ramdas (2021), while van der Laan et al. (2023b); van der Laan and Alaa (2025) establish asymptotic distribution-free guarantees for isotonic calibration. Most closely related for our purposes is recent work connecting calibration to semiparametric inference by formulating calibration through score equations and using it to support debiased estimation of statistical functionals (van der Laan et al., 2024c, 2023a, 2025b). In particular, van der Laan et al. (2024c) studies calibrated DML for doubly robust inference, which in the present setting reduces to calibrating the outcome regression and exploiting the resulting plug-in representation, while van der Laan et al. (2025b) studies calibrated plug-in regression estimators in reinforcement learning. More broadly, van der Laan et al. (2023a) studies plug-in calibration for both regression and treatment effect estimation and shows how calibration can support superefficient inference through dimension reduction.

2 Two-Sample Semisupervised Setup

2.1 Data structure and notation

We observe two independent datasets:

$$\mathcal{D}_L = \{(X_i, Y_i)\}_{i=1}^n \qquad \text{and} \qquad \mathcal{D}_U = \{\widetilde{X}_j\}_{j=1}^N.$$

The labeled sample is i.i.d. from a distribution P_0 on $\mathcal{X} \times \mathbb{R}$, while the unlabeled sample is i.i.d. from the marginal distribution $P_{0,X}$ of X under P_0. We target the population mean

$$\psi_0 := \mathbb{E}_{P_0}[Y], \qquad \mu_0(x) := \mathbb{E}_{P_0}[Y \mid X = x].$$

It is helpful to keep the joint two-sample experiment explicit. For each pair (n, N), the data law is

$$\mathbb{D}_0^{(n,N)} := P_0^{\otimes n} \otimes P_{0,X}^{\otimes N},$$

and the target parameter is the functional

$$\Psi\left(\mathbb{D}_0^{(n,N)}\right) := \psi_0 = \int y \, dP_0(x, y) = \int \mu_0(x) \, dP_{0,X}(x).$$

For measurable functions $a : \mathcal{X} \times \mathbb{R} \to \mathbb{R}$ and $b : \mathcal{X} \to \mathbb{R}$, define the empirical means

$$\mathbb{P}_n^L a := \frac{1}{n} \sum_{i=1}^n a(X_i, Y_i), \qquad \mathbb{P}_N^U b := \frac{1}{N} \sum_{j=1}^N b(\widetilde{X}_j).$$

For clarity, we will sometimes abuse notation and write, for example, $\mathbb{P}_n^L\{a(X, Y)\}$ for $\mathbb{P}_n^L a$ and, when a function depends only on X, write $\mathbb{P}_n^L\{b(X)\} := \frac{1}{n}\sum_{i=1}^n b(X_i)$. Throughout, we write $\rho_n = n/M$ for the labeled fraction, equivalently the known probability of observing the outcome, and use the identity

$$\frac{1}{n+N}\left\{\sum_{i=1}^n f(X_i) + \sum_{j=1}^N f(\widetilde{X}_j)\right\} = \rho_n \mathbb{P}_n^L f + (1 - \rho_n)\mathbb{P}_N^U f.$$

That is, the empirical mean of $f(X)$ over the pooled covariate sample is a convex combination of the empirical means in the labeled and unlabeled samples.

Assumption 1 (Two-sample design)**.** The following hold:

(i) $\mathcal{D}_L$ is i.i.d. from P_0, and $\mathcal{D}_U$ is i.i.d. from $P_{0,X}$.

(ii) $\mathcal{D}_L$ and $\mathcal{D}_U$ are independent.

(iii) Y has finite second moment under P_0.

(iv) With $M := n + N$, the labeled fraction $\rho_n := n/M$ satisfies $\rho_n \to \rho_0$ for some $\rho_0 \in (0, 1)$.

Remark 1. The labeled sample carries the outcome information, while the unlabeled sample sharpens estimation of the marginal covariate distribution. The pooled i.i.d. missing-data formulation in Appendix I is a special case of this two-sample setup.

2.2 Review of AIPW, PPI, and semiparametric efficiency

For later reference, we briefly review semiparametric efficiency in the unrestricted two-sample model (Bickel et al., 1993; Robins and Rotnitzky, 1995; van der Laan and Robins, 2003; Li and Luedtke, 2023), along with

the associated class of influence functions and AIPW estimators (Robins et al., 1994, 1995; Bang and Robins, 2005). We say that an estimator $\widehat{\psi}$ is asymptotically linear with *influence pair* (D^L, D^U) if

$$\sqrt{M}\,(\widehat{\psi} - \psi_0) = \frac{1}{\sqrt{M}} \sum_{i=1}^{n} D^L(X_i, Y_i) + \frac{1}{\sqrt{M}} \sum_{j=1}^{N} D^U(\widetilde{X}_j) + o_p(1),$$

where $\mathbb{E}_{P_0}[D^L(X,Y)] = 0$ and $\mathbb{E}_{P_{0,X}}[D^U(\widetilde{X})] = 0$. In other words, the error $\widehat{\psi} - \psi_0$ is asymptotically equivalent to the empirical mean $\frac{1}{M}\{\sum_{i=1}^{n} D^L(X_i, Y_i) + \sum_{j=1}^{N} D^U(\widetilde{X}_j)\}$. This is the natural two-sample analogue of the usual influence function in an i.i.d. model (Van der Vaart, 2000) and uniquely determines the corresponding influence function in the i.n.i.d. model (Bickel et al., 1993; Li and Luedtke, 2023). By the central limit theorem and the independence of the two datasets,

$$\sqrt{M}\,(\widehat{\psi} - \psi_0) \;\rightsquigarrow\; N\left(0,\, \rho_0 \mathrm{Var}_{P_0}\left(D^L(X,Y)\right) + (1-\rho_0)\mathrm{Var}_{P_{0,X}}\left(D^U(\widetilde{X})\right)\right).$$

Thus, the influence pair (D^L, D^U) determines the asymptotic variance of $\widehat{\psi}$.

The following result characterizes the class of regular estimators[1] (Van der Vaart, 2000) in this model through their influence functions and shows that each is asymptotically equivalent to an AIPW estimator. This provides the benchmark class for our later analysis, where we place calibrated plug-in estimators within the standard AIPW framework and characterize their efficiency. The result follows from Robins et al. (1994) together with the statistical equivalence between the two-sample model and the single-sample i.i.d. model established, for example, by Li and Luedtke (2023).

Proposition 1 (AIPW class of influence pairs)**.** *Under Assumption 1, a mean-zero pair (D^L, D^U) is the influence pair of a regular estimator of ψ_0 in the unrestricted two-sample model if and only if there exists a square-integrable function $f : \mathcal{X} \to \mathbb{R}$ such that, with*

$$\tilde{f}(X) := f(X) - \mathbb{E}_{P_{0,X}}[f(X)] + \psi_0,$$

the pair takes the form

$$D_f^L(X,Y) = \tilde{f}(X) - \psi_0 + \rho_0^{-1}\{Y - \tilde{f}(X)\}, \tag{1}$$

$$D_f^U(\widetilde{X}) = \tilde{f}(\widetilde{X}) - \psi_0. \tag{2}$$

Furthermore, for any such $f : \mathcal{X} \to \mathbb{R}$, a corresponding estimator with this influence pair is given by the AIPW estimator

$$\widehat{\psi}(f) := \rho_n \mathbb{P}_n^L\{f(X)\} + (1-\rho_n)\mathbb{P}_N^U\{f(\widetilde{X})\} + \mathbb{P}_n^L\{Y - f(X)\}. \tag{3}$$

AIPW, PPI, PPI++, and TMLE. Proposition 1 implies that, for any fixed choice of $f : \mathcal{X} \to \mathbb{R}$, the estimator $\widehat{\psi}(f)$ in (3) is unbiased in finite samples and asymptotically linear with the corresponding influence pair. A classical choice is $f(X) = m(X)$, where $m(X)$ denotes the prediction score, which yields the usual AIPW estimator (Robins et al., 1994):

$$\widehat{\psi}_{\mathrm{AIPW}} = \rho_n \mathbb{P}_n^L\{m(X)\} + (1-\rho_n)\mathbb{P}_N^U\{m(\widetilde{X})\} + \mathbb{P}_n^L\{Y - m(X)\}.$$

This estimator can also be written in the combined-sample form

$$\widehat{\psi}_{\mathrm{AIPW}} = \frac{1}{M}\left\{\sum_{i=1}^{n} m(X_i) + \sum_{i=1}^{N} m(\widetilde{X}_i) + \frac{1}{\rho_n}\sum_{i=1}^{n}\{Y_i - m(X_i)\}\right\},$$

where ρ_n^{-1} is the known inverse probability of observing the outcome. Because outcome missingness is completely at random with known constant probability ρ_n, the AIPW estimator remains valid even if m is

[1] Regularity means that the estimator's limiting distribution is invariant under sampling from local alternatives; it excludes pathological estimators, such as Hodges' estimator, that may be superefficient at specific laws but perform poorly in a local asymptotic sense (Le Cam, 1953; Van der Vaart, 2000).

misspecified; the prediction score affects only efficiency. Thus, AIPW is "safe" in this setting (Xu et al., 2025). This is a special case of double robustness[2] (Bang and Robins, 2005).

An alternative choice takes $f(X) = m(X)/(1-\rho_n)$, which yields

$$\widehat{\psi}_{\mathrm{PPI}} = \mathbb{P}_N^U\{m(\widetilde{X})\} + \mathbb{P}_n^L\{Y - m(X)\}.$$

This choice is used implicitly in Angelopoulos et al. (2023a), where it is termed PPI[3]. Relative to $\widehat{\psi}_{\mathrm{AIPW}}$, $\widehat{\psi}_{\mathrm{PPI}}$ omits the labeled-sample average of $m(X)$ from the plug-in term. It is therefore typically less efficient, especially when m is reasonably accurate; in particular, it is inefficient when $m = \mu_0$. When the labeled fraction ρ_n is small and m is inaccurate, however, the AIPW estimator need not be more efficient than the PPI estimator (Han, 2012; Rotnitzky et al., 2012), particularly when $\mathrm{Cov}(m(X), Y - m(X))$ is sufficiently positive, so that the correlation between the plug-in term and the residual-correction term increases variance rather than reducing it.

Because both $\widehat{\psi}_{\mathrm{AIPW}}$ and $\widehat{\psi}_{\mathrm{PPI}}$ are special cases of the scaling class $\{\lambda m(X) : \lambda \in \mathbb{R}\}$, one can construct an estimator that is asymptotically at least as efficient as both by minimizing the empirical influence-function variance over this class:

$$\sigma_n^2(f) := \rho_n \mathbb{P}_n^L\left[\left\{f(X) - \widehat{\psi}(f) + \rho_n^{-1}(Y - f(X))\right\}^2\right] + (1-\rho_n)\mathbb{P}_N^U\left[\left\{f(\widetilde{X}) - \widehat{\psi}(f)\right\}^2\right].$$

Equivalently, one chooses λ by empirical efficiency maximization over $\{\lambda m(X) : \lambda \in \mathbb{R}\}$ (Rubin and van der Laan, 2008; van der Laan and Robins, 2003); related ideas were used by Rotnitzky et al. (2012) to construct doubly robust estimators with improved efficiency under misspecification. The resulting estimator, called PPI++ by Angelopoulos et al. (2023b), is simply AIPW with an empirically rescaled prediction score $\hat{\lambda} m(X)$, where $\hat{\lambda}$ estimates the population-optimal scaling coefficient

$$\lambda^\star = \frac{\mathrm{Cov}\big(Y, m(X)\big)}{\mathrm{Var}\big(m(X)\big)}. \tag{4}$$

Efficiency can be improved further by minimizing the empirical efficiency criterion $\sigma_n^2(f)$ over richer classes, such as

$$\{\lambda m(X) + \beta_{\mathrm{lin}}^\top X : \lambda \in \mathbb{R},\ \beta_{\mathrm{lin}} \in \mathbb{R}^d\},$$

over monotone transformations of the score, or over a reproducing kernel Hilbert space (Appendix A). More generally, the empirical efficiency criterion can be used in a cross-validation procedure to select among different estimator classes (Rubin and van der Laan, 2008).

The semiparametrically efficient benchmark is attained by the unrestricted choice $f(X) = \mu_0(X)$, which must typically be estimated from the data. From the balancing perspective, this shows that efficiency generally requires using the full covariate information X, or at least a summary rich enough to recover the true regression function $\mu_0(X)$ (Robins et al., 1994). The next result identifies the efficient influence pair, and thus the most efficient regular estimator, in this model (Bickel et al., 1993). Accordingly, efficiency is optimized by using a prediction model $m(X)$ that approximates $\mu_0(X)$ as closely as possible (Moore and van der Laan, 2009a,b). Modern debiased machine learning methods implement this principle using flexible black-box learners to estimate nuisance functions (van der Laan and Robins, 2003; Moore and van der Laan, 2009a; van der Laan and Rose, 2011; Chernozhukov et al., 2018a).

Proposition 2 (Efficient influence pair in the full model)**.** *Under Assumption 1, the asymptotic variance over the class equations (1) and (2) is minimized at $f = \mu_0$. The efficient influence pair for $\psi_0 = \mathbb{E}_{P_0}[Y]$ in the unrestricted two-sample model is therefore*

$$D_{\mathrm{eff}}^L(X, Y) = \mu_0(X) - \psi_0 + \rho_0^{-1}\{Y - \mu_0(X)\}, \tag{5}$$

$$D_{\mathrm{eff}}^U(\widetilde{X}) = \mu_0(\widetilde{X}) - \psi_0. \tag{6}$$

[2]In more general missing-at-random settings with unknown, covariate-dependent missingness probabilities, the analogous AIPW estimator remains doubly robust asymptotically, but no longer enjoys the same finite-sample unbiasedness as in the present known-ρ_n setting (Seaman and Vansteelandt, 2018).

[3]The PPI and PPI++ estimators are algebraically equivalent to the model-assisted survey-design estimators of Cassel et al. (1976) and Särndal et al. (2003), respectively; see Mozer (2026) for a review.

To conclude, we illustrate how regression adjustment, balancing-weight adjustment, and AIPW estimation are closely connected (Rotnitzky et al., 2025). We also introduce TMLE (Van Der Laan and Rubin, 2006), which provides an alternative framework for debiased plug-in estimation without explicit augmentation.

Remark 2 (Plug-in approaches to bias correction and TMLE)**. Regression plug-in estimators.** A particularly simple realization of this first-order class is obtained through plug-in regression adjustment (Scharfstein et al., 1999; Van Der Laan and Rubin, 2006; Rosenblum and van der Laan, 2009; Rotnitzky et al., 2025). Let

$$m_n(X) = \hat{a}_{\text{lin}}\, m(X) + \hat{b}_{\text{lin}}, \qquad (\hat{a}_{\text{lin}}, \hat{b}_{\text{lin}}) = \arg\min_{a,b} \sum_{i=1}^{n} \{Y_i - a\, m(X_i) - b\}^2.$$

The corresponding plug-in estimator is

$$\rho_n \mathbb{P}_n^L\{m_n(X)\} + (1-\rho_n)\mathbb{P}_N^U\{m_n(\widetilde{X})\}.$$

For any $\lambda \in \mathbb{R}$, this estimator also admits the AIPW representation

$$\rho_n \mathbb{P}_n^L\{m_n(X)\} + (1-\rho_n)\mathbb{P}_N^U\{m_n(\widetilde{X})\} + \lambda \mathbb{P}_n^L\big[Y - m_n(X)\big],$$

because the augmentation term $\mathbb{P}_n^L[Y - m_n(X)]$ vanishes by the least-squares normal equations. Thus, when m_n converges to the best linear predictor of Y given $m(X)$, this plug-in estimator is typically at least as efficient as $\widehat{\psi}_{\text{AIPW}}$ and $\widehat{\psi}_{\text{PPI}}$, since it corresponds to an influence function with smaller or equal variance.

Balancing-weight estimators. Dually, the same estimator can be expressed through balancing weights. Let $w_n(X) = \hat{a}_{\text{bal}} + \hat{b}_{\text{bal}}\, m(X)$ be affine weights that balance $(1, m(X))$, in the sense that

$$\mathbb{P}_n^L\{w_n(X)\} = 1 \quad \text{and} \quad \mathbb{P}_n^L\big[m(X) w_n(X)\big] = \rho_n \mathbb{P}_n^L\{m(X)\} + (1-\rho_n)\mathbb{P}_N^U\{m(\widetilde{X})\}.$$

Then the corresponding balancing-weight estimator satisfies

$$\mathbb{P}_n^L\big\{w_n(X)Y\big\} = \rho_n \mathbb{P}_n^L\{m(X)\} + (1-\rho_n)\mathbb{P}_N^U\{m(\widetilde{X})\} + \mathbb{P}_n^L\big[w_n(X)\{Y - m(X)\}\big].$$

In fact, the plug-in and balancing-weight estimators are algebraically equivalent (Chattopadhyay and Zubizarreta, 2023; Bruns-Smith et al., 2025):

$$\mathbb{P}_n^L\big\{w_n(X)Y\big\} = \rho_n \mathbb{P}_n^L\{m_n(X)\} + (1-\rho_n)\mathbb{P}_N^U\{m_n(\widetilde{X})\}.$$

TMLE. More broadly, such debiased plug-in estimators can be viewed as instances of TMLE (Van Der Laan and Rubin, 2006; van der Laan and Rose, 2011; Højbjerre-Frandsen and Schuler, 2026), a general framework for debiased and efficient plug-in estimation that targets nuisance components to eliminate first-order bias. In the present setting, either the weights or the regression model are updated so that the corresponding augmentation term vanishes. In weighting-based TMLE, the update balances on the prediction model $m(X)$ (e.g., Hejazi and van der Laan 2023); one may either include the known weight $1/\rho_n$ as an offset or balance on the intercept as well. In regression-based TMLE, the prediction model $m(X)$ may be included either as a covariate or as an offset (e.g., Lendle et al. 2015), and the intercept adjustment corrects for the known inverse missingness-probability weight $1/\rho_n$. The targeting step is ultimately a practical choice, provided it solves the relevant estimating equation and removes the augmentation term; different choices may also yield stronger properties, such as additional bias reduction or doubly robust inference (Van der Laan, 2014; Carone et al., 2014; Díaz and van der Laan, 2017; Benkeser et al., 2017; van der Laan et al., 2021).

3 Calibration-Based Plug-In Estimation

The first subsection introduces the general calibrated plug-in estimator and shows that it admits an equivalent AIPW representation. The second subsection then describes concrete ways to enforce empirical calibration, focusing on isotonic regression. We next discuss linear calibration as a simpler adjustment step when data are especially scarce. The final subsection presents a cross-fitted version of the same construction, which allows the predictive score to be learned from the labeled data.

3.1 General estimator and exact AIPW representation

Let $m : \mathcal{X} \to \mathbb{R}$ denote a black-box regression fit on the labeled data. A calibration procedure takes the labeled scores $m(X_1), \ldots, m(X_n)$ and outcomes $Y_1, \ldots, Y_n$, fits a post-hoc map f_n, and outputs the calibrated regression

$$m_n^\star := f_n \circ m.$$

Thus, the fitted values on the two samples are $m_n^\star(X_i)$ for $i = 1, \ldots, n$ and $m_n^\star(\widetilde{X}_j)$ for $j = 1, \ldots, N$.

Fix a linear class $\mathcal{F}$ of real-valued functions $f : \mathbb{R} \to \mathbb{R}$ that contains the constant functions $t \mapsto c$ and the identity map $t \mapsto t$. Informally, calibration is a reliability property stating that the prediction model $m_n^\star$ cannot be improved by post-processing its output using transformations in $\mathcal{F}$. We say that $m_n^\star$ is *empirically calibrated over* $\mathcal{F}$ if no admissible transformation of its output improves the empirical squared risk, that is, if for every $f \in \mathcal{F}$,

$$\mathbb{P}_n^L\big[(Y - m_n^\star(X))^2\big] \le \mathbb{P}_n^L\big[(Y - f(m_n^\star(X)))^2\big], \tag{7}$$

where $\mathbb{P}_n^L$ denotes the empirical mean over the labeled sample. Equivalently, the corresponding first-order optimality conditions imply

$$\mathbb{P}_n^L\big[f(m_n^\star(X))\{Y - m_n^\star(X)\}\big] = 0, \qquad f \in \mathcal{F}. \tag{8}$$

Because $\mathcal{F}$ contains the constant functions and the identity map, empirical calibration implies, in particular, the mean-calibration and linear-calibration identities

$$\mathbb{P}_n^L\{Y - m_n^\star(X)\} = 0, \qquad \mathbb{P}_n^L\big[m_n^\star(X)\{Y - m_n^\star(X)\}\big] = 0.$$

The empirically calibrated plug-in estimator is

$$\widehat{\psi}_{\mathrm{cal}} := \frac{1}{n+N}\left\{\sum_{i=1}^{n} m_n^\star(X_i) + \sum_{j=1}^{N} m_n^\star(\widetilde{X}_j)\right\}. \tag{9}$$

Equivalently,

$$\widehat{\psi}_{\mathrm{cal}} = \rho_n \mathbb{P}_n^L\{m_n^\star(X)\} + (1-\rho_n)\mathbb{P}_N^U\{m_n^\star(\widetilde{X})\}.$$

We focus on two cases. In the nonparametric case, $\mathcal{F}$ is the class of all transformations $f : \mathbb{R} \to \mathbb{R}$, which we refer to as *full calibration*; in practice, this can be implemented by regressing Y on $m(X)$ using histogram regression or isotonic regression (Section 3.2). In the parametric case, $\mathcal{F}$ is restricted to affine transformations $t \mapsto at + b$, which we refer to as *linear calibration*; this amounts to fitting the working linear model $a\, m(X) + b$ for Y (Section 3.3). More generally, if $\mathcal{F}$ is finite-dimensional with basis $\{\varphi_1, \ldots, \varphi_p\}$, then an empirically calibrated predictor m_n^* can be constructed by regressing Y on $\{\varphi_1(m(X)), \ldots, \varphi_p(m(X))\}$.[4]

The key property of calibrated plug-in estimators is the following.

Theorem 1 (Exact AIPW representation under empirical calibration)**.** *Suppose the calibrated predictions satisfy equation (8), where $\mathcal{F}$ is a linear class containing the constant functions and the identity map. Then, for any empirical balancing weight $\widehat{w} \in \mathcal{F}$, the calibrated plug-in estimator admits the exact AIPW representation*

$$\widehat{\psi}_{\mathrm{cal}} = \rho_n \mathbb{P}_n^L\{m_n^\star(X)\} + (1-\rho_n)\mathbb{P}_N^U\{m_n^\star(\widetilde{X})\} + \rho_n \mathbb{P}_n^L\big[\widehat{w}(m_n^\star(X))\{Y - m_n^\star(X)\}\big]. \tag{10}$$

Moreover, if $\widehat{w}_n^* \in \mathcal{F}$ balances $\mathcal{F}$ in the sense that

$$\mathbb{P}_n^L\big[\widehat{w}_n^*(m_n^\star(X))\, f(m_n^\star(X))\big] = \rho_n \mathbb{P}_n^L\{f(m_n^\star(X))\} + (1-\rho_n)\mathbb{P}_N^U\{f(m_n^\star(\widetilde{X}))\}, \qquad f \in \mathcal{F}, \tag{11}$$

then $\widehat{\psi}_{\mathrm{cal}}$ admits the weighted-outcome representation

$$\widehat{\psi}_{\mathrm{cal}} = \mathbb{P}_n^L\big[\widehat{w}_n^*(m_n^\star(X))Y\big]. \tag{12}$$

[4] If $\mathcal{F}$ does not contain the identity map, one can instead include $m(X)$ as an offset, and (8) still holds.

Theorem 1 shows that the calibrated plug-in estimator $\widehat{\psi}_{\mathrm{cal}}$ is exactly equal to the AIPW estimator with $m_n^\star$ as the prediction score. It also yields a dual representation of $\widehat{\psi}_{\mathrm{cal}}$ as a balancing-weighted outcome estimator, where $\widehat{w}_n^*$ is chosen to balance moments of functions in $\mathcal{F}$ between the labeled and unlabeled samples (Chattopadhyay et al., 2020). More broadly, $\widehat{\psi}_{\mathrm{cal}}$ may also be viewed as a TMLE, with calibration playing the role of the targeting step.

Sketch of asymptotic theory. We now sketch the asymptotic theory for $\widehat{\psi}_{\mathrm{cal}}$. Suppose the estimated calibrated score $m_n^\star$ converges to a limit m_0, typically a transformation of m that minimizes a population risk over a suitable class. This limit is typically calibrated at the population level in the sense that

$$\mathbb{E}_{P_0}\big[f\{m_0(X)\}\{\mu_0(X) - m_0(X)\}\big] = 0, \qquad f \in \mathcal{F}.$$

If $\mathcal{F}$ contains all real-valued functions, then

$$m_0(X) = \mathbb{E}_{P_0}\left[Y \mid m_0(X)\right],$$

so the predictive performance of m_0 cannot be improved by any transformation of its output. Taking $\hat{w} = 1$ in (10), $\widehat{\psi}_{\mathrm{cal}}$ can be viewed as a plug-in version of the usual AIPW estimator with $m_n^\star$ as the regression adjustment. Therefore, if $m_n^\star$ is sufficiently close to m_0, then $\widehat{\psi}_{\mathrm{cal}}$ should behave asymptotically like the corresponding oracle AIPW estimator:

$$\begin{aligned}\widehat{\psi}_{\mathrm{cal}} - \psi_0 = \rho_n(\mathbb{P}_n^L - P_0)\Big\{m_0(X) - \psi_0 + \rho_n^{-1}\big(Y - m_0(X)\big)\Big\} \\ + (1-\rho_n)(\mathbb{P}_N^U - P_{0,X})\big\{m_0(\widetilde{X}) - \psi_0\big\} + o_p(M^{-1/2}).\end{aligned}$$

In particular, $\widehat{\psi}_{\mathrm{cal}}$ is asymptotically linear with influence pair

$$D_{m_0}^L(X,Y) = m_0(X) - \psi_0 + \rho_0^{-1}\{Y - m_0(X)\}, \qquad D_{m_0}^U(\widetilde{X}) = m_0(\widetilde{X}) - \psi_0.$$

In Section 4, we establish this result for the isotonic-calibrated plug-in estimator introduced next.

3.2 Full calibration via isotonic regression

In practice, full empirical calibration can be achieved by fitting a second-stage calibrator to the labeled pairs $\{(m(X_i), Y_i)\}_{i=1}^n$ (van der Laan and Alaa, 2025). Two especially simple options are histogram binning and isotonic calibration (Zadrozny and Elkan, 2001, 2002; Niculescu-Mizil and Caruana, 2005), both of which yield exact empirical calibration. Histogram binning fits a piecewise-constant map on a prespecified partition of the score (Gupta and Ramdas, 2021), whereas isotonic calibration estimates the best monotone transformation of the score and adaptively determines the effective bins using the pooled-adjacent-violators algorithm (Barlow and Brunk, 1972). We focus on isotonic regression because it is lightweight, tuning-free, and easy to implement using widely available software. Moreover, because the identity map is monotone, isotonic calibration can only improve the empirical mean squared error over $\mathcal{D}_L$ relative to the original scores.

The isotonic-calibrated plug-in estimator is

$$\widehat{\psi}_{\mathrm{iso}} := \frac{1}{n+N}\left\{\sum_{i=1}^n m_{n,\mathrm{iso}}^\star(X_i) + \sum_{j=1}^N m_{n,\mathrm{iso}}^\star(\widetilde{X}_j)\right\},$$

where

$$m_{n,\mathrm{iso}}^\star(X) = \hat{f}\{m(X)\}, \qquad \hat{f} \in \arg\min_{f \in \mathcal{F}_{\mathrm{iso}}} \sum_{i=1}^n \big\{Y_i - f(m(X_i))\big\}^2,$$

and $\mathcal{F}_{\mathrm{iso}}$ denotes the class of nondecreasing real-valued functions. The fitted values $m_{n,\mathrm{iso}}^\star(X_i) = \hat{f}(m(X_i))$ are blockwise empirical means determined by the pooled-adjacent-violators algorithm. This blockwise structure implies full empirical calibration, even though the calibration step is restricted to isotonic functions.

Algorithm 1 Isotonic-calibrated plug-in estimator

Require: $\mathcal{D}_L = \{(X_i, Y_i)\}_{i=1}^n$, $\mathcal{D}_U = \{\widetilde{X}_j\}_{j=1}^N$, score m, level $1-\alpha$

1: Compute $T_i = m(X_i)$ and $\widetilde{T}_j = m(\widetilde{X}_j)$

2: Fit $\hat{f} \in \arg\min_{f \in \mathcal{F}_{\text{iso}}} \sum_{i=1}^n \{Y_i - f(T_i)\}^2$

3: Set $m^\star_{n,\text{iso}}(X_i) = \hat{f}(T_i)$ and $m^\star_{n,\text{iso}}(\widetilde{X}_j) = \hat{f}(\widetilde{T}_j)$

4: Form $\widehat{\psi}_{\text{iso}} \leftarrow (n+N)^{-1}\{\sum_{i=1}^n m^\star_{n,\text{iso}}(X_i) + \sum_{j=1}^N m^\star_{n,\text{iso}}(\widetilde{X}_j)\}$

5: Set $\widehat{D}_i^L = m^\star_{n,\text{iso}}(X_i) - \widehat{\psi}_{\text{iso}} + \rho_n^{-1}\{Y_i - m^\star_{n,\text{iso}}(X_i)\}$ and $\widehat{D}_j^U = m^\star_{n,\text{iso}}(\widetilde{X}_j) - \widehat{\psi}_{\text{iso}}$

6: Set $\widehat{\text{SE}}^2_{\text{iso}} = (n+N)^{-2}\{\sum_{i=1}^n (\widehat{D}_i^L)^2 + \sum_{j=1}^N (\widehat{D}_j^U)^2\}$

7: Return $\widehat{\psi}_{\text{iso}} \pm z_{1-\alpha/2}\widehat{\text{SE}}_{\text{iso}}$

Proposition 3 (Isotonic calibration is fully calibrated)**.** *For every function $h : \mathbb{R} \to \mathbb{R}$,*

$$\mathbb{P}_n^L\big[h(m^\star_{n,\text{iso}}(X))\{Y - m^\star_{n,\text{iso}}(X)\}\big] = 0. \tag{13}$$

Moreover, it holds that

$$\mathbb{P}_n^L\big[(Y - m^\star_{n,\text{iso}}(X))^2\big] \le \mathbb{P}_n^L\big[(Y - m(X))^2\big].$$

Consequently, the AIPW representation of Theorem 1 holds with $\mathcal{F}$ equal to the class of all real-valued functions. In the context of treatment effect calibration, Proposition 3 was derived by van der Laan et al. (2023b).

Remark 3 (Implementation)**.** Isotonic regression is widely available in standard software. A standard algorithm is the pooled-adjacent-violators algorithm (PAVA) (Barlow and Brunk, 1972). Although it is generally fast, it may become slow at ultra-large sample sizes. A scalable alternative is to use regression tree software with monotonicity constraints, such as XGBoost or LightGBM (see Appendix K). In one dimension, isotonic regression is equivalent to fitting a regression tree of infinite depth under a monotonicity constraint. An advantage of tree-based software is that the fit can be regularized by constraining or tuning the number of leaves, the maximum tree depth, or the minimum number of observations per leaf.

3.3 Linear calibration via linear regression

For small labeled samples, one may prefer smoother parametric calibration rules to histogram binning or isotonic regression. Here we focus on linear calibration, which enforces that the prediction score is linearly uncorrelated with its residuals and is implemented by regressing the outcome on the prediction score (Mincer and Zarnowitz, 1969). Relative to isotonic regression, this imposes a more restrictive parametric form, akin in spirit to Platt scaling and temperature scaling (Platt et al., 1999; Guo et al., 2017), but it may be more stable when the labeled sample is small. Let

$$(\hat{a}_{\text{lin}}, \hat{b}_{\text{lin}}) \in \arg\min_{a,b \in \mathbb{R}} \sum_{i=1}^n \{Y_i - a\, m(X_i) - b\}^2, \qquad m^\star_{n,\text{lin}}(X) := \hat{a}_{\text{lin}}\, m(X) + \hat{b}_{\text{lin}}. \tag{14}$$

Define the associated linearly calibrated plug-in estimator[5] by

$$\widehat{\psi}_{\text{lin}} := \frac{1}{n+N}\left\{\sum_{i=1}^n m^\star_{n,\text{lin}}(X_i) + \sum_{j=1}^N m^\star_{n,\text{lin}}(\widetilde{X}_j)\right\}. \tag{15}$$

More generally, one can include additional covariates in the linear regression adjustment to further improve precision, provided there is sufficient labeled data (Algorithm 3). Because the affine class includes the constant function, the least-squares fit automatically enforces empirical mean calibration. As a result, the linearly calibrated plug-in estimator admits an exact AIPW representation.

[5]We already introduced this estimator in Remark 2. It can equivalently be implemented by balancing on $(1, m(X))$.

Algorithm 2 Linearly calibrated plug-in estimator

Require: $\mathcal{D}_L = \{(X_i, Y_i)\}_{i=1}^n$, $\mathcal{D}_U = \{\widetilde{X}_j\}_{j=1}^N$, score m, level $1-\alpha$

1: Compute $T_i = m(X_i)$ and $\widetilde{T}_j = m(\widetilde{X}_j)$

2: Fit $(\hat{a}, \hat{b}) \in \arg\min_{a,b} \sum_{i=1}^n \{Y_i - (aT_i + b)\}^2$

3: Set $m^\star_{n,\mathrm{lin}}(X_i) = \hat{a}T_i + \hat{b}$ and $m^\star_{n,\mathrm{lin}}(\widetilde{X}_j) = \hat{a}\widetilde{T}_j + \hat{b}$

4: Form $\widehat{\psi}_{\mathrm{lin}} \leftarrow (n+N)^{-1}\{\sum_{i=1}^n m^\star_{n,\mathrm{lin}}(X_i) + \sum_{j=1}^N m^\star_{n,\mathrm{lin}}(\widetilde{X}_j)\}$

5: Set $\widehat{D}_i^L = m^\star_{n,\mathrm{lin}}(X_i) - \widehat{\psi}_{\mathrm{lin}} + \rho_n^{-1}\{Y_i - m^\star_{n,\mathrm{lin}}(X_i)\}$ and $\widehat{D}_j^U = m^\star_{n,\mathrm{lin}}(\widetilde{X}_j) - \widehat{\psi}_{\mathrm{lin}}$

6: Set $\widehat{\mathrm{SE}}^2_{\mathrm{lin}} = (n+N)^{-2}\{\sum_{i=1}^n (\widehat{D}_i^L)^2 + \sum_{j=1}^N (\widehat{D}_j^U)^2\}$

7: Return $\widehat{\psi}_{\mathrm{lin}} \pm z_{1-\alpha/2}\widehat{\mathrm{SE}}_{\mathrm{lin}}$

Proposition 4 (Linear calibration as AIPW)**.** *The normal equations for equation (14) imply*

$$\mathbb{P}_n^L\{Y - m^\star_{n,\mathrm{lin}}(X)\} = 0 \qquad \textit{and} \qquad \mathbb{P}_n^L\big[m^\star_{n,\mathrm{lin}}(X)\{Y - m^\star_{n,\mathrm{lin}}(X)\}\big] = 0.$$

Hence, for any $\hat{w}(X) = a + b\, m^\star_{n,\mathrm{lin}}(X)$,

$$\widehat{\psi}_{\mathrm{lin}} = \rho_n \mathbb{P}_n^L\{m^\star_{n,\mathrm{lin}}(X)\} + (1-\rho_n)\mathbb{P}_N^U\{m^\star_{n,\mathrm{lin}}(\widetilde{X})\} + \mathbb{P}_n^L\big[\hat{w}(X)\{Y - m^\star_{n,\mathrm{lin}}(X)\}\big]. \tag{16}$$

The following proposition shows that, in the mean-estimation problem, PPI++ and linear calibration are first-order equivalent, although they are not generally identical in finite samples. In particular, linear calibration asymptotically attains the same efficiency-maximizing choice over the scaling class $\{\lambda m(X) : \lambda \in \mathbb{R}\}$ as PPI++. Let $\widehat{\psi}_{++}(\widehat{\lambda}_{++})$ denote the PPI++ estimator of Angelopoulos et al. (2023b), that is, the AIPW estimator based on the rescaled score $\widehat{\lambda}_{++} m(X)$, where $\widehat{\lambda}_{++}$ is obtained by empirical efficiency maximization over $\{\lambda m(X) : \lambda \in \mathbb{R}\}$ (Rubin and van der Laan, 2008). The key point is that $\widehat{\lambda}_{++}$ estimates λ_0 in (4), which is precisely the population slope in the linear regression of Y on $m(X)$.

Proposition 5 (First-order equivalence of PPI++ and linear calibration)**.** *Assume that* $\mathrm{Var}\{m(X)\} > 0$, $\mathbb{E}[Y^2] < \infty$, *and* $\mathbb{E}[m(X)^2] < \infty$. *Then*

$$\widehat{\psi}_{++}(\hat{\lambda}_{++}) - \widehat{\psi}_{\mathrm{lin}} = (1-\rho_n)(\hat{\lambda}_{++} - \hat{a}_{\mathrm{lin}})\left\{\mathbb{P}_N^U m(\widetilde{X}) - \mathbb{P}_n^L m(X)\right\}.$$

Moreover, $\hat{\lambda}_{++} - \hat{a}_{\mathrm{lin}} = O_p(n^{-1/2})$, *and, under* $\rho_n \to \rho_0 \in (0,1)$,

$$\widehat{\psi}_{++}(\hat{\lambda}_{++}) - \widehat{\psi}_{\mathrm{lin}} = O_p\left(\frac{1-\rho_n}{\sqrt{\rho_n}}\,(n+N)^{-1}\right) = O_p\big((n+N)^{-1}\big).$$

In particular,

$$\widehat{\psi}_{++}(\hat{\lambda}_{++}) - \widehat{\psi}_{\mathrm{lin}} = o_p\big(M^{-1/2}\big).$$

Remark 4 (Relation to prognostic-score adjustment)**.** Linear calibration can be viewed as prognostic-score regression adjustment, a classical approach to improving precision through covariate adjustment in small-sample settings (Hansen, 2008; Rosenblum and van der Laan, 2009; Lin, 2013; Schuler et al., 2022; Balzer et al., 2024; Højbjerre-Frandsen et al., 2025). It can also be viewed as a special case of TMLE (Højbjerre-Frandsen and Schuler, 2026). Proposition 5 shows that PPI++ has the same first-order interpretation.

Remark 5 (Platt scaling adjustment)**.** Linear calibration can also be achieved using losses other than squared error, such as logistic loss, Poisson loss, or, more generally, losses from canonical-link generalized linear models (Rosenblum and van der Laan, 2009). In particular, when $Y \in \{0,1\}$ and $m(X) \in (0,1)$ is an initial predictor, one may apply Platt scaling (Cox, 1958; Platt et al., 1999) by fitting a logistic regression of Y on the logit of $m(X)$:

$$(\hat{a}_{\mathrm{platt}}, \hat{b}_{\mathrm{platt}}) \in \arg\min_{a,b\in\mathbb{R}} \sum_{i=1}^n \Big[-Y_i \log\sigma\big(a\,\mathrm{logit}(m(X_i)) + b\big) - (1-Y_i)\log\Big(1 - \sigma\big(a\,\mathrm{logit}(m(X_i)) + b\big)\Big)\Big], \tag{17}$$

Algorithm 3 Linearly calibrated plug-in estimator with covariate adjustment

Require: $\mathcal{D}_L = \{(X_i, Y_i)\}_{i=1}^n$, $\mathcal{D}_U = \{\widetilde{X}_j\}_{j=1}^N$, score m, level $1-\alpha$

1: Compute $T_i = m(X_i)$ and $\widetilde{T}_j = m(\widetilde{X}_j)$

2: Fit $(\hat{a}, \hat{\beta}, \hat{b}) \in \arg\min_{a,\beta,b} \sum_{i=1}^n \{Y_i - (a + X_i^\top \beta + bT_i)\}^2$

3: Set $m^\star_{n,\mathrm{lin}}(X_i) = \hat{a} + X_i^\top \hat{\beta} + \hat{b}T_i$ and $m^\star_{n,\mathrm{lin}}(\widetilde{X}_j) = \hat{a} + \widetilde{X}_j^\top \hat{\beta} + \hat{b}\widetilde{T}_j$

4: Form $\widehat{\psi}_{\mathrm{lin}} \leftarrow (n+N)^{-1}\{\sum_{i=1}^n m^\star_{n,\mathrm{lin}}(X_i) + \sum_{j=1}^N m^\star_{n,\mathrm{lin}}(\widetilde{X}_j)\}$

5: Set $\widehat{D}_i^L = m^\star_{n,\mathrm{lin}}(X_i) - \widehat{\psi}_{\mathrm{lin}} + \rho_n^{-1}\{Y_i - m^\star_{n,\mathrm{lin}}(X_i)\}$ and $\widehat{D}_j^U = m^\star_{n,\mathrm{lin}}(\widetilde{X}_j) - \widehat{\psi}_{\mathrm{lin}}$

6: Set $\widehat{\mathrm{SE}}^2_{\mathrm{lin}} = (n+N)^{-2}\{\sum_{i=1}^n (\widehat{D}_i^L)^2 + \sum_{j=1}^N (\widehat{D}_j^U)^2\}$

7: Return $\widehat{\psi}_{\mathrm{lin}} \pm z_{1-\alpha/2}\widehat{\mathrm{SE}}_{\mathrm{lin}}$

with calibrated predictor

$$m^\star_{n,\mathrm{platt}}(X) := \sigma\big(\hat{a}_{\mathrm{platt}} \operatorname{logit}(m(X)) + \hat{b}_{\mathrm{platt}}\big),$$

where $\sigma(t) := (1+e^{-t})^{-1}$. Platt scaling may be preferred over linear regression calibration when predicted probabilities exhibit over- or under-confidence, in which case the calibration curve is often approximately sigmoidal. More generally, outcomes can be rescaled to $[0,1]$ before applying this procedure.

3.4 Cross-Fitting and Cross-Calibration

In the main exposition, we assume for simplicity that the initial black-box predictor is fixed. In practice, however, it may be trained on the labeled sample itself, in which case the same labeled observations are reused both to fit the predictor and to calibrate its scores. A cross-fitted version avoids this reuse by replacing the single prediction function with out-of-fold predictions and then fitting the calibrator on the resulting labeled out-of-fold scores.

1. Partition the labeled sample indices into folds $I_1, \ldots, I_K$.
2. For each fold k, train the black-box regression $m^{(-k)}$ on the labeled observations outside I_k, compute out-of-fold predictions $m^{(-k)}(X_i)$ for $i \in I_k$, and compute predictions $m^{(-k)}(\widetilde{X}_j)$ on the unlabeled sample.
3. Pool the labeled out-of-fold predictions $\{m^{(-k(i))}(X_i)\}_{i=1}^n$ and fit a calibrator f_n to the labeled pairs $(m^{(-k(i))}(X_i), Y_i)$.
4. Form calibrated predictions
$$m^\star_n(X_i) = f_n(m^{(-k(i))}(X_i))$$
and
$$m^\star_n(\widetilde{X}_j) = \frac{1}{K}\sum_{k=1}^K f_n(m^{(-k)}(\widetilde{X}_j)),$$
and then compute the plug-in estimator equation (9).

This cross-fitted modification is the natural DML-style implementation when one wishes to reduce overfitting from estimating the initial regression on the same labeled data later used for calibration and inference (Zheng and Van Der Laan, 2010; van der Laan and Rose, 2011; Chernozhukov et al., 2018a). This procedure has been used for debiased inference in van der Laan et al. (2024b,c); Rabenseifner et al. (2025), and a related cross-calibration variant appears in van der Laan et al. (2023b).

4 Theory for isotonic calibration

We study the isotonic-calibrated plug-in estimator $\widehat{\psi}_{\mathrm{iso}}$. We first establish its asymptotic linearity, asymptotic normality, and validity of Wald inference. We then study its efficiency, showing that $\widehat{\psi}_{\mathrm{iso}}$ is efficient in a reduced data model and giving conditions under which it is also efficient in the full model, that is, attains the minimum asymptotic variance among regular estimators (Bickel et al., 1993; Van der Vaart, 2000). To keep the exposition light, we focus on the non-cross-fitted estimator and treat the initial black-box score m as fixed.

4.1 Asymptotic linearity and inference

We now state the main asymptotic expansion for $\widehat{\psi}_{\mathrm{iso}}$. Throughout this subsection, let f_0 denote a population isotonic calibration map for the fixed score m, defined by

$$f_0 \in \arg\min_{f \in \mathcal{F}_{\mathrm{iso}}} \mathbb{E}_{P_0}\big[\{Y - f(m(X))\}^2\big], \qquad m_0 = f_0 \circ m. \tag{18}$$

Then m_0 is the population L^2-optimal monotone transformation of $m(X)$.

Assumption 2 (Regularity conditions for isotonic calibration)**.** Assume:

(i) **Boundedness:** There exists $C_0 < \infty$ such that $|\mathbb{E}[Y \mid m(X)]| \le C_0$ almost surely and $\sup_x |m^\star_{n,\mathrm{iso}}(x)| = O_p(1)$.

(ii) **Controlled tails:** The conditional error $Y - \mathbb{E}[Y \mid m(X)]$ is sub-Gaussian or subexponential under P_0.

Condition (i) requires only that the regression target be bounded and that the isotonic estimator remain bounded in probability. Condition (ii) allows Y to be unbounded, covering both Gaussian-like tails and somewhat heavier exponential-type tails. In particular, it holds for binary, Gaussian, Poisson, and bounded outcomes. Theory on isotonic regression under heavy-tailed errors suggests some degree of robustness to violations of these conditions (Han and Wellner, 2018, 2019).

The target of the asymptotic analysis is the population isotonic limit m_0, which need not equal the true regression μ_0. Assumption 2 ensures that the classical isotonic rate

$$\|m^\star_{n,\mathrm{iso}} - m_0\|^2_{2,P_{0,X}} = O_p(n^{-2/3})$$

is attained; see, for example, Chatterjee et al. (2015); Yang and Barber (2018).

Theorem 2 (Asymptotic linearity under the two-sample model)**.** *Suppose Assumptions 1 and 2 hold and the calibration score equations contain the constant score $h \equiv 1$. Then*

$$\begin{aligned}\widehat{\psi}_{\mathrm{iso}} - \psi_0 = \rho_n(\mathbb{P}^L_n - P_0)\Big\{m_0 - \psi_0 + \rho_n^{-1}(Y - m_0)\Big\} \\ + (1-\rho_n)(\mathbb{P}^U_N - P_{0,X})(m_0 - \psi_0) + R_{n,N},\end{aligned} \tag{19}$$

where $R_{n,N} = o_p(M^{-1/2})$ satisfies

$$R_{n,N} = O_p(n^{-2/3}) + O_p\big(n^{-1/3}N^{-1/2}\big)$$

Thus, under the $\sqrt{M}$ scaling, the isotonic calibrated plug-in estimator is asymptotically linear with influence pair $(D^L_{m_0}, D^U_{m_0})$ given by

$$D^L_{m_0}(X,Y) = m_0(X) - \psi_0 + \rho_0^{-1}\{Y - m_0(X)\}, \qquad D^U_{m_0}(\widetilde{X}) = m_0(\widetilde{X}) - \psi_0.$$

When $N > n$, the finite linear approximation error satisfies $R_{n,N} = O_p(n^{-2/3})$, where this rate arises from the nonparametric complexity of the isotonic calibrator f_n; under linear calibration, the corresponding remainder is typically $O_p(n^{-1})$.

Corollary 1 (Asymptotic normality and Wald inference)**.** *Under the conditions of theorem 2,*

$$\sqrt{M}\,(\widehat{\psi}_{\text{iso}} - \psi_0) \xrightarrow{d} N(0, \sigma_0^2),$$

where

$$\sigma_0^2 := \rho_0 \operatorname{Var}\left\{m_0(X) - \psi_0 + \rho_0^{-1}(Y - m_0(X))\right\} + (1-\rho_0)\operatorname{Var}\left\{m_0(\widetilde{X}) - \psi_0\right\}.$$

If

$$\widehat{D}_i^L := m^\star_{n,\text{iso}}(X_i) - \widehat{\psi}_{\text{iso}} + \rho_n^{-1}\{Y_i - m^\star_{n,\text{iso}}(X_i)\}, \qquad \widehat{D}_j^U := m^\star_{n,\text{iso}}(\widetilde{X}_j) - \widehat{\psi}_{\text{iso}},$$

and

$$\widehat{\sigma}^2 := \rho_n \mathbb{P}_n^L\left[(\widehat{D}^L)^2\right] + (1-\rho_n)\mathbb{P}_N^U\left[(\widehat{D}^U)^2\right],$$

then $\widehat{\sigma}^2 \to_p \sigma_0^2$*, and standard Wald confidence intervals follow.*

The above theorem is largely a special case of the calibrated DML analysis in van der Laan et al. (2024c). The proof combines the AIPW representation in Theorem 1, the finite uniform entropy integral of the isotonic calibrator class to control empirical-process remainders, and standard DML/TMLE arguments (Bickel et al., 1993; van der Laan and Robins, 2003; Bang and Robins, 2005; van der Laan and Rose, 2011); see, for example, Kennedy (2024) for a review.

4.2 Constructing confidence intervals via the bootstrap

Confidence intervals may be constructed using the standard Wald approximation from Corollary 1, namely

$$\widehat{\psi}_{\text{iso}} \pm z_{1-\alpha/2}\,\frac{\hat{\sigma}}{\sqrt{n+N}}.$$

As an alternative, one may use the nonparametric bootstrap and refit the isotonic calibration step within each bootstrap replicate, as proposed and theoretically studied by van der Laan et al. (2024c); see also Tang and Westling (2024). This avoids direct estimation of the influence-function variance and automatically propagates uncertainty from estimation of the calibration map.

For each bootstrap replicate $b = 1, \ldots, B$, we sample with replacement from the labeled observations $\{(X_i, Y_i)\}_{i=1}^n$ and, independently, from the unlabeled observations $\{\widetilde{X}_j\}_{j=1}^N$, obtaining bootstrap samples

$$\{(X_i^{*(b)}, Y_i^{*(b)})\}_{i=1}^n, \qquad \{\widetilde{X}_j^{*(b)}\}_{j=1}^N.$$

We then refit the isotonic calibrator on the bootstrap labeled sample:

$$\hat{f}^{*(b)} \in \arg\min_{f \in \mathcal{F}_{\text{iso}}} \sum_{i=1}^n \left\{Y_i^{*(b)} - f(m(X_i^{*(b)}))\right\}^2,$$

and define

$$m_{\text{iso}}^{*(b)}(X) := \hat{f}^{*(b)}\{m(X)\}.$$

The corresponding bootstrap replicate of the estimator is

$$\widehat{\psi}_{\text{iso}}^{*(b)} := \frac{1}{n+N}\left\{\sum_{i=1}^n m_{\text{iso}}^{*(b)}(X_i^{*(b)}) + \sum_{j=1}^N m_{\text{iso}}^{*(b)}(\widetilde{X}_j^{*(b)})\right\}.$$

After computing B bootstrap replicates, one may form a percentile interval using the empirical $\alpha/2$ and $1-\alpha/2$ quantiles of

$$\{\widehat{\psi}_{\text{iso}}^{*(b)}\}_{b=1}^B.$$

Alternatively, one may use the bootstrap standard error

$$\widehat{\sigma}^2_{\mathrm{boot}} := \frac{1}{B-1}\sum_{b=1}^{B}\left(\widehat{\psi}^{*(b)}_{\mathrm{iso}} - \bar{\psi}^{*}_{\mathrm{iso}}\right)^2, \qquad \bar{\psi}^{*}_{\mathrm{iso}} := \frac{1}{B}\sum_{b=1}^{B}\widehat{\psi}^{*(b)}_{\mathrm{iso}},$$

and report the normal-approximation interval

$$\widehat{\psi}_{\mathrm{iso}} \pm z_{1-\alpha/2}\,\widehat{\sigma}_{\mathrm{boot}}.$$

The key implementation detail is that the isotonic regression step must be refit within each bootstrap replicate rather than held fixed. Otherwise, the resulting interval does not reflect uncertainty from estimation of the calibration map.

4.3 Efficiency considerations

We now study the efficiency of the isotonic-calibrated plug-in estimator $\widehat{\psi}_{\mathrm{iso}}$. By Theorem 2, its influence pair is

$$D^L_{m_0}(X,Y) = m_0(X) - \psi_0 + \rho_0^{-1}\{Y - m_0(X)\}, \qquad D^U_{m_0}(\widetilde{X}) = m_0(\widetilde{X}) - \psi_0.$$

The key point is that $\widehat{\psi}_{\mathrm{iso}}$ is asymptotically based on the covariates only through the one-dimensional calibrated score $m_0(X)$. Its natural efficiency benchmark is therefore the reduced two-sample model in which X is replaced by $m_0(X)$. In that reduced model, the influence pair above is efficient for ψ_0. We then compare this reduced-model efficiency bound with the full-model bound and characterize when the two coincide.

Proposition 6 (Efficient influence pair in the reduced model indexed by m_0)**.** *Consider the reduced two-sample experiment*

$$\mathcal{E}_{m_0} := \left(\{(m_0(X_i), Y_i)\}_{i=1}^{n},\ \{m_0(\widetilde{X}_j)\}_{j=1}^{N}\right),$$

obtained by replacing the full covariate X with the scalar summary $m_0(X)$. The efficient influence pair for $\psi(P) = \mathbb{E}_P[Y]$ in $\mathcal{E}_{m_0}$ under sampling from P_0 is

$$D^L_{m_0}(X,Y) = m_0(X) - \psi_0 + \rho_0^{-1}\{Y - m_0(X)\}, \tag{20}$$

$$D^U_{m_0}(\widetilde{X}) = m_0(\widetilde{X}) - \psi_0. \tag{21}$$

Corollary 2 (Efficiency regimes)**.** *Under the conditions of theorem 2, the following hold.*

(i) The influence pair in theorem 2 coincides with the efficient influence pair in proposition 6. Hence $\widehat{\psi}_{\mathrm{iso}}$ is efficient for ψ_0 in the coarsened experiment $\mathcal{E}_{m_0}$.

(ii) If $m_0 = \mu_0$ almost surely, then the influence pair in theorem 2 coincides with the efficient influence pair in proposition 2. Hence $\widehat{\psi}_{\mathrm{iso}}$ is efficient for ψ_0 in the full two-sample experiment.

(iii) If $\mathbb{E}[Y \mid m(X)]$ is monotone increasing in $m(X)$, then $m_0(X) = \mathbb{E}[Y \mid m(X)]$ almost surely. Hence $\widehat{\psi}_{\mathrm{iso}}$ is efficient for ψ_0 in the coarsened experiment $\mathcal{E}_m$.

Remark 6 (Practical takeaway)**.** From a practical perspective, isotonic calibration can improve efficiency whenever the original prediction score is informative but miscalibrated. The estimator remains a simple plug-in estimator, but behaves as if the original score $m(X)$ were replaced by its best monotone calibration $m_0(X)$. When $\mathbb{E}[Y \mid m(X)]$ is monotone, no regular estimator based only on the original score $m(X)$ can improve on the isotonic-calibrated estimator. In particular, it is at least as efficient as AIPW, PPI, PPI++, and linear calibration when those estimators are based on the same score.

Remark 7 (When full efficiency is attained)**.** Full efficiency requires that the calibrated target recover all relevant regression information. This holds if $m_0 = \mu_0$ almost surely, and more generally fails when the one-dimensional score $m(X)$ omits predictive information in X that cannot be recovered through monotone calibration.

5 Experiments

5.1 Simulation study

We begin with a controlled binary-outcome simulation designed to isolate score miscalibration. Let $S \sim N(0,1)$, define

$$\mu_0(S) = \{1 + \exp(-5S)\}^{-1},$$

and draw $Y \mid S \sim \text{Bernoulli}\{\mu_0(S)\}$. To induce substantial miscalibration, we apply an affine-biased monotone distortion on the logit scale,

$$m(X) = \text{clip}\left[-0.15 + 0.75\,\sigma\{0.8\,z(S) + 0.1\,z(S)^3 - 1\},\ 0.01,\ 0.99\right],$$

where $z(S) := \text{logit}\{\mu_0(S)\}$. This transformation preserves ranking information while introducing nonlinear distortion on the probability scale. We compare the labeled-only mean (Naive), PPI, PPI++, AIPW, linear calibration (LinearCal), monotone spline calibration (MonoSpline) (Ramsay, 1988; Jiang et al., 2011), isotonic calibration (IsoCal), and an adaptive selector (Auto) that chooses among AIPW and the main calibration rules by cross-validated empirical efficiency. The labeled sample size ranges over $n \in \{50, 100, 200, 400, 800, 1200, 2400\}$, with $N/n = 1$ and $N/n = 16$. All figures and summaries are based on 500 Monte Carlo repetitions.

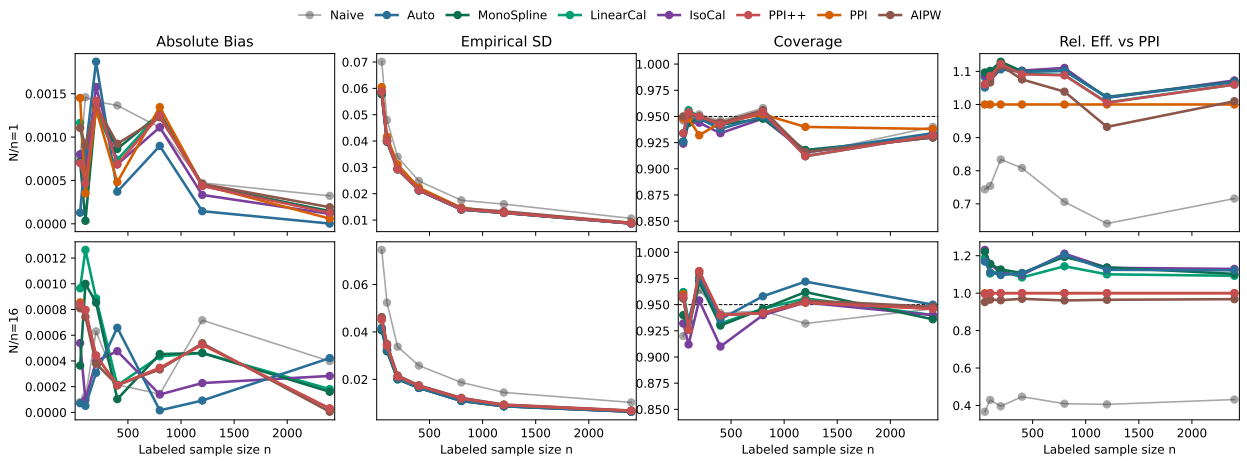

Figure 1: Synthetic simulation study with a poorly calibrated score in two unlabeled-sample regimes. The top row shows the balanced case $N/n = 1$, and the bottom row shows the large-unlabeled case $N/n = 16$. The four columns report absolute bias, empirical Monte Carlo standard deviation, coverage, and relative efficiency versus PPI, computed from empirical Monte Carlo variance. The score-based comparators now include PPI++, Auto, and MonoSpline in addition to PPI, AIPW, and the fixed calibration rules.

figure 1 reports four diagnostics most directly tied to the theory in two representative unlabeled-sample regimes, $N/n = 1$ and $N/n = 16$: absolute bias, empirical Monte Carlo standard deviation, coverage, and relative efficiency versus PPI. All score-based AIPW-type estimators are essentially unbiased. When the original score is already well calibrated, the main practical difference is efficiency: PPI is less efficient than AIPW and the calibration-based estimators. Under meaningful miscalibration, the gains from calibration are most visible when the unlabeled sample is large. For example, at $n = 400$ and $N/n = 16$, the calibrated methods reduce RMSE from about 0.0172 for PPI to about 0.0164, and by $n = 1200$ in the same regime, RMSE is about 0.0087 for Auto, MonoSpline, and IsoCal, compared with 0.0092 for PPI.

The additional benchmarks reinforce the same regime-dependent pattern. In the more balanced regime, PPI++ often tracks LinearCal closely, consistent with the first-order equivalence developed in Section 3.3. In the large-unlabeled regime, however, PPI++ is nearly indistinguishable from PPI and remains clearly worse than LinearCal. A plausible explanation is implementation-specific: we benchmark PPI++ using the official

`ppi_py` routine, which clips the optimized power-tuning coefficient λ to $[0, 1]$. If the variance-optimal scaling in this regime exceeds 1, this clipping pulls PPI++ back toward plain PPI. AUTO stays close to the best fixed score-based option, while MONOSPLINE provides a smooth monotone alternative that is usually close to isotonic calibration in RMSE and can have slightly more stable coverage in smaller samples. Coverage is broadly reasonable overall, although the more aggressive nonlinear monotone fits can show mild undercoverage when n is small. Overall, the results support the paper's main qualitative message: little changes when the score is already well calibrated, whereas under genuine miscalibration, even simple post-hoc calibration can yield noticeable semisupervised efficiency gains.

5.2 Empirical Illustration

We evaluate the proposed calibration-based estimators on the mean-estimation benchmarks used in the original experiments of Angelopoulos et al. (2023a,b), as provided in the `ppi_py` package. Our goal is not to revisit the original benchmark comparison, but rather to understand when one-dimensional calibration of a prediction score improves semisupervised mean inference. We use the same `ppi_py` datasets and retain the labeled-sample-size grids from the original examples. For each labeled sample size n, we repeatedly split the fully labeled benchmark into a labeled subset of size n and an unlabeled subset of size N, apply all estimators to the same split, and treat the full-sample mean as the ground truth.[6] The current draft summaries are averaged over 500 random splits for each n.

Benchmarks and baselines. We study the `forest`, `galaxies`, and `census_income` semisupervised mean-estimation examples from the official benchmark suite. The retained baselines are the original labeled-only interval, the imputation benchmark on the binary-outcome datasets, PPI, the classical AIPW estimator based on the same prediction score, and its efficiency-maximized variant, PPI++. The imputation benchmark averages model predictions over the pooled sample without a labeled residual correction. AIPW is the corresponding pooled plug-in-plus-augmentation estimator based on the original score; PPI is a less efficient variant that discards labeled covariate information from the plug-in term; and PPI++ is the empirical-efficiency-maximized version of AIPW studied in the original PPI benchmark suite.

We also consider four fixed calibration-based comparators that fit directly within our framework: linear calibration (`LinearCal`) (Mincer and Zarnowitz, 1969), monotone spline calibration (`MonoSpline`) (Ramsay, 1988; Jiang et al., 2011), Platt scaling (`Platt`) for the binary-outcome datasets (Platt et al., 1999), and isotonic calibration with a fixed minimum of 10 observations per bin (`IsoCal`) (Zadrozny and Elkan, 2002). We further report an adaptive selector (`Auto`) that chooses among AIPW, linear calibration, monotone spline calibration, and isotonic calibration by cross-validated empirical efficiency maximization (Rubin and van der Laan, 2008). For the binary-outcome datasets, we also report the Venn–Abers shrinkage variant described in Appendix A. Platt scaling and Venn–Abers are omitted for `census_income` because its outcome is not binary.

Metrics and main findings. For each estimator, we report Monte Carlo bias, empirical variance, MSE, Wald-interval coverage, and relative efficiency relative to PPI, with relative efficiency computed from empirical Monte Carlo variances. Estimated standard errors are used only to construct Wald intervals; all reported variability summaries are empirical Monte Carlo quantities.

The results suggest a simple practical picture. AIPW is the natural estimator based on the original score, and it typically matches or outperforms PPI. In `forest`, where the original score is already strong, the score-based methods perform similarly. In `galaxies`, AIPW and PPI remain nearly tied across the grid, while `Auto` usually stays near the top. When the unlabeled sample is much larger than the labeled sample, AIPW and PPI become nearly indistinguishable because the labeled contribution to the pooled plug-in term is negligible. This is what we see in `census_income`, where the unlabeled sample is enormous.

[6] Because repeated sample splits of a fixed finite benchmark dataset are not i.i.d. draws from a superpopulation, asymptotic Wald confidence intervals can exhibit either undercoverage or overcoverage. This is not specific to our calibration extensions, but is a feature of evaluating Wald-type intervals under finite-population resampling.

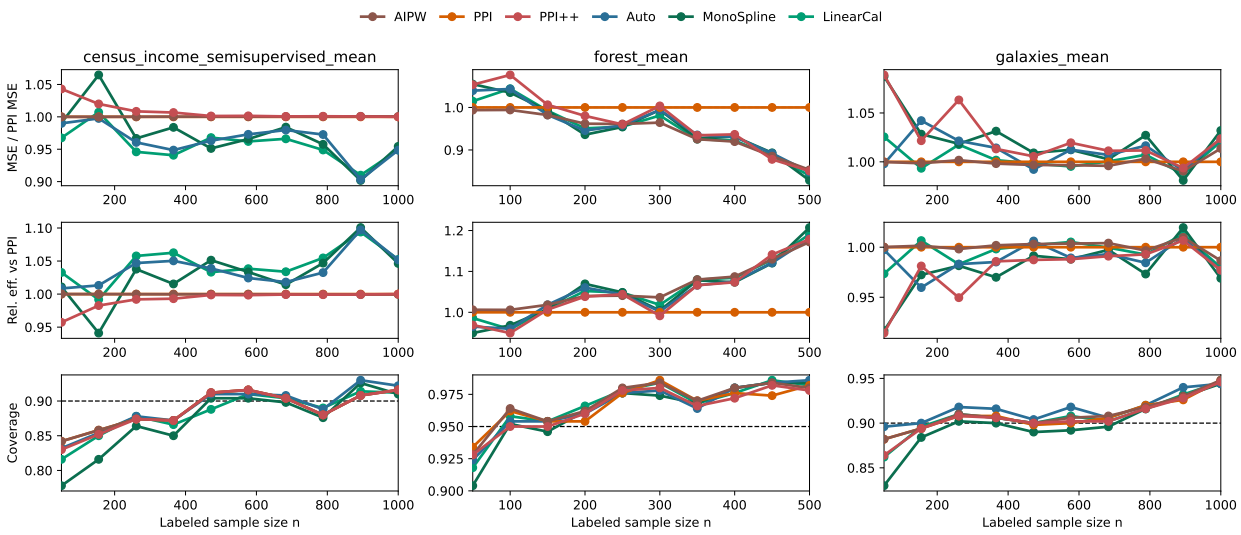

Figure 2: Main-text benchmark summary showing normalized MSE relative to PPI, relative efficiency versus PPI, and coverage across the reproduced PPI benchmarks, restricted to the primary score-based comparators: PPI, PPI++, AIPW, `Auto`, `MonoSpline`, and `LinearCal`. In the normalized-MSE panels, values below one indicate improvement over PPI. Relative efficiency is computed from empirical Monte Carlo variance, so values above one indicate improvement over PPI. The dashed horizontal line in the coverage panels marks the nominal target coverage, which is 0.95 for `forest` and 0.90 for `galaxies` and `census_income`.

Post-hoc calibration is most useful when the score is meaningfully miscalibrated and the labeled sample is large enough to estimate the second-stage calibrator stably. `Census_income` provides the clearest example: `LinearCal` performs best on average, with `IsoCal`, `Auto`, and `MonoSpline` often close behind. The added benchmarks sharpen this interpretation rather than changing it. `Auto` acts as a practical hedge that tends to stay close to the best fixed score-based method, while `MonoSpline` provides a smoother monotone alternative to fixed-bin isotonic calibration. Although PPI++ is first-order equivalent to linear calibration, in the reproduced real-data benchmarks it is generally less competitive than `LinearCal`, especially on `census_income` and in most settings for `galaxies` and `forest`. Overall, these results suggest little practical reason to prefer PPI over AIPW, and they indicate that simple calibration methods such as `LinearCal` can deliver modest but repeatable gains.

6 Conclusion and closing remarks

We studied semisupervised mean estimation with a black-box prediction score and proposed a simple calibrated plug-in procedure: fit the score, calibrate it on the labeled sample, and average the calibrated predictions over the pooled covariate sample. Our main observation is that this seemingly simple plug-in rule also admits an exact AIPW representation, placing it directly within the standard semiparametric framework for debiased estimation. In particular, isotonic calibration yields a lightweight, tuning-free estimator with asymptotic linearity, asymptotic normality, and valid Wald inference under weak conditions. Confidence intervals may be constructed from the influence function, or equivalently from the residual variance; alternatively, the bootstrap may better capture the finite-sample variability of the isotonic regression update.

From an efficiency perspective, the isotonic-calibrated estimator is efficient for the information retained by the calibrated score $m_0(X)$. It is nonparametrically efficient for the full covariate information in X when the calibrated limit m_0 coincides with the true regression function μ_0. Isotonic calibration improves efficiency over the simple AIPW estimator based on the original score $m(X)$, and typically also improves on linear calibration. Empirically, the practical message is simple: when the original score is already well calibrated, calibration changes little, whereas under meaningful miscalibration, even simple affine or monotone calibration

Dataset	Smaller labeled-sample regime	Larger labeled-sample regime
`forest`	AIPW and PPI are already hard to beat at the smallest n; the calibration-based alternatives are close but do not clearly improve on them in this strong-score regime.	At larger n, all of the score-based methods cluster tightly, with only modest gains from richer monotone calibration.
`galaxies`	AIPW, PPI, and `Auto` are very close at the smallest n, while PPI++ is somewhat less competitive in this noisier regime.	At larger n, the leading score-based methods remain tightly clustered, and no single calibration rule dominates uniformly.
`census_income`	PPI and AIPW are effectively indistinguishable because the unlabeled sample is enormous; `LinearCal` and `Auto` already offer modest gains at the smallest n, while PPI++ is somewhat less competitive.	`LinearCal` remains strongest on average, with `IsoCal`, `Auto`, and `MonoSpline` close behind as n grows, while PPI and AIPW remain nearly identical throughout.

Table 1: Compact summary of the reproduced PPI benchmarks across regimes. The main empirical contrast is between small labeled-sample regimes, where simple or adaptive low-complexity corrections are safer, and larger labeled-sample regimes, where richer monotone calibration can help when the score is miscalibrated.

can improve semisupervised efficiency while preserving a transparent plug-in form.

Positioning within existing semiparametric methods. More broadly, our results clarify that recent prediction-powered inference methods fit naturally within the standard semiparametric, debiased machine-learning, and flexible covariate-adjustment toolkit for randomized trials and missing-at-random settings. In particular, we show that PPI++ is AIPW with empirical efficiency maximization (Rubin and van der Laan, 2008) and is first-order equivalent to linear calibration and classical prognostic-score regression adjustment (Hansen, 2008; Rosenblum and van der Laan, 2009; Moore and van der Laan, 2009a; Lin, 2013). It is therefore closely related to randomized-trial approaches that learn prognostic scores from larger historical or auxiliary data sets and then use them for regression adjustment in smaller trials (Hansen, 2008; Schuler et al., 2022; Højbjerre-Frandsen et al., 2025). This perspective suggests that semisupervised inference may benefit from closer integration with the broader missing-data, causal-inference, and debiased machine-learning literatures. These lines of work address closely related versions of the same underlying problem, so further progress may come from translational work across them, especially in adapting tools developed for randomized trials and missing-at-random settings to semisupervised problems (Smith et al., 2023).

Beyond mean estimation. These connections extend beyond the semisupervised mean-estimation setting considered here. For example, when missingness affects both outcomes and covariates and may depend on observed covariates and outcomes, as well as in extensions to right-censored time-to-event outcomes and longitudinal settings, existing semiparametric methods can often be applied directly (Robins et al., 1994; van der Laan and Robins, 2003; Bang and Robins, 2005; Moore and van der Laan, 2009b; Rose and van der Laan, 2011; Van der Laan and Rose, 2018). Likewise, there is already a well-developed literature on machine-learning-assisted estimation of local and conditional regression targets, including dose-response curves and conditional average treatment effects, that can be used to extend inference beyond marginal mean estimation (Rubin and van der Laan, 2006; Díaz and van der Laan, 2013; Kennedy et al., 2017; Bibaut and van der Laan, 2017; van der Laan et al., 2018; Westling et al., 2020; Kennedy, 2023; Foster and Syrgkanis, 2023; Luedtke and Chung, 2024; Chernozhukov et al., 2024; Butzin-Dozier et al., 2024; Zhang et al., 2025). In particular, kernel-regression-based (Bibaut and van der Laan, 2017) and isotonic-regression-based (Westling and Carone, 2020) inference can be carried out in this setting using AIPW-score-based pseudo-outcomes, as in Van Der Laan and Dudoit (2003); Rubin and van der Laan (2006); Kennedy et al. (2017), or more generally using influence functions and Neyman-orthogonal losses (Bibaut and van der Laan, 2017; Foster and Syrgkanis, 2023; Chernozhukov et al., 2024). Finally, when outcomes are not missing completely at random, naively pooling labeled and unlabeled data can bias estimation because the missingness mechanism may remain confounded by unobserved outcomes. There is substantial work on combining gold-standard

and potentially biased auxiliary data more safely using semiparametric and machine-learning tools (Kallus et al., 2018; Rosenman, 2025; Dang et al., 2025; van der Laan et al., 2026). Likewise, when the unlabeled sample is selected in a data-dependent way, the problem is closely related to adaptive and selectively sampled designs, including two-phase and informative sampling (van der Laan, 2008; Chow and Chang, 2008; Rose and van der Laan, 2011; Malenica et al., 2021; Zhang and van der Laan, 2025; Liu et al., 2025). Extending these ideas to semisupervised settings is therefore a natural direction for future work.

Extensions of calibration. While we focus on calibration of regression functions and inference on means, the ideas here may extend to more general supervised learning problems defined through loss minimization, such as median and quantile regression (Noarov and Roth, 2023; Jung et al., 2021; Roth, 2022; Whitehouse et al., 2024). As a concrete example, Algorithm 3 extends our linearly calibrated mean estimator to linear regression by including the prediction score as a covariate, in the spirit of classical prognostic adjustment. Related validity results for such adjustments in generalized linear models are given by Rosenblum and van der Laan (2009). Similarly, for median and quantile regression, one could include prediction scores as covariates and use Wald-type inference for the corresponding target parameters. More generally, when the target parameter is defined by a moment equation, as in M-estimation, one could debias or calibrate the associated augmented moment equation (van der Laan and Robins, 2003); see Appendix C. A practical challenge in this setting is that augmented loss functions may be nonconvex, which can complicate optimization. In such cases, a more practical alternative may be a sieve-based, TMLE-style multiaccuracy adjustment rather than explicit augmentation, as in the Efficient Plug-in Learning framework of van der Laan et al. (2024a), which addresses this issue for nonconvex Neyman-orthogonal loss functions (Foster and Syrgkanis, 2023). Debiased inference and influence-function theory for generic smooth functionals of an M-estimand are studied by van der Laan et al. (2025a).

References

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

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

Daniele Ballinari and Nora Bearth. Improving the finite sample performance of double/debiased machine learning with propensity score calibration. *arXiv preprint arXiv:2409.04874*, 2024.

Laura B Balzer, Erica Cai, Lucas Godoy Garraza, and Pracheta Amaranath. Adaptive selection of the optimal strategy to improve precision and power in randomized trials. *Biometrics*, 80(1):ujad034, 2024.

Heejung Bang and James M. Robins. Doubly robust estimation in missing data and causal inference models. *Biometrics*, 61(4):962–973, 2005.

Richard E. Barlow and Hugh D. Brunk. The isotonic regression problem and its dual. *Journal of the American Statistical Association*, 67(337):140–147, 1972.

Antonio Bella, Cèsar Ferri, José Hernández-Orallo, and María José Ramírez-Quintana. Calibration of machine learning models. In *Handbook of Research on Machine Learning Applications and Trends: Algorithms, Methods, and Techniques*, pages 128–146. IGI Global, 2010.

David Benkeser, Marco Carone, MJ Van Der Laan, and Peter B Gilbert. Doubly robust nonparametric inference on the average treatment effect. *Biometrika*, 104(4):863–880, 2017.

Aurelien F Bibaut and Mark J van der Laan. Data-adaptive smoothing for optimal-rate estimation of possibly non-regular parameters. *arXiv preprint arXiv:1706.07408*, 2017.

Aurélien F. Bibaut and Mark J. van der Laan. Fast rates for empirical risk minimization over càdlàg functions with bounded sectional variation norm. *arXiv preprint arXiv:1907.09244*, 2019.

Peter J. Bickel, Chris A. J. Klaassen, Ya'acov Ritov, and Jon A. Wellner. *Efficient and Adaptive Estimation for Semiparametric Models*. Johns Hopkins University Press, 1993.

David Bruns-Smith, Oliver Dukes, Avi Feller, and Elizabeth L. Ogburn. Augmented balancing weights as linear regression. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, page qkaf019, 2025.

Zachary Butzin-Dozier, Sky Qiu, Alan E Hubbard, Junming Seraphina Shi, and Mark J van der Laan. Highly adaptive lasso: Machine learning that provides valid nonparametric inference in realistic models. *medRxiv*, 2024.

Marco Carone, Iván Díaz, and Mark J van der Laan. Higher-order targeted minimum loss-based estimation. 2014.

Claes M Cassel, Carl E Särndal, and Jan H Wretman. Some results on generalized difference estimation and generalized regression estimation for finite populations. *Biometrika*, 63(3):615–620, 1976.

Sabyasachi Chatterjee, Adityanand Guntuboyina, and Bodhisattva Sen. On risk bounds in isotonic and other shape restricted regression problems. *The Annals of Statistics*, 43(4):1774–1800, 2015.

Ambarish Chattopadhyay and José R. Zubizarreta. On the implied weights of linear regression for causal inference. *Biometrika*, 110(3):615–629, 2023.

Ambarish Chattopadhyay, Christopher H Hase, and José R Zubizarreta. Balancing vs modeling approaches to weighting in practice. *Statistics in Medicine*, 39(24):3227–3254, 2020.

Qizhao Chen, Vasilis Syrgkanis, and Morgane Austern. Debiased machine learning without sample-splitting for stable estimators. In *Advances in Neural Information Processing Systems*, volume 35, pages 30925–30937, 2022.

Song Xi Chen, Denis HY 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.

Yi-Hau Chen. Miscellanea. a robust imputation method for surrogate outcome data. *Biometrika*, 87(3): 711–716, 2000.

Yuxi Chen, Edward H Kennedy, and Sivaraman Balakrishnan. On the equivalence between neyman orthogonality and pathwise differentiability. *arXiv preprint arXiv:2603.15817*, 2026.

David Cheng, Ashwin N Ananthakrishnan, and Tianxi Cai. Robust and efficient semi-supervised estimation of average treatment effects with application to electronic health records data. *Biometrics*, 77(2):413–423, 2021.

Victor Chernozhukov, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen, Whitney Newey, and James Robins. Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal*, 21(1):C1–C68, 2018a.

Victor Chernozhukov, Mert Demirer, Esther Duflo, and Ivan Fernandez-Val. Generic machine learning inference on heterogeneous treatment effects in randomized experiments, with an application to immunization in india. Technical report, National Bureau of Economic Research, 2018b.

Victor Chernozhukov, Whitney K. Newey, and Rahul Singh. Automatic debiased machine learning of causal and structural effects. *Econometrica*, 90(3):967–1027, 2022.

Victor Chernozhukov, Whitney K Newey, and Vasilis Syrgkanis. Conditional influence functions. *arXiv preprint arXiv:2412.18080*, 2024.

Shein-Chung Chow and Mark Chang. Adaptive design methods in clinical trials–a review. *Orphanet journal of rare diseases*, 3(1):11, 2008.

David R Cox. Two further applications of a model for binary regression. *Biometrika*, 45(3/4):562–565, 1958.

Lauren Eyler Dang, Jens Magelund Tarp, Trine Julie Abrahamsen, Kajsa Kvist, John B Buse, Maya Petersen, and Mark van der Laan. Experiment-selector cross-validated targeted maximum likelihood estimator for hybrid rct-external data studies. *Journal of Causal Inference*, 13(1):20240041, 2025.

Ilker Demirel, Ahmed Alaa, Anthony Philippakis, and David Sontag. Prediction-powered generalization of causal inferences. *arXiv preprint arXiv:2406.02873*, 2024.

Shachi Deshpande and Volodymyr Kuleshov. Calibrated and conformal propensity scores for causal effect estimation. In *The 40th Conference on Uncertainty in Artificial Intelligence*, 2023.

Iván Díaz. Machine learning in the estimation of causal effects: targeted minimum loss-based estimation and double/debiased machine learning. *Biostatistics*, 21(2):353–358, 2020.

Iván Díaz and Mark J van der Laan. Targeted data adaptive estimation of the causal dose–response curve. *Journal of Causal Inference*, 1(2):171–192, 2013.

Iván Díaz and Mark J van der Laan. Doubly robust inference for targeted minimum loss–based estimation in randomized trials with missing outcome data. *Statistics in medicine*, 36(24):3807–3819, 2017.

Peng Ding and Fan Li. Causal inference: A missing data perspective, 2018. URL https://arxiv.org/abs/1712.06170.

Dylan J Foster and Vasilis Syrgkanis. Orthogonal statistical learning. *The Annals of Statistics*, 51(3):879–908, 2023.

Tilmann Gneiting, Fadoua Balabdaoui, and Adrian E Raftery. Probabilistic forecasts, calibration and sharpness. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 69(2):243–268, 2007.

Ellen Graham, Marco Carone, and Andrea Rotnitzky. Towards a unified theory for semiparametric data fusion with individual-level data. *arXiv preprint arXiv:2409.09973*, 2024.

Susan Gruber and Mark J Van Der Laan. Targeted maximum likelihood estimation: A gentle introduction. 2009.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. On calibration of modern neural networks. In *International conference on machine learning*, pages 1321–1330. PMLR, 2017.

Chirag Gupta and Aaditya Ramdas. Distribution-free calibration guarantees for histogram binning without sample splitting. In *International conference on machine learning*, pages 3942–3952. PMLR, 2021.

Rom Gutman, Ehud Karavani, and Yishai Shimoni. Propensity score models are better when post-calibrated. *arXiv preprint arXiv:2211.01221*, 2022.

Jens Hainmueller. Entropy balancing for causal effects: A multivariate reweighting method to produce balanced samples in observational studies. *Political Analysis*, 20(1):25–46, 2012.

Peisong Han. A note on improving the efficiency of inverse probability weighted estimator using the augmentation term. *Statistics & Probability Letters*, 82(12):2221–2228, 2012.

Qiyang Han and Jon A Wellner. Robustness of shape-restricted regression estimators: An envelope perspective. *arXiv preprint arXiv:1805.02542*, 2018.

Qiyang Han and Jon A Wellner. Convergence rates of least squares regression estimators with heavy-tailed errors. 2019.

Ben B Hansen. The prognostic analogue of the propensity score. *Biometrika*, 95(2):481–488, 2008.

Rafail Z Hasminskii and Ildar A Ibragimov. On the nonparametric estimation of functionals. In *Proceedings of the Second Prague Symposium on Asymptotic Statistics*, volume 473, pages 474–482. North-Holland Amsterdam, 1979.

Nima S. Hejazi and Mark J. van der Laan. Revisiting the propensity score's central role: Towards bridging balance and efficiency in the era of causal machine learning. *Observational Studies*, 9(1):23–34, 2023.

Nima S Hejazi, Mark J van der Laan, Holly E Janes, Peter B Gilbert, and David C Benkeser. Efficient nonparametric inference on the effects of stochastic interventions under two-phase sampling, with applications to vaccine efficacy trials. *Biometrics*, 77(4):1241–1253, 2021.

Katherine Hoffman. An illustrated guide to tmle, part i: Introduction and motivation, December 2020. URL https://www.khstats.com/blog/tmle/tutorial.

Emilie Højbjerre-Frandsen and Alejandro Schuler. "within-trial" prognostic score adjustment is targeted maximum likelihood estimation. *Pharmaceutical Statistics*, 25(2):e70080, 2026.

Emilie Højbjerre-Frandsen, Mark J van der Laan, and Alejandro Schuler. Powering rcts for marginal effects with glms using prognostic score adjustment. *arXiv preprint arXiv:2503.22284*, 2025.

Kosuke Imai and Marc Ratkovic. Covariate balancing propensity score. *Journal of the Royal Statistical Society Series B*, 76(1):243–263, 2014.

Wenlong Ji, Lihua Lei, and Tijana Zrnic. Predictions as surrogates: Revisiting surrogate outcomes in the age of ai. *arXiv preprint arXiv:2501.09731*, 2025.

Xiaoqian Jiang, Melanie Osl, Jihoon Kim, and Lucila Ohno-Machado. Smooth isotonic regression: a new method to calibrate predictive models. *AMIA Summits on Translational Science Proceedings*, 2011:16, 2011.

Christopher Jung, Changhwa Lee, Mallesh Pai, Aaron Roth, and Rakesh Vohra. Moment multicalibration for uncertainty estimation. In *Conference on Learning Theory*, pages 2634–2678. PMLR, 2021.

Nathan Kallus and Xiaojie Mao. On the role of surrogates in the efficient estimation of treatment effects with limited outcome data. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 87(2): 480–509, 2025.

Nathan Kallus, Aahlad Manas Puli, and Uri Shalit. Removing hidden confounding by experimental grounding. *Advances in neural information processing systems*, 31, 2018.

Edward H Kennedy. Towards optimal doubly robust estimation of heterogeneous causal effects. *Electronic Journal of Statistics*, 17(2):3008–3049, 2023.

Edward H Kennedy. Semiparametric doubly robust targeted double machine learning: a review. *Handbook of statistical methods for precision medicine*, pages 207–236, 2024.

Edward H Kennedy, Zongming Ma, Matthew D McHugh, and Dylan S Small. Non-parametric methods for doubly robust estimation of continuous treatment effects. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 79(4):1229–1245, 2017.

Chris AJ Klaassen. Consistent estimation of the influence function of locally asymptotically linear estimators. *The Annals of Statistics*, 15(4):1548–1562, 1987.

Nicolas S Lambert. Elicitation and evaluation of statistical forecasts. *Preprint*, 3(5):18, 2011.

Lucien Le Cam. On some asymptotic properties of maximum likelihood estimates and related bayes' estimates. *Univ. Calif. Publ. in Statist.*, 1:277–330, 1953.

Donghwan Lee, Xinmeng Huang, Hamed Hassani, and Edgar Dobriban. T-cal: An optimal test for the calibration of predictive models. *Journal of Machine Learning Research*, 24(335):1–72, 2023.

Se Yoon Lee and Jae-Kwang Kim. Mec: Machine-learning-assisted generalized entropy calibration for semi-supervised mean estimation. *arXiv preprint arXiv:2604.05446*, 2026.

Samuel D. Lendle, Bruce Fireman, and Mark J. van der Laan. Balancing score adjusted targeted minimum loss-based estimation. *Journal of Causal Inference*, 3(2):139–155, 2015.

Yan Leng and Drew Dimmery. Calibration of heterogeneous treatment effects in random experiments. *Available at SSRN 3875850*, 2021.

Boris Yakovlevich Levit. On efficiency of a class of non-parametric estimates. *Teoriya Veroyatnostei i ee Primeneniya*, 20(4):738–754, 1975.

Sijia Li and Alex Luedtke. Efficient estimation under data fusion. *Biometrika*, 110(4):1041–1054, 2023.

Sijia Li, Peter B Gilbert, Rui Duan, and Alex Luedtke. Data fusion using weakly aligned sources. *Journal of the American Statistical Association*, 120(552):2569–2579, 2025.

Sarah Lichtenstein, Baruch Fischhoff, and Lawrence D Phillips. Calibration of probabilities: The state of the art. In *Decision Making and Change in Human Affairs: Proceedings of the Fifth Research Conference on Subjective Probability, Utility, and Decision Making, Darmstadt, 1–4 September, 1975*, pages 275–324. Springer, 1977.

Winston Lin. Agnostic notes on regression adjustments to experimental data: Reexamining freedman's critique. *The Annals of Applied Statistics*, pages 295–318, 2013.

Jiajun Liu, Ke Zhu, Shu Yang, and Xiaofei Wang. Robust estimation and inference in hybrid controlled trials for binary outcomes: A case study on non-small cell lung cancer. *arXiv preprint arXiv:2505.00217*, 2025.

Alex Luedtke and Incheoul Chung. One-step estimation of differentiable hilbert-valued parameters. *The Annals of Statistics*, 52(4):1534–1563, 2024.

Ivana Malenica, Aurelien Bibaut, and Mark J van der Laan. Adaptive sequential design for a single time-series. *arXiv preprint arXiv:2102.00102*, 2021.

Jacob A Mincer and Victor Zarnowitz. The evaluation of economic forecasts. In *Economic forecasts and expectations: Analysis of forecasting behavior and performance*, pages 3–46. NBER, 1969.

Kelly L Moore and Mark J van der Laan. Covariate adjustment in randomized trials with binary outcomes: targeted maximum likelihood estimation. *Statistics in medicine*, 28(1):39–64, 2009a.

Kelly L Moore and Mark J van der Laan. Increasing power in randomized trials with right censored outcomes through covariate adjustment. *Journal of biopharmaceutical statistics*, 19(6):1099–1131, 2009b.

Kelly L Moore, Romain Neugebauer, Thamban Valappil, and Mark J van der Laan. Robust extraction of covariate information to improve estimation efficiency in randomized trials. *Statistics in medicine*, 30(19): 2389–2408, 2011.

Reagan Mozer. Ppi is the difference estimator: Recognizing the survey sampling roots of prediction-powered inference. *arXiv preprint arXiv:2603.19160*, 2026.

Alexandru Niculescu-Mizil and Rich Caruana. Predicting good probabilities with supervised learning. In *Proceedings of the 22nd International Conference on Machine Learning*, pages 625–632, 2005.

Georgy Noarov and Aaron Roth. The statistical scope of multicalibration. In *International Conference on Machine Learning*, pages 26283–26310. PMLR, 2023.

Margaret Sullivan Pepe. Inference using surrogate outcome data and a validation sample. *Biometrika*, 79(2): 355–365, 1992.

Margaret Sullivan Pepe, Marie Reilly, and Thomas R Fleming. Auxiliary outcome data and the mean score method. *Journal of Statistical Planning and Inference*, 42(1-2):137–160, 1994.

J Pfanzagl and W Wefelmeyer. Contributions to a general asymptotic statistical theory. *Statistics & Risk Modeling*, 3(3-4):379–388, 1985.

John Platt et al. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. *Advances in large margin classifiers*, 10(3):61–74, 1999.

Pierre-Emmanuel Poulet, Maylis Tran, Sophie Tezenas du Montcel, Bruno Dubois, Stanley Durrleman, Bruno Jedynak, and Alzheimer's Disease Neuroimaging Initiative. Prediction-powered inference for clinical trials: application to linear covariate adjustment. *BMC Medical Research Methodology*, 25(1):204, 2025.

Sky Qiu, Susan Gruber, Pamela A Shaw, Brian D Williamson, and Mark J van der Laan. Efficient targeted maximum likelihood estimators for two-phase design problems. *arXiv preprint arXiv:2602.24131*, 2026.

Jan Rabenseifner, Sven Klaassen, Jannis Kueck, and Philipp Bach. Calibration strategies for robust causal estimation: Theoretical and empirical insights on propensity score based estimators. *arXiv preprint arXiv:2503.17290*, 2025.

James O Ramsay. Monotone regression splines in action. *Statistical science*, pages 425–441, 1988.

James M Robins and Andrea Rotnitzky. Semiparametric efficiency in multivariate regression models with missing data. *Journal of the American Statistical Association*, 90(429):122–129, 1995.

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.

James M. Robins, Andrea Rotnitzky, and Lue Ping Zhao. Analysis of semiparametric regression models for repeated outcomes in the presence of missing data. *Journal of the American Statistical Association*, 90 (429):106–121, 1995.

Sherri Rose and Mark J van der Laan. A targeted maximum likelihood estimator for two-stage designs. *The international journal of biostatistics*, 7(1):17, 2011.

Michael Rosenblum and Mark J van der Laan. Using regression models to analyze randomized trials: Asymptotically valid hypothesis tests despite incorrectly specified models. *Biometrics*, 65(3):937–945, 2009.

Evan TR Rosenman. Methods for combining observational and experimental causal estimates: A review. *Wiley Interdisciplinary Reviews: Computational Statistics*, 17(2):e70027, 2025.

Rachael K Ross, Lina M Montoya, Dana E Goin, Iván Díaz, and Audrey Renson. Constructing targeted minimum loss/maximum likelihood estimators: a simple illustration to build intuition. *American journal of epidemiology*, page kwaf261, 2025.

Aaron Roth. Uncertain: Modern topics in uncertainty estimation. *Unpublished Lecture Notes*, 11(30-31):4, 2022.

Andrea Rotnitzky, Quanhong Lei, Mariela Sued, and James M Robins. Improved double-robust estimation in missing data and causal inference models. *Biometrika*, 99(2):439–456, 2012.

Andrea Rotnitzky, Ezequiel Smucler, and James M Robins. Characterization of parameters with a mixed bias property. *Biometrika*, 108(1):231–238, 2021.

Andrea Rotnitzky, Ezequiel Smucler, and James M Robins. A note on the relation between one-step, outcome regression and ipw-type estimators of parameters with the mixed bias property. *arXiv preprint arXiv:2509.22452*, 2025.

Daniel Rubin and Mark J van der Laan. Doubly robust censoring unbiased transformations. 2006.

Daniel B Rubin and Mark J van der Laan. Empirical efficiency maximization: improved locally efficient covariate adjustment in randomized experiments and survival analysis. *The International Journal of Biostatistics*, 4(1), 2008.

Donald B. Rubin. Inference and missing data. *Biometrika*, 63(3):581–592, 1976.

Donald B Rubin. Bayesian inference for causal effects: The role of randomization. *The Annals of statistics*, pages 34–58, 1978.

Carl-Erik Särndal, Bengt Swensson, and Jan Wretman. *Model assisted survey sampling*. Springer Science & Business Media, 2003.

Daniel O Scharfstein, Andrea Rotnitzky, and James M Robins. Adjusting for nonignorable drop-out using semiparametric nonresponse models. *Journal of the American Statistical Association*, 94(448):1096–1120, 1999.

Anton Schick. On asymptotically efficient estimation in semiparametric models. *The Annals of Statistics*, pages 1139–1151, 1986.

Alejandro Schuler, David Walsh, Diana Hall, Jon Walsh, Charles Fisher, Critical Path for Alzheimer's Disease, Alzheimer's Disease Neuroimaging Initiative, and Alzheimer's Disease Cooperative Study. Increasing the efficiency of randomized trial estimates via linear adjustment for a prognostic score. *The International Journal of Biostatistics*, 18(2):329–356, 2022.

Andrew J Scott and D Holt. The effect of two-stage sampling on ordinary least squares methods. *Journal of the American Statistical Association*, 77(380):848–854, 1982.

Shaun R Seaman and Stijn Vansteelandt. Introduction to double robust methods for incomplete data. *Statistical science: a review journal of the Institute of Mathematical Statistics*, 33(2):184, 2018.

Matthew J Smith, Rachael V Phillips, Miguel Angel Luque-Fernandez, and Camille Maringe. Application of targeted maximum likelihood estimation in public health and epidemiological studies: a systematic review. *Annals of epidemiology*, 86:34–48, 2023.

Yilin Song, Dan M Kluger, Harsh Parikh, and Tian Gu. Demystifying prediction powered inference. *arXiv preprint arXiv:2601.20819*, 2026.

Zhou Tang and Ted Westling. Consistency of the bootstrap for asymptotically linear estimators based on machine learning. *arXiv preprint arXiv:2404.03064*, 2024.

Anastasios A Tsiatis. *Semiparametric theory and missing data*. Springer, 2006.

Lars van der Laan and Ahmed Alaa. Generalized venn and venn-abers calibration with applications in conformal prediction. *arXiv preprint arXiv:2502.05676*, 2025.

Lars van der Laan and Ahmed M Alaa. Self-calibrating conformal prediction. *Advances in Neural Information Processing Systems*, 37:107138–107170, 2024.

Lars van der Laan and Nathan Kallus. Bellman calibration for v-learning in offline reinforcement learning. *arXiv preprint arXiv:2512.23694*, 2025.

Lars van der Laan, Marco Carone, Alex Luedtke, and Mark van der Laan. Adaptive debiased machine learning using data-driven model selection techniques. *arXiv preprint arXiv:2307.12544*, 2023a.

Lars van der Laan, Ernesto Ulloa-Pérez, Marco Carone, and Alex Luedtke. Causal isotonic calibration for heterogeneous treatment effects. In *International Conference on Machine Learning*, pages 34831–34854, 2023b.

Lars van der Laan, Marco Carone, and Alex Luedtke. Combining t-learning and dr-learning: a framework for oracle-efficient estimation of causal contrasts. *arXiv preprint arXiv:2402.01972*, 2024a.

Lars van der Laan, Ziming Lin, Marco Carone, and Alex Luedtke. Stabilized inverse probability weighting via isotonic calibration. *arXiv preprint arXiv:2411.06342*, 2024b.

Lars van der Laan, Alex Luedtke, and Marco Carone. Doubly robust inference via calibration. *arXiv preprint arXiv:2411.02771*, 2024c.

Lars van der Laan, Aurelien Bibaut, Nathan Kallus, and Alex Luedtke. Automatic debiased machine learning for smooth functionals of nonparametric m-estimands. *arXiv preprint arXiv:2501.11868*, 2025a.

Lars van der Laan, David Hubbard, Allen Tran, Nathan Kallus, and Aurélien Bibaut. Semiparametric double reinforcement learning with applications to long-term causal inference. *arXiv preprint arXiv:2501.06926*, 2025b.

Mark van der Laan. Cv-tmle and double machine learning, December 2019. URL https://vanderlaan-lab.org/2019/12/24/cv-tmle-and-double-machine-learning/.

Mark van der Laan, Zeyi Wang, and Lars van der Laan. Higher order targeted maximum likelihood estimation. *arXiv preprint arXiv:2101.06290*, 2021.

Mark van der Laan, Sky Qiu, Jens Magelund Tarp, and Lars van der Laan. Adaptive-tmle for the average treatment effect based on randomized controlled trial augmented with real-world data. *Journal of Causal Inference*, 14(1):20240025, 2026.

Mark J van der Laan. The construction and analysis of adaptive group sequential designs. 2008.

Mark J Van der Laan. Targeted estimation of nuisance parameters to obtain valid statistical inference. *The international journal of biostatistics*, 10(1):29–57, 2014.

Mark J Van Der Laan and Sandrine Dudoit. Unified cross-validation methodology for selection among estimators and a general cross-validated adaptive epsilon-net estimator: Finite sample oracle inequalities and examples. 2003.

Mark J. van der Laan and James M. Robins. *Unified Methods for Censored Longitudinal Data and Causality*. Springer, 2003.

Mark J. van der Laan and Sherri Rose. *Targeted Learning: Causal Inference for Observational and Experimental Data*. Springer, 2011.

Mark J Van der Laan and Sherri Rose. *Targeted learning in data science*. Springer, 2018.

Mark J Van Der Laan and Daniel Rubin. Targeted maximum likelihood learning. 2006.

Mark J van der Laan, Aurélien Bibaut, and Alexander R Luedtke. Cv-tmle for nonpathwise differentiable target parameters. In *Targeted Learning in Data Science: Causal Inference for Complex Longitudinal Studies*, pages 455–481. Springer, 2018.

Aad W Van der Vaart. *Asymptotic statistics*, volume 3. Cambridge university press, 2000.

Adriaan W. van der Vaart and Jon A. Wellner. *Weak Convergence and Empirical Processes*. Springer, 1996.

Vladimir Vovk and Ivan Petej. Venn-abers predictors. *arXiv preprint arXiv:1211.0025*, 2012.

Vladimir Vovk, Alexander Gammerman, and Glenn Shafer. *Algorithmic learning in a random world*. Springer, 2005.

Volodimir G Vovk. Universal forecasting algorithms. *Information and Computation*, 96(2):245–277, 1992.

Cheng Wang. Calibration in deep learning: A survey of the state-of-the-art. *arXiv preprint arXiv:2308.01222*, 2023.

Ted Westling and Marco Carone. A unified study of nonparametric inference for monotone functions. *Annals of statistics*, 48(2):1001, 2020.

Ted Westling, Peter Gilbert, and Marco Carone. Causal isotonic regression. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 82(3):719–747, 2020.

J Whitehouse, C Jung, V Syrgkanis, B Wilder, and ZS Wu. Orthogonal causal calibration. *arXiv preprint arXiv:2406.01933*, 2024.

Zichun Xu, Daniela Witten, and Ali Shojaie. A unified framework for semiparametrically efficient semi-supervised learning. *arXiv preprint arXiv:2502.17741*, 2025.

Fan Yang and Rina Foygel Barber. Contraction and uniform convergence of isotonic regression. *arXiv preprint arXiv:1706.01852*, 2018.

Bianca Zadrozny and Charles Elkan. Obtaining calibrated probability estimates from decision trees and naive bayesian classifiers. In *Proceedings of the Eighteenth International Conference on Machine Learning*, pages 609–616, 2001.

Bianca Zadrozny and Charles Elkan. Transforming classifier scores into accurate multiclass probability estimates. In *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 694–699, 2002.

Wenxin Zhang and Mark van der Laan. Efficient statistical estimation for sequential adaptive experiments with implications for adaptive designs. *arXiv preprint arXiv:2508.09135*, 2025.

Wenxin Zhang, Junming Shi, Alan Hubbard, and Mark van der Laan. Constructing confidence intervals for infinite-dimensional functional parameters by highly adaptive lasso. *arXiv preprint arXiv:2507.10511*, 2025.

Wenjing Zheng and Mark J Van Der Laan. Asymptotic theory for cross-validated targeted maximum likelihood estimation. 2010.

Tijana Zrnic and Emmanuel J. Candès. Cross-prediction-powered inference. *Proceedings of the National Academy of Sciences*, 121(15):e2322083121, 2024.

José R Zubizarreta. Stable weights that balance covariates for estimation with incomplete outcome data. *Journal of the American Statistical Association*, 110(511):910–922, 2015.

Contents

A Additional remarks

Remark 8 (Empirical efficiency maximization as a learning objective)**.** We review Rubin and van der Laan (2008). The empirical variance criterion from Section 2.2,

$$\sigma_n^2(f) := \rho_n \mathbb{P}_n^L \left[\left\{ f(X) - \widehat{\psi}(f) + \rho_n^{-1}(Y - f(X)) \right\}^2 \right] + (1-\rho_n)\mathbb{P}_N^U \left[\left\{ f(\widetilde{X}) - \widehat{\psi}(f) \right\}^2 \right],$$

is not tied to the one-dimensional scaling class $\{\lambda m(X) : \lambda \in \mathbb{R}\}$. For any candidate class $\mathcal{G} \subset L^2(P_{0,X})$, one may instead define

$$\hat{f} \in \arg\min_{f \in \mathcal{G}} \sigma_n^2(f), \qquad \widehat{\psi}_{\text{eem}} := \widehat{\psi}(\hat{f}),$$

so empirical efficiency maximization learns the augmentation $f(X)$ directly.

At the population level, let

$$V(f) := \rho_0 \text{Var}\{D_f^L(X,Y)\} + (1-\rho_0)\text{Var}\{D_f^U(\widetilde{X})\}$$

denote the asymptotic variance of $\widehat{\psi}(f)$. Using equations (1) and (2),

$$V(f) = \mathrm{Var}(Y) + \frac{1-\rho_0}{\rho_0}\mathrm{Var}\{Y - f(X)\}.$$

Equivalently, with $\tilde{f}(X) := f(X) - \mathbb{E}_{P_{0,X}}\{f(X)\} + \psi_0$,

$$V(f) = \mathrm{Var}(Y) + \frac{1-\rho_0}{\rho_0}\mathbb{E}\big[(Y - \tilde{f}(X))^2\big].$$

Thus the efficiency-optimal choice is the $L^2(P_{0,X})$-projection of μ_0 onto the normalized class $\{\tilde{f} : f \in \mathcal{G}\}$. If $\mathcal{G}$ is closed under additive constants, then minimizing $V(f)$ is equivalent to ordinary least-squares regression of Y on X over $\mathcal{G}$. In particular, rather than optimizing the empirical efficiency criterion $\sigma_n^2(f)$ directly, one can equivalently optimize the usual empirical mean-squared-error objective on the labeled sample.

This makes richer efficiency-maximizing classes immediate. For example, if $\mathcal{G} = \{b + h : h \in \mathcal{H},\ b \in \mathbb{R}\}$ for an RKHS $\mathcal{H}$ with kernel K, then a penalized version of empirical efficiency maximization is

$$(\hat{b}_\lambda, \hat{h}_\lambda) \in \arg\min_{b\in\mathbb{R},\ h\in\mathcal{H}} \left\{\mathbb{P}_n^L\big[(Y - b - h(X))^2\big] + \lambda\|h\|_{\mathcal{H}}^2\right\},$$

which is just kernel ridge regression on the labeled sample, up to the irrelevant additive constant in the variance criterion. By the representer theorem,

$$\hat{h}_\lambda(\cdot) = \sum_{i=1}^{n} \hat{\alpha}_i K(X_i, \cdot),$$

and, after centering the kernel or including the intercept explicitly,

$$\hat{\alpha} = (K_n + n\lambda I_n)^{-1}(Y - \hat{b}_\lambda \mathbf{1}_n).$$

Plugging $\hat{f}_\lambda = \hat{b}_\lambda + \hat{h}_\lambda$ into $\widehat{\psi}(\hat{f}_\lambda)$ yields the corresponding efficiency-maximized AIPW estimator.

For tuning and class selection, it is preferable to evaluate the same criterion out of fold. One may split the labeled sample into folds, fit f^{-k} on the training part of fold k, compute the held-out criterion $\sigma_{n,k}^2(f^{-k})$, and average over folds. The selected class or tuning parameter can then be used in a cross-fitted or cross-averaged AIPW estimator. This gives a direct cross-validated empirical-efficiency objective for choosing among classes such as linear, spline, isotonic, or RKHS-based regressors while preserving the usual sample-splitting protection against in-sample overfitting.

Remark 9 (Shrinkage via Venn–Abers)**.** When the labeled sample is small, isotonic calibration may be unstable. A natural alternative is Venn–Abers calibration, originally introduced for binary classification (Vovk and Petej, 2012); see van der Laan and Alaa (2024, 2025) for generalized versions beyond the binary setting. Assume $Y \in [0,1]$, or rescale otherwise. For each x, augment the labeled calibration sample with the hypothetical point $(m(x), y)$ for $y \in \{0,1\}$, fit the resulting isotonic calibrators $f_n^{(x,0)}$ and $f_n^{(x,1)}$, and form the interval-valued prediction

$$\big[f_n^{(x,0)}\{m(x)\},\ f_n^{(x,1)}\{m(x)\}\big].$$

Following Vovk and Petej (2012), one may convert this interval to a point prediction by shrinking its midpoint toward the (unadjusted) AIPW estimator:

$$m^*_{n,\mathrm{VA}}(x) := m^*_{n,\mathrm{mid}}(x) + \big\{f_n^{(x,1)}\{m(x)\} - f_n^{(x,0)}\{m(x)\}\big\}\big\{\widehat{\psi}_{AIPW} - m^*_{n,\mathrm{mid}}(x)\big\},$$

where

$$m^*_{n,\mathrm{mid}}(x) := \frac{f_n^{(x,1)}\{m(x)\} + f_n^{(x,0)}\{m(x)\}}{2}.$$

Since the shrinkage magnitude is proportional to the width of the Venn–Abers interval, this construction can be more stable than plain isotonic calibration in small samples.

Remark 10 (Relation between linear calibration, PPI, and prognostic-score adjustment)**.** Linear calibration can be viewed as prognostic-score regression adjustment, a classical approach to improving precision through flexible covariate adjustment in small-sample settings (Hansen, 2008; Rosenblum and van der Laan, 2009; Lin, 2013; Schuler et al., 2022; Balzer et al., 2024; Højbjerre-Frandsen et al., 2025). Proposition 5 shows that the PPI++ estimator admits the same first-order interpretation. Likewise, the PPI estimator $\widehat{\psi}_{\mathrm{PPI}}$ is algebraically equivalent to the unlabeled-sample plug-in estimator $\mathbb{P}_N^U\{m^\star_{n,\mathrm{const}}(X)\}$ based on the intercept-only adjustment

$$\hat{a}_{\mathrm{const}} \in \arg\min_{a\in\mathbb{R}} \sum_{i=1}^{n}\{Y_i - a - m(X_i)\}^2, \qquad m^\star_{n,\mathrm{const}}(X) := \hat{a}_{\mathrm{const}} + m(X), \tag{22}$$

which calibrates only the mean. By contrast, the standard AIPW estimator $\widehat{\psi}_{\mathrm{AIPW}}$ is equivalent to the pooled-sample mean $\rho_n\mathbb{P}_n^L\{m^\star_{n,\mathrm{const}}(X)\} + (1-\rho_n)\mathbb{P}_N^U\{m^\star_{n,\mathrm{const}}(X)\}$.

B Prediction-Powered Causal Inference via Calibration (i.e., Calibrated DML)

This appendix records the direct translation of the two-sample missing-outcome setup to randomized experiments. The central point is that the semisupervised problem studied in the main text and two-arm causal inference in a randomized trial are the same statistical problem once the unobserved potential outcome in one arm is interpreted as the missing outcome. We then illustrate how our calibrated plug-in estimators apply in this setting, where they arise as special cases of calibrated DML (van der Laan et al., 2024c).

We use the standard potential-outcomes notation. Each unit has two potential outcomes, $Y(1)$ under treatment and $Y(0)$ under control, and the observed outcome satisfies

$$Y = AY(1) + (1-A)Y(0),$$

where $A \in \{0,1\}$ is the treatment assignment. The mean potential outcomes are

$$\mu_1 := \mathbb{E}\{Y(1)\}, \qquad \mu_0 := \mathbb{E}\{Y(0)\},$$

and the average treatment effect is $\tau := \mu_1 - \mu_0$. In a randomized trial, identification follows from the usual causal conditions: consistency and no interference, exchangeability induced by randomization,

$$A \perp (X, Y(1), Y(0)),$$

and positivity, so that each arm has positive assignment probability.

Now partition the observed data into the two arm-specific samples

$$\mathcal{D}_{n,1} := \{(X_i, Y_i) : A_i = 1\}, \qquad \mathcal{D}_{n,0} := \{(\widetilde{X}_j, \widetilde{Y}_j) : A_j = 0\},$$

with sample sizes $n_1 := |\mathcal{D}_{n,1}|$, $n_0 := |\mathcal{D}_{n,0}|$, and $n = n_1 + n_0$. Let $\mathbb{P}_{n,a}$ denote the empirical mean over arm a, and let $\hat{\pi}_a := n_a/n$. To estimate μ_1, one treats $\mathcal{D}_{n,1}$ as the labeled sample for $Y(1)$ and uses the covariates in $\mathcal{D}_{n,0}$ as the unlabeled sample. To estimate μ_0, one reverses the roles of the two arms. Thus each mean potential outcome is a two-sample missing-outcome estimand of exactly the form analyzed in the main text.

For $a \in \{0,1\}$, let $m_a(X)$ be an arm-specific prediction score fit using only observations in arm a. The standard arm-specific AIPW estimator is

$$\hat{\mu}_a^{\mathrm{AIPW}} := \mathbb{P}_n\{m_a(X)\} + \mathbb{P}_n\left[\frac{\mathbb{1}(A=a)}{\hat{\pi}_a}\{Y - m_a(X)\}\right].$$

For $a \in \{0,1\}$, let $m^*_{n,a}(X)$ denote the calibrated version of the score $m_a(X)$. The calibrated estimator of μ_a is then

$$\hat{\mu}_a^{\mathrm{cal}} := \mathbb{P}_n\{m^*_{n,a}(X)\} = \hat{\pi}_a\,\mathbb{P}_{n,a}\{m^*_{n,a}(X)\} + (1-\hat{\pi}_a)\,\mathbb{P}_{n,1-a}\{m^*_{n,a}(X)\}. \tag{23}$$

This is exactly the pooled plug-in estimator from the two-sample formulation, with one arm contributing outcomes and the other contributing only covariates. The same estimator also has the exact AIPW form

$$\hat{\mu}_a^{\mathrm{cal}} = \mathbb{P}_n\{m_{n,a}^*(X)\} + \mathbb{P}_n\left[\frac{\mathbb{1}(A=a)}{\hat{\pi}_a}\{Y - m_{n,a}^*(X)\}\right]. \tag{24}$$

Because

$$\mathbb{P}_n\left[\frac{\mathbb{1}(A=a)}{\hat{\pi}_a}\{Y - m_{n,a}^*(X)\}\right] = \mathbb{P}_{n,a}\{Y - m_{n,a}^*(X)\},$$

the residual correction vanishes exactly whenever the calibrator includes the constant score and therefore enforces mean calibration within arm a. This is the randomized-trial analogue of the exact augmented representation in theorem 1.

The average treatment effect is then estimated by

$$\hat{\tau}^{\mathrm{cal}} := \hat{\mu}_1^{\mathrm{cal}} - \hat{\mu}_0^{\mathrm{cal}}. \tag{25}$$

Equivalently,

$$\hat{\tau}^{\mathrm{cal}} = \mathbb{P}_n\{\tilde{m}_1(X) - \tilde{m}_0(X)\} + \mathbb{P}_n\left[\frac{A}{\hat{\pi}_1}\{Y - \tilde{m}_1(X)\} - \frac{1-A}{\hat{\pi}_0}\{Y - \tilde{m}_0(X)\}\right].$$

Its influence function is the difference of the two arm-specific AIPW influences,

$$D_\tau(O) = \left\{\tilde{m}_1(X) - \mu_1 + \frac{A}{\pi_1}\big(Y - \tilde{m}_1(X)\big)\right\} - \left\{\tilde{m}_0(X) - \mu_0 + \frac{1-A}{\pi_0}\big(Y - \tilde{m}_0(X)\big)\right\}.$$

Accordingly, Wald inference for the ATE may be based on the empirical variance of the estimated influence values divided by n, or on the nonparametric bootstrap refitting the calibration step within each bootstrap sample.

The practical implementation is therefore immediate. One splits the trial into $\mathcal{D}_{n,1}$ and $\mathcal{D}_{n,0}$, treats one arm as labeled and the other as unlabeled when estimating a given mean potential outcome, fits and calibrates an arm-specific score, averages the calibrated predictions over the pooled covariate sample, and finally subtracts the two arm-specific estimates to obtain the ATE. In this way, the semisupervised mean estimators in the main text become direct estimators of mean potential outcomes and treatment effects in randomized experiments.

C General Moment Equations Under MCAR

We briefly review the general AIPW class for moment-equation targets under missing completely at random; see, for example, Robins et al. (1995); van der Laan and Robins (2003). Let

$$O = (X, R, RY),$$

where $R \in \{0,1\}$ indicates whether Y is observed, and suppose

$$R \perp Y \mid X, \qquad \pi_0(X) := \mathbb{P}(R=1 \mid X) > 0 \text{ a.s.}$$

Let the target $\theta_0 \in \Theta \subseteq \mathbb{R}^d$ be identified by

$$\mathbb{E}[\varphi(Y, \theta_0)] = 0,$$

where $\varphi : \mathcal{Y} \times \Theta \to \mathbb{R}^d$ is a known identifying function. Define

$$A_0 := \partial_\theta \mathbb{E}[\varphi(Y,\theta)]\big|_{\theta=\theta_0},$$

and assume A_0 is nonsingular.

Theorem 3 (AIPW class for general moment equations under MCAR)**.** *For any square-integrable function $m : \mathcal{X} \times \Theta \to \mathbb{R}^d$, define*

$$\psi(O; \theta, m, \pi_0) := m(X, \theta) + \frac{R}{\pi_0(X)}\{\varphi(Y, \theta) - m(X, \theta)\}. \tag{26}$$

Then

$$\mathbb{E}[\psi(O; \theta_0, m, \pi_0)] = 0.$$

Hence any estimator $\hat{\theta}$ solving the empirical estimating equation

$$\mathbb{P}_n \psi(O; \hat{\theta}, m, \pi_0) = 0$$

is a valid AIPW-style estimator of θ_0.

Moreover, the corresponding influence function is

$$D_m(O) = -A_0^{-1}\psi(O; \theta_0, m, \pi_0) = -A_0^{-1}\left[m(X, \theta_0) + \frac{R}{\pi_0(X)}\{\varphi(Y, \theta_0) - m(X, \theta_0)\}\right]. \tag{27}$$

Thus the class of regular influence-function-based estimators is indexed by the choice of augmentation function m.

The efficient choice in the full MCAR model is

$$m_0(X, \theta_0) := \mathbb{E}[\varphi(Y, \theta_0) \mid X], \tag{28}$$

which yields the efficient influence function

$$D_{\text{eff}}(O) = -A_0^{-1}\left[m_0(X, \theta_0) + \frac{R}{\pi_0(X)}\{\varphi(Y, \theta_0) - m_0(X, \theta_0)\}\right]. \tag{29}$$

Corollary 3 (Two-sample/PPI form)**.** *Suppose the labeled and unlabeled samples arise from the semisupervised model with constant labeling probability ρ_0, and let $S = f(X)$ be a chosen score. Replacing X by S gives the reduced AIPW class*

$$\psi(O; \theta, m, \rho_0) = m(S, \theta) + \frac{R}{\rho_0}\{\varphi(Y, \theta) - m(S, \theta)\}.$$

Equivalently, the sample estimating equation can be written as

$$0 = \rho_n \mathbb{P}_n^L\{m(S, \theta)\} + (1 - \rho_n)\mathbb{P}_N^U\{m(S, \theta)\} + \mathbb{P}_n^L\{\varphi(Y, \theta) - m(S, \theta)\}.$$

The efficient choice within the reduced score model is

$$m_0(S, \theta_0) = \mathbb{E}[\varphi(Y, \theta_0) \mid S].$$

D Proof of the representation theorem

Proof of theorem 1. Because equation (8) holds for every $h \in \mathcal{F}$, any empirical balancing weight $\widehat{w}(m_n^\star(X)) \in \mathcal{F}$ satisfies

$$\mathbb{P}_n^L\left[\widehat{w}(m_n^\star(X))\{Y - m_n^\star(X)\}\right] = 0.$$

Adding this zero term to the plug-in estimator equation (9) yields equation (10).

For the second claim, since the identity map belongs to $\mathcal{F}$, taking $f(t) = t$ in equation (11) gives

$$\mathbb{P}_n^L\left[\widehat{w}_n^*(m_n^\star(X))\, m_n^\star(X)\right] = \rho_n \mathbb{P}_n^L\{m_n^\star(X)\} + (1 - \rho_n)\mathbb{P}_N^U\{m_n^\star(\widetilde{X})\} = \widehat{\psi}_{\text{cal}}.$$

Since $\widehat{w}_n^* \in \mathcal{F}$, equation (8) also gives

$$\mathbb{P}_n^L\left[\widehat{w}_n^*(m_n^\star(X))\{Y - m_n^\star(X)\}\right] = 0.$$

Therefore,

$$\mathbb{P}_n^L\big[\widehat{w}_n^*(m_n^\star(X))Y\big] = \mathbb{P}_n^L\big[\widehat{w}_n^*(m_n^\star(X))\, m_n^\star(X)\big] + \mathbb{P}_n^L\big[\widehat{w}_n^*(m_n^\star(X))\{Y - m_n^\star(X)\}\big] = \widehat{\psi}_{\mathrm{cal}},$$

which is equation (12). □

Lemma 1 (Blockwise residual orthogonality)**.** *Let $B_1, \ldots, B_J$ be the pooled adjacent violator blocks of the isotonic fit, and let $m_{n,\mathrm{iso}}^\star(X_i) = c_j$ for $i \in B_j$. Then*

$$\sum_{i \in B_j} (Y_i - c_j) = 0, \qquad j = 1, \ldots, J.$$

Consequently, for every measurable function $h : \mathbb{R} \to \mathbb{R}$ such that $h(m_{n,\mathrm{iso}}^\star(X_i))$ is finite for all i,

$$\sum_{i=1}^n h(m_{n,\mathrm{iso}}^\star(X_i))\{Y_i - m_{n,\mathrm{iso}}^\star(X_i)\} = 0.$$

Proof. The fitted values from isotonic regression are constant on pooled adjacent violator blocks. The KKT conditions imply that residuals sum to zero on each block. Since $h(m_{n,\mathrm{iso}}^\star(X_i)) = h(c_j)$ for all $i \in B_j$,

$$\sum_{i=1}^n h(m_{n,\mathrm{iso}}^\star(X_i))\{Y_i - m_{n,\mathrm{iso}}^\star(X_i)\} = \sum_{j=1}^J h(c_j) \sum_{i \in B_j} (Y_i - c_j) = 0.$$

□

Proof of proposition 3. The orthogonality claim equation (13) is exactly the conclusion of Lemma 1. For the risk comparison, note that the identity map belongs to $\mathcal{F}_{\mathrm{iso}}$. Since $\hat{f}$ minimizes the empirical squared loss over $\mathcal{F}_{\mathrm{iso}}$,

$$\mathbb{P}_n^L\big[(Y - m_{n,\mathrm{iso}}^\star(X))^2\big] \le \mathbb{P}_n^L\big[(Y - \hat{f}_{\mathrm{id}}(m(X)))^2\big] = \mathbb{P}_n^L\big[(Y - m(X))^2\big],$$

where $\hat{f}_{\mathrm{id}}(t) = t$. □

Proof of proposition 4. The normal equations for equation (14) are

$$\mathbb{P}_n^L\{Y - m_{n,\mathrm{lin}}^\star(X)\} = 0 \qquad \text{and} \qquad \mathbb{P}_n^L\big[m(X)\{Y - m_{n,\mathrm{lin}}^\star(X)\}\big] = 0.$$

Because $m_{n,\mathrm{lin}}^\star(X) = \hat{a}_{\mathrm{lin}} m(X) + \hat{b}_{\mathrm{lin}}$, these imply

$$\mathbb{P}_n^L\big[m_{n,\mathrm{lin}}^\star(X)\{Y - m_{n,\mathrm{lin}}^\star(X)\}\big] = 0.$$

Hence, for any $\hat{w}(X) = a + b\, m_{n,\mathrm{lin}}^\star(X)$,

$$\mathbb{P}_n^L\big[\hat{w}(X)\{Y - m_{n,\mathrm{lin}}^\star(X)\}\big] = a\, \mathbb{P}_n^L\{Y - m_{n,\mathrm{lin}}^\star(X)\} + b\, \mathbb{P}_n^L\big[m_{n,\mathrm{lin}}^\star(X)\{Y - m_{n,\mathrm{lin}}^\star(X)\}\big] = 0.$$

Adding this zero term to equation (15) yields equation (16). □

E Proofs of first-order equivalence between PPI++ and linear calibration

Recall that

$$(\hat{a}_{\mathrm{lin}}, \hat{b}_{\mathrm{lin}}) \in \arg\min_{a,b \in \mathbb{R}} \sum_{i=1}^n \{Y_i - a\, m(X_i) - b\}^2,$$

and define the corresponding linear-calibration estimator by

$$\widehat{\psi}_{\mathrm{lin}} = \rho_n \mathbb{P}_n^L\{\hat{a}_{\mathrm{lin}} m(X) + \hat{b}_{\mathrm{lin}}\} + (1 - \rho_n)\mathbb{P}_N^U\{\hat{a}_{\mathrm{lin}} m(\widetilde{X}) + \hat{b}_{\mathrm{lin}}\}.$$

Also let $\widehat{\psi}_{++}$ denote the PPI++ estimator obtained by minimizing the empirical influence-function variance over the scaling class $\{\lambda m(X) : \lambda \in \mathbb{R}\}$, and let $\hat{\lambda}_{++}$ be the resulting selected coefficient. Define

$$\Delta_n := \mathbb{P}_N^U m(\widetilde{X}) - \mathbb{P}_n^L m(X),$$

and

$$a_0 := \frac{\mathrm{Cov}\{Y, m(X)\}}{\mathrm{Var}\{m(X)\}}.$$

Proof of Proposition 5. The normal equations for the linear regression fit are

$$\mathbb{P}_n^L\{Y - \hat{a}_{\mathrm{lin}} m(X) - \hat{b}_{\mathrm{lin}}\} = 0$$

and

$$\mathbb{P}_n^L\big[m(X)\{Y - \hat{a}_{\mathrm{lin}} m(X) - \hat{b}_{\mathrm{lin}}\}\big] = 0.$$

The first equation gives

$$\hat{b}_{\mathrm{lin}} = \mathbb{P}_n^L Y - \hat{a}_{\mathrm{lin}} \mathbb{P}_n^L m(X).$$

Substituting this into the second equation yields

$$\hat{a}_{\mathrm{lin}} = \frac{\mathbb{P}_n^L\big[(m(X) - \mathbb{P}_n^L m(X))(Y - \mathbb{P}_n^L Y)\big]}{\mathbb{P}_n^L\big[(m(X) - \mathbb{P}_n^L m(X))^2\big]}.$$

Hence $\hat{a}_{\mathrm{lin}}$ is the ordinary least-squares slope, and under finite second moments,

$$\hat{a}_{\mathrm{lin}} = a_0 + O_p(n^{-1/2}).$$

Using

$$\hat{b}_{\mathrm{lin}} = \mathbb{P}_n^L Y - \hat{a}_{\mathrm{lin}} \mathbb{P}_n^L m(X),$$

we obtain

$$\begin{aligned}
\widehat{\psi}_{\mathrm{lin}} &= \rho_n \mathbb{P}_n^L\{\hat{a}_{\mathrm{lin}} m(X) + \hat{b}_{\mathrm{lin}}\} + (1-\rho_n)\mathbb{P}_N^U\{\hat{a}_{\mathrm{lin}} m(\widetilde{X}) + \hat{b}_{\mathrm{lin}}\} \\
&= \hat{b}_{\mathrm{lin}} + \hat{a}_{\mathrm{lin}}\Big[\rho_n \mathbb{P}_n^L m(X) + (1-\rho_n)\mathbb{P}_N^U m(\widetilde{X})\Big] \\
&= \mathbb{P}_n^L Y - \hat{a}_{\mathrm{lin}} \mathbb{P}_n^L m(X) + \hat{a}_{\mathrm{lin}}\Big[\rho_n \mathbb{P}_n^L m(X) + (1-\rho_n)\mathbb{P}_N^U m(\widetilde{X})\Big] \\
&= \mathbb{P}_n^L Y + (1-\rho_n)\hat{a}_{\mathrm{lin}}\big\{\mathbb{P}_N^U m(\widetilde{X}) - \mathbb{P}_n^L m(X)\big\} \\
&= \mathbb{P}_n^L Y + (1-\rho_n)\hat{a}_{\mathrm{lin}}\Delta_n.
\end{aligned}$$

On the other hand,

$$\begin{aligned}
\widehat{\psi}_{++}(\hat{\lambda}_{++}) &= \rho_n \mathbb{P}_n^L\{\hat{\lambda}_{++} m(X)\} + (1-\rho_n)\mathbb{P}_N^U\{\hat{\lambda}_{++} m(\widetilde{X})\} + \mathbb{P}_n^L\{Y - \hat{\lambda}_{++} m(X)\} \\
&= \mathbb{P}_n^L Y + (1-\rho_n)\hat{\lambda}_{++}\big\{\mathbb{P}_N^U m(\widetilde{X}) - \mathbb{P}_n^L m(X)\big\} \\
&= \mathbb{P}_n^L Y + (1-\rho_n)\hat{\lambda}_{++}\Delta_n.
\end{aligned}$$

Therefore,

$$\widehat{\psi}_{++}(\hat{\lambda}_{++}) - \widehat{\psi}_{\mathrm{lin}} = (1-\rho_n)\big(\hat{\lambda}_{++} - \hat{a}_{\mathrm{lin}}\big)\Delta_n.$$

Under the empirical PPI++ variance criterion over the class $\{\lambda m(X) : \lambda \in \mathbb{R}\}$, the selected coefficient satisfies

$$\hat{\lambda}_{++} = \frac{\mathbb{P}_n^L\big[(Y - \mathbb{P}_n^L Y)(m(X) - \mathbb{P}_n^L m(X))\big]}{(1-\rho_n)\mathbb{P}_n^L\big[(m(X) - \mathbb{P}_n^L m(X))^2\big] + \rho_n \mathbb{P}_N^U\big[(m(\widetilde{X}) - \mathbb{P}_N^U m(\widetilde{X}))^2\big]}.$$

Its population limit is

$$\lambda_0 = \frac{\mathrm{Cov}\{Y, m(X)\}}{\mathrm{Var}\{m(X)\}} = a_0,$$

so standard M-estimation arguments yield

$$\hat{\lambda}_{++} = a_0 + O_p(n^{-1/2}).$$

Combining this with $\hat{a}_{\mathrm{lin}} = a_0 + O_p(n^{-1/2})$ gives

$$\hat{\lambda}_{++} - \hat{a}_{\mathrm{lin}} = O_p(n^{-1/2}).$$

Next, since $\mathbb{P}_n^L m(X)$ and $\mathbb{P}_N^U m(\widetilde{X})$ are sample means under the same marginal law of X,

$$\Delta_n = O_p(n^{-1/2} + N^{-1/2}).$$

Under the standing assumption $\rho_n = n/(n+N) \to \rho_0 \in (0,1)$, we may write

$$n^{-1/2} = \rho_n^{-1/2}(n+N)^{-1/2}, \qquad N^{-1/2} = (1-\rho_n)^{-1/2}(n+N)^{-1/2},$$

and hence

$$\Delta_n = O_p\Big(\Big\{\rho_n^{-1/2} + (1-\rho_n)^{-1/2}\Big\}(n+N)^{-1/2}\Big).$$

Therefore,

$$\begin{aligned}
\widehat{\psi}_{++}(\hat{\lambda}_{++}) - \widehat{\psi}_{\mathrm{lin}} &= (1-\rho_n)\big(\hat{\lambda}_{++} - \hat{a}_{\mathrm{lin}}\big)\Delta_n \\
&= (1-\rho_n)\, O_p(n^{-1/2})\, O_p\Big(\Big\{\rho_n^{-1/2} + (1-\rho_n)^{-1/2}\Big\}(n+N)^{-1/2}\Big) \\
&= O_p\Big((1-\rho_n)\rho_n^{-1/2}\Big\{\rho_n^{-1/2} + (1-\rho_n)^{-1/2}\Big\}(n+N)^{-1}\Big).
\end{aligned}$$

Since

$$(1-\rho_n)\rho_n^{-1/2}\Big\{\rho_n^{-1/2} + (1-\rho_n)^{-1/2}\Big\} = \frac{1-\rho_n}{\rho_n} + \sqrt{\frac{1-\rho_n}{\rho_n}},$$

It follows that

$$\widehat{\psi}_{++}(\hat{\lambda}_{++}) - \widehat{\psi}_{\mathrm{lin}} = O_p\left(\left\{\frac{1-\rho_n}{\rho_n} + \sqrt{\frac{1-\rho_n}{\rho_n}}\right\}(n+N)^{-1}\right).$$

In particular, this implies the simpler bound

$$\widehat{\psi}_{++}(\hat{\lambda}_{++}) - \widehat{\psi}_{\mathrm{lin}} = O_p\left(\frac{1-\rho_n}{\rho_n}(n+N)^{-1}\right).$$

Therefore, under $\rho_n \to \rho_0 \in (0,1)$,

$$\widehat{\psi}_{++}(\hat{\lambda}_{++}) - \widehat{\psi}_{\mathrm{lin}} = O_p\big((n+N)^{-1}\big) = o_p\big((n+N)^{-1/2}\big).$$

This proves the claim. □

Proposition 7 (PPI and AIPW as intercept-only calibration)**.** *Let*

$$\hat{a}_{\mathrm{const}} \in \arg\min_{a\in\mathbb{R}} \sum_{i=1}^{n} \{Y_i - a - m(X_i)\}^2, \qquad m^{\star}_{n,\mathrm{const}}(X) := \hat{a}_{\mathrm{const}} + m(X).$$

Then

$$\hat{a}_{\mathrm{const}} = \mathbb{P}_n^L\{Y - m(X)\}.$$

Consequently, the PPI estimator

$$\widehat{\psi}_{\mathrm{PPI}} := \mathbb{P}_N^U\{m(\widetilde{X})\} + \mathbb{P}_n^L\{Y - m(X)\}$$

satisfies

$$\widehat{\psi}_{\mathrm{PPI}} = \mathbb{P}_N^U\{m^\star_{n,\mathrm{const}}(\widetilde{X})\},$$

while the standard AIPW estimator

$$\widehat{\psi}_{\mathrm{AIPW}} := \rho_n \mathbb{P}_n^L\{m(X)\} + (1-\rho_n)\mathbb{P}_N^U\{m(\widetilde{X})\} + \mathbb{P}_n^L\{Y - m(X)\}$$

satisfies

$$\widehat{\psi}_{\mathrm{AIPW}} = \rho_n \mathbb{P}_n^L\{m^\star_{n,\mathrm{const}}(X)\} + (1-\rho_n)\mathbb{P}_N^U\{m^\star_{n,\mathrm{const}}(\widetilde{X})\}.$$

Hence PPI is the unlabeled-only plug-in estimator based on the mean-calibrated score $m^\star_{n,\mathrm{const}}$, whereas AIPW is the corresponding pooled plug-in estimator.

Proof. The objective in the definition of $\hat{a}_{\mathrm{const}}$ is quadratic in a, and its first-order condition is

$$0 = \mathbb{P}_n^L\{Y - a - m(X)\}.$$

Therefore

$$\hat{a}_{\mathrm{const}} = \mathbb{P}_n^L\{Y - m(X)\}.$$

Using $m^\star_{n,\mathrm{const}}(\widetilde{X}) = m(\widetilde{X}) + \hat{a}_{\mathrm{const}}$, we obtain

$$\mathbb{P}_N^U\{m^\star_{n,\mathrm{const}}(\widetilde{X})\} = \mathbb{P}_N^U\{m(\widetilde{X})\} + \hat{a}_{\mathrm{const}} = \mathbb{P}_N^U\{m(\widetilde{X})\} + \mathbb{P}_n^L\{Y - m(X)\} = \widehat{\psi}_{\mathrm{PPI}}.$$

Likewise,

$$\begin{aligned}
\rho_n \mathbb{P}_n^L\{m^\star_{n,\mathrm{const}}(X)\} + (1-\rho_n)\mathbb{P}_N^U\{m^\star_{n,\mathrm{const}}(\widetilde{X})\} &= \rho_n \mathbb{P}_n^L\{m(X) + \hat{a}_{\mathrm{const}}\} + (1-\rho_n)\mathbb{P}_N^U\{m(\widetilde{X}) + \hat{a}_{\mathrm{const}}\} \\
&= \rho_n \mathbb{P}_n^L\{m(X)\} + (1-\rho_n)\mathbb{P}_N^U\{m(\widetilde{X})\} + \hat{a}_{\mathrm{const}} \\
&= \rho_n \mathbb{P}_n^L\{m(X)\} + (1-\rho_n)\mathbb{P}_N^U\{m(\widetilde{X})\} + \mathbb{P}_n^L\{Y - m(X)\} \\
&= \widehat{\psi}_{\mathrm{AIPW}}.
\end{aligned}$$

This proves the claim. □

F Proofs of the efficient influence pairs

Proof of proposition 2. Under Assumption 1, a straightforward calculation shows that the asymptotic variance over the class equations (1) and (2) is minimized at $f = \mu_0$. We now show that the resulting efficient influence pair aligns with the classical efficient influence function, or canonical gradient, in the two-sample model (Bickel et al., 1993; Kennedy, 2024).

Fix P_0, and write $P_{0,X}$ for its X-marginal and $\mu_0(x) := \mathbb{E}_{P_0}[Y \mid X = x]$. The observed two-sample experiment is

$$\mathbb{D}^{(n,N)}(P) = P^{\otimes n} \otimes P_X^{\otimes N}.$$

In the local asymptotic limit with labeled fraction $\rho_0 \in (0,1)$, the relevant Hilbert space is

$$\mathbb{H} := L_0^2(P_0) \times L_0^2(P_{0,X}),$$

equipped with inner product

$$\big\langle (u^L, u^U), (v^L, v^U) \big\rangle = \rho_0\, \mathbb{E}_{P_0}[u^L v^L] + (1-\rho_0)\, \mathbb{E}_{P_{0,X}}[u^U v^U].$$

Let P_ε be a regular parametric submodel through P_0, and factorize

$$p_\varepsilon(x,y) = p_{X,\varepsilon}(x)\, p_{Y|X,\varepsilon}(y \mid x).$$

Write

$$a(x) := \partial_\varepsilon \log p_{X,\varepsilon}(x)\big|_{\varepsilon=0}, \qquad b(x,y) := \partial_\varepsilon \log p_{Y|X,\varepsilon}(y \mid x)\big|_{\varepsilon=0}.$$

Then

$$\mathbb{E}_{P_{0,X}}[a(X)] = 0, \qquad \mathbb{E}_{P_0}[b(X,Y) \mid X] = 0,$$

and the induced score in the two-sample experiment is

$$\big(a(X) + b(X,Y),\, a(\widetilde{X})\big).$$

Hence the tangent space is

$$\mathcal{T}_{\text{full}} = \Big\{(a+b,\ a) : a \in L_0^2(P_{0,X}),\ b \in L_0^2(P_0),\ \mathbb{E}[b(X,Y) \mid X] = 0\Big\}.$$

The parameter is

$$\psi(P) = \mathbb{E}_P[Y] = \mathbb{E}_{P_X}[\mu_P(X)].$$

Its pathwise derivative along P_ε is

$$\partial_\varepsilon \psi(P_\varepsilon)\big|_{\varepsilon=0} = \mathbb{E}_{P_0}[Y\{a(X) + b(X,Y)\}].$$

Using the tower property,

$$\mathbb{E}_{P_0}[Y a(X)] = \mathbb{E}_{P_{0,X}}[\mu_0(X) a(X)],$$

so

$$\dot{\psi}_{P_0}(a,b) = \mathbb{E}_{P_{0,X}}[\mu_0(X)a(X)] + \mathbb{E}_{P_0}[Y\, b(X,Y)].$$

We claim that the Riesz representer of this linear functional on $\mathcal{T}_{\text{full}}$ is

$$D_{\text{eff}}^L(X,Y) = \mu_0(X) - \psi_0 + \rho_0^{-1}\{Y - \mu_0(X)\}, \qquad D_{\text{eff}}^U(\widetilde{X}) = \mu_0(\widetilde{X}) - \psi_0.$$

Indeed, both components are mean zero, and for any $(a+b, a) \in \mathcal{T}_{\text{full}}$,

$$\begin{aligned}
&\big\langle (D_{\text{eff}}^L, D_{\text{eff}}^U), (a+b, a) \big\rangle \\
&= \rho_0\, \mathbb{E}_{P_0}\big[\big\{\mu_0(X) - \psi_0 + \rho_0^{-1}(Y - \mu_0(X))\big\}\,\{a(X) + b(X,Y)\}\big] + (1-\rho_0)\, \mathbb{E}_{P_{0,X}}[(\mu_0(X) - \psi_0)a(X)].
\end{aligned}$$

Now

$$\mathbb{E}_{P_0}[(\mu_0(X) - \psi_0) b(X,Y)] = 0$$

because $\mathbb{E}[b(X,Y) \mid X] = 0$, and

$$\mathbb{E}_{P_0}[(Y - \mu_0(X))a(X)] = 0$$

because $\mathbb{E}[Y - \mu_0(X) \mid X] = 0$. Therefore

$$\begin{aligned}
\big\langle (D_{\text{eff}}^L, D_{\text{eff}}^U), (a+b, a) \big\rangle &= \mathbb{E}_{P_{0,X}}[(\mu_0(X) - \psi_0)a(X)] + \mathbb{E}_{P_0}[(Y - \mu_0(X))b(X,Y)] \\
&= \mathbb{E}_{P_{0,X}}[\mu_0(X)a(X)] + \mathbb{E}_{P_0}[Y\, b(X,Y)] \\
&= \dot{\psi}_{P_0}(a,b).
\end{aligned}$$

Thus $(D_{\text{eff}}^L, D_{\text{eff}}^U)$ is the canonical gradient, hence the efficient influence pair. This proves equations (5) and (6). □

Proof of proposition 6. Let

$$S := m_0(X), \qquad \eta_0(s) := \mathbb{E}[Y \mid S = s].$$

Write Q_0 for the law of (S, Y), and $Q_{0,S}$ for the marginal law of S. In the reduced two-sample experiment, the labeled sample is i.i.d. from Q_0 and the unlabeled sample is i.i.d. from $Q_{0,S}$. The parameter remains

$$\psi(Q) = \mathbb{E}_Q[Y].$$

As in the full-model case, the Hilbert space is

$$\mathbb{H}_{\text{red}} := L_0^2(Q_0) \times L_0^2(Q_{0,S}),$$

with inner product

$$\big\langle (u^L, u^U), (v^L, v^U) \big\rangle = \rho_0 \, \mathbb{E}_{Q_0}[u^L v^L] + (1 - \rho_0) \, \mathbb{E}_{Q_{0,S}}[u^U v^U].$$

Let Q_ε be a regular submodel through Q_0, and factorize

$$q_\varepsilon(s, y) = q_{S,\varepsilon}(s) \, q_{Y|S,\varepsilon}(y \mid s).$$

Define the marginal and conditional scores

$$a(s) := \partial_\varepsilon \log q_{S,\varepsilon}(s)\big|_{\varepsilon=0}, \qquad b(s, y) := \partial_\varepsilon \log q_{Y|S,\varepsilon}(y \mid s)\big|_{\varepsilon=0}.$$

Then

$$\mathbb{E}_{Q_{0,S}}[a(S)] = 0, \qquad \mathbb{E}_{Q_0}[b(S, Y) \mid S] = 0,$$

and the tangent space is

$$\mathcal{T}_{\text{red}} = \Big\{ (a + b, \; a) : a \in L_0^2(Q_{0,S}), \; b \in L_0^2(Q_0), \; \mathbb{E}[b(S, Y) \mid S] = 0 \Big\}.$$

The pathwise derivative of $\psi(Q) = \mathbb{E}_Q[Y]$ is

$$\partial_\varepsilon \psi(Q_\varepsilon)|_{\varepsilon=0} = \mathbb{E}_{Q_0}[Y\{a(S) + b(S, Y)\}] = \mathbb{E}_{Q_{0,S}}[\eta_0(S) a(S)] + \mathbb{E}_{Q_0}[Y \, b(S, Y)].$$

Therefore the canonical gradient in the reduced model is

$$D^L_{\text{red}}(S, Y) = \eta_0(S) - \psi_0 + \rho_0^{-1}\{Y - \eta_0(S)\}, \qquad D^U_{\text{red}}(\widetilde{S}) = \eta_0(\widetilde{S}) - \psi_0.$$

The verification is identical to the full-model case: for any $(a + b, a) \in \mathcal{T}_{\text{red}}$,

$$\begin{aligned}
&\big\langle (D^L_{\text{red}}, D^U_{\text{red}}), (a + b, a) \big\rangle \\
&= \rho_0 \, \mathbb{E}_{Q_0}\big[\big\{\eta_0(S) - \psi_0 + \rho_0^{-1}(Y - \eta_0(S))\big\} \{a(S) + b(S, Y)\}\big] + (1 - \rho_0) \, \mathbb{E}_{Q_{0,S}}[(\eta_0(S) - \psi_0) a(S)] \\
&= \mathbb{E}_{Q_{0,S}}[\eta_0(S) a(S)] + \mathbb{E}_{Q_0}[Y \, b(S, Y)],
\end{aligned}$$

since

$$\mathbb{E}[(\eta_0(S) - \psi_0) b(S, Y)] = 0, \qquad \mathbb{E}[(Y - \eta_0(S)) a(S)] = 0.$$

This equals the pathwise derivative, so the displayed pair is the efficient influence pair in the reduced model.

Finally, at the calibrated truth we have

$$\eta_0(S) = \mathbb{E}[Y \mid S] = S = m_0(X) \qquad \text{a.s.}$$

Substituting $S = m_0(X)$ gives

$$D^L_{m_0}(X, Y) = m_0(X) - \psi_0 + \rho_0^{-1}\{Y - m_0(X)\}, \qquad D^U_{m_0}(\widetilde{X}) = m_0(\widetilde{X}) - \psi_0,$$

which is exactly equations (20) and (21). □

Lemma 2 (Empirical calibration centers the limit)**.** *Suppose Assumptions 1 and 2 hold and the calibration score equations contain the constant score $h \equiv 1$. Then*

$$\mathbb{E}_{P_{0,X}}[m_0(X)] = \psi_0.$$

Proof. Let $T := m(X)$, and write $m_0 = f_0(T)$, where

$$f_0 \in \arg\min_{f \in \mathcal{F}_{\text{iso}}} \mathbb{E}_{P_0}\big[(Y - f(T))^2\big].$$

Because $\mathcal{F}_{\text{iso}}$ is closed under addition of constants, for every $c \in \mathbb{R}$ the function $f_0 + c$ also belongs to $\mathcal{F}_{\text{iso}}$. Hence the function

$$g(c) := \mathbb{E}_{P_0}\big[(Y - f_0(T) - c)^2\big]$$

is minimized at $c = 0$. Since g is differentiable,

$$0 = g'(0) = -2\,\mathbb{E}_{P_0}[Y - f_0(T)].$$

Therefore

$$\mathbb{E}_{P_0}[Y] = \mathbb{E}_{P_0}[f_0(T)].$$

Recalling that $m_0(X) = f_0(m(X)) = f_0(T)$ and $\psi_0 = \mathbb{E}_{P_0}[Y]$, we conclude

$$\mathbb{E}_{P_{0,X}}[m_0(X)] = \psi_0.$$

□

Proof convention. In the following proofs, fix $\delta > 0$. By $\sup_x |m^\star_{n,\text{iso}}(x)| = O_p(1)$, there exist a deterministic constant $C = C_\delta \geq C_0$ and $n_0 < \infty$ such that the event

$$A_{n,C} := \Big\{\sup_x |m^\star_{n,\text{iso}}(x)| \leq C\Big\}$$

satisfies $P_0(A_{n,C}) \geq 1 - \delta$ for all $n \geq n_0$. We carry out the deterministic bounded-class arguments on $A_{n,C}$, where all implicit constants may depend on C but not on n. Since δ is arbitrary, the resulting rates and distributional statements hold unconditionally.

Lemma 3 (Isotonic L^2 rate)**.** *Suppose Assumptions 1 and 2 holds. Then*

$$\|m^\star_{n,\text{iso}} - m_0\|^2_{2,P_{0,X}} = O_p(n^{-2/3}).$$

Proof of Lemma 3. Let $T := m(X)$ and let

$$\mathcal{M}_C := \{f(T) : f \in \mathcal{F}_{\text{iso}},\ \|f(T)\|_\infty \leq C\}.$$

By Assumption 2, $g_0(T)$ is bounded by $C_0 \leq C$, and on $A_{n,C}$ we have $m^\star_{n,\text{iso}} \in \mathcal{M}_C$. Since clipping any monotone candidate to $[-C, C]$ preserves monotonicity and can only decrease squared risk against $g_0(T)$, the population isotonic projection m_0 also belongs to $\mathcal{M}_C$. Thus the problem is a one-dimensional isotonic least-squares regression of Y on the score T. Let

$$g_0(T) := \mathbb{E}[Y \mid T].$$

Then, for every $m \in \mathcal{M}_C$,

$$P_0\{(Y - m)^2\} = P_0\{(Y - g_0)^2\} + P_0\{(g_0 - m)^2\},$$

which shows that m_0 is the $L^2(P_T)$-projection of g_0 onto the closed convex set $\mathcal{M}_C$.

Now define the centered class

$$\mathcal{F}_0 := \{g = m - m_0 :\ m \in \mathcal{M}_C\}.$$

For $g \in \mathcal{F}_0$, let $\ell_g(Z) := (Y - m_0(T) - g(T))^2 - (Y - m_0(T))^2$, where $Z = (X, Y)$. Since $m^\star_{n,\text{iso}}$ minimizes the empirical squared risk over $\mathcal{F}_{\text{iso}}$, on $A_{n,C}$ the random element $\hat{g} := m^\star_{n,\text{iso}} - m_0$ belongs to $\mathcal{F}_0$ and satisfies

$$\mathbb{P}^L_n \ell_{\hat{g}} \leq 0 = \mathbb{P}^L_n \ell_0.$$

Moreover,

$$P_0\ell_g = \|g\|^2_{2,P_{0,X}} - 2P_0\left[(Y - m_0)g\right].$$

Because g is measurable with respect to T, the tower property gives

$$P_0\left[(Y - m_0)g\right] = P_0\left[(g_0 - m_0)g\right].$$

Since m_0 is the $L^2(P_T)$-projection of g_0 onto the convex set $\mathcal{M}_C$, the Hilbert-space projection inequality yields

$$P_0\left[(g_0 - m_0)(m - m_0)\right] \le 0 \qquad \text{for every } m \in \mathcal{M}_C.$$

Hence

$$P_0\ell_g \ge \|g\|^2_{2,P_{0,X}},$$

so squared loss has the required quadratic margin around m_0.

We next bound the modulus of continuity of the empirical process indexed by the localized classes

$$\mathcal{F}_\delta := \{g \in \mathcal{F}_0 : \|g\|_{2,P_{0,X}} \le \delta\}.$$

Because every $m \in \mathcal{M}_C$ is bounded by C, every $g \in \mathcal{F}_\delta$ is uniformly bounded, and the standard bracketing bound for one-dimensional monotone classes gives

$$\log N_{[]}(\varepsilon, \mathcal{F}_\delta, L_2(P_{0,X})) \lesssim \delta/\varepsilon, \qquad 0 < \varepsilon < \delta.$$

Therefore

$$J_{[]}(\delta, \mathcal{F}_\delta, L_2(P_{0,X})) \lesssim \delta^{1/2};$$

see, e.g., van der Vaart and Wellner (1996).

Write $\xi := Y - m_0(T)$, so that $\ell_g = g^2 - 2\xi g$. The class $\{g^2 : g \in \mathcal{F}_\delta\}$ is a bounded Lipschitz image of $\mathcal{F}_\delta$, so by Lemma 3.4.2 of van der Vaart and Wellner (1996) and the same entropy bound,

$$E^* \sup_{g\in\mathcal{F}_\delta} \sqrt{n}\,|(\mathbb{P}_n^L - P_0)(g^2)| \lesssim \delta^{1/2}.$$

For the multiplier term, Assumption 2 implies that $Y - g_0(T)$ is sub-Gaussian or subexponential. Since both $g_0(T)$ and $m_0(T)$ are bounded by C, the difference $g_0(T) - m_0(T)$ is bounded, and therefore $\xi = Y - m_0(T)$ is also sub-Gaussian or subexponential. Thus the product class $\xi\mathcal{F}_\delta := \{\xi g : g \in \mathcal{F}_\delta\}$ satisfies the same localized bracketing bound in the Bernstein-type norm used in the subexponential least-squares proof of Bibaut and van der Laan (2019), yielding

$$E^* \sup_{g\in\mathcal{F}_\delta} \sqrt{n}\,|(\mathbb{P}_n^L - P_0)(\xi g)| \lesssim \delta^{1/2}.$$

Combining the previous two displays,

$$E^* \sup_{g\in\mathcal{F}_\delta} \sqrt{n}\,|(\mathbb{P}_n^L - P_0)\ell_g| \lesssim \delta^{1/2}.$$

We may therefore apply Theorem 3.4.1 of van der Vaart and Wellner (1996) with $d(g) = \|g\|_{2,P_{0,X}}$, quadratic margin $P_0\ell_g \gtrsim d^2(g)$, and modulus $\phi_n(\delta) \asymp \delta^{1/2}$. The fixed-point condition $r_n^2\phi_n(r_n^{-1}) \lesssim \sqrt{n}$ gives $r_n \asymp n^{1/3}$, and hence

$$\|m^\star_{n,\mathrm{iso}} - m_0\|_{2,P_{0,X}} = O_p(n^{-1/3}), \qquad \|m^\star_{n,\mathrm{iso}} - m_0\|^2_{2,P_{0,X}} = O_p(n^{-2/3}).$$

The preceding display is obtained on $A_{n,C}$. Since $P_0(A^c_{n,C}) \le \delta$ for all large n, and $\delta > 0$ was arbitrary, the same rate holds unconditionally. This is the classical one-dimensional isotonic rate, stated directly in the random-design $L^2(P_{0,X})$ norm relevant for the theorem. □

Proof of Lemma 2. Because $h \equiv 1$ belongs to the calibration score class, equation (8) gives

$$\mathbb{P}_n^L\{Y - m^\star_{n,\mathrm{iso}}(X)\} = 0.$$

Hence

$$\mathbb{P}_n^L Y - \mathbb{P}_n^L m_0(X) = \mathbb{P}_n^L\{m^\star_{n,\mathrm{iso}}(X) - m_0(X)\}.$$

By Cauchy–Schwarz and Lemma 3,

$$P_{0,X}\left|m^\star_{n,\mathrm{iso}} - m_0\right| \le \|m^\star_{n,\mathrm{iso}} - m_0\|_{2,P_{0,X}} = O_p(n^{-1/3}),$$

and Lemma 4 gives

$$(\mathbb{P}_n^L - P_{0,X})\{m^\star_{n,\mathrm{iso}} - m_0\} = O_p(n^{-2/3}).$$

Therefore

$$\mathbb{P}_n^L\{m^\star_{n,\mathrm{iso}}(X) - m_0(X)\} = P_{0,X}\{m^\star_{n,\mathrm{iso}} - m_0\} + (\mathbb{P}_n^L - P_{0,X})\{m^\star_{n,\mathrm{iso}} - m_0\} = o_p(1).$$

By Assumption 1, the law of large numbers yields $\mathbb{P}_n^L Y \to_p \psi_0$ and $\mathbb{P}_n^L m_0(X) \to_p \mathbb{E}_{P_{0,X}}[m_0(X)]$. Therefore $\mathbb{E}_{P_{0,X}}[m_0(X)] = \psi_0$. □

Lemma 4 (Centered empirical-process bounds for isotonic calibration)**.** *Suppose Assumptions 1 and 2 holds. Then*

$$(\mathbb{P}_n^L - P_{0,X})\{m^\star_{n,\mathrm{iso}}(X) - m_0(X)\} = O_p(n^{-2/3}), \qquad (\mathbb{P}_N^U - P_{0,X})\{m^\star_{n,\mathrm{iso}}(\widetilde{X}) - m_0(\widetilde{X})\} = O_p\big(n^{-1/3}N^{-1/2}\big).$$

Proof of Lemma 4. By Lemma 3, the isotonic fit satisfies

$$\|m^\star_{n,\mathrm{iso}} - m_0\|^2_{2,P_{0,X}} = O_p(n^{-2/3}).$$

To control the centered empirical-process terms, consider the localized class

$$\mathcal{F}_{n,C} := \{g = f \circ m - m_0 : \ f \in \mathcal{F}_{\mathrm{iso}}, \ \|f \circ m\|_\infty \le C, \ \|g\|^2_{2,P_{0,X}} \lesssim n^{-2/3}\}.$$

On $A_{n,C}$, the function $m^\star_{n,\mathrm{iso}} - m_0$ belongs to $\mathcal{F}_{n,C}$. Moreover, $\mathcal{F}_{n,C}$ has a bounded envelope because both $f \circ m$ and m_0 are bounded by C. Because m is fixed, $\mathcal{F}_{n,C}$ is a localized class of bounded monotone transformations of a one-dimensional score, and standard entropy bounds for monotone classes imply a finite entropy integral. Applying a local maximal inequality for empirical processes, such as Lemma 3.4.2 of van der Vaart and Wellner (1996), at localization radius $n^{-1/3}$ then gives, on $A_{n,C}$,

$$|(\mathbb{P}_n^L - P_{0,X})\{m^\star_{n,\mathrm{iso}} - m_0\}| = O_p(n^{-2/3}).$$

Since $P_0(A^c_{n,C}) \le \delta$ for all large n, and $\delta > 0$ was arbitrary, the same bound holds unconditionally. For the unlabeled sample, conditional on the labeled data, $m^\star_{n,\mathrm{iso}} - m_0$ is fixed and the unlabeled sample is independent, so

$$(\mathbb{P}_N^U - P_{0,X})\{m^\star_{n,\mathrm{iso}} - m_0\} = O_p\left(N^{-1/2}\|m^\star_{n,\mathrm{iso}} - m_0\|_{2,P_{0,X}}\right) = O_p\big(n^{-1/3}N^{-1/2}\big)$$

by Lemma 3. □

G Proof of asymptotic linearity

Proof of theorem 2. By Lemma 2, $P_{0,X} m_0 = \psi_0$. Using theorem 1 with $\widehat{w} \equiv 1$,

$$\widehat{\psi}_{\mathrm{iso}} = \rho_n \mathbb{P}_n^L\{m^\star_{n,\mathrm{iso}}(X)\} + (1-\rho_n)\mathbb{P}_N^U\{m^\star_{n,\mathrm{iso}}(\widetilde{X})\} + \rho_n \mathbb{P}_n^L\{Y - m^\star_{n,\mathrm{iso}}(X)\}.$$

Add and subtract the AIPW form built from m_0:

$$\widehat{\psi}_{\text{iso}} = \rho_n \mathbb{P}_n^L \left\{ m_0 + \rho_n^{-1}(Y - m_0) \right\} + (1-\rho_n)\mathbb{P}_N^U\{m_0(\widetilde{X})\} + (1-\rho_n)(\mathbb{P}_N^U - \mathbb{P}_n^L)\{m_{n,\text{iso}}^\star - m_0\}.$$

Subtracting ψ_0 gives

$$\begin{aligned}\widehat{\psi}_{\text{iso}} - \psi_0 = \rho_n(\mathbb{P}_n^L - P_0)\Big\{ m_0 - \psi_0 + \rho_n^{-1}(Y - m_0) \Big\} \\ + (1-\rho_n)(\mathbb{P}_N^U - P_{0,X})(m_0 - \psi_0) \\ + (1-\rho_n)(\mathbb{P}_N^U - \mathbb{P}_n^L)\{m_{n,\text{iso}}^\star - m_0\}.\end{aligned}$$

To replace ρ_n^{-1} by ρ_0^{-1}, note that

$$\rho_n(\mathbb{P}_n^L - P_0)\Big\{ (\rho_n^{-1} - \rho_0^{-1})(Y - m_0) \Big\} = \Big(1 - \frac{\rho_n}{\rho_0}\Big)(\mathbb{P}_n^L - P_0)(Y - m_0).$$

The right-hand side is $o_p(M^{-1/2})$ because $\rho_n \to \rho_0$ and $(\mathbb{P}_n^L - P_0)(Y - m_0) = O_p(n^{-1/2}) = O_p(M^{-1/2})$. Also,

$$(1-\rho_n)(\mathbb{P}_N^U - \mathbb{P}_n^L)\{m_{n,\text{iso}}^\star - m_0\} = (1-\rho_n)(\mathbb{P}_N^U - P_{0,X})\{m_{n,\text{iso}}^\star - m_0\} - (1-\rho_n)(\mathbb{P}_n^L - P_{0,X})\{m_{n,\text{iso}}^\star - m_0\},$$

so the labeled term is $O_p(n^{-2/3})$ and the unlabeled term is $O_p(n^{-1/3}N^{-1/2})$ by Lemma 4. Thus the remainder $R_{n,N}$, defined as the difference between $\widehat{\psi}_{\text{iso}} - \psi_0$ and the two leading empirical-process terms in equation (19), satisfies the displayed bound in theorem 2. Since each component is $o_p(M^{-1/2})$, this also yields equation (19). □

Proof of corollary 1. By theorem 2,

$$\sqrt{M}\,(\widehat{\psi}_{\text{iso}} - \psi_0) = \frac{1}{\sqrt{M}} \sum_{i=1}^{n} D_{m_0}^L(X_i, Y_i) + \frac{1}{\sqrt{M}} \sum_{j=1}^{N} D_{m_0}^U(\widetilde{X}_j) + o_p(1).$$

The labeled and unlabeled samples are independent, each summand has mean zero, and $\rho_n \to \rho_0 \in (0,1)$ by Assumption 1. A triangular-array central limit theorem therefore gives the stated asymptotic normal law with variance σ_0^2.

For the variance estimator, let $g_n := m_{n,\text{iso}}^\star - m_0$. By Lemma 3,

$$P_{0,X}(g_n^2) = \|g_n\|_{2,P_{0,X}}^2 = O_p(n^{-2/3}).$$

Moreover, the proof of Lemma 3 already established the localized empirical-process bound

$$\sup_{g \in \mathcal{F}_\delta} |(\mathbb{P}_n^L - P_{0,X})(g^2)| = O_p(n^{-1/2}\delta^{1/2}),$$

for the squared class $\{g^2 : g \in \mathcal{F}_\delta\}$. Taking $\delta \asymp n^{-1/3}$, which matches the localization radius of g_n, gives

$$(\mathbb{P}_n^L - P_{0,X})(g_n^2) = O_p(n^{-2/3}).$$

Hence

$$\mathbb{P}_n^L(g_n^2) = O_p(n^{-2/3}) = o_p(1).$$

On the boundedness event $A_{n,C}$, conditional on the labeled data, g_n is fixed and $|g_n| \le 2C$, so

$$(\mathbb{P}_N^U - P_{0,X})(g_n^2) = O_p\left(N^{-1/2}\|g_n^2\|_{2,P_{0,X}}\right) = O_p\left(C\,N^{-1/2}\|g_n\|_{2,P_{0,X}}\right) = O_p\big(n^{-1/3}N^{-1/2}\big).$$

Therefore

$$\mathbb{P}_N^U(g_n^2) = O_p(n^{-2/3}) + O_p\big(n^{-1/3}N^{-1/2}\big) = o_p(1).$$

Since $P_0(A_{n,C}^c) \le \delta$ for all large n and arbitrary $\delta > 0$, these bounds hold unconditionally. Also, $\widehat{\psi}_{\text{iso}} \to_p \psi_0$ follows from theorem 2. Finally,

$$\widehat{D}^L - D_{m_0}^L = (1 - \rho_n^{-1})g_n - (\widehat{\psi}_{\text{iso}} - \psi_0) + (\rho_n^{-1} - \rho_0^{-1})(Y - m_0),$$

and

$$\widehat{D}^U - D^U_{m_0} = g_n(\widetilde{X}) - (\widehat{\psi}_{\text{iso}} - \psi_0).$$

Because $\rho_n \to \rho_0$, $Y - m_0$ has finite second moment, and $\mathbb{P}^L_n(g_n^2), \mathbb{P}^U_N(g_n^2) = o_p(1)$, we obtain

$$\widehat{D}^L_i - D^L_{m_0}(X_i, Y_i) = o_p(1) \qquad \text{and} \qquad \widehat{D}^U_j - D^U_{m_0}(\widetilde{X}_j) = o_p(1)$$

in empirical L^2, which yields $\widehat{\sigma}^2 \to_p \sigma_0^2$. The Wald interval claim then follows by Slutsky's theorem. □

H Proof of reduced-model efficiency

Proof of corollary 2. Part (i) follows because, under the conditions of theorem 2, the influence pair in theorem 2 is exactly the pair in proposition 6, noting that

$$\psi_0 = \mathbb{E}_{P_0}[Y] = \mathbb{E}_{P_0}[m_0(X)],$$

since m_0 is the population mean calibration. Part (ii) follows because, if $m_0 = \mu_0$ almost surely, then the same influence pair also coincides with the efficient influence pair in proposition 2. Part (iii) follows because monotonicity of $\mathbb{E}[Y \mid m(X)]$ in $m(X)$ identifies the score-based regression with the isotonic target m_0, so the reduced model indexed by m_0 is the same as the model indexed by the original score $m(X)$. □

I Pooled i.i.d. Missing-Data Formulation

The joint two-sample setup can be embedded into a pooled i.i.d. model by introducing a sample-membership indicator $R \in \{0, 1\}$, where $R = 1$ denotes labeled observations and $R = 0$ denotes unlabeled observations. In that formulation one observes i.i.d. data $O = (X, R, RY)$ with $\mathbb{P}(R = 1) = \rho_0$, and the calibrated estimator becomes

$$\widehat{\psi}_{\text{cal}} = \mathbb{P}_M(m_n^\star) + \mathbb{P}_M\left[\frac{R}{\rho_0}\{Y - m_n^\star\}\right]$$

up to the same calibration identity as in the main text. The corresponding observed-data efficient influence function is

$$m_0(X) - \psi_0 + \frac{R}{\rho_0}\{Y - m_0(X)\},$$

which is the single-sample analogue of the reduced-model influence pair in proposition 6. Thus the usual i.i.d. missing-data formulation is a convenient special case, but the main text keeps the two-dataset structure explicit.

J Additional empirical details

This appendix records the benchmark construction used in section 5.2. We use the official `ppi_py` real-data mean-estimation examples and preserve their labeled-sample-size grids. The `forest` experiment uses $n \in \{50, 100, \ldots, 500\}$. The `galaxies` and semisupervised `census_income` benchmarks use $n \in \{50, 155, 261, 366, 472, 577, 683, 788, 894, 1000\}$. For each n, the remaining observations form the unlabeled sample. In the current draft we rerandomize this split 500 times at each n, always evaluating all estimators on the same split and taking the full-sample mean as the benchmark truth.

The estimator set mirrors the original PPI comparison wherever possible. In the paper-facing summaries, we report the labeled-only estimator, the imputation benchmark on binary-outcome tasks, PPI computed using the official `ppi_py` implementation, PPI++, and the classical AIPW estimator based on the same prediction score. We then add five calibration-oriented comparators: linear calibration; smooth monotone spline calibration; Platt scaling on the binary-outcome benchmarks; isotonic calibration with a fixed minimum bin size of 10; and an adaptive `Auto` selector that chooses among AIPW, linear calibration, monotone spline calibration,

and isotonic calibration by cross-validated empirical efficiency. For the binary-outcome benchmarks, we also report Venn–Abers shrinkage. Platt scaling and Venn–Abers are omitted on `census_income` because that outcome is not binary. Although PPI++ and prognostic-score adjustment lie on the same first-order correction path as linear calibration in this one-dimensional score setting, the current reproduced real-data results show a more mixed finite-sample picture, so we retain PPI++ in the displayed summaries primarily as a direct empirical reference point.

For the fixed calibration-based plug-in estimators, we fit the calibrator on the labeled pairs $\{(m(X_i), Y_i)\}_{i=1}^n$, transform both the labeled and unlabeled scores, and then compute the pooled plug-in mean together with the same semisupervised influence-style Wald interval used throughout the benchmark code. AIPW uses the same score but averages it over the pooled covariate sample before adding the labeled residual correction, whereas PPI uses only the unlabeled score average. In theory this retains less labeled-sample score information, though in the finite-sample benchmark results the empirical variance gap is often small and not uniform across all datasets. Linear calibration regresses Y on an intercept and the score, then clips the fitted values to the range of predicted outcomes observed in the labeled data to prevent extrapolation on the unlabeled data. `MonoSpline` instead fits a smooth nondecreasing spline of Y on the score, using monotonicity constraints together with a mild roughness penalty, and then applies the same clipped pooled plug-in construction. Platt scaling fits a logistic regression of the binary outcome on a stabilized logit transform of the score, while the fixed-bin isotonic estimator fits a stepwise monotone recalibration map and also clips to the labeled outcome range. The `Auto` estimator is a selector rather than a new calibrator: in the experiments it compares AIPW, linear calibration, monotone spline calibration, and fixed-bin isotonic calibration using 20-fold cross-validation and the estimated influence-function variance as the selection criterion. To keep this step inexpensive when the unlabeled sample is very large, the foldwise criterion is evaluated on an unlabeled subsample of size $\min(N, 10n)$. The reported summaries are Monte Carlo averages of bias, empirical variance, MSE, interval coverage, and relative efficiency versus PPI across the repeated random splits. The complete numerical outputs, including dataset-level tables and per-metric plots, are generated automatically by the accompanying reproduction pipeline.

K `Python` Code

Code availability. A public repository containing the `ppi_aipw` package, experiment scripts, and paper assets is available at github.com/Larsvanderlaan/ppi-aipw. Package documentation and worked examples are hosted at larsvanderlaan.github.io/ppi-aipw/.

Below, we provide a minimal, self-contained implementation of the method introduced in this paper.

K.1 Isotonic regression via XGBoost

```
import numpy as np
import xgboost as xgb

def isoreg_with_xgboost(x, y, max_depth=15, min_child_weight=20, weights=None):
    """
    Fit a monotone calibrator f so that f(x) is nondecreasing in x.

    Parameters
    ----------
    x : array-like, shape (n,) or (n, 1)
        Predictor used for calibration. In the PPI application this is usually
        the score m(X), so x is typically one-dimensional.
    y : array-like, shape (n,)
        Labeled outcomes.
    max_depth : int, default=15
        Maximum depth for the one-round XGBoost fit.
    min_child_weight : float, default=20
```

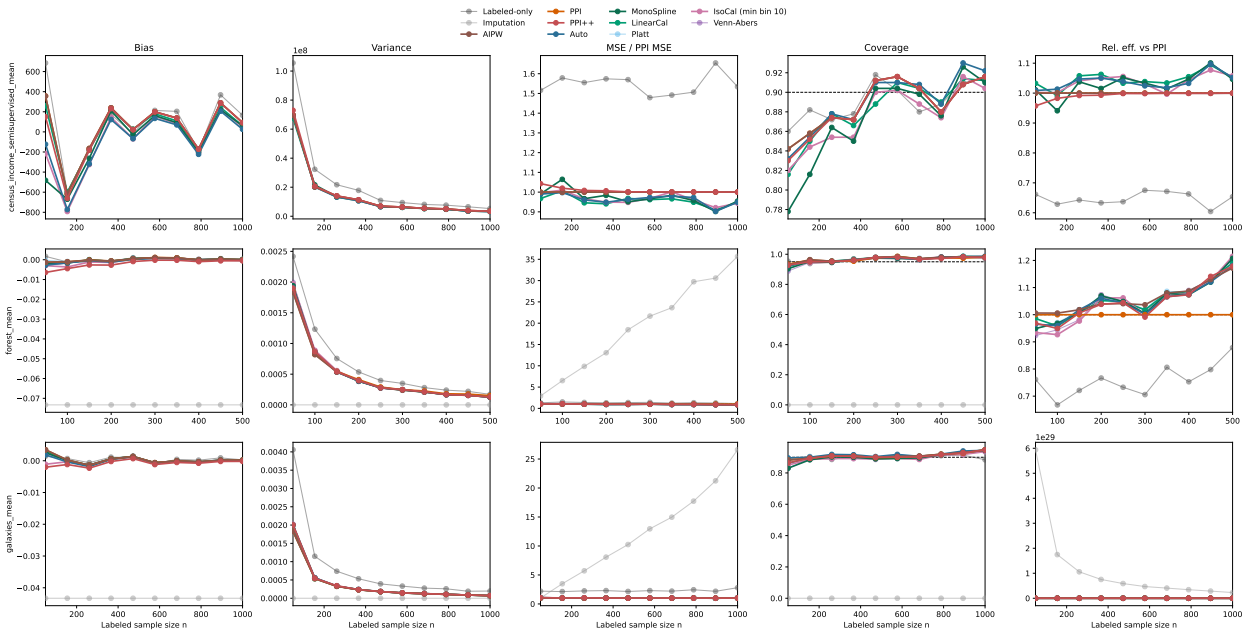

Figure 3: Full diagnostic grid for the reproduced PPI benchmarks, showing bias, empirical variance, normalized MSE relative to PPI, coverage, and relative efficiency versus PPI across labeled-sample-size regimes. In the normalized-MSE and relative-efficiency panels, the dashed horizontal line marks parity with PPI; in the coverage panels it marks the nominal target coverage $1 - \alpha$. `Auto` and `MonoSpline` appear alongside PPI, PPI++, AIPW, and the fixed calibration rules, while Platt scaling and Venn–Abers appear only on the binary-outcome datasets where they are well defined.

```
    Minimum total weight in a terminal node.
weights : array-like, optional
    Optional observation weights.

Returns
-------
predict_fn : callable
    Function mapping new x values to calibrated predictions f(x).
"""
x = np.asarray(x)
y = np.asarray(y).reshape(-1)
x = x.reshape(len(y), -1)

data = xgb.DMatrix(data=x, label=y, weight=weights)
params = {
    "max_depth": max_depth,
    "min_child_weight": min_child_weight,
    "monotone_constraints": "(" + ",".join(["1"] * x.shape[1]) + ")",
    "eta": 1.0,
    "gamma": 0.0,
    "lambda": 0.0,
    "objective": "reg:squarederror",
    "verbosity": 0,
}
iso_fit = xgb.train(params=params, dtrain=data, num_boost_round=1)

def predict_fn(x_new):
    x_new = np.asarray(x_new)
```

```
        if x_new.ndim == 1:
            x_new = x_new.reshape(-1, 1)
        data_pred = xgb.DMatrix(data=x_new)
        return iso_fit.predict(data_pred)

    return predict_fn
```

K.2 Isotonic-calibrated plug-in

```
import numpy as np
from statistics import NormalDist


def _resolve_scores(
    Y_labeled,
    X_labeled=None,
    X_unlabeled=None,
    score_fn=None,
    score_labeled=None,
    score_unlabeled=None,
):
    Y_labeled = np.asarray(Y_labeled).reshape(-1)

    if score_fn is not None:
        if X_labeled is None or X_unlabeled is None:
            raise ValueError(
                "If score_fn is provided, then X_labeled and X_unlabeled "
                "must also be provided."
            )
        score_labeled = score_fn(X_labeled)
        score_unlabeled = score_fn(X_unlabeled)
    elif score_labeled is None or score_unlabeled is None:
        raise ValueError(
            "Provide either score_fn together with X_labeled/X_unlabeled, "
            "or provide score_labeled and score_unlabeled."
        )

    score_labeled = np.asarray(score_labeled).reshape(-1)
    score_unlabeled = np.asarray(score_unlabeled).reshape(-1)

    if len(score_labeled) != len(Y_labeled):
        raise ValueError("Y_labeled and score_labeled must have the same length.")

    return Y_labeled, score_labeled, score_unlabeled


def isotonic_calibrated_plugin(
    Y_labeled,
    X_labeled=None,
    X_unlabeled=None,
    score_fn=None,
    score_labeled=None,
    score_unlabeled=None,
    alpha=0.05,
    min_obs=20,
):
    Y_labeled, score_labeled, score_unlabeled = _resolve_scores(
        Y_labeled=Y_labeled,
        X_labeled=X_labeled,
```

```
        X_unlabeled=X_unlabeled,
        score_fn=score_fn,
        score_labeled=score_labeled,
        score_unlabeled=score_unlabeled,
    )

    n = len(Y_labeled)
    N = len(score_unlabeled)
    M = n + N
    rho = n / M

    calibrator = isoreg_with_xgboost(
        x=score_labeled,
        y=Y_labeled,
        min_child_weight=min_obs,
    )

    m_labeled = calibrator(score_labeled)
    m_unlabeled = calibrator(score_unlabeled)

    psi_hat = (n * m_labeled.mean() + N * m_unlabeled.mean()) / M

    if_labeled = m_labeled - psi_hat + (Y_labeled - m_labeled) / rho
    if_unlabeled = m_unlabeled - psi_hat
    var_hat = (
        rho * np.var(if_labeled, ddof=1)
        + (1 - rho) * np.var(if_unlabeled, ddof=1)
    )
    se_hat = np.sqrt(var_hat / M)

    z = NormalDist().inv_cdf(1 - alpha / 2)
    ci = (psi_hat - z * se_hat, psi_hat + z * se_hat)

    return {
        "estimate": psi_hat,
        "standard_error": se_hat,
        "confidence_interval": ci,
    }


# Example:
# fit = isotonic_calibrated_plugin(
#     Y_labeled=Y,
#     X_labeled=X,
#     X_unlabeled=X_tilde,
#     score_fn=m,
# )
```

K.3 Linear calibration

```
import numpy as np
from statistics import NormalDist


def _resolve_scores(
    Y_labeled,
    X_labeled=None,
    X_unlabeled=None,
    score_fn=None,
```

```
    score_labeled=None,
    score_unlabeled=None,
):
    Y_labeled = np.asarray(Y_labeled).reshape(-1)

    if score_fn is not None:
        if X_labeled is None or X_unlabeled is None:
            raise ValueError(
                "If score_fn is provided, then X_labeled and X_unlabeled "
                "must also be provided."
            )
        score_labeled = score_fn(X_labeled)
        score_unlabeled = score_fn(X_unlabeled)
    elif score_labeled is None or score_unlabeled is None:
        raise ValueError(
            "Provide either score_fn together with X_labeled/X_unlabeled, "
            "or provide score_labeled and score_unlabeled."
        )

    score_labeled = np.asarray(score_labeled).reshape(-1)
    score_unlabeled = np.asarray(score_unlabeled).reshape(-1)

    if len(score_labeled) != len(Y_labeled):
        raise ValueError("Y_labeled and score_labeled must have the same length.")

    return Y_labeled, score_labeled, score_unlabeled


def linear_calibrated_plugin(
    Y_labeled,
    X_labeled=None,
    X_unlabeled=None,
    score_fn=None,
    score_labeled=None,
    score_unlabeled=None,
    alpha=0.05,
):
    Y_labeled, score_labeled, score_unlabeled = _resolve_scores(
        Y_labeled=Y_labeled,
        X_labeled=X_labeled,
        X_unlabeled=X_unlabeled,
        score_fn=score_fn,
        score_labeled=score_labeled,
        score_unlabeled=score_unlabeled,
    )

    n = len(Y_labeled)
    N = len(score_unlabeled)
    M = n + N
    rho = n / M

    X_design = np.column_stack([np.ones(n), score_labeled])
    beta_hat, _, _, _ = np.linalg.lstsq(X_design, Y_labeled, rcond=None)

    def calibrator(score):
        score = np.asarray(score).reshape(-1)
        return beta_hat[0] + beta_hat[1] * score

    m_labeled = calibrator(score_labeled)
    m_unlabeled = calibrator(score_unlabeled)
```

```
    psi_hat = (n * m_labeled.mean() + N * m_unlabeled.mean()) / M

    if_labeled = m_labeled - psi_hat + (Y_labeled - m_labeled) / rho
    if_unlabeled = m_unlabeled - psi_hat
    var_hat = (
        rho * np.var(if_labeled, ddof=1)
        + (1 - rho) * np.var(if_unlabeled, ddof=1)
    )
    se_hat = np.sqrt(var_hat / M)

    z = NormalDist().inv_cdf(1 - alpha / 2)
    ci = (psi_hat - z * se_hat, psi_hat + z * se_hat)

    return {
        "estimate": psi_hat,
        "standard_error": se_hat,
        "confidence_interval": ci,
    }


# Example:
# fit = linear_calibrated_plugin(
#     Y_labeled=Y,
#     score_labeled=mX,
#     score_unlabeled=mX_tilde,
# )
```

Dataset	n	N	Estimator	Bias	Variance	MSE	Coverage	RelEff
census_income_semisupervised_mean	50	380041	AIPW	356.4905	69836665.4410	69963750.9366	0.842	1.000
census_income_semisupervised_mean	50	380041	Auto	-123.2501	69242944.1124	69258134.7078	0.832	1.009
census_income_semisupervised_mean	50	380041	Labeled-only	685.3532	105558402.2427	106028111.3137	0.860	0.662
census_income_semisupervised_mean	50	380041	IsoCal (min bin 10)	-215.3832	69697400.2434	69743790.1536	0.820	1.002
census_income_semisupervised_mean	50	380041	LinearCal	260.9294	67622601.4006	67690685.5430	0.816	1.033
census_income_semisupervised_mean	50	380041	MonoSpline	-482.0477	69082916.1270	69315286.1508	0.778	1.011
census_income_semisupervised_mean	50	380041	PPI	356.4473	69834865.9486	69961920.5977	0.842	1.000
census_income_semisupervised_mean	50	380041	PPI++	150.8664	72953053.8328	72975814.4980	0.830	0.957
census_income_semisupervised_mean	577	379514	AIPW	199.8136	6353884.5833	6393810.0648	0.916	1.000
census_income_semisupervised_mean	577	379514	Auto	134.3746	6201029.6459	6219086.1828	0.910	1.024
census_income_semisupervised_mean	577	379514	Labeled-only	213.2465	9407595.6062	9453069.6946	0.902	0.675
census_income_semisupervised_mean	577	379514	IsoCal (min bin 10)	155.0503	6162964.7883	6187005.3977	0.902	1.031
census_income_semisupervised_mean	577	379514	LinearCal	179.8005	6116832.9552	6149161.1594	0.910	1.038
census_income_semisupervised_mean	577	379514	MonoSpline	162.1952	6146069.6864	6172376.9541	0.904	1.033
census_income_semisupervised_mean	577	379514	PPI	199.7932	6351836.6166	6391753.9369	0.916	1.000
census_income_semisupervised_mean	577	379514	PPI++	198.7409	6362047.6345	6401545.5819	0.916	0.998
census_income_semisupervised_mean	1000	379091	AIPW	92.1091	3462051.4638	3470535.5425	0.916	0.999
census_income_semisupervised_mean	1000	379091	Auto	25.8207	3289467.9320	3290134.6381	0.922	1.052
census_income_semisupervised_mean	1000	379091	Labeled-only	165.4763	5293296.2822	5320678.6872	0.914	0.654
census_income_semisupervised_mean	1000	379091	IsoCal (min bin 10)	51.2697	3271622.4413	3274251.0223	0.904	1.058
census_income_semisupervised_mean	1000	379091	LinearCal	60.7501	3287796.5296	3291487.1058	0.912	1.052
census_income_semisupervised_mean	1000	379091	MonoSpline	60.9626	3306575.6859	3310292.1289	0.910	1.046
census_income_semisupervised_mean	1000	379091	PPI	91.9155	3460071.7259	3468520.1894	0.916	1.000
census_income_semisupervised_mean	1000	379091	PPI++	91.9155	3460071.7259	3468520.1894	0.916	1.000
forest_mean	50	1546	AIPW	-0.0012	0.0018	0.0018	0.928	1.006
forest_mean	50	1546	Auto	-0.0028	0.0019	0.0019	0.924	0.965
forest_mean	50	1546	Labeled-only	0.0017	0.0024	0.0024	0.954	0.762
forest_mean	50	1546	Imputation	-0.0733	0.0000	0.0054	0.000	inf
forest_mean	50	1546	IsoCal (min bin 10)	-0.0030	0.0020	0.0020	0.930	0.934
forest_mean	50	1546	LinearCal	-0.0016	0.0019	0.0019	0.918	0.986
forest_mean	50	1546	MonoSpline	-0.0021	0.0019	0.0019	0.904	0.949
forest_mean	50	1546	Platt	-0.0003	0.0019	0.0019	0.900	0.948
forest_mean	50	1546	PPI	-0.0013	0.0018	0.0018	0.934	1.000
forest_mean	50	1546	PPI++	-0.0064	0.0019	0.0019	0.928	0.969
forest_mean	50	1546	Venn-Abers	-0.0027	0.0020	0.0020	0.888	0.923
forest_mean	300	1296	AIPW	0.0011	0.0002	0.0002	0.984	1.036
forest_mean	300	1296	Auto	0.0006	0.0002	0.0002	0.978	1.002
forest_mean	300	1296	Labeled-only	0.0007	0.0003	0.0004	0.968	0.706
forest_mean	300	1296	Imputation	-0.0733	0.0000	0.0054	0.000	inf
forest_mean	300	1296	IsoCal (min bin 10)	0.0010	0.0002	0.0002	0.978	1.000
forest_mean	300	1296	LinearCal	0.0010	0.0002	0.0002	0.984	1.018
forest_mean	300	1296	MonoSpline	0.0010	0.0002	0.0002	0.974	1.004
forest_mean	300	1296	Platt	0.0010	0.0002	0.0002	0.982	1.020
forest_mean	300	1296	PPI	0.0011	0.0002	0.0002	0.986	1.000
forest_mean	300	1296	PPI++	-0.0002	0.0002	0.0002	0.980	0.991
forest_mean	300	1296	Venn-Abers	0.0011	0.0002	0.0002	0.978	1.015
forest_mean	500	1096	AIPW	0.0001	0.0001	0.0001	0.980	1.172
forest_mean	500	1096	Auto	-0.0003	0.0001	0.0001	0.986	1.180
forest_mean	500	1096	Labeled-only	0.0001	0.0002	0.0002	0.972	0.879
forest_mean	500	1096	Imputation	-0.0733	0.0000	0.0054	0.000	inf
forest_mean	500	1096	IsoCal (min bin 10)	-0.0000	0.0001	0.0001	0.986	1.215
forest_mean	500	1096	LinearCal	0.0001	0.0001	0.0001	0.982	1.191
forest_mean	500	1096	MonoSpline	0.0000	0.0001	0.0001	0.984	1.207
forest_mean	500	1096	Platt	0.0001	0.0001	0.0001	0.984	1.203
forest_mean	500	1096	PPI	0.0000	0.0002	0.0002	0.982	1.000
forest_mean	500	1096	PPI++	-0.0006	0.0001	0.0001	0.978	1.179
forest_mean	500	1096	Venn-Abers	-0.0001	0.0001	0.0001	0.978	1.206
galaxies_mean	50	16693	AIPW	0.0035	0.0018	0.0018	0.882	1.000
galaxies_mean	50	16693	Auto	0.0019	0.0018	0.0018	0.896	0.997
galaxies_mean	50	16693	Labeled-only	0.0014	0.0041	0.0041	0.890	0.451
galaxies_mean	50	16693	Imputation	-0.0433	0.0000	0.0019	0.000	594072276610619419150610595840.000
galaxies_mean	50	16693	IsoCal (min bin 10)	-0.0011	0.0020	0.0020	0.850	0.908
galaxies_mean	50	16693	LinearCal	0.0030	0.0019	0.0019	0.862	0.973
galaxies_mean	50	16693	MonoSpline	0.0026	0.0020	0.0020	0.830	0.917
galaxies_mean	50	16693	Platt	0.0033	0.0019	0.0019	0.856	0.966
galaxies_mean	50	16693	PPI	0.0035	0.0018	0.0018	0.882	1.000
galaxies_mean	50	16693	PPI++	-0.0020	0.0020	0.0020	0.864	0.914
galaxies_mean	50	16693	Venn-Abers	0.0022	0.0020	0.0020	0.830	0.907
galaxies_mean	577	16166	AIPW	-0.0006	0.0001	0.0001	0.906	1.004
galaxies_mean	577	16166	Auto	-0.0008	0.0001	0.0001	0.918	0.989
galaxies_mean	577	16166	Labeled-only	-0.0011	0.0003	0.0003	0.910	0.436
galaxies_mean	577	16166	Imputation	-0.0433	0.0000	0.0019	0.000	468490623120291195838330306560.000
galaxies_mean	577	16166	IsoCal (min bin 10)	-0.0007	0.0001	0.0001	0.894	0.986
galaxies_mean	577	16166	LinearCal	-0.0007	0.0001	0.0001	0.908	1.005
galaxies_mean	577	16166	MonoSpline	-0.0006	0.0001	0.0001	0.892	0.988
galaxies_mean	577	16166	Platt	-0.0007	0.0001	0.0001	0.896	1.002
galaxies_mean	577	16166	PPI	-0.0006	0.0001	0.0001	0.900	1.000
galaxies_mean	577	16166	PPI++	-0.0012	0.0001	0.0001	0.902	0.988
galaxies_mean	577	16166	Venn-Abers	-0.0006	0.0001	0.0001	0.892	0.986
galaxies_mean	1000	15743	AIPW	0.0002	0.0001	0.0001	0.948	0.987
galaxies_mean	1000	15743	Auto	0.0001	0.0001	0.0001	0.944	0.977
galaxies_mean	1000	15743	Labeled-only	0.0003	0.0002	0.0002	0.884	0.357
galaxies_mean	1000	15743	Imputation	-0.0433	0.0000	0.0019	0.000	228666320208537975913615196l6.000
galaxies_mean	1000	15743	IsoCal (min bin 10)	0.0001	0.0001	0.0001	0.936	0.951
galaxies_mean	1000	15743	LinearCal	0.0002	0.0001	0.0001	0.946	0.980
galaxies_mean	1000	15743	MonoSpline	0.0002	0.0001	0.0001	0.944	0.969
galaxies_mean	1000	15743	Platt	0.0002	0.0001	0.0001	0.942	0.979
galaxies_mean	1000	15743	PPI	0.0002	0.0001	0.0001	0.948	1.000
galaxies_mean	1000	15743	PPI++	-0.0002	0.0001	0.0001	0.946	0.977
galaxies_mean	1000	15743	Venn-Abers	0.0002	0.0001	0.0001	0.942	0.952

Table 2: Representative numerical summary at the smallest, middle, and largest labeled sample sizes for each reproduced benchmark. The reported quantities are Monte Carlo bias, empirical variance of the point estimator, MSE, Wald-interval coverage, and relative efficiency versus PPI for the displayed benchmark set, including PPI++, `Auto`, and `MonoSpline`.