TRANSFER CALIBRATED PREDICTION POWERED INFERENCE

Anonymous authors
Paper under double-blind review

ABSTRACT

Prediction-powered inference (PPI), and its power-tuned extension (PPI++), improve confidence intervals by combining a small gold-standard labeled sample with a large AI model's predictions. Its efficiency gain relies on low residual variance, which may not hold if the predictor is pre-trained on a different source domain. We propose Transfer Calibrated Prediction-Powered Inference (TC-PPI), adapting the source-domain predictor to the target domain using gold-standard samples through cross-fitting. This approach supports various adaptation methods, such as sparse linear calibration, LoRA, and fine-tuning. Our jointly tuned cross-fit estimator, Joint-TC-Cross-PPI++, maintains unbiasedness and is simultaneously at least as efficient as classical inference, PPI, and PPI++, therefore protecting against negative transfer. We provide high-dimensional MSE bounds for calibration and show empirical improvements over baseline methods across various real-world applications.

1 INTRODUCTION

Reliable statistical inference is often difficult in domains with scarce amounts of "gold-standard" (i.e., labeled) data. In such settings, one may have a small set of labeled data (X_i, Y_i) along with a much larger set of N unlabeled data points, $\tilde{X}_j$, and a pretrained black-box AI model to produce predictions. With just the gold samples, we can perform valid inference, but this can lead to larger errors or more conservative intervals, due to the small sample size. Conversely, relying on the unlabeled data and the model's predictions with them is invalid due to potential bias or inaccurate predictions. A popular framework for performing reliable inference using both types of data is the Prediction-Powered Inference (PPI) framework (Angelopoulos et al., 2023a;b; Zrnic & Candès, 2024; Boyeau et al., 2024; Cowen-Breen et al., 2026; Mani et al., 2026; Nath et al., 2025) and closely related semi-supervised inference literature (Zhang & Bradic, 2022). PPI uses the gold-standard data to rectify the model's predictions on the larger unlabeled dataset, enabling more efficient inference while maintaining validity. For mean estimation, the PPI estimator is

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

If X_i and $\tilde{X}_i$ are i.i.d., $\hat{\theta}^{\text{PP}}$ is unbiased regardless of f, and has smaller variance than the classical estimator whenever f predicts Y reasonably well, since $N \gg n$ and $Y_i - f(X_i)$ has smaller variance than Y_i. This benefit can deteriorate in the presence of domain shifts.

Often, the available predictor is a source-domain model f_s trained on data distinct from the target population. While the PPI estimator remains valid when adjusted with target gold data, its efficiency can decline significantly if the residual variance $\text{Var}_t(Y - f_s(X))$ is large. To redress, we

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

introduce the *Transfer Calibrated Prediction-Powered Inference* (TC-PPI) framework. TC-PPI employs transfer-learning (Zhuang et al., 2020; Tripuraneni et al., 2020; Day & Khoshgoftaar, 2017) to calibrate the source predictor using target-domain data.

Core Methodology. We develop in stages, beginning with TC-PPI and its power-tuned version (Angelopoulos et al., 2023b), TC-PPI++, which reduces PPI residual variance via transfer calibration. Next, we introduce a joint power-tuning step that keeps the source predictor and transfer correction separate, ensuring robustness against ineffective corrections. We prove a "no-harm" guarantee (Theorem 3), showing that the Joint TC estimator with an optimal tuning parameter is at least as effective as classical methods. Finally, we integrate joint tuning with cross-fitting, utilizing all gold observations for rectification. This yields our Joint TC–Cross-PPI++ estimator, which significantly improves performance, as using the full gold sample is crucial for effective rectification.

Key Differentiators. Unlike recent studies (Bastani, 2021; Li et al., 2022; Chang et al., 2024), which explore transfer learning in high-dimensional linear regression and require individual-level data from the source domain, our approach learns the sparse linear correction term using Lasso (Tibshirani, 1996) *with features available only in the target domain*. While our methodology and theory allow any suitable transfer or fine-tuning procedure (e.g., last-layer fine-tuning, LoRA), sparse linear calibration lets us *obtain explicit high-dimensional convergence rates*. Applying transfer learning or fine-tuning in PPI is challenging: using the same gold data for calibration and rectification leads to bias, and naive sample splitting reduces performance. Our cross-fitting formulation (Section 2), addresses these issues. Our joint power-tuning method (Section 3), *helps mitigate negative transfer* (Zhang et al., 2023). We prove that our estimators are *minimax optimal* (Section 4). Our framework effectively works across various estimation targets (e.g., mean, OLS, GLM, Bradley-Terry coefficients) and correction methods, *demonstrating its versatility across diverse domains* (Section 5).

The original PPI framework (Angelopoulos et al., 2023a) provides valid confidence intervals when a predictor is already available. Cross-prediction-powered inference (Zrnic & Candès, 2024) instead learns the predictor directly from the labeled sample via cross-fitting, but this requires enough target-domain data to train a good model from scratch. Instead, our proposal *adapts a pretrained model using the available gold data through a cross-fitting step*. PPI++ (Angelopoulos et al., 2023b) addresses poor predictors with a power-tuning parameter, while (Mani et al., 2026) notes that near-oracle conditions may not always hold (i.e., *no free lunch*). Our joint power-tuning approach aims for a *cheap lunch*, improving a predictor with the same gold data while ensuring no harm with high probability. Other approaches address predictor heterogeneity structurally. StratPPI partitions the population into strata of homogeneous predictor quality (Appendix F). Multi-Source PPI (Li et al., 2026) aggregates pseudo-labeled sources by minimizing asymptotic confidence-region volume, while PPI by Mixture of Experts (Gu et al., 2026) learns a mixture of predictors with a best-expert guarantee. They utilize structure across strata, sources, or experts, whereas *our transfer calibration improves a single-source predictor's target-domain alignment using the* **same** *gold-standard data*.

2 Transfer calibration and PPI/PPI++ rectification

PPI under Domain Shift. Let P_t denote the target-domain joint distribution of (X, Y) and P_s denote the source-domain joint distribution. In the target domain, we observe two datasets,

$$\mathcal{D}_t^{\text{gold}} = \{(X_i, Y_i) : i = 1, \ldots, n\} \overset{\text{iid}}{\sim} P_t, \qquad \mathcal{D}_t^{\text{pred}} = \{\widetilde{X}_j : j = 1, \ldots, N\} \overset{\text{iid}}{\sim} P_{t,X},$$

with $N \gg n$. We assume that a predictor $f_s : \mathcal{X} \to \mathcal{Y}$ has been learned from a vast source corpus: $\mathcal{D}_s = \{(X_i^s, Y_i^s) : i = 1, \ldots, n_s\} \overset{\text{iid}}{\sim} P_s$. We do not observe individual-level data from this corpus.

We allow this predictor to be misspecified for the response in the target domain P_t. The inferential target is target-domain parameter θ_t^*. We use the mean, $\mathbb{E}_t[Y] \coloneqq \int y dP_t(x, y)$ to develop key ideas, with additional results on linear regression, generalized linear models, and low-dimensional M-estimation in the Appendix. We treat the source predictor f_s as a black-box machine-learning model (e.g., tree-based, deep neural network, or large language model). We condition on the source data $\mathcal{D}_s$ and consequently treat f_s as fixed. Let $f_t(x) \coloneqq \mathbb{E}_t[Y \mid X = x]$ denote the target regression function and define the oracle target-domain correction

$$\Delta^*(x) \coloneqq f_t(x) - f_s(x). \tag{1}$$

Thus the target mean function can be written as $f_t(x) = f_s(x) + \Delta^*(x)$. The role of transfer learning in our framework is to estimate Δ^*, or, more generally, to construct a useful target-adapted prediction function, from limited gold-standard target data. As stated before, we consider the domain difference Δ^* to be a simple function or a low-dimensional function that scientists and domain researchers can easily train with limited data and limited resources. A transfer-calibration algorithm $\mathcal{A}_{\text{cal}}$ applied to a subset of the gold data produces a learned correction $\widehat{\Delta}$ and hence $\widehat{f}_t(x) = f_s(x) + \widehat{\Delta}(x)$. The calibration algorithm may be sparse linear regression (lasso) on the residuals $Y_i - f_s(X_i)$ either using the raw features or using embedding from an open-weight AI model, parameter-efficient fine-tuning methods on deep neural networks such as adapter (Houlsby et al., 2019) or LoRA (Hu et al., 2021), or another low-complexity transfer procedure. If a fine-tuning procedure directly produces a predictor $\hat{f}_t$, we define $\widehat{\Delta} = \widehat{f}_t - f_s$ after fitting. Thus, the additive notation does not restrict the architecture or parameterization of the transfer calibration. Define the target noise $\varepsilon = Y - f_t(X)$. Then, by definition of $f_t(X)$, we have $\mathbb{E}_t[\varepsilon \mid X] = 0$, and $\mathbb{E}_t[\varepsilon] = 0$. While some of our theoretical results rely on a sparse linear correction Δ^*, the validity results below require no structural assumptions on Δ^*, such as linearity or sparsity. Such assumptions enter only when we seek explicit rates for learning the transfer correction.

Calibration and Rectification. A naive construction, where transfer calibration is trained on one part of the data and PPI rectification is obtained on the other, is given in the Appendix, along with proofs that such an estimator is conditionally unbiased and its asymptotic limiting distribution, along with a derivation of the conditional variance. While the conditional variance cleanly decomposes in this construction (independent), it wastes gold samples during rectification, reducing performance. We use cross-fitting (Zrnic & Candès, 2024; Mani et al., 2026) to recover all labels for rectification.

We partition the labeled gold sample into K equal folds $I_1, \ldots, I_K$, with fixed K and $|I_k| = n/K$. For each fold k, we fit a transfer correction using combined gold data from the other folds,

$$\widehat{\Delta}^{(-k)} = \mathcal{A}_{\text{cal}}(f_s, \{(X_i, Y_i) : i \notin I_k\}), \qquad \widehat{f}_t^{(-k)}(x) = f_s(x) + \widehat{\Delta}^{(-k)}(x).$$

For $i \in I_k$, we define the out-of-fold prediction as $\widehat{f}_i^{\text{oof}} = \widehat{f}_t^{(-k)}(X_i)$ and the averaged unlabeled prediction as $\overline{f}_t(\widetilde{X}_j) = \frac{1}{K}\sum_{k=1}^K \widehat{f}_t^{(-k)}(\widetilde{X}_j)$. The new cross-fitted estimator of mean is

$$\widehat{\mu}_{\text{TC-Cross}} = \frac{1}{K}\sum_{k=1}^{K}\left[\frac{1}{N}\sum_{j=1}^{N}\widehat{f}_t^{(-k)}(\widetilde{X}_j) + \frac{K}{n}\sum_{i\in I_k}\{Y_i - \widehat{f}_t^{(-k)}(X_i)\}\right]. \tag{2}$$

The averaged unlabeled predictor $\overline{f}_t$ and the out-of-fold predictions are not independent, as observation i was used to train one of the predictors in the average. However, the estimator averages fold-specific PPI estimators, where the same predictor $\widehat{f}_t^{(-k)}$ is used for both the unlabeled prediction and rectification on its held-out fold, and each held-out observation is excluded from fitting the predictor used in its own rectification term.

Proposition 1 (Unbiasedness)**.** *Suppose the folds are fixed independently of the observations and have equal size* $|I_k| = n/K$*. Under Assumption 1,* $\mathbb{E}[\widehat{\mu}_{\text{TC-Cross}} \mid \mathcal{D}_s] = \theta_t^*$.

The cross-fitted estimator differs from the split-sample TC-PPI estimator as it lacks a straightforward conditional variance decomposition after using a single calibration sample. The overlap in training samples and shared unlabeled samples among fold-specific fitted predictors results in nonzero cross-fold covariance terms. Theorem 1 addresses this dependence with an asymptotic linear representation, and the estimation errors from fold-specific calibration functions are collected in a remainder that is $o_p(n^{-1/2})$, if the transfer calibration is consistent. Define $\mathcal{D}_{-k} \coloneqq \{(X_i, Y_i) : i \notin I_k\}$ and

$$e_k(x) \coloneqq \widehat{f}_t^{(-k)}(x) - f_t(x), \qquad a_n \coloneqq \max_{1 \le k \le K} \mathbb{E}_t[e_k(X)^2 \mid \mathcal{D}_s, \mathcal{D}_{-k}].$$

Then $e_k(x)$ is the error that the predictor trained on all folds except for k fold makes in approximating $f_t(x)$. The quantity a_n is the maximum expected squared error over all folds.

Theorem 1 (TC–Cross-PPI for the target mean)**.** *Suppose* K *is fixed, the folds have equal size* $|I_k| = n/K$*,* $n/N \to \rho \in [0, \infty)$*, and, conditionally on the source data,* $Y, f_t(X), \hat{f}_t(X)$ *have finite* $(2+\eta)$ *moments for some* $\eta > 0$*. Assume* $a_n = o_p(1)$*. Then*

$$\widehat{\mu}_{\text{TC-Cross}} - \theta_t^* = \frac{1}{n}\sum_{i=1}^{n} \varepsilon_i + \frac{1}{N}\sum_{j=1}^{N} \{f_t(\widetilde{X}_j) - \mathbb{E}_t f_t(X)\} + o_p(n^{-1/2}),$$

and consequently, $\sqrt{n}\big(\widehat{\mu}_{\text{TC-Cross}} - \theta_t^*\big) \xrightarrow{D} N(0, \text{Var}_t(\varepsilon) + \rho \text{Var}_t\{f_t(X)\})$.

Sparse linear calibration: high-dimensional lasso rates. The above theorem is agnostic to the form of the transfer correction and instead assumes the conditional mean squared error of the correction is $o_p(1)$. To obtain an explicit rate, we now specialize to a sparse linear correction as follows,

$$\Delta^*(x) = x^\top \delta^*, \quad \delta^* \in \mathbb{R}^p, \quad \|\delta^*\|_0 \le s, \tag{3}$$

for some $s > 0$. We assume the features are high-dimensional, i.e., $p \gg n$. We estimate δ^* by the lasso on the calibration data for each fold as,

$$\widehat{\delta}^{(-k)} \in \arg\min_{\delta \in \mathbb{R}^p} \left\{ \frac{1}{2n_{-k}} \sum_{i \notin I_k} \big(Y_i - f_s(X_i) - X_i^\top \delta\big)^2 + \lambda \|\delta\|_1 \right\}, \qquad n_{-k} = \frac{K-1}{K} n.$$

Corollary 1 specializes Theorem 1 to the case of sparse linear calibration estimated through lasso.

Corollary 1 (Sparse calibration rate for TC–Cross-PPI)**.** *Consider the sparse linear calibration model and assume the lasso related conditions in Appendix B hold on every training complement. Then, uniformly over all folds (fixed),* $(\widehat{\delta}^{(-k)} - \delta^*)^\top \Sigma_t (\widehat{\delta}^{(-k)} - \delta^*) = O_p\Big(\frac{s \log p}{(K-1)n/K}\Big)$ *. Consequently, if* $\frac{s \log p}{(K-1)n/K} \to 0$*,, then Theorem 1 holds. For fixed* K*, it is equivalent to* $s \log p / n \to 0$*.*

Remark 1 (Comparison with vanilla PPI)**.** Using all n gold observations, the vanilla PPI estimator based on the unadapted source predictor f_s has $\sqrt{n}$-scale asymptotic variance of $V_{\text{PPI}} = \rho \text{Var}_t\{f_s(X)\} + \text{Var}_t(\varepsilon) + \text{Var}_t\{\Delta^*(X)\}$. Under Theorem 1, TC–Cross-PPI also uses all n gold observations and has $\sqrt{n}$-scale asymptotic variance $V_{\text{TC-Cross}} = \rho \text{Var}_t\{f_t(X)\} + \text{Var}_t(\varepsilon)$. Thus

$$V_{\text{PPI}} - V_{\text{TC-Cross}} = \text{Var}_t\{\Delta^*(X)\} + \rho\left[\text{Var}_t\{f_s(X)\} - \text{Var}_t\{f_t(X)\}\right].$$

In the regime where $N \gg n$, we have $\rho = 0$, and consistent transfer calibration removes the source-to-target drift variance $\text{Var}_t\{\Delta^*(X)\}$, making TC-Cross-PPI more efficient than vanilla PPI.

Transfer Calibrated Extension for PPI++. (Angelopoulos et al., 2023b) note that if the AI model poorly predicts responses in the gold sample, PPI may underperform compared to classical inference. They show that inference can be improved by interpolating between classical and prediction-

powered approaches using a tuning parameter λ. This concept applies here as even a calibrated predictor may be imperfect. We define the λ-weighted transfer-calibrated estimator as:

$$\widehat{\mu}_{\text{TC-Cross},\lambda} = \bar{Y}_n + \lambda \left\{ \frac{1}{N}\sum_{j=1}^{N} \overline{f}_t(\widetilde{X}_j) - \frac{1}{n}\sum_{k=1}^{K}\sum_{i\in I_k} \widehat{f}_t^{(-k)}(X_i) \right\}. \tag{4}$$

This estimator equals classical inference at $\lambda = 0$ and ordinary TC–Cross-PPI at $\lambda = 1$.

Theorem 2 (TC–Cross-PPI++)**.** *Suppose the fold and moment conditions of Theorem 1 hold. Assume there exists a square-integrable function $f_\dagger$ such that, uniformly over the fixed number of folds, $\mathbb{E}_t\Big[\{\widehat{f}_t^{(-k)}(X) - f_\dagger(X)\}^2 \mid \mathcal{D}_s, \mathcal{D}_{-k}\Big] = o_p(1)$. Then, for every fixed λ, we have*

$$\sqrt{n}\big(\widehat{\mu}_{\text{TC-Cross},\lambda} - \theta_t^*\big) \xrightarrow{D} N\big(0, \mathrm{Var}_t(Y) - 2\lambda\mathrm{Cov}_t\{Y, f_\dagger(X)\} + (1+\rho)\lambda^2\mathrm{Var}_t\{f_\dagger(X)\}\big)\,.$$

If $\mathrm{Var}_t\{f_\dagger(X)\} > 0$, the oracle limiting minimizer is

$$\lambda_\dagger^* = \frac{\mathrm{Cov}_t\{Y, f_\dagger(X)\}}{(1+\rho)\mathrm{Var}_t\{f_\dagger(X)\}}, \tag{5}$$

and the corresponding minimized variance is $\mathrm{Var}_t(Y) - \frac{\mathrm{Cov}_t\{Y,f_\dagger(X)\}^2}{(1+\rho)\mathrm{Var}_t\{f_\dagger(X)\}}$. If in addition, the transfer-calibrated predictor is consistent for the oracle target predictor, $f_\dagger(X) = f_t(X)$, and then $\lambda_t^* = \frac{1}{1+\rho}$, with the minimized variance being $\mathrm{Var}_t(Y) - \frac{\mathrm{Var}_t\{f_t(X)\}}{1+\rho}$. If a feasible estimator $\widehat{\lambda}$ satisfies $\widehat{\lambda} - \lambda_\dagger^* = o_p(1)$, using $\widehat{\lambda}$ in the estimator gives the same asymptotic limiting distribution.

In the above theorem, we let the transfer-corrected predictors from the folds converge to a function $f_\dagger(X)$, instead of the true target domain predictor function $f_t(X)$. This removes the assumption that, post-transfer learning we are able to correctly approximate the true target domain conditional expectation of Y given X. The oracle parameter $\lambda_\dagger^*$ in equation 5 depends on unknown target-domain moments. A plug-in estimator for $\lambda_\dagger^*$ is $\widehat{\lambda} = \frac{\widehat{\mathrm{Cov}}_{\text{oof}}(Y,\widehat{f})}{(1+n/N)\widehat{\mathrm{Var}}_{\text{oof}}(\widehat{f})}$, when the denominator is positive, with $\widehat{\lambda} = 0$ otherwise. One may also clip $\widehat{\lambda}$ to $[0,1]$ for stability. Under the foldwise prediction-consistency and moment conditions of Theorem 2, $\widehat{\lambda} - \lambda_\dagger^* = o_p(1)$.

3 Joint power tuning: negative transfer protection

The TC–Cross-PPI++ estimator in equation 4 controls the influence of the transfer-calibrated predictor, ensuring its variance does not exceed that of the classical-only variance. However, as target samples are limited, transfer calibration can be weak or unstable. We thus also retain the original source predictor instead of replacing it with the calibrated predictor. For $i \in I_k$, define $\widehat{\Delta}_i^{\text{oof}} \coloneqq \widehat{f}_t^{(-k)}(X_i) - f_s(X_i)$ and $\widehat{H}_i^{\text{oof}} \coloneqq \begin{pmatrix} f_s(X_i) \\ \widehat{\Delta}_i^{\text{oof}} \end{pmatrix}$, and the averaged unlabeled vector

$$\widetilde{H}_N^{\text{CF}} \coloneqq \left(\frac{1}{N}\sum_{j=1}^{N} f_s(\widetilde{X}_j) \qquad \frac{1}{KN}\sum_{k=1}^{K}\sum_{j=1}^{N} \left\{ \widehat{f}_t^{(-k)}(\widetilde{X}_j) - f_s(\widetilde{X}_j) \right\} \right)^T.$$

For $\boldsymbol{\lambda} = (\lambda_s, \lambda_\Delta)^\top \in \mathbb{R}^2$, define the Joint TC-Cross-PPI++ estimator as

$$\widehat{\mu}_{\text{JTC-Cross},\boldsymbol{\lambda}} = \overline{Y}_n + \boldsymbol{\lambda}^\top \left\{ \widetilde{H}_N^{\text{CF}} - \frac{1}{n}\sum_{i=1}^{n} \widehat{H}_i^{\text{oof}} \right\}. \tag{6}$$

Special Cases: $(0,0)^\top$: classical inference, $(1,0)^\top$: Source PPI using f_s, $(1,1)^\top$: TC–Cross-PPI, $(\lambda,0)^\top$: Source PPI++ using f_s, $(\lambda,\lambda)^\top$: TC–Cross-PPI++.

Optimizing over the vector $\boldsymbol{\lambda}$ allows one to ignore the transfer correction, use it fully, or combine it with the original source predictor in ways that neither vanilla PPI++ nor TC–Cross-PPI++ can represent. Proposition 3 (see Appendix) shows the estimator is unbiased for any fixed $\boldsymbol{\lambda}$.

Theorem 3 (Joint TC–Cross-PPI++ and asymptotic no-harm)**.** *Suppose the fold and moment conditions of Theorem 1 hold. Assume there exists a square-integrable function $\Delta^\dagger$ such that $\max_{1\le k\le K}\mathbb{E}_t\Big[\{\widehat{\Delta}^{(-k)}(X)-\Delta^\dagger(X)\}^2 \mid \mathcal{D}_s,\mathcal{D}_{-k}\Big] = O_p(r_n)$, with $r_n = o(1)$. Define $H_\dagger(x) \coloneqq \begin{pmatrix} f_s(x) & \Delta^\dagger(x)\end{pmatrix}^\top$, and $f_\dagger(x) \coloneqq f_s(x)+\Delta^\dagger(x)$. Then for every fixed $\boldsymbol{\lambda}\in\mathbb{R}^2$,*

$$\sqrt{n}\left(\widehat{\mu}_{\text{JTC-Cross},\boldsymbol{\lambda}} - \theta_t^*\right) \xrightarrow{D} N\big(0, \mathrm{Var}_t\{Y-\boldsymbol{\lambda}^\top H_\dagger(X)\} + \rho\mathrm{Var}_t\{\boldsymbol{\lambda}^\top H_\dagger(X)\}\big)\,.$$

Further, let, $\Sigma_\dagger \coloneqq \mathrm{Var}_t\{H_\dagger(X)\}$, and $c_\dagger \coloneqq \mathrm{Cov}_t\{H_\dagger(X),Y\}$. If $\Sigma_\dagger$ is nonsingular, the unique oracle limiting weight is $\boldsymbol{\lambda}^*_{\text{Jcross}} = \frac{1}{1+\rho}\Sigma_\dagger^{-1}c_\dagger$. If $\Sigma_\dagger$ is singular, a minimum-norm oracle solution is obtained by replacing $\Sigma_\dagger^{-1}$ with its Moore–Penrose pseudoinverse $\Sigma_\dagger^\dagger$. Now, let $V^{\dagger,*}_{\text{Jcross}} \coloneqq \inf_{\boldsymbol{\lambda}\in\mathbb{R}^2} V^\dagger_{\text{Jcross}}(\boldsymbol{\lambda})$. Then the minimized variance is given by, $V^{\dagger,*}_{\text{Jcross}} = \mathrm{Var}_t(Y) - \frac{c_\dagger^\top\Sigma_\dagger^\dagger c_\dagger}{1+\rho}$. Denote the optimized source PPI++ variance by $V^{\text{cross}}_{s,++}$ and the optimized scalar TC–Cross-PPI++ variance associated with the limiting transfer predictor $f_\dagger$ by $V^{\text{cross}}_{\dagger,++}$. Because the joint class contains source-PPI++, TC–Cross-PPI++, and classical,

$$V^{\dagger,*}_{\text{Jcross}} \le \min\left\{V^{\text{cross}}_{s,++}, V^{\text{cross}}_{\dagger,++}, \mathrm{Var}_t(Y)\right\}. \tag{7}$$

The asymptotic no-harm property relative to source PPI++ does not require the transfer learner to be correctly specified. It suffices for fold-specific estimators to converge to a stable, square-integrable limiting function $\Delta^\dagger$. If in addition, the transfer learner is consistent for the oracle correction, so that $\Delta^\dagger = \Delta^*$, then $V^{\dagger,*}_{\text{Jcross}} = \mathrm{Var}_t(Y) - \frac{\mathrm{Var}_t\{f_t(X)\}}{1+\rho}$. This equals the oracle TC–Cross-PPI++ variance in Theorem 2, and is smaller than the variance of classical inference. While convergence to a stable transfer function ensures the no-harm safeguard, consistency for the oracle correction guarantees greater efficiency than classical inference. Ordinary source PPI++ has limiting variance $\mathrm{Var}_t(Y) - \frac{\mathrm{Cov}_t\{Y,f_s(X)\}^2}{(1+\rho)\mathrm{Var}_t\{f_s(X)\}}$. Since $f_t(X) = \mathbb{E}_t[Y \mid X]$, we have, $\mathrm{Cov}_t\{Y,f_s(X)\} = \mathrm{Cov}_t\{f_t(X),f_s(X)\}$, and the covariance-variance inequality gives, $\frac{\mathrm{Cov}_t\{Y,f_s(X)\}^2}{\mathrm{Var}_t\{f_s(X)\}} \le \mathrm{Var}_t\{f_t(X)\}$. Therefore, TC–Cross-PPI++, with oracle power tuning parameters and consistent transfer correction, is guaranteed to be no less efficient than source PPI++. Theorem 3 strengthens this comparison by retaining source PPI++ as a nested subfamily even when the transfer learner converges to a misspecified direction.

If a feasible weight estimator satisfies $\widehat{\boldsymbol{\lambda}} - \boldsymbol{\lambda}^*_{\text{Jcross}} = o_p(1)$, substitution of $\widehat{\boldsymbol{\lambda}}$ provides the same asymptotic limiting distribution. We can estimate $\boldsymbol{\lambda}$ using the plug-in estimator (with a small ridge regularization when needed), $\widehat{\boldsymbol{\lambda}} = \frac{1}{1+n/N}(\widehat{\Sigma}_H + \tau I_2)^{-1}\widehat{c}_H$, with $\tau \ge 0$. Estimating the weights $\boldsymbol{\lambda}$ using the above estimator incurs an additional inflation of $O(n^{-1}+r_n+\tau^2)$ to the $\sqrt{n}$ scale variance (Proposition 4, Appendix B). The empirical estimators for the asymptotic variances in the above theorems, along with results on their consistency and coverage rates of the corresponding Wald intervals, are provided in Appendix B.3. Unlike Zrnic & Candès (2024), we avoid expensive bootstrap to estimate standard errors and obtain confidence intervals. Instead, we show that despite using cross-fitting, we have the correct asymptotic coverage rate in Lemma 3.

4 Nonasymptotic rates and Minimax Optimality

In this section, we specialize to the sparse linear calibration model and study the finite-sample cost of estimating the transfer correction. For balanced K-fold cross-fitting, each correction is estimated on $n_{\text{tr}} \coloneqq \frac{K-1}{K}n$ observations, while all n gold observations contribute to rectification. Define $r_n \coloneqq \max_{1\le k\le K}\mathbb{E}\big[\mathbb{E}_t\{e_k(X)^2 \mid \mathcal{D}_s,\mathcal{D}_{-k}\} \mid \mathcal{D}_s\big]$. Under sparse linear calibration with lasso,

$r_n \lesssim \frac{\sigma_\epsilon^2 s \log p}{n_{\mathrm{tr}}}$ (Lemma 4, Appendix), up to constants determined by the design and noise conditions (Appendix B). Now, we can state the non-asymptotic MSE bound theorem as follows.

Theorem 4 (Nonasymptotic MSE bound)**.** *For fixed, balanced folds, and finite second moments, the TC-Cross PPI satisfies*

$$\mathrm{MSE}(\widehat{\mu}_{\text{TC-Cross}} \mid \mathcal{D}_s) \le \frac{\sigma_\epsilon^2}{n} + \frac{2V_t}{N} + \left(\frac{2}{N} + \frac{K}{n}\right) r_n. \tag{8}$$

Further consider the TC–Cross-PPI++ estimator with $\lambda_0 \coloneqq \frac{1}{1+n/N}$ *as the tuning parameter. Then*

$$\mathrm{MSE}(\widehat{\mu}_{\text{TC-Cross},\lambda_0} \mid \mathcal{D}_s) \le \frac{\sigma_\epsilon^2}{n} + \frac{V_t}{n+N} + K\frac{N}{n(n+N)}\, r_n. \tag{9}$$

The difference between the two bounds clarifies the role of power tuning. Power tuning is not only a safeguard against a poor predictor. When the target predictor is consistently learned, the finite-sample choice $\lambda_0 = N/(n+N)$ changes the oracle signal term from V_t/N to $V_t/(n+N)$ and cancels the leading interaction between the oracle prediction fluctuation and the calibration error. Consequently, the $O(\sqrt{r_n}/N)$ term appearing in 8 disappears and the calibration cost in 9 is proportional to r_n. We next give a matching information-theoretic lower bound. Consider the Gaussian sparse-transfer subclass $\mathcal{G}_s(V_s, V_\delta, \sigma_\epsilon^2)$ defined in Appendix A.3, and let $\psi_{n,p,s}(V_\delta, \sigma_\epsilon^2) \coloneqq \min\left\{V_\delta,\ \frac{\sigma_\epsilon^2(s-1)\log\{e(p-1)/(s-1)\}}{n-1}\right\}$.

Theorem 5 (Minimax lower bound with an unknown sparse transfer direction)**.** *Assume* $n \ge 3$, $N \ge 1$, *and* $2 \le s \le p/2$. *There exists a universal constant* $c > 0$ *such that*

$$\inf_{\widehat{\theta} \in \mathfrak{T}_{n,N}} \sup_{F \in\, \mathcal{G}_s(V_s, V_\delta, \sigma_\epsilon^2)} \mathbb{E}_F\left[(\widehat{\theta} - \theta)^2\right] \ \ge\ \frac{\sigma_\epsilon^2}{n} + \frac{V_t}{n+N} + c\,\frac{N}{n(n+N)}\psi_{n,p,s}(V_\delta, \sigma_\epsilon^2). \tag{10}$$

Remark 2 (Minimax optimality)**.** In the estimable regime above and on the Gaussian class, both the upper bound 9 and the lower bound 10 have the oracle rate $\sigma_\epsilon^2/n + V_t/(n+N)$ and a calibration cost of order $\frac{N}{n(n+N)} \cdot \frac{s}{n}$. The two bounds differ only by a multiplicative factor independent of n, N, p, or s, and by the logarithmic factor whenever $s \ll p$; the lasso-r_n upper bound carries $\log p$, while the lower bound carries $\log(ep/s)$. TC–Cross-PPI++ at $\lambda_0 = N/(n+N)$ is thus minimax optimal.

5 Empirical Experiments and Results

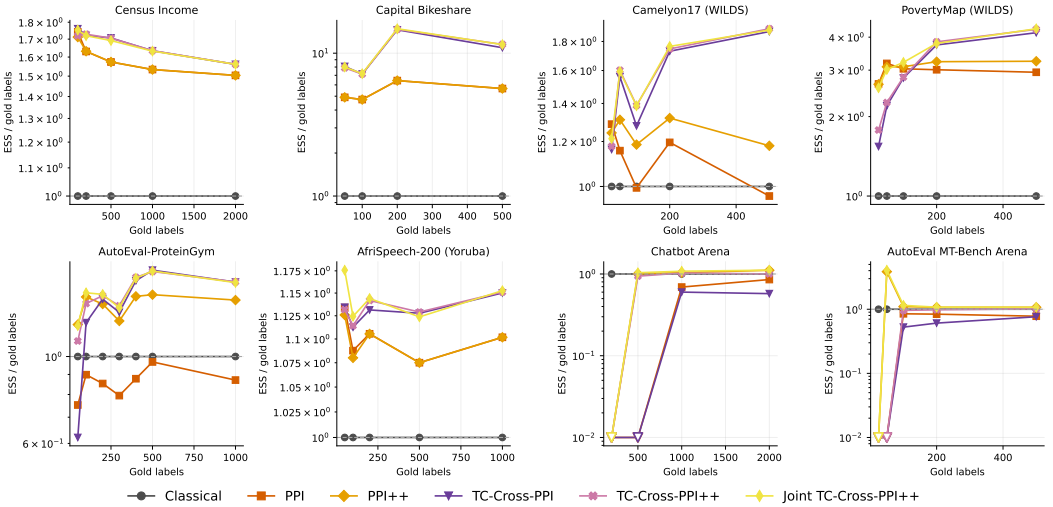


Figure 1: Empirical effective sample size (ESS) divided by the number of gold labels, on a logarithmic y-axis. The dashed line at 1 is classical inference. Values above one are MSE **improvements** over the gold-only estimator. Values below 0.01 are drawn at 0.01 as open downward triangles.

We evaluate transfer-calibrated PPI variants across a wide range of datasets. We use all the target-domain observations for evaluation, with n sampled points serving as "gold labels" and the rest as prediction-rich unlabeled samples. The ground truth is the parameter estimated using the entire target-domain data. We set the nominal coverage to 95%, except for the two AutoEval replications which match the original coverage of 90% (Appendix E). In addition to reporting classical inference, PPI, and PPI++, we report metrics for the cross-fitted and joint-power tuned variants of TC-PPI [1]. We include some results with other PPI variants in Appendix F.5. Metrics reported include MSE, interval width, and Effective Sample Size (ESS) defined as $\text{ESS} := n \cdot (\text{MSE}_{\text{classical}}/\text{MSE}_{\text{method}})$. LoRA-based calibration for protein and LLM-arena applications is in Appendix E and F.

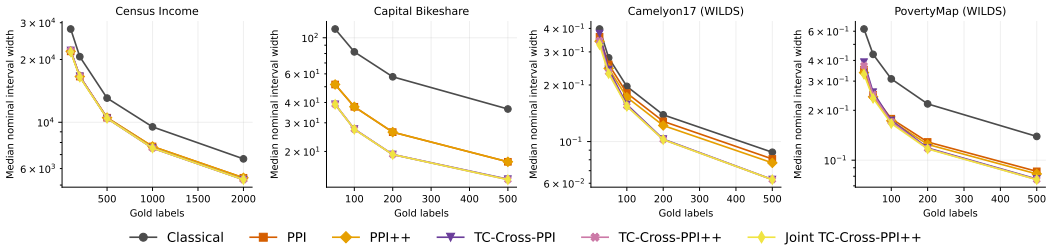


Figure 2: Median (over trials) width (in estimand's natural units (e.g., dollars for Census)) of nominal 95% confidence intervals, on a logarithmic y-axis. Remaining datasets are in Appendix F.

Census income. For mean income, the source XGBoost predictions are informative for the target year as PPI/PPI++ and the cross-fitted TC-PPI methods all substantially improve on classical inference, with relative MSE of 0.58–0.61 at $n = 200$ and ESS gains of 63–73%. TC–Cross-PPI++ and Joint TC–Cross-PPI++ are tied for best. Sparse calibration changes the residual variance by only a few percent (Table 4 in Appendix F), so most of this gain comes from the source predictor itself.

Temporal transfer of bicycle demand (Bikeshare). We train a gradient-boosted source predictor on 2011 data and transfer it to 2012. The inferential target is the mean 2012 hourly rental demand, 234.67. The source predictions correlate strongly with target outcomes (0.953) but their mean is only 150.27, and the error varies with calendar and weather covariates rather than being a constant offset. Transfer calibration roughly halves the residual variance (Table 4), producing some of the largest gains we observe. At $n = 100$, relative MSE ranges from 0.21 for PPI++ down to 0.14 for the cross-fitted methods, corresponding to ESS gains of 370–620%. Joint TC–Cross-PPI++ is best.

AfriSpeech-200 (Yoruba). We estimate the mean word error rate (WER) for Yoruba-accented English speech (Olatunji et al., 2023) using `wav2vec2-base-960h` (Baevski et al., 2020) as the source ASR model. The gains are modest but consistent across all n: every method improves on classical inference at every budget, TC–Cross-PPI's ESS gains range from 11.3% at $n = 100$ to 15.0% at $n = 1000$, and coverage stays within 0.91–0.98 at every n and method.

Distribution shift in the wild. The two WILDS benchmarks (Koh et al., 2020) resemble Bikeshare: the source predictions are correlated with the target, but their errors differ across target sub-populations (hospitals, slides, countries) in ways the correction can learn. Camelyon17 (Bándi et al., 2019) has histopathology patches from hospitals with different staining protocols; we transfer a DenseNet-121 to two held-out hospitals to estimate tumor prevalence (0.500). PovertyMap (Yeh et al., 2020) pairs satellite imagery with survey-based wealth measurements; we transfer an 8-band ResNet18 to ten held-out countries to estimate mean wealth (-0.049). Both sources degrade out of domain. Camelyon17's accuracy falls from 0.966 in distribution to 0.787 and 0.673 on the two held-out hospitals, PovertyMap's correlation falls from 0.868 to 0.830, and their mean predictions

[1] Split TC-PPI(++) performs poorly due to half the data not being used for rectification and is omitted.

are 0.339 and 0.115 against true values of 0.500 and -0.049. On PovertyMap, cross-fitted methods reach relative MSEs of 0.23–0.27 for $n \geq 200$, versus 0.31–0.34 for PPI/PPI++; on Camelyon17 they reach 0.53–0.58 versus 0.76–1.04, with PPI trailing classical at several budgets. Coverage stays between 0.88 and 0.99.

Table 1: Representative small-label results for different datasets and the inferential target in parentheses. MSE and interval width are normalized by classical inference at the same label budget.

Method	*Tabular (Mean)*		*Speech (WER)*	*Protein Language Models (Mean Mutational Fitness)*
	Census Income $n = 200$	Capital Bikeshare $n = 100$	AfriSpeech-200 (Yoruba), $n = 500$	AutoEval-ProteinGym $n = 500$
Classical	(1.000, 1.000)	(1.000, 1.000)	(1.000, 1.000)	(1.000, 1.000)
Joint TC-Cross-PPI++	(0.581, **0.798**)	(**0.139**, **0.334**)	(0.890, **0.946**)	(**0.605**, **0.806**)
PPI++	(0.613, 0.803)	(0.211, 0.460)	(0.930, 0.963)	(0.695, 0.850)
TC-Cross-PPI++	(**0.578**, 0.806)	(0.140, 0.335)	(**0.886**, 0.950)	(0.607, 0.809)
Method	*WILDS (Mean)*		*LLM Arena (Bradley–Terry)*	
	Camelyon17 (WILDS) $n = 100$	PovertyMap (WILDS) $n = 200$	Chatbot Arena $n = 500$	AutoEval MT-Bench Arena $n = 500$
Classical	(1.000, 1.000)	(1.000, 1.000)	(1.000, 1.000)	(1.000, 1.000)
Joint TC-Cross-PPI++	(0.721, **0.784**)	(0.264, **0.535**)	(**0.965**, **0.946**)	(0.933, **0.959**)
PPI++	(0.844, 0.875)	(0.310, 0.569)	(0.996, 0.962)	(**0.932**, 0.965)
TC-Cross-PPI++	(**0.720**, 0.788)	(**0.260**, 0.538)	(1.063, 1.009)	(0.987, 0.976)

AutoEval-ProteinGym (SPG1 replication). We reproduce AutoEval's (Boyeau et al., 2024) released SPG1 assay comparison, in which seven candidate zero-shot models (see Appendix E) provide source/annotator scores. PPI++ has relative MSE of 0.70–0.74 across $n = 200$–500. The cross-fitted methods improve further as n grows, reaching an MSE of 0.61, or ESS gains of up to 65%; Joint TC–Cross-PPI++ is consistently the best performing. We also study annotator-quality sensitivity across eight annotators (the seven candidates plus VESPA) and repeat it for the standard ProteinGym BLAT assay with three ESM-2 annotators (see Appendix F).

Chatbot Arena. We keep the 20 most compared models and use Meta-Llama-3-8B-Instruct as the source judge. With few labeled points, the 19-parameter Bradley–Terry model is underdetermined, and every method except Classical suffers quasi-separation. Near separation, the power-weight estimate becomes ill-conditioned, which is a finite-sample effect, not a violation of the asymptotic no-harm guarantee. (Theorem 3). From $n \geq 500$, PPI++ and Joint TC–Cross-PPI++ match or beat Classical, reaching 0.90 relative MSE (ESS $+11\%$) at $n = 2000$.

AutoEval MT-Bench Arena. We reproduce AutoEval's MT-Bench experiment, matching human ratings to GPT-4 pairwise judgments across six systems, calibrated with frozen MiniLM sentence embeddings. Like Chatbot Arena, the power-tuned methods help reliably. PPI++ and Joint TC–Cross-PPI++ have relative MSE of 0.910 and 0.870 at $n = 100$, or ESS gains of 9.9% and 14.9%, and 0.932 and 0.933 at $n = 500$, both $+7.2\%$. The fixed-λ PPI and TC–Cross-PPI stay worse than classical throughout, and TC–Cross-PPI++ becomes beneficial at $n = 500$ (0.988, $+1.3\%$).

6 Concluding Remarks

We introduce Transfer-Calibrated Prediction Powered Inference (TC-PPI), a framework that adapts a pretrained source predictor to a target domain using gold-standard data from PPI's rectification step. This approach avoids the inefficiencies of training a target predictor from scratch and provably addresses source-predictor misalignment issues. We highlight the challenges of transfer learning in PPI, where using the gold sample for both calibration and rectification can introduce bias. To mitigate, we propose cross-fitting and a joint power-tuning approach to prevent negative transfer. We recommend that researchers always adapt a pre-trained AI/ML model to their data using our TC-PPI framework for potentially significant gains with asymptotically negligible downside risk.

AI USE STATEMENT

For quick initial prototyping of our framework code, we relied on Codex/Claude Code. Subsequent versions relied on manual corrections and interventions. The authors wrote all proofs and the original draft manually. We used Grammarly for grammar corrections and, in a few cases, to improve the prose for clarity. We did not use AI tools for the proofs or references. We used Google's PAT to identify gaps, which we addressed in the current version of the draft.

We, the authors, take responsibility for the final content of this work.

REPRODUCIBILITY STATEMENT

In the spirit of reproducibility, proofs are available for the theoretical aspects of our work, and details of the experiments, including datasets and models, can be found in Appendix E. The source code is available in the supplemental material.

REFERENCES

Ethan C Alley, Grigory Khimulya, Surojit Biswas, Mohammed AlQuraishi, and George M Church. Unified rational protein engineering with sequence-based deep representation learning. *Nature Methods*, 16(12):1315–1322, 2019. doi: 10.1038/s41592-019-0598-1.

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.

Alexei Baevski, Henry Zhou, Abdelrahman Mohamed, and Michael Auli. wav2vec 2.0: A framework for self-supervised learning of speech representations, 2020. URL `https://arxiv.org/abs/2006.11477`.

Hamsa Bastani. Predicting with proxies: Transfer learning in high dimension. *Management Science*, 67(5):2964–2984, 2021.

Pierre Boyeau, Anastasios N Angelopoulos, Nir Yosef, Jitendra Malik, and Michael I Jordan. Autoeval done right: Using synthetic data for model evaluation. *International Conference on Machine Learning*, 2024. doi: 10.48550/arXiv.2403.07008.

Péter Bándi, Oscar G. F. Geessink, Quirine F. Manson, Marcory van Dijk, Maschenka C. A. Balkenhol, M. Hermsen, Babak Ehteshami Bejnordi, Byungjae Lee, K. Paeng, Aoxiao Zhong, et al. From detection of individual metastases to classification of lymph node status at the patient level: The CAMELYON17 challenge. *IEEE Transactions on Medical Imaging*, 38(2):550–560, 2019. doi: 10.1109/TMI.2018.2867350.

Jae Ho Chang, Massimiliano Russo, and Subhadeep Paul. Heterogeneous transfer learning for high-dimensional regression with feature mismatch. *arXiv preprint arXiv:2412.18081*, 2024.

Tianqi Chen and Carlos Guestrin. XGBoost: A scalable tree boosting system. In *Knowledge Discovery and Data Mining*, pp. 785–794, 2016. doi: 10.1145/2939672.2939785.

Wei-Lin Chiang, Lianmin Zheng, Ying Sheng, Anastasios Nikolas Angelopoulos, Tianle Li, Dacheng Li, Banghua Zhu, Hao Zhang, Michael Jordan, Joseph E Gonzalez, et al. Chatbot

arena: An open platform for evaluating LLMs by human preference. In *International Conference on Machine Learning*, volume 235, pp. 8359–8388, 2024. doi: 10.48550/arXiv.2403.04132.

Charlie Cowen-Breen, Alekh Agarwal, Stephen Bates, William W. Cohen, Jacob Eisenstein, Amir Globerson, and Adam Fisch. Multiple-prediction-powered inference. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=gJZ5rf2bS4.

Oscar Day and Taghi M Khoshgoftaar. A survey on heterogeneous transfer learning. *Journal of Big Data*, 4:1–42, 2017.

Frances Ding, Moritz Hardt, John Miller, and Ludwig Schmidt. Retiring adult: New datasets for fair machine learning. *Advances in neural information processing systems*, 34:6478–6490, 2021.

Hadi Fanaee-T and Joao Gama. Event labeling combining ensemble detectors and background knowledge. *Progress in Artificial Intelligence*, 2(2-3):113–127, 2014. doi: 10.1007/s13748-013-0040-3.

Adam Fisch, Joshua Maynez, R. Alex Hofer, Bhuwan Dhingra, Amir Globerson, and William W. Cohen. Stratified prediction-powered inference for hybrid language model evaluation. In *Advances in Neural Information Processing Systems 37*, volume 37, pp. 111489–111514. Neural Information Processing Systems Foundation, Inc. (NeurIPS), 2024. doi: 10.52202/079017-3541.

Yanwu Gu, Linglong Kong, and Dong Xia. Prediction-powered inference by mixture of experts. *arXiv.org*, 2026. doi: 10.48550/arXiv.2604.27892.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Computer Vision and Pattern Recognition*, pp. 770–778. IEEE, 2015. doi: 10.1109/cvpr.2016.90.

Daniel Hesslow, Niccolò Zanichelli, Pascal Notin, Iacopo Poli, and Debora Marks. RITA: a study on scaling up generative protein sequence models. *arXiv.org*, 2022. doi: 10.48550/arXiv.2205.05789.

Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. Parameter-efficient transfer learning for nlp. In *International conference on machine learning*, pp. 2790–2799. PMLR, 2019.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Liang Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *International Conference on Learning Representations*, 1(2):3, 2021.

Gao Huang, Zhuang Liu, and Kilian Q. Weinberger. Densely connected convolutional networks. In *Computer Vision and Pattern Recognition*, pp. 2261–2269. IEEE, 2016. doi: 10.1109/CVPR.2017.243.

Pang Wei Koh, Shiori Sagawa, Henrik Marklund, Sang Michael Xie, Marvin Zhang, Akshay Balsubramani, Weihua Hu, Michihiro Yasunaga, Richard Lanas Phillips, Irena Gao, et al. WILDS: A benchmark of in-the-wild distribution shifts. In *International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 5637–5664, 2020.

Sai Li, T Tony Cai, and Hongzhe Li. Transfer learning for high-dimensional linear regression: Prediction, estimation and minimax optimality. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 84(1):149–173, 2022.

Wenhui Li, Fen Jiang, and Xinyu Zhang. Multi-source prediction-powered inference. *arXiv*, 2026.

Zeming Lin, Halil Akin, Roshan Rao, Brian L. Hie, Zhong-Kai Zhu, Wen-Ting Lu, Nikita Smetanin, Robert Verkuil, Ori Kabeli, Y. Shmueli, et al. Evolutionary-scale prediction of atomic-level protein structure with a language model. *bioRxiv*, 379(6637):1123–1130, 2022. doi: 10.1126/science.ade2574.

Pranav Mani, Peng Xu, Zachary Chase Lipton, and Michael Oberst. No free lunch: Non-asymptotic analysis of prediction-powered inference. In *Forty-third International Conference on Machine Learning*, 2026. URL https://openreview.net/forum?id=zM5MbWc6wc.

Céline Marquet, Michael Heinzinger, Tobias Olenyi, Christian Dallago, Kyra Erckert, Michael Bernhofer, Dmitrii Nechaev, and Burkhard Rost. Embeddings from protein language models predict conservation and variant effects. *Human Genetics*, 140(10):1629–1644, 2021. doi: 10.1007/s00439-021-02411-y.

Joshua Meier, Roshan Rao, Robert Verkuil, Jason Liu, Tom Sercu, and Alexander Rives. Language models enable zero-shot prediction of the effects of mutations on protein function. *bioRxiv*, 34: 29287–29303, 2021. doi: 10.1101/2021.07.09.450648.

Shanjukta Nath, Jiwon Hong, Jae Ho Chang, Keith Warren, and Subhadeep Paul. Recidivism prediction, peer effect estimation, and prediction-powered inference with llm text measures. *arXiv preprint arXiv:2509.20634*, 2025.

Erik Nijkamp, Jeffrey A Ruffolo, Eli N Weinstein, Nikhil Naik, and Ali Madani. ProGen2: Exploring the boundaries of protein language models. *Cell Systems*, 14(11):968–978.e3, 2023. doi: 10.1016/j.cels.2023.10.002.

Pascal Notin, Aaron W. Kollasch, D. Ritter, Lood van Niekerk, Steffanie Paul, Han Spinner, Nathan J. Rollins, Ada Y. Shaw, Rose Orenbuch, Ruben Weitzman, et al. ProteinGym: Large-scale benchmarks for protein fitness prediction and design. In *Neural Information Processing Systems*, volume 36, pp. 64331–64379. openRxiv, 2023. doi: 10.1101/2023.12.07.570727.

Tobi Olatunji, Tejumade Afonja, Aditya Yadavalli, Chris Chinenye Emezue, Sahib Singh, Bonaventure F. P. Dossou, Joanne Osuchukwu, Salomey Osei, A. Tonja, N. Etori, et al. Afrispeech-200: Pan-african accented speech dataset for clinical and general domain asr. *Transactions of the Association for Computational Linguistics*, 11:1669–1685, 2023. doi: 10.1162/tacl_a_00627.

Garvesh Raskutti, Martin J Wainwright, and Bin Yu. Restricted eigenvalue properties for correlated gaussian designs. *The Journal of Machine Learning Research*, 11:2241–2259, 2010.

Alexander Rives, Siddharth Goyal, Joshua Meier, Demi Guo, Myle Ott, C. L. Zitnick, Jerry Ma, and R. Fergus. Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. *Proceedings of the National Academy of Sciences of the United States of America*, 118(15):e2016239118, 2020. doi: 10.1073/pnas.2016239118.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. In *NeurIPS EMC^2 Workshop*, 2019.

Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 58(1):267–288, 1996.

Nilesh Tripuraneni, Michael Jordan, and Chi Jin. On the theory of transfer learning: The importance of task diversity. *Advances in neural information processing systems*, 33:7852–7862, 2020.

Kevin K Yang, Nicolo Fusi, and Alex X Lu. Convolutions are competitive with transformers for protein sequence pretraining. *bioRxiv*, 15(3):286–294.e2, 2024. doi: 10.1101/2022.05.19.492714.

Christopher Yeh, Anthony Perez, Anne Driscoll, George Azzari, Zhongyi Tang, David Lobell, Stefano Ermon, and Marshall Burke. Using publicly available satellite imagery and deep learning to understand economic well-being in Africa. *Nature Communications*, 11(1):2583, 2020. doi: 10.1038/s41467-020-16185-w.

Wen Zhang, Lingfei Deng, Lei Zhang, and Dongrui Wu. A survey on negative transfer. *IEEE/CAA Journal of Automatica Sinica*, 10(2):305–329, 2023. ISSN 2329-9274. doi: 10.1109/jas.2022.106004. URL http://dx.doi.org/10.1109/JAS.2022.106004.

Yuqian Zhang and Jelena Bradic. High-dimensional semi-supervised learning: in search of optimal inference of the mean. *Biometrika*, 109(2):387–403, 2022.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P Xing, et al. Judging LLM-as-a-Judge with MT-Bench and chatbot arena. In *Neural Information Processing Systems*, volume 36, pp. 46595–46623. Neural Information Processing Systems Foundation, Inc. (NeurIPS), 2023. doi: 10.52202/075280-2020.

Fuzhen Zhuang, Zhiyuan Qi, Keyu Duan, Dongbo Xi, Yongchun Zhu, Hengshu Zhu, Hui Xiong, and Qing He. A comprehensive survey on transfer learning. *Proceedings of the IEEE*, 109(1): 43–76, 2020.

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

A NONASYMPTOTIC AND MINIMAX RESULTS FOR SPARSE LINEAR CALIBRATION

A.1 NONASYMPTOTIC MSE UPPER BOUNDS

Recall, cross-fitting uses all n gold observations for rectification, while each fold-specific transfer correction is estimated on a complementary training sample of size $n_{\mathrm{tr}} \coloneqq \frac{K-1}{K} n$, under balanced K-fold cross-fitting. Further recall that we condition on the source data $\mathcal{D}_s$ and treat the source predictor f_s as fixed. The target domain model with sparse linear discrepancy assumption is,

$$Y = f_s(X) + X^\top \delta^* + \epsilon, \qquad \mathbb{E}_t[\epsilon \mid X] = 0, \qquad \|\delta^*\|_0 \le s,$$

and define

$$f_t(X) \coloneqq f_s(X) + X^\top \delta^* = \mathbb{E}_t[Y \mid X].$$

We use the following short-hand notations,

$$\sigma_\epsilon^2 \coloneqq \mathrm{Var}_t(\epsilon), \qquad V_t \coloneqq \mathrm{Var}_t\{f_t(X)\}, \qquad V_s \coloneqq \mathrm{Var}_t\{f_s(X)\}.$$

For simplicity, assume the predictors are centered, i.e., $\mathbb{E}_t[X] = 0$. If this does not hold, then the quadratic forms below should be interpreted with the centered covariance matrix. The variance of the source-to-target drift term is

$$R_s \coloneqq \mathrm{Var}_t\{X^\top \delta^*\} = (\delta^*)^\top \Sigma_t \delta^*.$$

For each fold k, let $\widehat{\delta}^{(-k)}$ denote the lasso calibration estimator fitted on the training complement $\mathcal{D}_{-k}$ and define

$$e_k(x) \coloneqq x^\top (\widehat{\delta}^{(-k)} - \delta^*),$$

as the prediction error for the calibration estimator. Let

$$a_{n,k} \coloneqq \mathbb{E}_t\left[e_k(X)^2 \mid \mathcal{D}_s, \mathcal{D}_{-k}\right], \qquad r_n \coloneqq \max_{1 \le k \le K} \mathbb{E}[a_{n,k} \mid \mathcal{D}_s]\,. \tag{11}$$

Then the quantity r_n measures the maximum foldwise prediction risk of the transfer-calibration step. For the lasso based transfer calibration, this prediction risk from proposition 7,

$$r_n \lesssim \frac{s \log p}{n_{\mathrm{tr}}} = \frac{s \log p}{(K-1)n/K}.$$

Recall the decomposition from Theorem 1,

$$\widehat{\mu}_{\text{TC-Cross}} - \theta_t^* = \frac{1}{n}\sum_{i=1}^{n} \epsilon_i + \frac{1}{N}\sum_{j=1}^{N} \{f_t(\widetilde{X}_j) - \mathbb{E}_t f_t(X)\} + R_n,$$

where

$$R_n = \frac{1}{K}\sum_{k=1}^{K} \left[\frac{1}{N}\sum_{j=1}^{N} e_k(\widetilde{X}_j) - \frac{K}{n}\sum_{i \in I_k} e_k(X_i)\right].$$

The first two terms form the oracle PPI estimator based on f_t, while R_n contains the error from estimating the transfer correction.

The bound in Theorem 4 separates the oracle PPI variance (which requires knowing f_t),

$$\frac{\sigma_\epsilon^2}{n} + \frac{V_t}{N},$$

from the finite-sample cost of estimating the fold-specific transfer corrections. In particular, if

$$r_n \le C_{\mathrm{cal}} \frac{s \log p}{n_{\mathrm{tr}}}, \tag{12}$$

then

$$\mathrm{MSE}(\widehat{\mu}_{\text{TC-Cross}} \mid \mathcal{D}_s) \le \frac{\sigma_\epsilon^2}{n} + \frac{V_t}{N} + \frac{2\sqrt{C_{\text{cal}} V_t}}{N}\sqrt{\frac{s\log p}{n_{\text{tr}}}} + C_{\text{cal}}\left(\frac{1}{N}+\frac{K}{n}\right)\frac{s\log p}{n_{\text{tr}}}. \tag{13}$$

For fixed K and $N \gtrsim n$, since $n_{\text{tr}} \asymp n$, the excess over the oracle MSE is therefore at most of order

$$O\left(\frac{\sqrt{s\log p}}{n^{3/2}} + \frac{s\log p}{n^2}\right).$$

When the unlabeled sample is sufficiently large, the first term becomes negligible and the contribution of transfer-calibration estimation to the MSE is of order $s\log p/n^2$.

A.2 Proof of Theorem 4

Proof of Theorem 4. Condition throughout on the source data $\mathcal{D}_s$. Define

$$A_n := \frac{1}{n}\sum_{i=1}^{n}\epsilon_i, \qquad B_N := \frac{1}{N}\sum_{j=1}^{N}\{f_t(\widetilde{X}_j) - \mathbb{E}_t f_t(X)\},$$

and, for each fold k,

$$Z_{n,k} := \frac{1}{N}\sum_{j=1}^{N} e_k(\widetilde{X}_j) - \frac{K}{n}\sum_{i\in I_k} e_k(X_i).$$

Then

$$R_n = \frac{1}{K}\sum_{k=1}^{K} Z_{n,k}.$$

Therefore combining these components the error of the TC-Cross-PPI estimator can be written as,

$$\widehat{\mu}_{\text{TC-Cross}} - \theta_t^* = A_n + B_N + R_n.$$

By $\mathbb{E}_t[\epsilon \mid X] = 0$, we can write

$$\mathbb{E}[A_n \mid \mathcal{D}_s] = 0, \qquad \mathrm{Var}(A_n \mid \mathcal{D}_s) = \frac{\sigma_\epsilon^2}{n}.$$

Similarly, looking at B_N we have,

$$\mathbb{E}[B_N \mid \mathcal{D}_s] = 0, \qquad \mathrm{Var}(B_N \mid \mathcal{D}_s) = \frac{V_t}{N}.$$

The gold and unlabeled samples are independent, so

$$\mathbb{E}[A_n B_N \mid \mathcal{D}_s] = 0.$$

Hence the oracle part $A_n + B_N$ has the conditional MSE of

$$\mathbb{E}[(A_n + B_N)^2 \mid \mathcal{D}_s] = \frac{\sigma_\epsilon^2}{n} + \frac{V_t}{N}.$$

We next control the cross-fitting remainder. Recall that for each fixed fold k, $\mathcal{D}_{-k} = \{(X_i, Y_i) : i \notin I_k\}$. denotes the combined data from other folds. Conditional on $\mathcal{D}_s$ and $\mathcal{D}_{-k}$, the function e_k is fixed, the observations in I_k are independent of e_k. Further, the unlabeled sample is independent of both the training complement and the held-out fold. The two empirical averages defining $Z_{n,k}$ therefore have the same conditional population mean, so

$$\mathbb{E}[Z_{n,k} \mid \mathcal{D}_s, \mathcal{D}_{-k}] = 0.$$

Moreover,

$$\begin{aligned}\mathbb{E}[Z_{n,k}^2 \mid \mathcal{D}_s, \mathcal{D}_{-k}] &= \frac{1}{N}\mathrm{Var}_t\{e_k(X) \mid \mathcal{D}_s, \mathcal{D}_{-k}\} + \frac{K}{n}\mathrm{Var}_t\{e_k(X) \mid \mathcal{D}_s, \mathcal{D}_{-k}\} \\ &\le \left(\frac{1}{N}+\frac{K}{n}\right) a_{n,k}.\end{aligned}$$

Note that the above calculation so far is made fold-wise. Since the $Z_{n,k}$ need not be independent across folds, we use

$$R_n^2 = \left(\frac{1}{K}\sum_{k=1}^{K} Z_{n,k}\right)^2 \le \frac{1}{K}\sum_{k=1}^{K} Z_{n,k}^2.$$

Taking iterated conditional expectations gives

$$\begin{aligned}\mathbb{E}[R_n^2 \mid \mathcal{D}_s] &\le \frac{1}{K}\sum_{k=1}^{K}\mathbb{E}[Z_{n,k}^2 \mid \mathcal{D}_s] \le \left(\frac{1}{N}+\frac{K}{n}\right)\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}[a_{n,k} \mid \mathcal{D}_s] \\ &\le \left(\frac{1}{N}+\frac{K}{n}\right) r_n. \end{aligned} \tag{14}$$

The above display then provides the conditional MSE of the remainder term. The only thing left to examine is the covariance between the oracle part and R_n. First,

$$\mathbb{E}[A_n R_n \mid \mathcal{D}_s] = 0.$$

To see this, it is enough to show $\mathbb{E}[A_n Z_{n,k} \mid \mathcal{D}_s] = 0$ for every k. Conditional on $\mathcal{D}_s$ and $\mathcal{D}_{-k}$, decompose

$$A_n = \frac{1}{n}\sum_{i\notin I_k}\epsilon_i + \frac{1}{n}\sum_{i\in I_k}\epsilon_i.$$

The first term is fixed under this conditioning and has zero covariance with $Z_{n,k}$ because $\mathbb{E}[Z_{n,k} \mid \mathcal{D}_s, \mathcal{D}_{-k}] = 0$. For the second term, the unlabeled part of $Z_{n,k}$ is independent of the held-out observations, while

$$\mathbb{E}_t[\epsilon\, e_k(X) \mid \mathcal{D}_s, \mathcal{D}_{-k}] = \mathbb{E}_t[e_k(X)\mathbb{E}_t(\epsilon \mid X) \mid \mathcal{D}_s, \mathcal{D}_{-k}] = 0.$$

Independence across distinct held-out observations then gives

$$\mathbb{E}\left[\left(\frac{1}{n}\sum_{i\in I_k}\epsilon_i\right) Z_{n,k}\,\middle|\, \mathcal{D}_s, \mathcal{D}_{-k}\right] = 0.$$

Iterated expectation proves the claim.

The covariance of B_N with the held-out component of $Z_{n,k}$ is zero, because B_N depends only on the unlabeled sample. Conditional on $\mathcal{D}_s$ and $\mathcal{D}_{-k}$, the covariance with the unlabeled component is

$$\mathbb{E}[B_N Z_{n,k} \mid \mathcal{D}_s, \mathcal{D}_{-k}] = \frac{1}{N}\mathrm{Cov}_t\{f_t(X), e_k(X) \mid \mathcal{D}_s, \mathcal{D}_{-k}\}.$$

Therefore, by Cauchy–Schwarz,

$$|\mathbb{E}[B_N Z_{n,k} \mid \mathcal{D}_s, \mathcal{D}_{-k}]| \le \frac{\sqrt{V_t}}{N}\sqrt{a_{n,k}}.$$

Averaging over folds, applying iterated expectation, and using Jensen's inequality gives

$$\begin{aligned}|\mathbb{E}[B_N R_n \mid \mathcal{D}_s]| &\le \frac{\sqrt{V_t}}{KN}\sum_{k=1}^{K}\mathbb{E}[\sqrt{a_{n,k}} \mid \mathcal{D}_s] \\ &\le \frac{\sqrt{V_t}}{KN}\sum_{k=1}^{K}\sqrt{\mathbb{E}[a_{n,k} \mid \mathcal{D}_s]} \\ &\le \frac{\sqrt{V_t r_n}}{N}. \end{aligned} \tag{15}$$

By Proposition 1, $\widehat{\mu}_{\text{TC-Cross}}$ is conditionally unbiased given $\mathcal{D}_s$, so its conditional MSE is its conditional second moment around θ_t^*. Combining the preceding calculations,

$$\begin{aligned}
&\text{MSE}(\widehat{\mu}_{\text{TC-Cross}} \mid \mathcal{D}_s) - \left(\frac{\sigma_\epsilon^2}{n} + \frac{V_t}{N}\right) \\
&\quad = 2\mathbb{E}[(A_n + B_N)R_n \mid \mathcal{D}_s] + \mathbb{E}[R_n^2 \mid \mathcal{D}_s] \\
&\quad = 2\mathbb{E}[B_N R_n \mid \mathcal{D}_s] + \mathbb{E}[R_n^2 \mid \mathcal{D}_s].
\end{aligned}$$

Using equation 14 and equation 15 yields

$$\left|\text{MSE}(\widehat{\mu}_{\text{TC-Cross}} \mid \mathcal{D}_s) - \left(\frac{\sigma_\epsilon^2}{n} + \frac{V_t}{N}\right)\right| \leq \frac{2\sqrt{V_t r_n}}{N} + \left(\frac{1}{N} + \frac{K}{n}\right) r_n,$$

The upper bound equation 8 follows immediately after applying the simple fact; $\sqrt{2ab} \leq a + b$ for $a, b > 0$.

We now establish the second claim of the theorem for the TC–Cross-PPI++ estimator with the fixed power parameter

$$\lambda_0 = \frac{1}{1 + n/N} = \frac{N}{n+N}.$$

In addition to B_N, define the centered gold-sample average

$$B_n^{\text{L}} \coloneqq \frac{1}{n}\sum_{i=1}^{n}\{f_t(X_i) - \mathbb{E}_t f_t(X)\}.$$

Recall that for $i \in I_k$,

$$\widehat{f}_t^{(-k)}(X_i) = f_t(X_i) + e_k(X_i),$$

and similarly, for an unlabeled observation,

$$\frac{1}{K}\sum_{k=1}^{K}\widehat{f}_t^{(-k)}(\widetilde{X}_j) = f_t(\widetilde{X}_j) + \frac{1}{K}\sum_{k=1}^{K} e_k(\widetilde{X}_j).$$

Therefore, for any fixed λ,

$$\widehat{\mu}_{\text{TC-Cross},\lambda} - \theta_t^* = A_n + (1-\lambda)B_n^{\text{L}} + \lambda B_N + \lambda R_n. \tag{16}$$

The first three terms in equation 16 are components of the error of the oracle PPI++ estimator based on the knowledge of f_t, while λR_n is the contribution from estimating the transfer correction.

For convenience, write

$$C_{n,N}(\lambda) \coloneqq (1-\lambda)B_n^{\text{L}} + \lambda B_N.$$

Since the labeled and unlabeled samples are independent,

$$\mathbb{E}[B_n^{\text{L}} B_N \mid \mathcal{D}_s] = 0.$$

Moreover, because $\mathbb{E}_t[\epsilon \mid X] = 0$,

$$\mathbb{E}[A_n B_n^{\text{L}} \mid \mathcal{D}_s] = 0, \qquad \mathbb{E}[A_n B_N \mid \mathcal{D}_s] = 0.$$

Thus

$$\mathbb{E}\big[\{A_n + C_{n,N}(\lambda)\}^2 \mid \mathcal{D}_s\big] = \frac{\sigma_\epsilon^2}{n} + \left\{\frac{(1-\lambda)^2}{n} + \frac{\lambda^2}{N}\right\} V_t. \tag{17}$$

At $\lambda = \lambda_0 = N/(n+N)$,

$$\frac{(1-\lambda_0)^2}{n} + \frac{\lambda_0^2}{N} = \frac{1}{n+N},$$

and therefore

$$\mathbb{E}\big[\{A_n + C_{n,N}(\lambda_0)\}^2 \mid \mathcal{D}_s\big] = \frac{\sigma_\epsilon^2}{n} + \frac{V_t}{n+N}. \tag{18}$$

It remains to control the interaction term between the oracle PPI++ part (the first 3 terms) and the calibration remainder (the last term). We already showed above that

$$\mathbb{E}[A_n R_n \mid \mathcal{D}_s] = 0.$$

We next show that the choice λ_0 also makes $C_{n,N}(\lambda_0)$ orthogonal to R_n.

Fix a fold k and condition on $\mathcal{D}_s$ and $\mathcal{D}_{-k}$. The part of B_n^{L} involving observations outside I_k is then fixed, and hence has zero expectation when multiplied by $Z_{n,k}$ because

$$\mathbb{E}[Z_{n,k} \mid \mathcal{D}_s, \mathcal{D}_{-k}] = 0.$$

For the held-out part, the unlabeled component of $Z_{n,k}$ is independent of the observations in I_k, while independence across held-out observations gives

$$\mathbb{E}[B_n^{\mathrm{L}} Z_{n,k} \mid \mathcal{D}_s, \mathcal{D}_{-k}] = -\frac{1}{n}\mathrm{Cov}_t\{f_t(X), e_k(X) \mid \mathcal{D}_s, \mathcal{D}_{-k}\}. \tag{19}$$

On the other hand, as derived above,

$$\mathbb{E}[B_N Z_{n,k} \mid \mathcal{D}_s, \mathcal{D}_{-k}] = \frac{1}{N}\mathrm{Cov}_t\{f_t(X), e_k(X) \mid \mathcal{D}_s, \mathcal{D}_{-k}\}. \tag{20}$$

Combining equation 19 and equation 20, for a general fixed λ,

$$\mathbb{E}[C_{n,N}(\lambda) Z_{n,k} \mid \mathcal{D}_s, \mathcal{D}_{-k}] = \left\{\frac{\lambda}{N} - \frac{1-\lambda}{n}\right\} \mathrm{Cov}_t\{f_t(X), e_k(X) \mid \mathcal{D}_s, \mathcal{D}_{-k}\}. \tag{21}$$

Now setting $\lambda = \lambda_0 = \frac{N}{n+N}$, in the above equation, we have

$$\frac{\lambda_0}{N} = \frac{1-\lambda_0}{n} = \frac{1}{n+N},$$

and consequently, the right-hand side of equation 21 is identically zero. Averaging over the folds and applying iterated expectation yields

$$\mathbb{E}[C_{n,N}(\lambda_0) R_n \mid \mathcal{D}_s] = 0. \tag{22}$$

Combining equation 16, equation 18, equation 22, and the previously established identity $\mathbb{E}[A_n R_n \mid \mathcal{D}_s] = 0$, we obtain

$$\mathrm{MSE}(\widehat{\mu}_{\text{TC-Cross},\lambda_0} \mid \mathcal{D}_s) = \frac{\sigma_\epsilon^2}{n} + \frac{V_t}{n+N} + \lambda_0^2 \mathbb{E}[R_n^2 \mid \mathcal{D}_s].$$

Applying equation 14,

$$\begin{aligned}\mathrm{MSE}(\widehat{\mu}_{\text{TC-Cross},\lambda_0} \mid \mathcal{D}_s) &\le \frac{\sigma_\epsilon^2}{n} + \frac{V_t}{n+N} + \lambda_0^2 \left(\frac{1}{N} + \frac{K}{n}\right) r_n \\ &= \frac{\sigma_\epsilon^2}{n} + \frac{V_t}{n+N} + \frac{N(n+KN)}{n(n+N)^2}\, r_n.\end{aligned}$$

Finally, since $K \ge 1$,

$$n + KN \le K(n+N),$$

and hence

$$\frac{N(n+KN)}{n(n+N)^2} \le \frac{KN}{n(n+N)}.$$

Therefore,

$$\mathrm{MSE}(\widehat{\mu}_{\text{TC-Cross},\lambda_0} \mid \mathcal{D}_s) \le \frac{\sigma_\epsilon^2}{n} + \frac{V_t}{n+N} + K\frac{N}{n(n+N)} r_n,$$

which proves equation 9 and completes the proof.

□

A.3 MINIMAX LOWER BOUND.

Recall the settings: Define the class $\mathcal{G}_s(V_s, V_\delta, \sigma_\epsilon^2)$ as follows. For parameters

$$\theta \in \mathbb{R}, \qquad \mu_s \in \mathbb{R}, \qquad \mu \in \mathbb{R}^p, \qquad \delta \in \mathbb{R}^p,$$

with

$$\|\delta\|_0 \le s, \qquad \|\delta\|_2^2 = V_\delta, \tag{23}$$

let

$$S = \mu_s + \xi, \qquad \xi \sim N(0, V_s), \tag{24}$$

$$X = \mu + W, \qquad W \sim N_p(0, I_p), \tag{25}$$

$$Y = \theta + \xi + W^\top \delta + \varepsilon, \qquad \varepsilon \sim N(0, \sigma_\epsilon^2), \tag{26}$$

where ξ, W, ε are mutually independent. Thus $\mu_s = \mathbb{E}(S)$. The target regression on the observed features is

$$f_t(s, x) \coloneqq \mathbb{E}_t[Y \mid S = s, X = x] = \theta + (s - \mu_s) + (x - \mu)^\top \delta. \tag{27}$$

Using the observed score S itself as the available source prediction $f_s = S$, the transfer residual is

$$f_t(S, X) - S = \theta - \mu_s - \mu^\top \delta + X^\top \delta.$$

Moreover,

$$\mathbb{E}[Y] = \theta, \qquad \mathrm{Var}\{f_t(S, X)\} = V_s + V_\delta =: V_t, \qquad \mathrm{Var}(Y \mid S, X) = \sigma_\epsilon^2.$$

With n iid samples $\xi_i, \tilde{\xi}_i$ from $N(0, V_s)$, we observe n independent labeled draws

$$\{(S_i, X_i, Y_i) : i = 1, \dots, n\}$$

and, independently, N unlabeled target draws

$$\{(\widetilde{S}_j, \widetilde{X}_j) : j = 1, \dots, N\}.$$

Write

$$M \coloneqq n + N.$$

Let $\mathfrak{T}_{n,N}$ denote the class of *all measurable estimators* of θ based on these observations.

Define

$$\psi_{n,p,s}(V_\delta, \sigma_\epsilon^2) \coloneqq \min\left\{ V_\delta, \frac{\sigma_\epsilon^2 (s-1) \log\left(\frac{e(p-1)}{s-1}\right)}{n-1} \right\}. \tag{28}$$

We first separate the information about the unknown means from the information about the sparse slope. Let

$$\bar{X}_L = \frac{1}{n} \sum_{i=1}^{n} X_i, \qquad \bar{X}_U = \frac{1}{N} \sum_{j=1}^{N} \widetilde{X}_j, \qquad \bar{X}_M = \frac{n\bar{X}_L + N\bar{X}_U}{M},$$

and define the labeled–unlabeled contrast

$$D_X \coloneqq \bar{X}_L - \bar{X}_U.$$

Then

$$\bar{X}_M \sim N_p\left(\mu, \frac{1}{M} I_p\right), \qquad D_X \sim N_p\left(0, \left(\frac{1}{n} + \frac{1}{N}\right) I_p\right), \tag{29}$$

and $\bar{X}_M$ and D_X are independent. The analogous statements hold for S.

Let $Q \in \mathbb{R}^{n \times (n-1)}$ have orthonormal columns satisfying

$$Q^\top Q = I_{n-1}, \qquad Q^\top \mathbf{1}_n = 0.$$

Define the labeled residual vector

$$U_L = (Y_1 - S_1, \ldots, Y_n - S_n)^\top.$$

From equation 26,

$$Y_i - S_i = \theta - \mu_s + (X_i - \mu)^\top \delta + \varepsilon_i.$$

Hence

$$Q^\top U_L = (Q^\top X_L)\delta + Q^\top \varepsilon. \tag{30}$$

Because the rows of X_L are Gaussian with covariance I_p, the transformed design $Q^\top X_L$ has $n-1$ independent $N_p(0, I_p)$ rows, and $Q^\top \varepsilon \sim N_{n-1}(0, \sigma_\epsilon^2 I_{n-1})$. Moreover, the centered regression experiment equation 30 is independent of the sample means and the labeled–unlabeled mean contrasts. Thus the unknown transfer slope δ is learned from an ordinary sparse Gaussian linear-regression experiment with effective sample size $n-1$.

The next lemma is the main bridge between sparse-regression difficulty and target-mean difficulty.

Lemma 1 (Bayes decomposition)**.** *Let π be any prior supported on sparse vectors satisfying $\|\delta\|_2^2 = V_\delta$ almost surely. Put independent Gaussian priors*

$$\theta \sim N(0, T^2), \qquad \mu_s \sim N(0, T^2), \qquad \mu \sim N_p(0, T^2 I_p),$$

independently of $\delta \sim \pi$. Let

$$R_B(T, \pi) = \inf_{\widehat{\theta}} \mathbb{E}_{\pi, \mathcal{D}} \left[(\widehat{\theta}(\mathcal{D}) - \theta)^2 \right] \tag{31}$$

be the Bayes risk for estimating θ under squared error from the full labeled and unlabeled experiment. Then

$$\lim_{T \to \infty} R_B(T, \pi) = \frac{\sigma_\epsilon^2}{n} + \frac{V_s + V_\delta}{M} + \frac{N}{nM} R_\delta(\pi), \tag{32}$$

where

$$R_\delta(\pi) \coloneqq \inf_{\widehat{\delta}} \mathbb{E}_\pi \left\| \widehat{\delta} - \delta \right\|_2^2 \tag{33}$$

is the Bayes risk for estimating δ from the centered regression experiment equation 30.

Proof. By definition of the Bayes risk under squared error loss,

$$R_B(T, \pi) = \inf_{\widehat{\theta}} \mathbb{E}_{\mathcal{D}} \left[\mathbb{E}_{\theta|\mathcal{D}} \left[(\widehat{\theta}(\mathcal{D}) - \theta)^2 \mid \mathcal{D} \right] \right].$$

Because the outer expectation over $\mathcal{D}$ does not constrain the pointwise choice of estimate for each realized sample $\mathcal{D}$, the infimum passes inside the expectation:

$$R_B(T, \pi) = \mathbb{E}_{\mathcal{D}} \left[\inf_{c \in \mathbb{R}} \mathbb{E}_{\theta|\mathcal{D}} \left[(c - \theta)^2 \mid \mathcal{D} \right] \right].$$

For any random variable $\theta \mid \mathcal{D}$ with finite second moment, the scalar constant c minimizing the mean squared error is the conditional posterior mean:

$$c^* = \mathbb{E}[\theta \mid \mathcal{D}].$$

Substituting this minimizer $\widehat{\theta}_{\text{Bayes}}(\mathcal{D}) = \mathbb{E}[\theta \mid \mathcal{D}]$ into the inner expectation yields the conditional posterior variance:

$$\inf_{c \in \mathbb{R}} \mathbb{E}_{\theta|\mathcal{D}} \left[(c - \theta)^2 \mid \mathcal{D} \right] = \mathbb{E}_{\theta|\mathcal{D}} \left[(\theta - \mathbb{E}[\theta \mid \mathcal{D}])^2 \mid \mathcal{D} \right] = \mathrm{Var}(\theta \mid \mathcal{D}).$$

Taking the remaining expectation over the marginal distribution of $\mathcal{D}$ gives:

$$R_B(T, \pi) = \mathbb{E}_{\mathcal{D}} \left[\mathrm{Var}(\theta \mid \mathcal{D}) \right]. \tag{34}$$

With this observation, we now give the diffuse-prior calculation, which is the limit of the proper Gaussian-prior calculation as $T \to \infty$. Under a flat prior for μ_s and μ, Gaussian mean–residual

independence gives

$$\mu_s \mid \mathcal{D} \sim N\left(\bar{S}_M, \frac{V_s}{M}\right), \qquad \mu \mid \mathcal{D} \sim N_p\left(\bar{X}_M, \frac{1}{M} I_p\right).$$

The posterior for δ depends only on the centered regression data in equation 30 and is independent of these mean posteriors and of D_X. Write

$$m_\delta \coloneqq \mathbb{E}[\delta \mid \mathcal{D}_c], \qquad C_\delta \coloneqq \mathrm{Var}(\delta \mid \mathcal{D}_c),$$

where $\mathcal{D}_c$ denotes the centered regression data.

Now let

$$\bar{S}_L = \frac{1}{n}\sum_{i=1}^{n} S_i, \quad \bar{Y}_L = \frac{1}{n}\sum_{i=1}^{n} Y_i.$$

The labeled mean satisfies

$$\overline{Y}_L = \theta + (\overline{S}_L - \mu_s) + (\overline{X}_L - \mu)^\top \delta + \bar{\varepsilon}, \qquad \bar{\varepsilon} \sim N\left(0, \frac{\sigma_\epsilon^2}{n}\right).$$

Under the diffuse prior for θ, conditional on (μ_s, μ, δ) and the observed data,

$$\theta \mid \mathcal{D}, \mu_s, \mu, \delta \sim N\left(m_\theta(\mu_s, \mu, \delta), \frac{\sigma_\epsilon^2}{n}\right),$$

where

$$m_\theta(\mu_s, \mu, \delta) = \overline{Y}_L - (\overline{S}_L - \mu_s) - (\overline{X}_L - \mu)^\top \delta.$$

Therefore, by the law of total variance,

$$\begin{aligned}\mathrm{Var}(\theta \mid \mathcal{D}) = \mathbb{E}[\mathrm{Var}(\theta \mid \mathcal{D}, \mu_s, \mu, \delta) \mid \mathcal{D}] \\ + \mathrm{Var}(\mathbb{E}[\theta \mid \mathcal{D}, \mu_s, \mu, \delta] \mid \mathcal{D}).\end{aligned} \tag{35}$$

The first term in equation 35 is simply

$$\frac{\sigma_\epsilon^2}{n}.$$

To evaluate the second term, write

$$A_s \coloneqq \mu_s - \overline{S}_M, \qquad B_\mu \coloneqq \mu - \overline{X}_M, \qquad a_N \coloneqq \frac{N}{M}.$$

Since

$$\overline{X}_L - \mu = \frac{N}{M} D_X - (\mu - \overline{X}_M) = a_N D_X - B_\mu,$$

the conditional posterior mean can be written as

$$m_\theta(\mu_s, \mu, \delta) = C_\mathcal{D} + A_s - a_N D_X^\top \delta + B_\mu^\top \delta,$$

where $C_\mathcal{D}$ depends only on the observed data and therefore does not contribute to the conditional variance.

Under the diffuse-prior posterior,

$$A_s \mid \mathcal{D} \sim N\left(0, \frac{V_s}{M}\right), \qquad B_\mu \mid \mathcal{D} \sim N_p\left(0, \frac{1}{M} I_p\right),$$

and these variables are independent of each other and of $\delta \mid \mathcal{D}_c$. Moreover, D_X is fixed when conditioning on $\mathcal{D}$. Hence all cross-covariance terms vanish and

$$\begin{aligned}\mathrm{Var}(m_\theta(\mu_s, \mu, \delta) \mid \mathcal{D}) = \frac{V_s}{M} + \left(\frac{N}{M}\right)^2 D_X^\top C_\delta D_X \\ + \mathrm{Var}\left(B_\mu^\top \delta \mid \mathcal{D}\right),\end{aligned}$$

where

$$C_\delta = \mathrm{Var}(\delta \mid \mathcal{D}_c).$$

Finally, using the conditional independence of B_μ and δ,

$$\begin{aligned}\mathrm{Var}\big(B_\mu^\top \delta \mid \mathcal{D}\big) &= \mathbb{E}\big[(B_\mu^\top \delta)^2 \mid \mathcal{D}\big] \\ &= \mathbb{E}\big[\delta^\top \mathbb{E}(B_\mu B_\mu^\top \mid \mathcal{D})\delta \mid \mathcal{D}\big] \\ &= \frac{1}{M}\mathbb{E}\big[\|\delta\|_2^2 \mid \mathcal{D}_c\big]\,.\end{aligned}$$

Combining these terms gives

$$\begin{aligned}\mathrm{Var}(\theta \mid \mathcal{D}) = \frac{\sigma_\epsilon^2}{n} + \frac{V_s}{M} + \frac{1}{M}\mathbb{E}\big[\|\delta\|_2^2 \mid \mathcal{D}_c\big] \\ + \left(\frac{N}{M}\right)^2 D_X^\top C_\delta D_X. \end{aligned} \tag{36}$$

By assumption, $\|\delta\|_2^2 = V_\delta$ almost surely under π. Also D_X is independent of $\mathcal{D}_c$ and

$$\mathbb{E}[D_X D_X^\top] = \left(\frac{1}{n} + \frac{1}{N}\right) I_p = \frac{M}{nN} I_p.$$

Averaging equation 36 therefore gives

$$\begin{aligned}\mathbb{E}\mathrm{Var}(\theta \mid \mathcal{D}) &= \frac{\sigma_\epsilon^2}{n} + \frac{V_s + V_\delta}{M} + \left(\frac{N}{M}\right)^2 \frac{M}{nN}\,\mathbb{E}\,\mathrm{Tr}(C_\delta) \\ &= \frac{\sigma_\epsilon^2}{n} + \frac{V_s + V_\delta}{M} + \frac{N}{nM}\mathbb{E}\,\mathrm{Tr}(C_\delta).\end{aligned}$$

For squared-error loss, the posterior mean is Bayes and its Bayes risk is the expected posterior variance. Likewise,

$$\mathbb{E}\,\mathrm{Tr}(C_\delta) = \inf_{\widehat{\delta}} \mathbb{E}_\pi \left\|\widehat{\delta} - \delta\right\|_2^2 = R_\delta(\pi).$$

The same identity is obtained as the limit of the proper Gaussian-prior Bayes risks as $T \to \infty$, proving equation 32. □

We next construct a finite prior on an exact V_δ shell whose Bayes risk has the sparse-regression minimax order.

Lemma 2 (Sparse-regression Bayes lower bound)**.** *Suppose $s \geq 1$. Under the conditions of Theorem 5, there exists a prior π_* supported on*

$$\left\{\delta : \|\delta\|_0 \leq s,\ \|\delta\|_2^2 = V_\delta\right\}$$

and a universal constant $c_0 > 0$ such that the Bayes risk in the centered Gaussian regression experiment equation 30 satisfies

$$R_\delta(\pi_*) \geq c_0\, \psi_{n,p,s}(V_\delta, \sigma_\epsilon^2). \tag{37}$$

Proof. Set

$$k = s - 1, \qquad q = p - 1.$$

Since $s \leq p/2$, we have $q \geq 2k$. A constant-weight Varshamov–Gilbert construction yields a collection

$$\mathcal{V} \subset \{v \in \{0,1\}^q : \|v\|_0 = k\}$$

with

$$\log|\mathcal{V}| \geq c_1 k \log\left(\frac{eq}{k}\right) \tag{38}$$

and pairwise Hamming distance

$$d_H(v, v') \geq c_2 k, \qquad v \neq v'. \tag{39}$$

Here and below $c_1, c_2 > 0$ are universal constants.

Choose a sufficiently small universal $c_3 > 0$ and set

$$r_* := c_3 \min\left\{V_\delta, \frac{\sigma_\epsilon^2 k \log(eq/k)}{n-1}\right\}. \tag{40}$$

Let

$$a^2 = \frac{r_*}{k}, \qquad b^2 = V_\delta - r_*.$$

For each $v \in \mathcal{V}$, embed v in coordinates $2, \dots, p$ and define

$$\delta_v := be_1 + a(0, v^\top)^\top. \tag{41}$$

Then $\|\delta_v\|_0 \le s$ and $\|\delta_v\|_2^2 = V_\delta$ for every v. Moreover, $d_H(v, v') \le 2k$ by definition and

$$\|\delta_v - \delta_{v'}\|_2^2 = a^2 d_H(v, v') \ge c_2 r_* \qquad (v \ne v'). \tag{42}$$

Let P_v denote the law of the centered regression experiment under δ_v. Since the Gaussian design distribution is identical under all parameters, the KL divergence is

$$\begin{aligned} D_{\mathrm{KL}}(P_v \| P_{v'}) &= \frac{n-1}{2\sigma_\epsilon^2} \|\delta_v - \delta_{v'}\|_2^2 \\ &\le \frac{n-1}{\sigma_\epsilon^2} r_* \le c_3 k \log(eq/k). \end{aligned} \tag{43}$$

Now, for any estimator $\widehat{\delta}$, introduce a test $\tilde{v} := \arg\min_{v \in \mathcal{V}} \|\widehat{\delta} - \delta_v\|_2$. Let π_* be a uniform prior on $\delta_1, \dots \delta_V$. By equation 42 and triangle inequality, it holds that

$$\|\widehat{\delta} - \delta_v\|_2^2 \ge \frac{c_2 r_*}{4} \mathbb{I}(\tilde{v} \ne v) \text{ hence } R_\delta(\pi_*) \ge c_2 r_*/4 \inf_{\widehat{\delta}} P(\tilde{v} \ne v).$$

Choosing $c_3 \le c_1/4$ makes the average KL divergence at most a fixed fraction of $\log|\mathcal{V}|$. i.e.,

$$\frac{1}{|\mathcal{V}|^2} \sum_{v, v'} D_{\mathrm{KL}}(P_v \| P_{v'}) \le \alpha \log|\mathcal{V}| \text{ for some } \alpha \in (0, 1/4].$$

Fano's inequality gives

$$P(\tilde{v} \ne v) \ge 1 - \frac{\frac{1}{|\mathcal{V}|^2}\sum_{v,v'} D_{\mathrm{KL}}(P_v \| P_{v'}) + \log 2}{\log|\mathcal{V}|} \ge 1 - \alpha - \frac{\log 2}{\log|\mathcal{V}|} > 1/2$$

hence

$$R_\delta(\pi_*) \ge c_4 r_*$$

for a universal $c_4 > 0$. Combining this with equation 40 proves equation 37. □

Proof of Theorem 5. For any proper prior supported on the parameter class, minimax risk is at least Bayes risk. Use the prior π_* from Lemma 2 for δ, independent $N(0, T^2)$ priors for θ and μ_s, and an independent $N_p(0, T^2 I_p)$ prior for μ. Then

$$\inf_{\widehat{\theta} \in \mathfrak{T}_{n,N}} \sup_{P \in \mathcal{G}_s} \mathbb{E}_P(\widehat{\theta} - \theta)^2 \ge R_B(T, \pi_*)$$

for every $T < \infty$. Letting $T \to \infty$ and applying Lemmas 1 and 2 gives

$$\inf_{\widehat{\theta} \in \mathfrak{T}_{n,N}} \sup_{P \in \mathcal{G}_s} \mathbb{E}_P(\widehat{\theta} - \theta)^2 \ge \frac{\sigma_\epsilon^2}{n} + \frac{V_s + V_\delta}{M} + \frac{N}{nM} c_0 \psi_{n,p,s}(V_\delta, \sigma_\epsilon^2),$$

which is equation 10 because $V_t = V_s + V_\delta$ and $M = n + N$. □

B Additional theoretical details

Assumption 1. Conditionally on the source-trained predictor f_s, the variables Y, $f_t(X)$, and $\Delta^*(X)$ are square-integrable. Further, $\mathbb{E}_t[\widehat{f}_t(X)^2 \mid \mathcal{D}_s, \mathcal{D}_{t,1}^{\mathrm{gold}}] < \infty$ almost surely.

B.1 SPLIT TC-PPI

We first describe our procedure using an inefficient sample splitting approach which we will later improve through cross-fitting. We split the target gold sample into two disjoint sets $\mathcal{I}_1, \mathcal{I}_2 \subseteq [n]$ such that $|\mathcal{I}_1| = n_1$ and $|\mathcal{I}_2| = n_2$ with $n_1 + n_2 = n$. The first split is used for transfer calibration, and the second split is used only for PPI rectification. We denote the corresponding data splits by $\mathcal{D}^{\text{gold}}_{t,j} := \{(X_i, Y_i)_{i \in \mathcal{I}_j}\}$ for $j = 1, 2$.

On the calibration split $\mathcal{D}^{\text{gold}}_{t,1}$, we apply a transfer-calibration algorithm appropriate for the type of data and discrepancy model, to the source predictor and target gold data points in $\mathcal{I}_1$, and obtain, $\widehat{\Delta} = \mathcal{A}_{\text{cal}}\Big(f_s, \mathcal{D}^{\text{gold}}_{t,1}\Big)$. Using this we form the target-adapted predictor

$$\widehat{f}_t = f_s + \widehat{\Delta} \tag{44}$$

Once $\widehat{f}_t$ is obtained on the calibration split, using only the independent rectification split $\mathcal{I}_2$ and the unlabeled target features, define

$$\widehat{\theta}_{\text{TC}} = \frac{1}{N}\sum_{j=1}^{N} \widehat{f}_t(\widetilde{X}_j) + \frac{1}{n_2}\sum_{i \in \mathcal{I}_2} \big\{Y_i - \widehat{f}_t(X_i)\big\}. \tag{45}$$

This is the same as the PPI mean estimator, but with the calibrated target predictor $\widehat{f}_t$ in place of the original source predictor f_s. The following proposition shows the conditional unbiasedness of the estimator and an expression for conditional variance. All proofs are given in Appendix D.

Proposition 2. *Under Assumption 1,* $\mathbb{E}\big[\widehat{\theta}_{\text{TC}} \mid \mathcal{D}_s, \mathcal{D}^{\text{gold}}_{t,1}\big] = \theta^*_t$*, and consequently, unconditionally,* $\mathbb{E}[\widehat{\theta}_{\text{TC}}] = \theta^*_t$*. Further, with the notation* $V_t := \text{Var}_t(\cdot)$*,*

$$\text{Var}\big(\widehat{\theta}_{\text{TC}} \mid \mathcal{D}_s, \mathcal{D}^{\text{gold}}_{t,1}\big) = \frac{V_t\big(\widehat{f}_t(X) \mid \mathcal{D}_s, \mathcal{D}^{\text{gold}}_{t,1}\big)}{N} + \frac{V_t(\varepsilon) + V_t\big\{\widehat{\Delta}(X) - \Delta^*(X) \mid \mathcal{D}_s, \mathcal{D}^{\text{gold}}_{t,1}\big\}}{n_2}.$$

We state the asymptotic normality of the TC-PPI estimator and its limiting variance without specifying how the transfer correction is learned. Define the conditional target prediction error

$$b_{n_1} := \mathbb{E}_t\Big[\{\widehat{\Delta}(X) - \Delta^*(X)\}^2 \mid \mathcal{D}_s, \mathcal{D}^{\text{gold}}_{t,1}\Big]. \tag{46}$$

The next theorem states the limiting distribution of the TC-PPI estimator by imposing a condition on the accuracy of the target correction, b_{n_1}.

Theorem 6. *Assume Assumption 1. Suppose* Y*,* $f_t(X)$*, and* $\widehat{f}_t(X)$ *have conditional moments of order* $2 + \eta$ *exist uniformly with probability tending to one, for some* $\eta > 0$*, and*

$$b_{n_1} = o_p(1), \qquad \frac{n_2}{N} \to \rho_1 \in [0, \infty). \tag{47}$$

If the limiting variance below is positive, then, conditionally on $\mathcal{D}_s$ *and the calibration split,*

$$\frac{\sqrt{n_2}\,(\widehat{\theta}_{\text{TC}} - \theta^*_t)}{\sqrt{V_{n_1}}} \xrightarrow{D} N(0, 1), \tag{48}$$

where

$$V_{n_1} = \frac{n_2}{N}\text{Var}_t\{f_t(X)\} + \text{Var}_t(\varepsilon) + O_p\Big(\sqrt{b_{n_1}} + b_{n_1}\Big). \tag{49}$$

Consequently, $\sqrt{n_2}\,(\widehat{\theta}_{\text{TC}} - \theta^*_t) \xrightarrow{D} N(0, \rho_1 \text{Var}_t\{f_t(X)\} + \text{Var}_t(\varepsilon))$*. In the regime* $N \gg n_2$*, so that* $\rho_1 = 0$*, the asymptotic variance simplifies to* $\text{Var}_t(\varepsilon)$*.*

Sparse linear calibration: high-dimensional lasso rates The preceding construction is agnostic to the form of the transfer correction. To obtain an explicit rate, we now specialize to a sparse linear correction as follows,

$$\Delta^*(x) = x^\top \delta^*, \quad \delta^* \in \mathbb{R}^p. \tag{50}$$

We assume the features are high-dimensional, i.e., $p \gg n$, and the vector δ^* is s-sparse, i.e., $\|\delta^*\|_0 \leq s$ for some $s = s(n_1) > 0$. We estimate δ^* by the lasso on the calibration split:

$$\hat{\delta} \in \arg\min_{\delta \in \mathbb{R}^p} \left\{ \frac{1}{2n_1} \sum_{i \in \mathcal{I}_1} (Y_i - f_s(X_i) - X_i^\top \delta)^2 + \lambda \|\delta\|_1 \right\}. \tag{51}$$

The following corollary specializes Theorem 6 to the case of sparse linear calibration estimated through lasso.

Corollary 2 (Sparse linear calibration rate)**.** *Under the sparse linear calibration model equation 50 and the assumptions of Proposition 7 in the Appendix, we have,* $b_{n_1} = O_p\left(\frac{s \log p}{n_1}\right)$*. Consequently, if* $s \log p = o(n_1)$*, Theorem 6 applies and*

$$V_{n_1} = \frac{n_2}{N} \mathrm{Var}_t\{f_t(X)\} + \mathrm{Var}_t(\varepsilon) + O_p\left(\sqrt{\frac{s \log p}{n_1}}\right).$$

Thus sparse linear calibration is one of the approaches that satisfies the generic prediction-consistency condition equation 47. Our results therefore separate the role of the transfer correction learner from the validity mechanism of PPI. For a general fine-tuning procedure, it is enough for the learned correction to satisfy $b_{n_1} = o_p(1)$ in order for split TC-PPI to attain the desired asymptotic variance. In the sparse linear specialization, Proposition7 yields $b_{n_1} = O_p(s \log p / n_1)$, providing a necessary condition on the size of n_1.

We note, in the target domain, $Y - f_s(X) = \varepsilon + \Delta^*(X)$. Since $\Delta^*(X)$ is measurable in X and $\mathbb{E}_t[\varepsilon \mid X] = 0$, we have $\mathrm{Cov}_t\{\varepsilon, \Delta^*(X)\} = 0$. Hence $\mathrm{Var}_t\{Y - f_s(X)\} = \mathrm{Var}_t(\varepsilon) + \mathrm{Var}_t\{\Delta^*(X)\}$. Therefore, under the assumptions of Theorem 6, if the *same* n_2 *sized rectification sample* is used, the vanilla PPI estimator based on the unadapted source predictor f_s has asymptotic variance $V_{\mathrm{PPI}} = \rho_1 \mathrm{Var}_t\{f_s(X)\} + \mathrm{Var}_t(\varepsilon) + \mathrm{Var}_t\{\Delta^*(X)\}$. In comparison, the transfer-calibrated PPI estimator has asymptotic variance $V_{\mathrm{TC}} = \rho_1 \mathrm{Var}_t\{f_t(X)\} + \mathrm{Var}_t(\varepsilon) + O_p\left(\sqrt{b_{n_1}} + b_{n_1}\right)$. Transfer calibration replaces the variance of the source-to-target discrepancy function $\mathrm{Var}_t\{\Delta^*(X)\}$ by the variance of the calibration error $\widehat{\Delta}(X) - \Delta^*(X)$. Therefore, the rectification component improves whenever the learned correction is sufficiently more accurate (and asymptotically vanishes) than leaving the source-to-target discrepancy uncorrected. Under sparse linear calibration and mean-zero X, $\mathrm{Var}_t\{\Delta^*(X)\} = (\delta^*)^\top \Sigma_t \delta^*$. Therefore as n_1 increases, since $b_{n_1} = O_p\left(\frac{s \log p}{n_1}\right)$, the TC-PPI has a variance that is lower by $(\delta^*)^\top \Sigma_t \delta^*$ than the vanilla PPI with high probability, provided the rectification sample is of the same size.

Proposition 3 (Unbiasedness for fixed joint weights)**.** *Suppose the folds are fixed independently of the observations and have equal size* $|I_k| = n/K$*. Under Assumption 1, for every fixed* $\boldsymbol{\lambda} \in \mathbb{R}^2$*, we have* $\mathbb{E}[\widehat{\mu}_{\mathrm{JTC\text{-}Cross},\boldsymbol{\lambda}} \mid \mathcal{D}_s] = \theta_t^*$.

B.2 Variance inflation with estimated tuning vector

Theorem 3 provides a no-harm guarantee with oracle $\boldsymbol{\lambda}$. There is an additional finite-sample cost from estimating the joint power parameter. We analyze this cost next.

Let $\Delta^\dagger$ be the limiting function that the transfer correction converges to in Theorem 3, and recall

$$H_\dagger(X) = \begin{pmatrix} f_s(X) \\ \Delta^\dagger(X) \end{pmatrix}, \qquad \Sigma_\dagger = \mathrm{Var}_t\{H_\dagger(X)\}, \qquad c_\dagger = \mathrm{Cov}_t\{H_\dagger(X), Y\}.$$

For the finite sample ratio $\rho_n = n/N$, define

$$V_{\mathrm{J},n}^\dagger(\boldsymbol{\lambda}) \coloneqq \mathrm{Var}_t(Y) - 2\boldsymbol{\lambda}^\top c_\dagger + (1 + \rho_n)\boldsymbol{\lambda}^\top \Sigma_\dagger \boldsymbol{\lambda}. \tag{52}$$

For simplicity, we assume $\Sigma_\dagger$ is nonsingular. Then the minimizer of the above equation is

$$\boldsymbol{\lambda}^\dagger_{\mathrm{J},n} = \frac{1}{1+n/N}\Sigma_\dagger^{-1} c_\dagger. \tag{53}$$

Proposition 4 (Rate of feasible joint tuning)**.** *Suppose $\lambda_{\min}(\Sigma_\dagger) \geq \kappa_H > 0$, the relevant fourth moments are finite, K is fixed, $r_n = o(1)$, and the fold-specific transfer errors satisfy*

$$\max_{1\leq k\leq K} \mathbb{E}_t\Big[\{\widehat{\Delta}^{(-k)}(X) - \Delta^\dagger(X)\}^2 \mid \mathcal{D}_s, \mathcal{D}_{-k}\Big] = O_p(r_n).$$

Let $\widehat{\Sigma}_H$ and $\widehat{c}_H$ be the pooled out-of-fold estimators. Then

$$\|\widehat{\Sigma}_H - \Sigma_\dagger\|_{\mathrm{op}} + \|\widehat{c}_H - c_\dagger\|_2 = O_p\Big(n^{-1/2} + \sqrt{r_n}\Big).$$

Consequently, for the ridge estimator $\widehat{\boldsymbol{\lambda}} = \frac{1}{1+n/N}(\widehat{\Sigma}_H + \tau I_2)^{-1}\widehat{c}_H$, with $\tau = o(1)$,

$$\Big\|\widehat{\boldsymbol{\lambda}} - \boldsymbol{\lambda}^\dagger_{\mathrm{J},n}\Big\|_2 = O_p\Big(n^{-1/2} + \sqrt{r_n} + \tau\Big). \tag{54}$$

Moreover, because equation 52 is quadratic,

$$V^\dagger_{\mathrm{J},n}(\widehat{\boldsymbol{\lambda}}) - V^\dagger_{\mathrm{J},n}(\boldsymbol{\lambda}^\dagger_{\mathrm{J},n}) = (1+n/N)\left(\widehat{\boldsymbol{\lambda}} - \boldsymbol{\lambda}^\dagger_{\mathrm{J},n}\right)^\top \Sigma_\dagger \left(\widehat{\boldsymbol{\lambda}} - \boldsymbol{\lambda}^\dagger_{\mathrm{J},n}\right) = O_p\big(n^{-1} + r_n + \tau^2\big). \tag{55}$$

Proof. See Appendix D.

The above proposition shows that the excess risk of the Joint TC-Cross-PPI++ estimator with estimated power tuning parameter over the one with oracle power tuning parameter is given by

$$O_p\left(n^{-2} + \frac{r_n}{n} + \frac{\tau^2}{n}\right).$$

For the case of sparse linear calibration with fixed K, we have $r_n \asymp \frac{s\log p}{n_{\mathrm{tr}}}$. Therefore, when τ is negligible,

$$\Big\|\widehat{\boldsymbol{\lambda}} - \boldsymbol{\lambda}^\dagger_{\mathrm{J},n}\Big\|_2 = O_p\left(n^{-1/2} + \sqrt{\frac{s\log p}{n_{\mathrm{tr}}}}\right),$$

and the excess limiting variance is

$$O_p\left(\frac{1}{n} + \frac{s\log p}{n n_{\mathrm{tr}}}\right).$$

B.3 Variance estimation and confidence intervals

We now describe the variance estimators used to construct confidence intervals for the cross-fitted estimators in the experiments section. The formulas below estimate the asymptotic variances in Theorems 1, 2, and 3 from the available labeled and unlabeled samples. While our construction of the cross-fitted TC-Cross-PPI is similar to the Cross-PPI of Zrnic & Candès (2024), we do not employ the same computationally inefficient bootstrap strategy to estimate the limiting variance. In Zrnic & Candès (2024), the limiting variance is expressed in terms of an "average" model $\bar{f}$, which they propose to estimate through bootstrap. However, we show that under our assumptions we can obtain a consistent estimator without requiring bootstrap.

For a scalar sample $a_1, \ldots, a_m$, write the sample variance as,

$$\widehat{\mathrm{Var}}_m(a) := \frac{1}{m-1}\sum_{i=1}^m (a_i - \overline{a})^2, \qquad \overline{a} := \frac{1}{m}\sum_{i=1}^m a_i.$$

Likewise, for vectors $v_1, \ldots, v_m \in \mathbb{R}^d$, write the sample variance as

$$\widehat{\mathrm{Var}}_m(v) := \frac{1}{m-1}\sum_{i=1}^m (v_i - \overline{v})(v_i - \overline{v})^\top.$$

B.3.1 MEAN ESTIMATION

Recall that, for $i \in I_k$, the fitted transfer-calibrated predictor is

$$\widehat{f}_i^{\text{oof}} = \widehat{f}_t^{(-k)}(X_i).$$

For each unlabeled observation, define the fold-averaged prediction

$$\widetilde{f}_j^{\text{CF}} := \frac{1}{K}\sum_{k=1}^{K} \widehat{f}_t^{(-k)}(\widetilde{X}_j), \qquad j = 1, \dots, N.$$

Thus

$$\frac{1}{N}\sum_{j=1}^{N} \widetilde{f}_j^{\text{CF}} = \frac{1}{KN}\sum_{k=1}^{K}\sum_{j=1}^{N} \widehat{f}_t^{(-k)}(\widetilde{X}_j).$$

Then, replacing the components of the population variance with their sample versions, we can get an estimator for the variance of TC-Cross-PPI estimator with a fixed λ in the $\sqrt{n}$ scale as follows,

$$\widehat{V}_{\text{TC}}(\lambda) = \widehat{\text{Var}}_n\{Y_i - \lambda \widehat{f}_i^{\text{oof}}\} + \frac{n}{N}\widehat{\text{Var}}_N\{\lambda \widetilde{f}_j^{\text{CF}}\}. \tag{56}$$

For ordinary TC–Cross-PPI, one simply sets $\lambda = 1$. For TC–Cross-PPI++, we substitute with the estimates $\hat{\lambda}$.

The corresponding standard error is

$$\widehat{\text{SE}}_{\text{TC}} = \sqrt{\frac{\widehat{V}_{\text{TC}}(\widehat{\lambda})}{n}},$$

and the nominal $(1-\alpha)$ Wald interval is

$$\widehat{\mu}_{\text{TC-Cross},\widehat{\lambda}} \pm z_{1-\alpha/2}\widehat{\text{se}}_{\text{TC}}.$$

Joint TC–Cross-PPI++. Just as in the last part, it is useful to define the prediction-powered vector separately for each unlabeled observation. Recall that, for $i \in I_k$,

$$\widehat{H}_i^{\text{oof}} = \begin{pmatrix} f_s(X_i) \\ \widehat{\Delta}^{(-k)}(X_i) \end{pmatrix}.$$

For the unlabeled data points $j = 1, \dots, N$, define

$$\widetilde{H}_j^{\text{CF}} := \begin{pmatrix} f_s(\widetilde{X}_j) \\ \dfrac{1}{K}\displaystyle\sum_{k=1}^{K} \widehat{\Delta}^{(-k)}(\widetilde{X}_j) \end{pmatrix}.$$

Then the vector $\widetilde{H}_N^{\text{CF}}$ used in the main text satisfies

$$\widetilde{H}_N^{\text{CF}} = \frac{1}{N}\sum_{j=1}^{N} \widetilde{H}_j^{\text{CF}}.$$

For a fixed or estimated joint power vector $\boldsymbol{\lambda}$, the estimated $\sqrt{n}$-scale variance is

$$\widehat{V}_{\text{JTC}}(\boldsymbol{\lambda}) = \widehat{\text{Var}}_n\left\{Y_i - \boldsymbol{\lambda}^\top \widehat{H}_i^{\text{oof}}\right\} + \frac{n}{N}\widehat{\text{Var}}_N\left\{\boldsymbol{\lambda}^\top \widetilde{H}_j^{\text{CF}}\right\}. \tag{57}$$

We use the plug-in version $\widehat{V}_{\text{JTC}} := \widehat{V}_{\text{JTC}}(\widehat{\boldsymbol{\lambda}})$, where $\widehat{\boldsymbol{\lambda}}$ is the joint power estimate. Then the standard error is obtained as

$$\widehat{\text{SE}}_{\text{JTC}} = \sqrt{\frac{\widehat{V}_{\text{JTC}}}{n}},$$

and the corresponding Wald interval is

$$\widehat{\mu}_{\text{JTC-Cross},\widehat{\boldsymbol{\lambda}}} \pm z_{1-\alpha/2}\widehat{\text{SE}}_{\text{JTC}}.$$

Some intuition behind why the above estimators are reasonable (and indeed consistent) estimators of the limiting variances in the theorems are in order. Although the fold-specific predictors are dependent because their training samples overlap, this dependence affects only the nuisance-estimation remainder. Under foldwise consistency and fixed K conditions of the theorems, the cross-fitted estimator admits an asymptotic linear representation obtained by replacing every fitted fold predictor by its limiting function $f^\dagger$, with replacement error $o_p(n^{-1/2})$. The resulting leading terms are an empirical average over the i.i.d. gold sample and an independent empirical average over the i.i.d. unlabeled sample. Moreover, the sample variances computed using the out-of-fold predictions differ by $o_p(1)$ from those hypothetically computed using $f^\dagger$. Hence the labeled and unlabeled sample variances consistently estimate the two components of the first-order asymptotic variance.

Lemma 3 (Consistency of the variance estimators)**.** *Suppose the conditions of Theorem 2 hold. If* $\widehat{\lambda} - \lambda_\dagger^* = o_p(1)$*, then*

$$\widehat{V}_{\mathrm{TC}}(\widehat{\lambda}) \xrightarrow{P} V_{\mathrm{cross}}^\dagger(\lambda_\dagger^*).$$

Likewise, under the conditions of Theorem 3, if $\widehat{\boldsymbol{\lambda}} - \boldsymbol{\lambda}_{\mathrm{Jcross}}^* = o_p(1)$*, then*

$$\widehat{V}_{\mathrm{JTC}}(\widehat{\boldsymbol{\lambda}}) \xrightarrow{P} V_{\mathrm{Jcross}}^\dagger(\boldsymbol{\lambda}_{\mathrm{Jcross}}^*).$$

Consequently, if the corresponding limiting variance is positive, the associated Wald intervals have asymptotic coverage $1-\alpha$*.*

Proof of Lemma 3. We first consider scalar TC–Cross-PPI++. Recall that for $i \in I_k$,

$$e_k(X_i) = \widehat{f}_t^{(-k)}(X_i) - f^\dagger(X_i).$$

For each k, conditional on $\mathcal{D}_s$ and $\mathcal{D}_{-k}$, the observations in I_k are independent of e_k. Hence

$$\mathbb{E}\left[\frac{K}{n}\sum_{i\in I_k} e_k(X_i)^2 \middle| \mathcal{D}_s, \mathcal{D}_{-k}\right] = a_{n,k}.$$

Conditional Markov's inequality and $a_{n,k} = o_p(1)$ therefore give

$$\frac{K}{n}\sum_{i\in I_k} e_k(X_i)^2 = o_p(1).$$

Since K is fixed,

$$\frac{1}{n}\sum_{i=1}^{n}\left\{\widehat{f}_i^{\mathrm{oof}} - f^\dagger(X_i)\right\}^2 = o_p(1). \tag{58}$$

For the unlabeled sample, define

$$\overline{e}_n(X) := \frac{1}{K}\sum_{k=1}^{K}\left\{\widehat{f}_t^{(-k)}(X) - f^\dagger(X)\right\}.$$

By Jensen's inequality,

$$\mathbb{E}_t\{\overline{e}_n(X)^2\} \le \frac{1}{K}\sum_{k=1}^{K}\mathbb{E}_t\left[\{\widehat{f}_t^{(-k)}(X) - f^\dagger(X)\}^2 \mid \mathcal{D}_s, \mathcal{D}_{-k}\right] = o_p(1),$$

by assumptions on foldwise convergence and fixed number of folds. Conditional on the gold data, the unlabeled observations are i.i.d. and independent of all fitted fold-specific predictors. The conditional Markov's inequality therefore gives

$$\frac{1}{N}\sum_{j=1}^{N}\left\{\widetilde{f}_j^{\mathrm{CF}} - f^\dagger(\widetilde{X}_j)\right\}^2 = o_p(1). \tag{59}$$

Now define the oracle labeled and unlabeled influence contributions

$$\xi_i^{L,*} = Y_i - \lambda^* f^\dagger(X_i), \qquad \xi_j^{U,*} = \lambda^* f^\dagger(\widetilde{X}_j).$$

Their feasible counterparts that replaces λ^* by $\hat{\lambda}$ and $f^\dagger$ by $\widehat{f}_i^{\text{oof}}$ for the labeled samples and $\widetilde{f}_j^{\text{CF}}$ for the unlabeled samples are,

$$\widehat{\xi}_i^L = Y_i - \widehat{\lambda}\widehat{f}_i^{\text{oof}}, \qquad \widehat{\xi}_j^U = \widehat{\lambda}\widetilde{f}_j^{\text{CF}}.$$

Since $\widehat{\lambda} - \lambda^* = o_p(1)$, $\widehat{\lambda} = O_p(1)$. Using equation 58,

$$\frac{1}{n}\sum_{i=1}^n (\widehat{\xi}_i^L - \xi_i^{L,*})^2 \le 2(\widehat{\lambda} - \lambda^*)^2 \frac{1}{n}\sum_{i=1}^n f^\dagger(X_i)^2 + 2\widehat{\lambda}^2 \frac{1}{n}\sum_{i=1}^n \{\widehat{f}_i^{\text{oof}} - f^\dagger(X_i)\}^2 = o_p(1),$$

because $f^\dagger$ is square-integrable and hence $n^{-1}\sum_i f^\dagger(X_i)^2 = O_p(1)$. Similarly, by equation 59,

$$\frac{1}{N}\sum_{j=1}^N (\widehat{\xi}_j^U - \xi_j^{U,*})^2 = o_p(1).$$

We use the elementary fact that if

$$\frac{1}{m}\sum_{\ell=1}^m (A_\ell - B_\ell)^2 = o_p(1)$$

and

$$\frac{1}{m}\sum_{\ell=1}^m B_\ell^2 = O_p(1),$$

then

$$\widehat{\text{Var}}_m(A_\ell) - \widehat{\text{Var}}_m(B_\ell) = o_p(1).$$

Applying this fact to the labeled and unlabeled contributions gives

$$\widehat{\text{Var}}_n\left\{Y_i - \widehat{\lambda}\widehat{f}_i^{\text{oof}}\right\} = \widehat{\text{Var}}_n\left\{Y_i - \lambda^* f^\dagger(X_i)\right\} + o_p(1),$$

and

$$\widehat{\text{Var}}_N\left\{\widehat{\lambda}\widetilde{f}_j^{\text{CF}}\right\} = \widehat{\text{Var}}_N\left\{\lambda^* f^\dagger(\widetilde{X}_j)\right\} + o_p(1).$$

The quantities on the right are ordinary empirical variances based on the i.i.d. gold and unlabeled samples since a fixed function $f^\dagger$ is applied to them. Therefore, by the law of large numbers,

$$\widehat{\text{Var}}_n\left\{Y_i - \lambda^* f^\dagger(X_i)\right\} \xrightarrow{P} \text{Var}_t\{Y - \lambda^* f^\dagger(X)\},$$

and

$$\widehat{\text{Var}}_N\left\{\lambda^* f^\dagger(\widetilde{X}_j)\right\} \xrightarrow{P} (\lambda^*)^2 \text{Var}_t\{f^\dagger(X)\}.$$

Since $n/N \to \rho$, it follows that

$$\widehat{V}_{\text{TC}}(\widehat{\lambda}) \xrightarrow{P} \text{Var}_t\{Y - \lambda^* f^\dagger(X)\} + \rho(\lambda^*)^2 \text{Var}_t\{f^\dagger(X)\},$$

which is $V_{\text{cross}}^\dagger(\lambda_\dagger^*)$. The joint result follows by the same argument. Combining these results for consistency of the variance estimators with the asymptotic normality in Theorems 2 and 3, and applying Slutsky's theorem, gives the stated asymptotic coverage of the corresponding Wald intervals. □

C Extension to generalized linear models and M estimation

C.1 Generalized linear models

We first consider the problem of estimating the parameter $\beta_t^* \in \mathbb{R}^d$ in a generalized linear model. The true target parameter can be defined by the solution of the population moment equation as

follows,

$$\mathbb{E}_t[g_{\beta_t^*}(X,Y)] = 0, \qquad g_\beta(x,y) = x\{\mu(x^\top\beta) - y\},$$

where μ is a continuously differentiable inverse link function. Let $\hat{f}_t$ be the calibrated predictor constructed using the calibration split $\mathcal{D}_{t,1}^{\text{gold}}$. On the independent rectification split $\mathcal{I}_2$, define

$$\Psi_{n_2,N}(\beta; \hat{f}_t) = \frac{1}{n_2}\sum_{i\in\mathcal{I}_2} g_\beta(X_i, Y_i) - \lambda\frac{1}{n_2}\sum_{i\in\mathcal{I}_2} g_\beta(X_i, \hat{f}_t(X_i)) + \lambda\frac{1}{N}\sum_{j=1}^{N} g_\beta(\tilde{X}_j, \hat{f}_t(\tilde{X}_j)).$$

This is then a modified PPI sample moment equation. Define $\hat{\beta}_{\text{TC},\lambda}$ as any measurable solution of $\Psi_{n_2,N}(\hat{\beta}_{\text{TC},\lambda}; \hat{f}_t) = 0$.

Let f_t denote the oracle target-calibrated predictor and let $\rho = \lim n_2/N \in [0,\infty)$.

Theorem 7 (Conditional asymptotic linearity for GLMs)**.** *Assume: (i) d is fixed; (ii) μ is continuously differentiable in a neighborhood of β_t^* and $\mathbb{E}_t\Big[\sup_{\|\beta-\beta_t^*\|\le r}\|X\|^2|\mu'(X^\top\beta)|\Big] < \infty$ for some $r > 0$; (iii) $A_t := \mathbb{E}_t[\nabla_\beta g_{\beta_t^*}(X,Y)]$ is nonsingular; (iv) the relevant score variables have finite $(2+\eta)$ moments for some $\eta > 0$; and (v)*

$$\mathbb{E}_t\Big[\|X\|^2\{\hat{f}_t(X) - f_t(X)\}^2 \mid \mathcal{D}_s, \mathcal{D}_{t,1}^{\text{gold}}\Big] = o_p(1).$$

Then, conditionally on $(\mathcal{D}_s, \mathcal{D}_{t,1}^{\text{gold}})$,

$$\sqrt{n_2}(\hat{\beta}_{\text{TC},\lambda} - \beta_t^*) = -A_t^{-1}\Big[\frac{1}{\sqrt{n_2}}\sum_{i\in\mathcal{I}_2}\{g_{\beta_t^*}(X_i,Y_i) - \lambda g_{\beta_t^*}(X_i, f_t(X_i))\} + \frac{\lambda\sqrt{n_2}}{N}\sum_{j=1}^{N} g_{\beta_t^*}(\tilde{X}_j, f_t(\tilde{X}_j))\Big] + o_p(1).$$

Further,

$$\sqrt{n_2}(\hat{\beta}_{\text{TC},\lambda} - \beta_t^*) \xrightarrow{D} N\big(0, A_t^{-1}V_{\lambda,t}A_t^{-T}\big),$$

where

$$V_{\lambda,t} = \text{Var}_t\Big(g_{\beta_t^*}(X,Y) - \lambda g_{\beta_t^*}(X, f_t(X))\Big) + \lambda^2\rho\,\text{Var}_t\Big(g_{\beta_t^*}(X, f_t(X))\Big).$$

C.2 General convex Mestimation

We now formulate TC-PPI for a general low-dimensional convex M-estimation problem. The structure parallels the mean-estimation section: a source-domain predictor is first calibrated on one target split, and then the calibrated predictor is inserted into a prediction-powered estimating equation on an independent target split and the unlabeled target sample.

C.2.1 Target estimand and rectified estimating equation

Let $\Theta \subset \mathbb{R}^d$ be open, with fixed dimension d. Suppose the target parameter $\theta_t^* \in \Theta$ is defined by the moment condition

$$\mathbb{E}_t[g_{\theta_t^*}(X,Y)] = 0, \tag{60}$$

where $g_\theta(x,y) \in \mathbb{R}^d$ is the estimating function associated with a convex loss $\ell_\theta(x,y)$.

Let $\hat{f}_t$ be a calibrated predictor learned on the calibration split $\mathcal{D}_{t,1}^{\text{gold}}$, independent of the rectification split $\mathcal{I}_2$ and of the unlabeled target sample $\{\tilde{X}_j\}_{j=1}^N$.

For a fixed scalar $\lambda \in \mathbb{R}$, define the TC-PPI estimating equation

$$\Psi_{n_2,N}(\theta; \hat{f}_t) = \frac{1}{n_2}\sum_{i\in\mathcal{I}_2} g_\theta(X_i, Y_i) - \lambda\frac{1}{n_2}\sum_{i\in\mathcal{I}_2} g_\theta(X_i, \hat{f}_t(X_i)) + \lambda\frac{1}{N}\sum_{j=1}^{N} g_\theta(\tilde{X}_j, \hat{f}_t(\tilde{X}_j)). \tag{61}$$

Our proposed TC-PPI estimator is then any measurable solution

$$\hat{\theta}_{\mathrm{TC},\lambda} \in \{\theta \in \Theta : \Psi_{n_2,N}(\theta; \hat{f}_t) = 0\}. \tag{62}$$

We continue to use the same notations for the source and target predictors and the discrepancy function. Define the population rectified moment map for any predictor f as

$$\Psi_t(\theta; f) \coloneqq \mathbb{E}_t[g_\theta(X,Y)] - \lambda \mathbb{E}_t[g_\theta(X, f(X))] + \lambda \mathbb{E}_t[g_\theta(\tilde{X}, f(\tilde{X}))]. \tag{63}$$

Since X and $\tilde{X}$ have the same target covariate law,

$$\mathbb{E}_t[g_\theta(X, f(X))] = \mathbb{E}_t[g_\theta(\tilde{X}, f(\tilde{X}))],$$

and therefore

$$\Psi_t(\theta; f) = \mathbb{E}_t[g_\theta(X,Y)]. \tag{64}$$

In particular, $\Psi_t(\theta_t^*; f) = 0$ for every measurable predictor f for which the expectations exist.

C.3 Assumptions

Assumption 2 (Identification and local smoothness)**.** The parameter θ_t^* is the unique solution of equation 60 in a neighborhood $\mathcal{N}$ of θ_t^*. For almost every (x, y), the map $\theta \mapsto g_\theta(x, y)$ is continuously differentiable on $\mathcal{N}$.

Assumption 3 (Nonsingular Jacobian)**.** The matrix

$$A_t \coloneqq \mathbb{E}_t[\nabla_\theta g_{\theta_t^*}(X,Y)]$$

exists and is nonsingular.

Assumption 4 (Moment envelope)**.** There exists $\eta > 0$ and an integrable envelope $G(X,Y)$ such that

$$\sup_{\theta \in \mathcal{N}} \|g_\theta(X,Y)\| \le G(X,Y), \qquad \mathbb{E}_t[G(X,Y)^{2+\eta}] < \infty.$$

Also,

$$\mathbb{E}_t\Big[\sup_{\theta \in \mathcal{N}} \|\nabla_\theta g_\theta(X,Y)\|\Big] < \infty.$$

Analogous moment conditions hold with Y replaced by $f_t(X)$ and by $\hat{f}_t(X)$.

Assumption 5 (Calibration consistency in score norm)**.** Conditionally on $\mathcal{D}_s$ and the calibration split $\mathcal{D}_{t,1}^{\mathrm{gold}}$,

$$\mathbb{E}_t\Big[\|g_{\theta_t^*}(X, \hat{f}_t(X)) - g_{\theta_t^*}(X, f_t(X))\|^2\Big] = o_p(1),$$

and

$$\mathbb{E}_t\Big[\sup_{\theta \in \mathcal{N}} \|\nabla_\theta g_\theta(X, \hat{f}_t(X)) - \nabla_\theta g_\theta(X, f_t(X))\|\Big] = o_p(1).$$

Assumption 5 is the natural high-level calibration condition. For the mean-estimation problem with sparse linear discrepancy and lasso calibration, it follows from the rate

$$(\hat{\delta} - \delta^*)^\top \Sigma_t (\hat{\delta} - \delta^*) = O_p\Big(\frac{s \log p}{n_1}\Big)$$

together with the condition $s \log p / n_1 \to 0$. For smooth GLMs and other smooth estimating equations, the same assumptions can often be verified by Lipschitz continuity of g_θ in its second argument.

C.4 Results for split TC PPI M estimator

Proposition 5 (Exact conditional unbiasedness)**.** *Under Assumption 4, for every fixed $\theta \in \Theta$,*

$$\mathbb{E}\Big[\Psi_{n_2,N}(\theta; \hat{f}_t) \mid \mathcal{D}_s, \mathcal{D}_{t,1}^{\mathrm{gold}}\Big] = \mathbb{E}_t[g_\theta(X,Y)].$$

In particular,

$$\mathbb{E}\Big[\Psi_{n_2,N}(\theta_t^*; \hat{f}_t) \mid \mathcal{D}_s, \mathcal{D}_{t,1}^{\text{gold}}\Big] = 0.$$

Theorem 8 (Consistency)**.** *Suppose Assumptions 2–5 hold and $n_2 \to \infty$, $N \to \infty$. Then*

$$\hat{\theta}_{\text{TC},\lambda} \xrightarrow{P} \theta_t^*.$$

C.5 Asymptotic linearity and asymptotic normality

Define the oracle influence contribution

$$\zeta_{\lambda,t}(X, Y, \tilde{X}) = g_{\theta_t^*}(X, Y) - \lambda g_{\theta_t^*}(X, f_t(X)) + \lambda g_{\theta_t^*}(\tilde{X}, f_t(\tilde{X})). \tag{65}$$

Let $\rho = \lim n_2/N \in [0, \infty)$ and define

$$V_{\lambda,t} = \text{Var}_t\big(g_{\theta_t^*}(X, Y) - \lambda g_{\theta_t^*}(X, f_t(X))\big) + \lambda^2 \rho \,\text{Var}_t\big(g_{\theta_t^*}(X, f_t(X))\big). \tag{66}$$

Theorem 9 (Asymptotic linearity and normality)**.** *Suppose Assumptions 2–5 hold and $n_2/N \to \rho \in [0, \infty)$. Then*

$$\sqrt{n_2}\,(\hat{\theta}_{\text{TC},\lambda} - \theta_t^*) = -A_t^{-1}\Bigg[\frac{1}{\sqrt{n_2}} \sum_{i \in \mathcal{I}_2} \big\{g_{\theta_t^*}(X_i, Y_i) - \lambda g_{\theta_t^*}(X_i, f_t(X_i))\big\} + \frac{\lambda\sqrt{n_2}}{N} \sum_{j=1}^{N} g_{\theta_t^*}(\tilde{X}_j, f_t(\tilde{X}_j))\Bigg] + o_p(1). \tag{67}$$

Further,

$$\sqrt{n_2}\,(\hat{\theta}_{\text{TC},\lambda} - \theta_t^*) \xrightarrow{d} N\big(0,\, A_t^{-1} V_{\lambda,t} A_t^{-T}\big).$$

C.6 Cross-fitted extension for general Mestimation

Partition the gold sample into K equal folds and let $\hat{f}_t^{(-k)}$ be trained without fold I_k. Define

$$\Psi_{\text{CF}}(\theta) = \frac{1}{n} \sum_{k=1}^{K} \sum_{i \in I_k} \{g_\theta(X_i, Y_i) - \lambda g_\theta(X_i, \hat{f}_t^{(-k)}(X_i))\} + \frac{\lambda}{KN} \sum_{k=1}^{K} \sum_{j=1}^{N} g_\theta(\widetilde{X}_j, \hat{f}_t^{(-k)}(\widetilde{X}_j)).$$

Let $\hat{\theta}_{\text{TC-Cross},\lambda}$ be a consistent root of $\Psi_{\text{CF}}(\theta) = 0$.

Theorem 10 (Cross-fitted TC-PPI for general M-estimation)**.** *Suppose K is fixed and the identification, smoothness, Jacobian, and moment conditions in Assumptions 2–4 hold. Suppose, uniformly over the fixed folds,*

$$\mathbb{E}_t\Big[\|g_{\theta_t^*}(X, \hat{f}_t^{(-k)}(X)) - g_{\theta_t^*}(X, f_t(X))\|^2 \mid \mathcal{D}_s, \mathcal{D}_{-k}\Big] = o_p(1),$$

with the analogous Jacobian consistency condition, and let $n/N \to \rho \in [0, \infty)$. Then

$$\sqrt{n}\,(\hat{\theta}_{\text{TC-Cross},\lambda} - \theta_t^*) = -A_t^{-1}\Bigg[\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \{g_{\theta_t^*}(X_i, Y_i) - \lambda g_{\theta_t^*}(X_i, f_t(X_i))\} + \frac{\lambda\sqrt{n}}{N} \sum_{j=1}^{N} g_{\theta_t^*}(\widetilde{X}_j, f_t(\widetilde{X}_j))\Bigg] + o_p(1),$$

and consequently

$$\sqrt{n}\,(\hat{\theta}_{\text{TC-Cross},\lambda} - \theta_t^*) \xrightarrow{d} N\big(0, A_t^{-1} V_{\lambda,t}^{\text{cross}} A_t^{-T}\big),$$

where

$$V_{\lambda,t}^{\text{cross}} = \text{Var}_t\{g_{\theta_t^*}(X, Y) - \lambda g_{\theta_t^*}(X, f_t(X))\} + \lambda^2 \rho \text{Var}_t\{g_{\theta_t^*}(X, f_t(X))\}.$$

Thus the split-sample result with scale n_2 is replaced by a full-sample result with scale n. This theorem covers TC–Cross-PPI++ by allowing $\lambda \neq 1$; a consistent data-dependent $\widehat{\lambda}$ can be substituted without changing the first-order limit under the usual stochastic equicontinuity condition.

C.7 PROOFS OF RESULTS FROM THIS SECTION

Proof of Theorem 7. A mean-value expansion of the estimating equation around β_t^* gives

$$\sqrt{n_2}(\hat{\beta}_{\mathrm{TC},\lambda} - \beta_t^*) = -A_t^{-1}\sqrt{n_2}\Psi_{n_2,N}(\beta_t^*; \hat{f}_t) + o_p(1),$$

by standard fixed-dimensional Z-estimation arguments and the Jacobian assumptions. It remains to replace $\hat{f}_t$ by f_t. Put

$$h_{\hat{f}}(X) = X\{\hat{f}_t(X) - f_t(X)\}.$$

The difference between the two scores is

$$\lambda\sqrt{n_2}\left\{\frac{1}{n_2}\sum_{i\in\mathcal{I}_2} h_{\hat{f}}(X_i) - \frac{1}{N}\sum_{j=1}^{N} h_{\hat{f}}(\widetilde{X}_j)\right\}.$$

Conditional on the calibration data, the two empirical averages have the same population mean, so the mean cancels exactly. Their conditional variances are bounded by the weighted L_2 calibration error times $1/n_2$ and $1/N$, respectively. Since $n_2/N = O(1)$, Chebyshev's inequality makes the displayed term $o_p(1)$. Substituting f_t therefore yields the claimed expansion.

Apply the multivariate Lindeberg–Feller central limit theorem to the two independent empirical averages in Theorem 7. The labeled contribution has covariance $\mathrm{Var}_t\{g_{\beta_t^*}(X,Y) - \lambda g_{\beta_t^*}(X, f_t(X))\}$, while the unlabeled contribution has covariance $\lambda^2(n_2/N)\mathrm{Var}_t\{g_{\beta_t^*}(X, f_t(X))\}$. Slutsky's theorem and $n_2/N \to \rho$ give the result. □

Proof of Proposition 5. Condition on $\mathcal{D}_s$ and $\mathcal{D}_{t,1}^{\mathrm{gold}}$, so that $\hat{f}_t$ is fixed. Because the rectification split $\{(X_i, Y_i)\}_{i\in\mathcal{I}_2}$ and the unlabeled sample $\{\tilde{X}_j\}_{j=1}^N$ are independent of the conditioning sigma-field and are sampled from the target-domain law,

$$\mathbb{E}\left[\frac{1}{n_2}\sum_{i\in\mathcal{I}_2} g_\theta(X_i, Y_i)\middle|\mathcal{D}_s, \mathcal{D}_{t,1}^{\mathrm{gold}}\right] = \mathbb{E}_t[g_\theta(X,Y)],$$

$$\mathbb{E}\left[\frac{1}{n_2}\sum_{i\in\mathcal{I}_2} g_\theta(X_i, \hat{f}_t(X_i))\middle|\mathcal{D}_s, \mathcal{D}_{t,1}^{\mathrm{gold}}\right] = \mathbb{E}_t[g_\theta(X, \hat{f}_t(X))],$$

and

$$\mathbb{E}\left[\frac{1}{N}\sum_{j=1}^{N} g_\theta(\tilde{X}_j, \hat{f}_t(\tilde{X}_j))\middle|\mathcal{D}_s, \mathcal{D}_{t,1}^{\mathrm{gold}}\right] = \mathbb{E}_t[g_\theta(X, \hat{f}_t(X))].$$

Substituting into equation 61 proves the claim. □

Proof of Theorem 8. By Proposition 5 and the uniform law of large numbers implied by Assumption 4,

$$\sup_{\theta\in\mathcal{N}}\left\|\Psi_{n_2,N}(\theta; \hat{f}_t) - \mathbb{E}_t[g_\theta(X,Y)]\right\| = o_p(1).$$

By Assumption 2, the map $\theta \mapsto \mathbb{E}_t[g_\theta(X,Y)]$ has a unique zero at θ_t^* in $\mathcal{N}$. Standard Z-estimation consistency then gives

$$\hat{\theta}_{\mathrm{TC},\lambda} \xrightarrow{P} \theta_t^*.$$

□

Proof of Theorem 9. Since $\hat{\theta}_{\mathrm{TC},\lambda}$ solves equation 62,

$$0 = \Psi_{n_2,N}(\hat{\theta}_{\mathrm{TC},\lambda}; \hat{f}_t).$$

Expand around θ_t^*:

$$0 = \Psi_{n_2,N}(\theta_t^*; \hat{f}_t) + \dot{\Psi}_{n_2,N}(\bar{\theta}; \hat{f}_t)(\hat{\theta}_{\mathrm{TC},\lambda} - \theta_t^*)$$

for some random $\bar{\theta}$ between $\hat{\theta}_{\mathrm{TC},\lambda}$ and θ_t^*. By Theorem 8, $\bar{\theta} \to \theta_t^*$ in probability. By Assumptions 4–5 and a uniform law of large numbers,

$$\dot{\Psi}_{n_2,N}(\bar{\theta}; \hat{f}_t) = A_t + o_p(1).$$

Hence,

$$\sqrt{n_2}(\hat{\theta}_{\mathrm{TC},\lambda} - \theta_t^*) = -A_t^{-1}\sqrt{n_2}\,\Psi_{n_2,N}(\theta_t^*; \hat{f}_t) + o_p(1).$$

It remains to replace $\hat{f}_t$ by f_t in the score at θ_t^*. Let

$$h_{\hat{f}}(X) = g_{\theta_t^*}(X, \hat{f}_t(X)) - g_{\theta_t^*}(X, f_t(X)).$$

The replacement error equals

$$\lambda\sqrt{n_2}\left\{\frac{1}{N}\sum_{j=1}^{N} h_{\hat{f}}(\widetilde{X}_j) - \frac{1}{n_2}\sum_{i\in\mathcal{I}_2} h_{\hat{f}}(X_i)\right\}.$$

Conditional on the calibration data, the two empirical averages have the same population mean, so that mean cancels exactly. Assumption 5 implies that the conditional variance of the displayed difference is $o_p(1)$ after multiplication by n_2, because $n_2/N = O(1)$. Hence Chebyshev's inequality gives

$$\sqrt{n_2}\left(\Psi_{n_2,N}(\theta_t^*; \hat{f}_t) - \Psi_{n_2,N}(\theta_t^*; f_t)\right) = o_p(1).$$

Substituting this into the previous display proves equation 67.

The bracketed term in equation 67 is a sum of two independent empirical averages, one based on the rectification split and one based on the unlabeled sample. By Assumption 4, the Lindeberg–Feller central limit theorem applies, yielding

$$\frac{1}{\sqrt{n_2}}\sum_{i\in\mathcal{I}_2}\left\{g_{\theta_t^*}(X_i, Y_i) - \lambda g_{\theta_t^*}(X_i, f_t(X_i))\right\} + \frac{\lambda\sqrt{n_2}}{N}\sum_{j=1}^{N} g_{\theta_t^*}(\tilde{X}_j, f_t(\tilde{X}_j)) \xrightarrow{d} N(0, V_{\lambda,t}).$$

Combine this with Theorem 9 and Slutsky's theorem. □

Proof of Theorem 10. Expand the cross-fitted estimating equation around θ_t^*. The Jacobian converges to A_t by the assumed foldwise Jacobian consistency and fixed K. For the nuisance replacement, define

$$h_k(X) = g_{\theta_t^*}(X, \hat{f}_t^{(-k)}(X)) - g_{\theta_t^*}(X, f_t(X)).$$

For each fold, conditional on its training complement, the held-out average of h_k and the unlabeled average of h_k have the same population mean; hence that mean cancels. The foldwise L_2 condition and Chebyshev's inequality imply that the sum of the resulting centered differences is $o_p(n^{-1/2})$ because K is fixed and $n/N = O(1)$. Replacing every fold predictor by f_t therefore yields the displayed full-sample asymptotic linear representation. The multivariate CLT for the labeled and unlabeled oracle empirical averages gives the stated covariance. □

D PROOFS OF MAIN RESULTS

D.1 SPLIT-TC-PPI PROOFS

Proof of Proposition 2. Condition on $\mathcal{D}_s$ and the calibration split $\mathcal{D}_{t,1}^{\mathrm{gold}}$, so that $\widehat{f}_t$ is fixed. Because the unlabeled sample $\tilde{X}$ and the rectification split $\{X_i, Y_i\}_{i\in\mathcal{I}_2}$ are independent of the conditioning sigma-field $\sigma(\mathcal{D}_s, \mathcal{D}_{t,1}^{\mathrm{gold}})$, and each has the target-domain covariate or joint law,

$$\mathbb{E}\left[\frac{1}{N}\sum_{j=1}^{N}\widehat{f}_t(\widetilde{X}_j)\middle|\mathcal{D}_s, \mathcal{D}_{t,1}^{\mathrm{gold}}\right] = \mathbb{E}_t[\widehat{f}_t(X)]$$

and

$$\mathbb{E}\left[\frac{1}{n_2}\sum_{i\in\mathcal{I}_2}\{Y_i-\widehat{f}_t(X_i)\}\middle|\mathcal{D}_s,\mathcal{D}_{t,1}^{\text{gold}}\right]=\mathbb{E}_t[Y-\widehat{f}_t(X)].$$

Summing gives $\mathbb{E}\big[\widehat{\theta}_{\text{TC}}\mid\mathcal{D}_s,\mathcal{D}_{t,1}^{\text{gold}}\big]=\theta_t^*$. The law of iterated expectations give the unconditional expectation.

Now, conditional on $\mathcal{D}_s$ and the calibration split $\mathcal{D}_{t,1}^{\text{gold}}$, the two empirical means which are being added in equation 45 are each computed from independent samples. Therefore they have covariance 0 and consequently, the variance of the TC-PPI estimator is the sum of their variances.

Now let,

$$e(X)\coloneqq\widehat{\Delta}(X)-\Delta^*(X),$$

denote the estimation error of the discrepancy. Combining equation 1, equation 44, and definition of ε, we get

$$Y-\widehat{f}_t(X)=\varepsilon-e(X).$$

Conditional on $\mathcal{D}_s$ and $\mathcal{D}_{t,1}^{\text{gold}}$, the function e is fixed and measurable in X. Since $\mathbb{E}_t[\varepsilon\mid X]=0$,

$$\text{Cov}_t\{\varepsilon,e(X)\mid\mathcal{D}_s,\mathcal{D}_{t,1}^{\text{gold}}\}=\mathbb{E}_t\Big[e(X)\mathbb{E}_t(\varepsilon\mid X)\mid\mathcal{D}_s,\mathcal{D}_{t,1}^{\text{gold}}\Big]=0.$$

Therefore

$$\text{Var}_t\{Y-\widehat{f}_t(X)\mid\mathcal{D}_s,\mathcal{D}_{t,1}^{\text{gold}}\}=\text{Var}_t(\varepsilon)+\text{Var}_t\{e(X)\mid\mathcal{D}_s,\mathcal{D}_{t,1}^{\text{gold}}\},$$

which is the expression for the conditional variance. □

Proof of Theorem 6. We again condition on $\mathcal{D}_s$ and the calibration split $\mathcal{D}_{t,1}^{\text{gold}}$. Then the quantities

$$A_j=\widehat{f}_t(\widetilde{X}_j)-\mathbb{E}_t\{\widehat{f}_t(X)\},\qquad j=1,\dots,N,$$

and

$$B_i=Y_i-\widehat{f}_t(X_i)-\mathbb{E}_t\{Y-\widehat{f}_t(X)\},\qquad i\in\mathcal{I}_2,$$

are independent, centered, and conditionally i.i.d. within each group. The stated conditional $(2+\eta)$ moment condition gives the Lindeberg condition, and therefore the triangular-array CLT yields the asymptotic normality equation 48 with conditional variance V_{n_1} as follows,

$$V_{n_1}=\frac{n_2}{N}\text{Var}_t\big(\widehat{f}_t(X)\mid\mathcal{D}_s,\mathcal{D}_{t,1}^{\text{gold}}\big)+\text{Var}_t\big(Y-\widehat{f}_t(X)\mid\mathcal{D}_s,\mathcal{D}_{t,1}^{\text{gold}}\big).\tag{68}$$

By Proposition 2,

$$\text{Var}_t\{Y-\widehat{f}_t(X)\mid\mathcal{D}_s,\mathcal{D}_{t,1}^{\text{gold}}\}=\text{Var}_t(\varepsilon)+\text{Var}_t\{e(X)\mid\mathcal{D}_s,\mathcal{D}_{t,1}^{\text{gold}}\}=\text{Var}_t(\varepsilon)+O_p(b_{n_1}),$$

since $\text{Var}_t\{e(X)\}\le\mathbb{E}_t[e(X)^2]=b_{n_1}$. Also since $\widehat{f}_t=f_t+e$, we have,

$$\begin{aligned}\text{Var}_t\{\widehat{f}_t(X)\mid\mathcal{D}_s,\mathcal{D}_{t,1}^{\text{gold}}\}=\text{Var}_t\{f_t(X)\}+2\text{Cov}_t\{f_t(X),e(X)\mid\mathcal{D}_s,\mathcal{D}_{t,1}^{\text{gold}}\}\\+\text{Var}_t\{e(X)\mid\mathcal{D}_s,\mathcal{D}_{t,1}^{\text{gold}}\}.\end{aligned}$$

By Cauchy–Schwarz inequality, the covariance term is $O_p(\sqrt{b_{n_1}})$, while the last term is $O_p(b_{n_1})$. This proves equation 49. Since $b_{n_1}=o_p(1)$ and $n_2/N\to\rho_1$, we have V_{n_1} converges in probability to $\rho_1\text{Var}_t\{f_t(X)\}+\text{Var}_t(\varepsilon)$. □

D.2 LASSO RESULTS

Assumption 6 (Target design and noise)**.** The target covariate $X\in\mathbb{R}^p$ is mean-zero and sub-Gaussian with covariance matrix Σ_t. There exist constants $0<\underline{\kappa}\le\overline{\kappa}<\infty$ such that

$$\underline{\kappa}\le\lambda_{\min}(\Sigma_t)\le\lambda_{\max}(\Sigma_t)\le\overline{\kappa}.$$

Further, the conditional distribution of noise $\varepsilon|X$, in $\varepsilon = Y - f_t(X)$, is sub-Gaussian with variance proxy σ_ε^2.

For $S \coloneqq \mathrm{supp}(\delta^*)$, let $s = |S|$ and let

$$\hat{\Sigma}_1 \coloneqq \frac{1}{n_1} \sum_{i \in \mathcal{I}_1} X_i X_i^\top, \quad \text{so that} \quad \mathbb{E}_t[\hat{\Sigma}_1] = \Sigma_t.$$

We will use the following standard compatibility condition on the sample Gram matrix.

Assumption 7 (Sparsity)**.** The vector δ^* is s-sparse. That is, $\|\delta^*\|_0 \le s$ for some $s = s(n_1) > 0$.

Assumption 8 (Compatibility)**.** Let $C_{n_1}(\phi) \coloneqq \left\{ \min_{u \neq 0;\ \|u_{S^c}\|_1 \le 3\|u_S\|_1} \frac{\sqrt{s}\,\|u\|_{\hat{\Sigma}_1}}{\|u_S\|_1} \ge \phi \right\}$ where $\|u\|_{\hat{\Sigma}_1}^2 \coloneqq u^\top \hat{\Sigma}_1 u$. There exists a constant $\phi_0 > 0$ such that

$$\mathbb{P}\Big[C_{n_1}(\phi_0)\Big] \overset{n_1 \to \infty}{\to} 1.$$

Assumption 9 (Score control)**.** For some deterministic sequence $\lambda = \lambda_{n_1,p} > 0$,

$$\mathbb{P}\left(\left\| \frac{1}{n_1} \sum_{i \in \mathcal{I}_1} X_i \varepsilon_i \right\|_\infty \le \frac{\lambda}{2} \,\middle|\, \mathcal{D}_s \right) \to 1.$$

Remark 3. Assumption 9 is a high-level condition. It can be verified under a variety of primitive assumptions. For example, it holds with $\lambda \asymp \sqrt{(\log p)/n_1}$ if the coordinates of X are sub-Gaussian and $X_j \varepsilon$ are uniformly sub-exponential. We do not require marginal sub-Gaussianity of ε, which would be too restrictive when f_s is an arbitrary black-box predictor.

We first state the deterministic oracle inequality.

Proposition 6 (Deterministic lasso bound)**.** *Suppose Assumption 7 holds. Then, on the events $C_{n_1}(\phi_0)$ and*

$$\mathcal{E}_\lambda \coloneqq \left\{ \left\| \frac{1}{n_1} \sum_{i \in \mathcal{I}_1} X_i \varepsilon_i \right\|_\infty \le \frac{\lambda}{2} \right\},$$

any solution $\hat{\delta}$ of equation 51 satisfies

$$\|\hat{\delta} - \delta^*\|_1 \le \frac{12 s \lambda}{\phi_0^2}, \tag{69}$$

$$\|\hat{\delta} - \delta^*\|_{\hat{\Sigma}_1}^2 \le \frac{9 s \lambda^2}{\phi_0^2}. \tag{70}$$

The next proposition gives the calibration rate used later in the TC-PPI mean analysis.

Proposition 7 (Calibration lasso rate)**.** *Suppose Assumptions 7–9 in the Appendix hold. In addition, assume that on the lasso cone $\mathcal{C}(S,3) = \{u : \|u_{S^c}\|_1 \le 3\|u_S\|_1\}$ there is a constant $C_{\mathrm{pop}} > 0$ such that, with probability tending to one,*

$$u^\top \Sigma_t u \le C_{\mathrm{pop}}\, u^\top \widehat{\Sigma}_1 u \qquad \textit{for all } u \in \mathcal{C}(S,3). \tag{71}$$

Then any lasso solution $\hat{\delta}$ of equation 51 satisfies

$$\|\hat{\delta} - \delta^*\|_1 = O_{\mathbb{P}}(s\lambda), \tag{72}$$

$$(\hat{\delta} - \delta^*)^\top \Sigma_t (\hat{\delta} - \delta^*) = O_{\mathbb{P}}(s\lambda^2). \tag{73}$$

In particular, if $\lambda \asymp \sqrt{(\log p)/n_1}$, then

$$\|\hat{\delta} - \delta^*\|_1 = O_{\mathbb{P}}\left(s\sqrt{\frac{\log p}{n_1}} \right), \quad (\hat{\delta} - \delta^*)^\top \Sigma_t (\hat{\delta} - \delta^*) = O_{\mathbb{P}}\left(\frac{s \log p}{n_1} \right).$$

Lemma 4 (Lasso calibration risk in expectation)**.** *Suppose Assumptions 6 and 7 hold and* $\|\delta^*\|_2 \le B$*. For fold* k*, let* $\widehat{\delta}^{(-k)}$ *be any solution of the lasso on* $\mathcal{D}_{-k}$*,*

$$\widehat{\delta}^{(-k)} \in \arg\min_{\delta\in\mathbb{R}^p}\Big\{\frac{1}{2n_{\mathrm{tr}}}\sum_{i\notin I_k}\big(Y_i - f_s(X_i) - X_i^\top\delta\big)^2 + \lambda\|\delta\|_1\Big\}, \qquad \lambda = A\sigma_\varepsilon\sqrt{\frac{\log p}{n_{\mathrm{tr}}}}.$$

Assume $p \ge \max\{3, n_{\mathrm{tr}}^{\upsilon}\}$ *for some* $\upsilon > 0$*. Then there exist constants* $A_0, c_0, C > 0$*, depending only on* $\underline{\kappa}, \overline{\kappa}, \upsilon$ *and* B/σ_ε*, such that if* $A \ge A_0$ *and* $s\log p/n_{\mathrm{tr}} \le c_0$ *(with* c_0 *allowed to depend on* A*), then for every* k*,*

$$\mathbb{E}\Big[(\widehat{\delta}^{(-k)} - \delta^*)^\top\Sigma_t(\widehat{\delta}^{(-k)} - \delta^*)\,\Big|\,\mathcal{D}_s\Big] \le C\,\frac{\sigma_\varepsilon^2\, s\log p}{n_{\mathrm{tr}}}.$$

In particular, this lemma implies that $r_n \le C\sigma_\varepsilon^2 s\log p/n_{\mathrm{tr}}$ under the lasso-calibration.

D.2.1 Proofs for this section

Proof of Propostion 6. Let $\Delta = \hat{\delta} - \delta^*$. By optimality of $\hat{\delta}$,

$$\frac{1}{2n_1}\|X\Delta - \varepsilon\|_2^2 + \lambda\|\hat{\delta}\|_1 \le \frac{1}{2n_1}\|\varepsilon\|_2^2 + \lambda\|\delta^*\|_1,$$

where X is the $n_1 \times p$ design matrix on the calibration split and $\varepsilon = (\varepsilon_i)_{i\in\mathcal{D}_{t,1}^{\mathrm{gold}}}$. Expanding and rearranging gives

$$\frac{1}{2n_1}\|X\Delta\|_2^2 \le \frac{1}{n_1}\varepsilon^\top X\Delta + \lambda(\|\delta^*\|_1 - \|\hat{\delta}\|_1).$$

On $\mathcal{E}_\lambda$,

$$\frac{1}{n_1}\varepsilon^\top X\Delta \le \Big\|\frac{1}{n_1}X^\top\varepsilon\Big\|_\infty\|\Delta\|_1 \le \frac{\lambda}{2}\|\Delta\|_1.$$

Since $\delta^*_{S^c} = 0$,

$$\|\delta^*\|_1 - \|\hat{\delta}\|_1 \le \|\Delta_S\|_1 - \|\Delta_{S^c}\|_1.$$

Hence

$$\frac{1}{2}\|\Delta\|_{\hat{\Sigma}_1}^2 \le \frac{\lambda}{2}(\|\Delta_S\|_1 + \|\Delta_{S^c}\|_1) + \lambda(\|\Delta_S\|_1 - \|\Delta_{S^c}\|_1) = \frac{3\lambda}{2}\|\Delta_S\|_1 - \frac{\lambda}{2}\|\Delta_{S^c}\|_1.$$

Therefore

$$\|\Delta_{S^c}\|_1 \le 3\|\Delta_S\|_1,$$

so Δ belongs to the compatibility cone. On the compatibility event in Assumption 8,

$$\|\Delta_S\|_1 \le \frac{\sqrt{s}}{\phi_0}\|\Delta\|_{\hat{\Sigma}_1}.$$

Combining with the previous display yields

$$\frac{1}{2}\|\Delta\|_{\hat{\Sigma}_1}^2 \le \frac{3\lambda\sqrt{s}}{2\phi_0}\|\Delta\|_{\hat{\Sigma}_1},$$

which implies

$$\|\Delta\|_{\hat{\Sigma}_1} \le \frac{3\lambda\sqrt{s}}{\phi_0}.$$

This gives equation 70. Then

$$\|\Delta\|_1 \le \|\Delta_S\|_1 + \|\Delta_{S^c}\|_1 \le 4\|\Delta_S\|_1 \le \frac{4\sqrt{s}}{\phi_0}\|\Delta\|_{\hat{\Sigma}_1},$$

which gives equation 69. □

Proof of Proposition 7. By Assumption 9, the score event $\mathcal{E}_\lambda$ holds with probability tending to one. Proposition 6 then gives the cone condition and

$$\|\hat{\delta} - \delta^*\|_{\widehat{\Sigma}_1}^2 \le Cs\lambda^2, \qquad \|\hat{\delta} - \delta^*\|_1 \le Cs\lambda.$$

Since $\Delta = \hat{\delta} - \delta^*$ belongs to $\mathcal{C}(S, 3)$, equation 71 implies

$$\Delta^\top \Sigma_t \Delta \le C_{\text{pop}} \Delta^\top \widehat{\Sigma}_1 \Delta = O_p(s\lambda^2),$$

hence the claimed rates follow. Under a sub-Gaussian random design, equation 71 is a standard restricted covariance comparison and holds under the usual scaling $n_1 \gtrsim s \log p$ with appropriate constants. □

Proof of Corollary 2. Under sparse linear calibration,

$$\widehat{\Delta}(X) - \Delta^*(X) = X^\top(\widehat{\delta} - \delta^*).$$

Because X is mean-zero under Assumption 6,

$$b_{n_1} = (\widehat{\delta} - \delta^*)^\top \Sigma_t (\widehat{\delta} - \delta^*).$$

Therefore, Proposition 7 therefore gives

$$b_{n_1} = O_p\left(\frac{s \log p}{n_1}\right).$$

Substitution into Theorem 6 proves the result. □

Proof of Lemma 4. Fix k and condition on $\mathcal{D}_s$ throughout, so that f_s is fixed; the target sample $\mathcal{D}_{-k}$ is independent of $\mathcal{D}_s$. Write $n \equiv n_{\text{tr}}$, let $X \in \mathbb{R}^{n \times p}$ be the design on $\mathcal{D}_{-k}$, $\widehat{\Sigma} = n^{-1} X^\top X$, $\tilde{y}_i = Y_i - f_s(X_i) = X_i^\top \delta^* + \varepsilon_i$, $\widehat{\delta} \equiv \widehat{\delta}^{(-k)}$ and $\Delta = \widehat{\delta} - \delta^*$. Let $S = \operatorname{supp}(\delta^*)$ and $\mathcal{C}(S, 3) = \{u : \|u_{S^c}\|_1 \le 3\|u_S\|_1\}$. Throughout, $c, c', c_0, c_1, c_2, \ldots$ and $C_1, C_2, \ldots$ denote generic positive constants.

By Raskutti et al. (2010, Theorem 1), if $s \log p / n \le c_0$,

$$\mathcal{R} := \left\{ u^\top \widehat{\Sigma} u \;\ge\; c\, u^\top \Sigma_t u \;\text{ for all } u \in \mathcal{C}(S, 3) \right\}$$

holds with $\mathbb{P}(\mathcal{R}^c) \le c_1 e^{-c_2 n}$. Also, taking $\lambda = A\sigma_\varepsilon \sqrt{\log p / n}$, we have

$$\mathbb{P}(\mathcal{S}^c) \le 2p^{\,1 - c_4 A^2}, \qquad \mathcal{S} := \left\{ \left\| n^{-1} X^\top \varepsilon \right\|_\infty \le \lambda/2 \right\}.$$

Let $\mathcal{E} := \mathcal{R} \cap \mathcal{S}$, so that

$$\mathbb{P}(\mathcal{E}^c) \le c_1 e^{-c_2 n} + 2p^{\,1 - c_4 A^2}. \tag{74}$$

The proof of Proposition 6 shows that on $\mathcal{S}$, $\Delta \in \mathcal{C}(S, 3)$ and $\frac{1}{2}\|\Delta\|_{\widehat{\Sigma}}^2 \le \frac{3\lambda}{2}\|\Delta_S\|_1$. On $\mathcal{R}$, $\|\Delta_S\|_1 \le \sqrt{s}\|\Delta\|_2 \le \sqrt{s/\underline{\kappa}}\,\|\Delta\|_{\Sigma_t} \le \sqrt{s/(\underline{\kappa} c)}\,\|\Delta\|_{\widehat{\Sigma}}$, so $\|\Delta\|_{\widehat{\Sigma}} \le c'\lambda\sqrt{s/(\underline{\kappa} c)}$ and

$$\Delta^\top \Sigma_t \Delta \;\le\; \frac{\|\Delta\|_{\widehat{\Sigma}}^2}{c} \;\le\; \frac{9A^2}{\underline{\kappa}\, c^2} \frac{\sigma_\varepsilon^2 s \log p}{n} \le c' \frac{A^2 s \sigma_\varepsilon^2 \log p}{n} \qquad \text{on } \mathcal{E}. \tag{75}$$

Since equation 75 holds pointwise on $\mathcal{E}$ and its right-hand side is deterministic given $\mathcal{D}_s$,

$$\mathbb{E}\big[\Delta^\top \Sigma_t \Delta\, \mathbf{1}_{\mathcal{E}} \mid \mathcal{D}_s\big] \le c' \frac{A^2 s \sigma_\varepsilon^2 \log p}{n}. \tag{76}$$

Next, comparing the lasso objective at $\widehat{\delta}$ and at 0,

$$\lambda \|\widehat{\delta}\|_1 \le \frac{1}{2n}\|\tilde{y}\|_2^2 =: \frac{T}{2}, \qquad \text{so} \qquad \|\widehat{\delta}\|_1 \le \frac{T}{2\lambda}.$$

Since $\|\delta^*\|_1 \le \sqrt{s}\,\|\delta^*\|_2 \le \sqrt{s}\,B$,

$$\Delta^\top \Sigma_t \Delta \le \overline{\kappa}\|\Delta\|_2^2 \le \overline{\kappa}\|\Delta\|_1^2 \le 2\overline{\kappa}\Big(\frac{T^2}{4\lambda^2} + sB^2\Big).$$

Since $x \mapsto x^4$ is convex, applying it to the empirical average gives $T^4 = \big(n^{-1}\sum_i \tilde{y}_i^2\big)^4 \le n^{-1}\sum_i \tilde{y}_i^8$. As the $\tilde{y}_i$ are i.i.d. given $\mathcal{D}_s$, $\mathbb{E}[T^4 \mid \mathcal{D}_s] \le \mathbb{E}[\tilde{y}_1^8 \mid \mathcal{D}_s]$. Since $\tilde{y} = X^\top \delta^* + \varepsilon$ is

sub-Gaussian with $\|\tilde{y}\|_{\psi_2}^2 \le C_2 M$, where $M := \sigma_\varepsilon^2 + \overline{\kappa} B^2$, we have $\mathbb{E}\tilde{y}^8 \le C_3 M^4$. Therefore

$$\big(\mathbb{E}[(\Delta^\top \Sigma_t \Delta)^2]\big)^{1/2} \le C_4\Big(\frac{M^2}{\lambda^2} + sB^2\Big) = C_4\Big(\frac{M^2 n}{A^2 \sigma_\varepsilon^2 \log p} + sB^2\Big) \le \gamma\, n,$$

where $\gamma := C_5\{M^2/\sigma_\varepsilon^2 + B^2\}$, using $\log p \ge 1$ and $s \le n$. By Cauchy–Schwarz and equation 74,

$$\mathbb{E}\big[\Delta^\top \Sigma_t \Delta\, \mathbf{1}_{\mathcal{E}^c}\big] \le \gamma n\, \mathbb{P}(\mathcal{E}^c)^{1/2} \le c'' \gamma n \Big(e^{-c_2 n/2} + p^{(1-c_4 A^2)/2}\Big).$$

Choose $A_0 \ge 1$ with $c_4 A_0^2 \ge 1 + 6/\upsilon$. Then for $A \ge A_0$ and $p \ge n^\upsilon$, $p^{(1-c_4A^2)/2} \le n^{-\upsilon(c_4A^2-1)/2} \le n^{-3}$. Also $ne^{-c_2 n/2} \le C_6 n^{-2}$. Hence

$$\mathbb{E}\big[\Delta^\top \Sigma_t \Delta\, \mathbf{1}_{\mathcal{E}^c}\big] \le C_6 \gamma\, n^{-2} \le \frac{C_6 \gamma}{\sigma_\varepsilon^2} \frac{\sigma_\varepsilon^2 s \log p}{n}, \tag{77}$$

where the last step uses $s \ge 1$ and $\log p \ge 1$.

Combining the bounds on $\mathcal{E}$ and $\mathcal{E}^c$ completes the proof. □

D.3 TC-CROSS-PPI PROOFS

Proof of Proposition 1. For each fold k, let

$$\mathcal{D}_{-k} = \{(X_i, Y_i) : i \notin I_k\}.$$

Condition on $\mathcal{D}_s$ and $\mathcal{D}_{-k}$. Then $\widehat{f}_t^{(-k)}$ is fixed, while the held-out observations $\{(X_i, Y_i) : i \in I_k\}$ are independent of $\mathcal{D}_{-k}$ and have law P_t. The unlabeled sample is also independent of the gold sample and has covariate law $P_{t,X}$. Therefore,

$$\mathbb{E}\left[\frac{1}{N}\sum_{j=1}^{N} \widehat{f}_t^{(-k)}(\widetilde{X}_j)\,\middle|\,\mathcal{D}_s, \mathcal{D}_{-k}\right] = \mathbb{E}_t\left[\widehat{f}_t^{(-k)}(X)\,\middle|\,\mathcal{D}_s, \mathcal{D}_{-k}\right],$$

whereas, since $|I_k| = n/K$,

$$\begin{aligned}\mathbb{E}\left[\frac{K}{n}\sum_{i\in I_k}\{Y_i - \widehat{f}_t^{(-k)}(X_i)\}\,\middle|\,\mathcal{D}_s, \mathcal{D}_{-k}\right] \\ = \mathbb{E}_t[Y] - \mathbb{E}_t\left[\widehat{f}_t^{(-k)}(X)\,\middle|\,\mathcal{D}_s, \mathcal{D}_{-k}\right].\end{aligned}$$

Hence the conditional expectation of the kth bracket in equation 2 is θ_t^*. By linearity and iterated expectation,

$$\begin{aligned}\mathbb{E}[\widehat{\mu}_{\text{TC-Cross}} \mid \mathcal{D}_s] &= \frac{1}{K}\sum_{k=1}^{K} \mathbb{E}[\mathbb{E}[T_k \mid \mathcal{D}_s, \mathcal{D}_{-k}]|\mathcal{D}_s] \\ &= \frac{1}{K}\sum_{k=1}^{K} \theta_t^* = \theta_t^*,\end{aligned}$$

where T_k denotes the kth bracket in equation 2. □

Proof of Theorem 1. Recall $e_k(x)$ is the error of the predictor trained on all folds except for the kth fold in approximating the target function

$$e_k(x) = \widehat{f}_t^{(-k)}(x) - f_t(x).$$

Now define for the kth fold,

$$Z_{n,k} = \frac{1}{N}\sum_{j=1}^{N} e_k(\widetilde{X}_j) - \frac{K}{n}\sum_{i\in I_k} e_k(X_i),$$

and let

$$R_n = \frac{1}{K}\sum_{k=1}^{K} Z_{n,k}.$$

Substituting $Y = f_t(x) + \varepsilon$ into the estimator gives

$$\widehat{\mu}_{\text{TC-Cross}} - \theta_t^* = \frac{1}{n}\sum_{i=1}^{n}\varepsilon_i + \frac{1}{N}\sum_{j=1}^{N}\{f_t(\widetilde{X}_j) - \mathbb{E}_t f_t(X)\} + R_n.$$

For each fixed k, let us condition on $\mathcal{D}_s$ and the training complement $\mathcal{D}_{-k}$. Then e_k is fixed, the observations in I_k are independent of e_k, and the unlabeled sample $\widetilde{X}$ is independent of both the training complement and the held-out fold. Therefore the two empirical averages defining $Z_{n,k}$ are independent, and it is easy to see that they have the same conditional population mean. Consequently,

$$\mathbb{E}[Z_{n,k} \mid \mathcal{D}_s, \mathcal{D}_{-k}] = 0,$$

and

$$\begin{aligned}\mathbb{E}[Z_{n,k}^2 \mid \mathcal{D}_s, \mathcal{D}_{-k}] &= \frac{1}{N}\text{Var}_t\{e_k(X) \mid \mathcal{D}_s, \mathcal{D}_{-k}\} + \frac{K}{n}\text{Var}_t\{e_k(X) \mid \mathcal{D}_s, \mathcal{D}_{-k}\} \\ &\le \left(\frac{1}{N} + \frac{K}{n}\right)\mathbb{E}_t[e_k(X)^2 \mid \mathcal{D}_s, \mathcal{D}_{-k}].\end{aligned}$$

The quantities $Z_{n,1}, \ldots, Z_{n,K}$ need not be independent because the training complements overlap and the same unlabeled observations are used for every fold. Since K is fixed,

$$R_n^2 = \left(\frac{1}{K}\sum_{k=1}^{K} Z_{n,k}\right)^2 \le \frac{1}{K}\sum_{k=1}^{K} Z_{n,k}^2.$$

Hence, if

$$\max_{1\le k\le K}\mathbb{E}_t[e_k(X)^2 \mid \mathcal{D}_s, \mathcal{D}_{-k}] = o_p(1),$$

conditional Chebyshev's inequality applied fold by fold, followed by the fixed K bound above, gives

$$\sqrt{n}\, R_n = o_p(1).$$

If the maximum foldwise prediction error is instead $O_p(r_n)$ for a deterministic sequence $r_n \to 0$, the same argument yields

$$\sqrt{n}\, R_n = O_p(\sqrt{r_n}).$$

Therefore,

$$\sqrt{n}(\widehat{\mu}_{\text{TC-Cross}} - \theta_t^*) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\varepsilon_i + \frac{\sqrt{n}}{N}\sum_{j=1}^{N}\{f_t(\widetilde{X}_j) - \mathbb{E}_t f_t(X)\} + o_p(1).$$

The two leading empirical averages are based on independent gold and unlabeled samples. The ordinary central limit theorem therefore gives

$$\sqrt{n}(\widehat{\mu}_{\text{TC-Cross}} - \theta_t^*) \xrightarrow{D} N(0, \text{Var}_t(\varepsilon) + \rho\text{Var}_t\{f_t(X)\}).$$

□

Proof of Corollary 1. Each training complement has size $(1-1/K)n$. Applying Proposition 7 with that training-sample size gives, for fixed K,

$$(\widehat{\delta}^{(-k)} - \delta^*)^\top \Sigma_t (\widehat{\delta}^{(-k)} - \delta^*) = O_p\left(\frac{s\log p}{n}\right)$$

for each fold. Because K is fixed, the maximum over folds has the same order. Thus $a_n = o_p(1)$ when $s\log p/n \to 0$, and Theorem 1 applies. □

D.4 Proofs for TC-Cross-PPI++ and Joint TC-Cross-PPI++

Proof of Theorem 2. Under the assumed foldwise mean squared error convergence to $f_\dagger$, the same cross-fitting remainder argument as in Theorem 1 gives

$$\widehat{\mu}_{\text{TC-Cross},\lambda} - \theta_t^* = \frac{1}{n}\sum_{i=1}^{n}[Y_i - \lambda f_\dagger(X_i) - \mathbb{E}_t\{Y - \lambda f_\dagger(X)\}]$$
$$+\frac{\lambda}{N}\sum_{j=1}^{N}\left[f_\dagger(\widetilde{X}_j) - \mathbb{E}_t f_\dagger(X)\right] + o_p(n^{-1/2}).$$

The first two terms again are centered empirical averages and are independent. Hence

$$\sqrt{n}\big(\widehat{\mu}_{\text{TC-Cross},\lambda} - \theta_t^*\big) \xrightarrow{D} N\big(0, V_{\text{cross}}^\dagger(\lambda)\big),$$

where

$$V_{\text{cross}}^\dagger(\lambda) = \text{Var}_t\{Y - \lambda f_\dagger(X)\} + \rho\lambda^2\text{Var}_t\{f_\dagger(X)\}.$$

Expanding gives

$$V_{\text{cross}}^\dagger(\lambda) = \text{Var}_t(Y) - 2\lambda\text{Cov}_t\{Y, f_\dagger(X)\} + (1+\rho)\lambda^2\text{Var}_t\{f_\dagger(X)\}.$$

Differentiating with respect to λ therefore yields

$$\lambda_\dagger^* = \frac{\text{Cov}_t\{Y, f_\dagger(X)\}}{(1+\rho)\text{Var}_t\{f_\dagger(X)\}}.$$

Substitution gives the minimized variance in the statement of the theorem.

If $f_\dagger = f_t = \mathbb{E}_t[Y \mid X]$, then

$$\text{Cov}_t\{Y, f_t(X)\} = \text{Cov}_t\{\mathbb{E}_t(Y \mid X), f_t(X)\} = \text{Var}_t\{f_t(X)\},$$

and consequently

$$\lambda_t^* = \frac{1}{1+\rho}.$$

The result with a consistent substitution of $\widehat{\lambda}$ follows from Slutsky's theorem and the fixed-dimensional foldwise law of large numbers. □

Proof of Proposition 3. For each fold k, define

$$\widehat{H}^{(-k)}(x) = \begin{pmatrix} f_s(x) \\ \widehat{\Delta}^{(-k)}(x) \end{pmatrix}.$$

Since the first coordinate does not depend on k, equation 6 can be written as

$$\widehat{\mu}_{\text{JTC-Cross},\boldsymbol{\lambda}} = \frac{1}{K}\sum_{k=1}^{K}\left[\frac{K}{n}\sum_{i\in I_k} Y_i + \boldsymbol{\lambda}^\top\left\{\frac{1}{N}\sum_{j=1}^{N}\widehat{H}^{(-k)}(\widetilde{X}_j) - \frac{K}{n}\sum_{i\in I_k}\widehat{H}^{(-k)}(X_i)\right\}\right].$$

For each fixed k, condition on $\mathcal{D}_s$ and the training complement $\mathcal{D}_{-k}$. Then $\widehat{H}^{(-k)}$ is fixed, while the held-out observations in I_k have law P_t and are independent of $\mathcal{D}_{-k}$. The unlabeled observations are also independent and have covariate law $P_{t,X}$. Therefore,

$$\mathbb{E}\left[\frac{1}{N}\sum_{j=1}^{N}\widehat{H}^{(-k)}(\widetilde{X}_j)\middle|\mathcal{D}_s, \mathcal{D}_{-k}\right] = \mathbb{E}_t\left[\widehat{H}^{(-k)}(X)\middle|\mathcal{D}_s, \mathcal{D}_{-k}\right],$$

and

$$\mathbb{E}\left[\frac{K}{n}\sum_{i\in I_k}\widehat{H}^{(-k)}(X_i)\middle|\mathcal{D}_s, \mathcal{D}_{-k}\right] = \mathbb{E}_t\left[\widehat{H}^{(-k)}(X)\middle|\mathcal{D}_s, \mathcal{D}_{-k}\right].$$

Also,

$$\mathbb{E}\left[\frac{K}{n}\sum_{i\in I_k} Y_i \middle| \mathcal{D}_s, \mathcal{D}_{-k}\right] = \theta_t^*.$$

Thus the conditional expectation of each fold-specific bracket is θ_t^*. Averaging over folds and applying iterated expectation yields

$$\mathbb{E}[\widehat{\mu}_{\text{JTC-Cross},\boldsymbol{\lambda}} \mid \mathcal{D}_s] = \theta_t^*.$$

□

Proof of Theorem 3. For $i \in I_k$, write

$$e_k(x) \coloneqq \widehat{\Delta}^{(-k)}(x) - \Delta^\dagger(x).$$

Then

$$\widehat{H}_i^{\text{oof}} = H_\dagger(X_i) + \begin{pmatrix} 0 \\ e_k(X_i) \end{pmatrix}.$$

Similarly, the averaged unlabeled vector can be written as

$$\widetilde{H}_N^{\text{CF}} = \frac{1}{N}\sum_{j=1}^N H_\dagger(\widetilde{X}_j) + \begin{pmatrix} 0 \\ \frac{1}{K}\sum_{k=1}^K \frac{1}{N}\sum_{j=1}^N e_k(\widetilde{X}_j) \end{pmatrix}.$$

Substituting these expressions into equation 6 gives

$$\begin{aligned}\widehat{\mu}_{\text{JTC-Cross},\boldsymbol{\lambda}} - \theta_t^* = \frac{1}{n}\sum_{i=1}^n \left[Y_i - \boldsymbol{\lambda}^\top H_\dagger(X_i) - \mathbb{E}_t\{Y - \boldsymbol{\lambda}^\top H_\dagger(X)\}\right] \\ + \frac{1}{N}\sum_{j=1}^N \left[\boldsymbol{\lambda}^\top H_\dagger(\widetilde{X}_j) - \mathbb{E}_t\{\boldsymbol{\lambda}^\top H_\dagger(X)\}\right] + R_n(\boldsymbol{\lambda}),\end{aligned} \tag{78}$$

where, writing $\boldsymbol{\lambda} = (\lambda_s, \lambda_\Delta)^\top$,

$$R_n(\boldsymbol{\lambda}) = \frac{\lambda_\Delta}{K}\sum_{k=1}^K \left[\frac{1}{N}\sum_{j=1}^N e_k(\widetilde{X}_j) - \frac{K}{n}\sum_{i\in I_k} e_k(X_i)\right]. \tag{79}$$

We first show that the cross-fitting remainder is negligible. Define

$$Z_{n,k} \coloneqq \frac{1}{N}\sum_{j=1}^N e_k(\widetilde{X}_j) - \frac{K}{n}\sum_{i\in I_k} e_k(X_i).$$

For each fixed k, condition on $\mathcal{D}_s$ and the training complement $\mathcal{D}_{-k}$. Then e_k is fixed, the observations in I_k are independent of e_k, and the unlabeled observations are independent of both the training complement and the held-out observations. Moreover, the two empirical averages defining $Z_{n,k}$ have the same conditional population mean. Hence

$$\mathbb{E}[Z_{n,k} \mid \mathcal{D}_s, \mathcal{D}_{-k}] = 0.$$

Let

$$a_{n,k} \coloneqq \mathbb{E}_t\left[e_k(X)^2 \mid \mathcal{D}_s, \mathcal{D}_{-k}\right].$$

Using independence of the held-out and unlabeled samples conditional on $\mathcal{D}_s, \mathcal{D}_{-k}$,

$$\begin{aligned}\mathrm{Var}(Z_{n,k} \mid \mathcal{D}_s, \mathcal{D}_{-k}) &= \frac{1}{N}\mathrm{Var}_t(e_k(X) \mid \mathcal{D}_s, \mathcal{D}_{-k}) \\ &\quad + \frac{K}{n}\mathrm{Var}_t(e_k(X) \mid \mathcal{D}_s, \mathcal{D}_{-k}) \\ &\leq \left(\frac{1}{N} + \frac{K}{n}\right) a_{n,k}.\end{aligned}$$

Since $n/N \to \rho < \infty$ and K is fixed,

$$n\left(\frac{1}{N} + \frac{K}{n}\right) = O(1).$$

Together with the assumption that fold-specific transfer corrections converge uniformly, conditional Chebyshev's inequality therefore gives

$$\sqrt{n}\, Z_{n,k} = o_p(1)$$

for every fixed k. Because K is fixed,

$$\sqrt{n}\, R_n(\boldsymbol{\lambda}) = \frac{\lambda_\Delta}{K}\sum_{k=1}^{K}\sqrt{n}\, Z_{n,k} = o_p(1)$$

for every fixed $\boldsymbol{\lambda}$. Notice that no independence across different folds is required for this argument; the training complements may overlap and the same unlabeled sample may be used in all folds.

It follows from equation 78 that

$$\begin{aligned}\sqrt{n}\,(\widehat{\mu}_{\text{JTC-Cross},\boldsymbol{\lambda}} - \theta_t^*) &= \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left[Y_i - \boldsymbol{\lambda}^\top H_\dagger(X_i) - \mathbb{E}_t\{Y - \boldsymbol{\lambda}^\top H_\dagger(X)\}\right] \\ &\quad + \frac{\sqrt{n}}{N}\sum_{j=1}^{N}\left[\boldsymbol{\lambda}^\top H_\dagger(\widetilde{X}_j) - \mathbb{E}_t\{\boldsymbol{\lambda}^\top H_\dagger(X)\}\right] + o_p(1).\end{aligned}$$

The two leading empirical averages are based on independent gold and unlabeled samples. Under the stated $(2+\eta)$ moment conditions, the Lindeberg–Feller central limit theorem therefore gives

$$\sqrt{n}\,(\widehat{\mu}_{\text{JTC-Cross},\boldsymbol{\lambda}} - \theta_t^*) \xrightarrow{D} N\Big(0, V^\dagger_{\text{Jcross}}(\boldsymbol{\lambda})\Big),$$

where

$$V^\dagger_{\text{Jcross}}(\boldsymbol{\lambda}) = \mathrm{Var}_t\{Y - \boldsymbol{\lambda}^\top H_\dagger(X)\} + \rho\mathrm{Var}_t\{\boldsymbol{\lambda}^\top H_\dagger(X)\}.$$

Now write

$$\Sigma_\dagger = \mathrm{Var}_t\{H_\dagger(X)\}, \qquad c_\dagger = \mathrm{Cov}_t\{H_\dagger(X), Y\}.$$

Expanding the preceding variance yields

$$V^\dagger_{\text{Jcross}}(\boldsymbol{\lambda}) = \mathrm{Var}_t(Y) - 2\boldsymbol{\lambda}^\top c_\dagger + (1+\rho)\boldsymbol{\lambda}^\top\Sigma_\dagger\boldsymbol{\lambda},$$

If $\Sigma_\dagger$ is nonsingular, differentiation with respect to $\boldsymbol{\lambda}$ gives

$$(1+\rho)\Sigma_\dagger\boldsymbol{\lambda}^*_{\text{Jcross}} = c_\dagger,$$

and hence

$$\boldsymbol{\lambda}^*_{\text{Jcross}} = \frac{1}{1+\rho}\Sigma_\dagger^{-1}c_\dagger.$$

Substitution into the variance criterion gives

$$V^{\dagger,*}_{\text{Jcross}} = \mathrm{Var}_t(Y) - \frac{c_\dagger^\top\Sigma_\dagger^{-1}c_\dagger}{1+\rho}.$$

If $\Sigma_\dagger$ is singular, let v belong to its null space. Then

$$\mathrm{Var}_t\{v^\top H_\dagger(X)\} = v^\top \Sigma_\dagger v = 0,$$

so $v^\top H_\dagger(X)$ is almost surely constant. Therefore

$$v^\top c_\dagger = \mathrm{Cov}_t\{v^\top H_\dagger(X), Y\} = 0.$$

Hence $c_\dagger$ belongs to the column space of $\Sigma_\dagger$, and the quadratic criterion has a minimizer. Its minimum-norm minimizer is

$$\boldsymbol{\lambda}^*_{\mathrm{Jcross}} = \frac{1}{1+\rho}\Sigma_\dagger^\dagger c_\dagger,$$

and the minimized variance is

$$V^{\dagger,*}_{\mathrm{Jcross}} = \mathrm{Var}_t(Y) - \frac{c_\dagger^\top \Sigma_\dagger^\dagger c_\dagger}{1+\rho}, \tag{80}$$

For the no-harm comparisons, the joint parameter space $\mathbb{R}^2$ contains the source-PPI++ line

$$\{(\lambda, 0)^\top : \lambda \in \mathbb{R}\},$$

for which

$$\boldsymbol{\lambda}^\top H_\dagger(X) = \lambda f_s(X),$$

and the scalar TC–Cross-PPI++ line

$$\{(\lambda, \lambda)^\top : \lambda \in \mathbb{R}\},$$

for which

$$\boldsymbol{\lambda}^\top H_\dagger(X) = \lambda\{f_s(X) + \Delta^\dagger(X)\} = \lambda f_\dagger(X).$$

The joint class also contains $(0,0)^\top$, which gives classical inference. Therefore minimizing over the full joint class cannot yield a larger first-order variance than minimizing over either scalar subfamily or choosing classical inference. This proves equation 7 without requiring $\Delta^\dagger = \Delta^*$.

Finally, suppose that

$$\Delta^\dagger = \Delta^*, \qquad f_\dagger = f_t = \mathbb{E}_t[Y \mid X].$$

Let

$$a = (1,1)^\top,$$

so that

$$f_t(X) = a^\top H_\dagger(X).$$

Because $H_\dagger(X)$ is measurable with respect to X and $f_t(X) = \mathbb{E}_t[Y \mid X]$,

$$c_\dagger = \mathrm{Cov}_t\{H_\dagger(X), Y\} = \mathrm{Cov}_t\{H_\dagger(X), f_t(X)\} = \Sigma_\dagger a.$$

Consequently,

$$\begin{aligned} c_\dagger^\top \Sigma_\dagger^\dagger c_\dagger &= a^\top \Sigma_\dagger \Sigma_\dagger^\dagger \Sigma_\dagger a \\ &= a^\top \Sigma_\dagger a \\ &= \mathrm{Var}_t\{f_t(X)\}. \end{aligned}$$

Substitution into equation 80 gives

$$V^{\dagger,*}_{\mathrm{Jcross}} = \mathrm{Var}_t(Y) - \frac{\mathrm{Var}_t\{f_t(X)\}}{1+\rho},$$

It remains to justify feasible substitution. Define

$$D_n \coloneqq \widetilde{H}_N^{\mathrm{CF}} - \frac{1}{n}\sum_{i=1}^n \widehat{H}_i^{\mathrm{oof}}.$$

The same decomposition used above gives

$$D_n = \left\{ \frac{1}{N} \sum_{j=1}^{N} H_\dagger(\widetilde{X}_j) - \frac{1}{n} \sum_{i=1}^{n} H_\dagger(X_i) \right\} + o_p(n^{-1/2}),$$

componentwise, and therefore

$$\sqrt{n}\, D_n = O_p(1).$$

Since the estimator in equation 6 is affine in $\boldsymbol{\lambda}$,

$$\begin{aligned} \sqrt{n}\Big\{ \widehat{\mu}_{\text{JTC-Cross},\widehat{\boldsymbol{\lambda}}} - \widehat{\mu}_{\text{JTC-Cross},\boldsymbol{\lambda}^*_{\text{Jcross}}} \Big\} &= \sqrt{n} \left(\widehat{\boldsymbol{\lambda}} - \boldsymbol{\lambda}^*_{\text{Jcross}} \right)^\top D_n \\ &= o_p(1) \end{aligned}$$

whenever

$$\widehat{\boldsymbol{\lambda}} - \boldsymbol{\lambda}^*_{\text{Jcross}} = o_p(1).$$

Thus, substituting a consistent feasible weight estimator leaves the same first-order limiting distribution and variance. □

Proof of Proposition 4. Write

$$H_i^\dagger \coloneqq H_\dagger(X_i), \qquad q_i \coloneqq \widehat{H}_i^{\text{oof}} - H_i^\dagger.$$

If $i \in I_k$, then

$$q_i = \begin{pmatrix} 0 \\ \widehat{\Delta}^{(-k)}(X_i) - \Delta^\dagger(X_i) \end{pmatrix} = \begin{pmatrix} 0 \\ e_k(X_i) \end{pmatrix}.$$

Let

$$\overline{H}_\dagger = \frac{1}{n} \sum_{i=1}^{n} H_i^\dagger, \qquad \overline{q} = \frac{1}{n} \sum_{i=1}^{n} q_i, \qquad \overline{H}_{\text{oof}} = \overline{H}_\dagger + \overline{q}.$$

We first control the empirical size of the foldwise transfer errors. For each fold k, conditional on $\mathcal{D}_s$ and $\mathcal{D}_{-k}$,

$$\mathbb{E}\left[\frac{K}{n} \sum_{i \in I_k} e_k(X_i)^2 \middle| \mathcal{D}_s, \mathcal{D}_{-k} \right] = \mathbb{E}_t[e_k(X)^2 \mid \mathcal{D}_s, \mathcal{D}_{-k}] = O_p(r_n).$$

Conditional Markov's inequality and fixed K therefore imply

$$\frac{1}{n} \sum_{i=1}^{n} \|q_i\|_2^2 = O_p(r_n).$$

Consequently,

$$\|\overline{q}\|_2 \le \left(\frac{1}{n} \sum_{i=1}^{n} \|q_i\|_2^2 \right)^{1/2} = O_p(\sqrt{r_n}). \tag{81}$$

Let

$$\widetilde{\Sigma}_H = \frac{1}{n-1} \sum_{i=1}^{n} (H_i^\dagger - \overline{H}_\dagger)(H_i^\dagger - \overline{H}_\dagger)^\top$$

denote the empirical covariance matrix computed from the limiting prediction vector $H_\dagger$. Since $H_\dagger(X)$ has fixed dimension and finite fourth moments,

$$\|\widetilde{\Sigma}_H - \Sigma_\dagger\|_{\text{op}} = O_p(n^{-1/2}). \tag{82}$$

Similarly, if

$$\widetilde{c}_H = \frac{1}{n-1} \sum_{i=1}^{n} (H_i^\dagger - \overline{H}_\dagger)(Y_i - \overline{Y}_n),$$

then

$$\|\widetilde{c}_H - c_\dagger\|_2 = O_p(n^{-1/2}). \tag{83}$$

We next compare the out-of-fold covariance estimators with these oracle empirical quantities. Since

$$\widehat{H}_i^{\mathrm{oof}} - \overline{H}_{\mathrm{oof}} = (H_i^\dagger - \overline{H}_\dagger) + (q_i - \overline{q}),$$

expansion gives

$$\begin{aligned}\widehat{\Sigma}_H - \widetilde{\Sigma}_H = \frac{1}{n-1}\sum_{i=1}^n (H_i^\dagger - \overline{H}_\dagger)(q_i - \overline{q})^\top \\ + \frac{1}{n-1}\sum_{i=1}^n (q_i - \overline{q})(H_i^\dagger - \overline{H}_\dagger)^\top \\ + \frac{1}{n-1}\sum_{i=1}^n (q_i - \overline{q})(q_i - \overline{q})^\top.\end{aligned}$$

By Cauchy–Schwarz,

$$\left\| \frac{1}{n-1}\sum_{i=1}^n (H_i^\dagger - \overline{H}_\dagger)(q_i - \overline{q})^\top \right\|_{\mathrm{op}} \le \left[\frac{1}{n-1}\sum_{i=1}^n \|H_i^\dagger - \overline{H}_\dagger\|_2^2\right]^{1/2} \left[\frac{1}{n-1}\sum_{i=1}^n \|q_i - \overline{q}\|_2^2\right]^{1/2}.$$

The first factor is $O_p(1)$ by the moment assumptions, while the second is $O_p(\sqrt{r_n})$ by the preceding bound. The transpose term has the same order, and

$$\left\| \frac{1}{n-1}\sum_{i=1}^n (q_i - \overline{q})(q_i - \overline{q})^\top \right\|_{\mathrm{op}} = O_p(r_n).$$

Hence

$$\|\widehat{\Sigma}_H - \widetilde{\Sigma}_H\|_{\mathrm{op}} = O_p(\sqrt{r_n} + r_n) = O_p(\sqrt{r_n}), \tag{84}$$

where the last equality uses $r_n = o(1)$.

For the covariance with the outcome,

$$\widehat{c}_H - \widetilde{c}_H = \frac{1}{n-1}\sum_{i=1}^n (q_i - \overline{q})(Y_i - \overline{Y}_n).$$

Again by Cauchy–Schwarz,

$$\begin{aligned}\|\widehat{c}_H - \widetilde{c}_H\|_2 &\le \left[\frac{1}{n-1}\sum_{i=1}^n \|q_i - \overline{q}\|_2^2\right]^{1/2} \left[\frac{1}{n-1}\sum_{i=1}^n (Y_i - \overline{Y}_n)^2\right]^{1/2} \\ &= O_p(\sqrt{r_n}).\end{aligned}$$

Combining this with equation 82–equation 83 gives

$$\|\widehat{\Sigma}_H - \Sigma_\dagger\|_{\mathrm{op}} + \|\widehat{c}_H - c_\dagger\|_2 = O_p\Big(n^{-1/2} + \sqrt{r_n}\Big).$$

Now put

$$\alpha_n := n^{-1/2} + \sqrt{r_n} + \tau.$$

Since $r_n = o(1)$ and $\tau = o(1)$,

$$\|\widehat{\Sigma}_H + \tau I_2 - \Sigma_\dagger\|_{\mathrm{op}} = O_p(\alpha_n) = o_p(1).$$

The assumption $\lambda_{\min}(\Sigma_\dagger) \ge \kappa_H > 0$ therefore implies

$$\|(\widehat{\Sigma}_H + \tau I_2)^{-1}\|_{\mathrm{op}} = O_p(1).$$

Using the resolvent identity,

$$(\widehat{\Sigma}_H + \tau I_2)^{-1} - \Sigma_\dagger^{-1} = (\widehat{\Sigma}_H + \tau I_2)^{-1}\{\Sigma_\dagger - \widehat{\Sigma}_H - \tau I_2\}\Sigma_\dagger^{-1},$$

and hence

$$\left\|(\widehat{\Sigma}_H + \tau I_2)^{-1} - \Sigma_\dagger^{-1}\right\|_{\mathrm{op}} = O_p(\alpha_n).$$

Therefore

$$\left\|\widehat{\boldsymbol{\lambda}} - \boldsymbol{\lambda}^{\dagger}_{\mathrm{J},n}\right\|_2 \le \frac{1}{1+n/N}\Big[\|(\widehat{\Sigma}_H + \tau I_2)^{-1}\|_{\mathrm{op}}\|\widehat{c}_H - c_\dagger\|_2 + \left\|(\widehat{\Sigma}_H + \tau I_2)^{-1} - \Sigma_\dagger^{-1}\right\|_{\mathrm{op}}\|c_\dagger\|_2\Big] = O_p\Big(n^{-1/2} + \sqrt{r_n} + \tau\Big),$$

which proves equation 54.

It remains to evaluate the excess population variance criterion. By equation 53,

$$c_\dagger = (1 + n/N)\Sigma_\dagger \boldsymbol{\lambda}^{\dagger}_{\mathrm{J},n}.$$

Completing the square in equation 52 therefore gives the exact identity

$$V^{\dagger}_{\mathrm{J},n}(\boldsymbol{\lambda}) - V^{\dagger}_{\mathrm{J},n}(\boldsymbol{\lambda}^{\dagger}_{\mathrm{J},n}) = (1+n/N)\left(\boldsymbol{\lambda} - \boldsymbol{\lambda}^{\dagger}_{\mathrm{J},n}\right)^{\top}\Sigma_\dagger\left(\boldsymbol{\lambda} - \boldsymbol{\lambda}^{\dagger}_{\mathrm{J},n}\right).$$

Taking $\boldsymbol{\lambda} = \widehat{\boldsymbol{\lambda}}$ and using the bounded eigenvalues of the fixed-dimensional $\Sigma_\dagger$ yields

$$V^{\dagger}_{\mathrm{J},n}(\widehat{\boldsymbol{\lambda}}) - V^{\dagger}_{\mathrm{J},n}(\boldsymbol{\lambda}^{\dagger}_{\mathrm{J},n}) = O_p\big(n^{-1} + r_n + \tau^2\big),$$

which proves equation 55. □

E Experimental Details

For each dataset, Table 2 gives the population size N from which gold labels are sampled, the swept range of gold-label counts n, and the repetition/cross-fitting protocol. Each repetition reveals n rows' labels uniformly at random and treats the remaining $N - n$ rows as the prediction-rich unlabeled pool. The unlabeled share $(N-n)/N$ at the largest gold-label budget swept ranges from about 78% (AutoEval MT-Bench Arena, $n = 750$ of $N = 3355$) to over 99% (Census, AutoEval-ProteinGym); it is smallest for the two applications where n is swept up to a sizable fraction of a comparatively small population (AutoEval MT-Bench Arena, ProteinGym).

Each application below gives its dataset, the source predictor that supplies the score, and the sparse calibration features appended to that score. Common setup details are in Appendix E.6.

Table 2: Population size ($N + n$), gold-label range (n), repetition count, and cross-fitting fold count (K). N rows are unlabeled in each repetition. Parenthesized values give the LoRA variant's count where it differs from the sparse-calibration default shown (if applicable).

Application	Population ($N+n$)	Gold labels (n)	Repetitions	K
Census mean / OLS	380,091	100–2000	200	5
Census healthcare	318,215	100–2000	100	5
Capital Bikeshare	8,734 [2]	50–500	100	5
ASR accent transfer	15,378	50–1000	250	5
Camelyon17 (WILDS)	119,958 [3]	25–500	200	5
PovertyMap (WILDS)	7,872 [4]	25–500	200	5
ProteinGym (BLAT)	4,783	100–1000	100 (20)	5
AutoEval-ProteinGym (SPG1)	536,962	50–1000	250 (20)	5 (2)
Chatbot Arena	14,379	200–2000	50 (20)	5 (2)
Chatbot Arena (judge quality)	14,379	500	250	5
AutoEval MT-Bench Arena	3,355	25–750	250 (25)	5 (3)

E.1 TABULAR

Census. 2019 American Community Survey microdata for California via the `folktables` package (Ding et al., 2021), covering three targets: mean income, an income-regression coefficient, and private health-insurance coverage.

Source predictor: XGBoost (Chen & Guestrin, 2016) predictions, precomputed and shipped by the upstream `ppi_py` benchmark rather than trained in this codebase.

Calibration features: raw tabular covariates plus the source prediction, together with degree-two interactions and squares of its two covariates (6 features total, shared across all three Census applications).

Capital Bikeshare. UCI's hourly Capital Bikeshare rental counts for Washington, D.C. (Fanaee-T & Gama, 2014); a temporal-transfer benchmark predicting 2012 rentals from a source trained only on 2011 data.

Source predictor: a gradient-boosted regressor trained on the 2011 data: learning rate 0.06, 300 boosting rounds, at most 31 leaf nodes per tree, ℓ_2 regularization 1.0, all other settings at defaults).

Calibration features: raw tabular covariates plus the source prediction, one-hot-encoding the categorical covariates alongside the standardized continuous ones with no polynomial expansion (60 features total).

E.2 SPEECH

AfriSpeech-200. African-accented English speech (Olatunji et al., 2023); we use the Yoruba-accented portion for a word-error-rate transfer task, pooling its train, dev and test splits into a single target population (Table 2).

Source predictor: wav2vec2-base-960h (Baevski et al., 2020) with zero-shot greedy CTC decoding (no beam search), trained on American-English speech and applied directly to the Yoruba-accented audio with no fine-tuning.

Calibration features: the source WER prediction, audio duration, hypothesis word count, speaking rate, and one-hot speaker gender (7 features total, three gender categories observed).

E.3 WILDS BENCHMARKS

E.3.1 CAMELYON17

The WILDS (Koh et al., 2020) version of the CAMELYON17 lymph-node histopathology challenge (Bándi et al., 2019): $455{,}954$ 96×96 RGB patches from 50 slides across five hospitals, labeled 1 if the patch's central 32×32 region contains tumor tissue. Three hospitals are used for training and the two held-out ones, the OOD validation and OOD test hospitals, are pooled into a single target population. Pooling is sound because we select the source checkpoint on in-distribution data, so neither held-out hospital informs training or model selection; it also makes the target a two-hospital mixture rather than a single unseen hospital. WILDS's own CodaLab download host

[2] Target-year (2012) population. The 8,645 source-year (2011) rows are used only to fit the source predictor.

[3] Pooled held-out-hospital population ($85{,}054$ test plus $34{,}904$ validation patches). The $335{,}996$ training-hospital patches, including the in-distribution validation split, are used only to fit and select the source predictor.

[4] Fold A's pooled OOD-country population (3,963 test plus 3,909 validation clusters). The 10,796 training- and in-distribution-validation-country clusters are used only to fit and select the source predictor.

was unreachable, so we load the `wltjr1007/Camelyon17-WILDS` Hugging Face mirror and assign hospitals from its per-patch `center` field, which reproduces WILDS's official split sizes exactly (302,436 training, 33,560 in-distribution validation, 34,904 OOD validation and 85,054 test patches).

Source predictor: DenseNet-121 (Huang et al., 2016) trained from scratch on the three training hospitals, following WILDS's ERM baseline (binary cross-entropy; SGD with momentum 0.9, learning rate 10^{-3}, weight decay 10^{-2}; batch size 32; 10 epochs; random flips). Its sigmoid probability is the source score. The selected checkpoint reaches 0.968 and 0.966 accuracy on the training hospitals and in-distribution validation, against 0.787 and 0.673 on the two held-out hospitals.

Calibration features: the source probability, the patch's (x, y) coordinates within its slide, and one-hot indicators for the 20 slides of the two target hospitals (23 features total). The slide indicators let the correction absorb slide-to-slide staining variation within and between the target hospitals.

E.3.2 PovertyMap

The WILDS (Koh et al., 2020) version of the PovertyMap benchmark (Yeh et al., 2020): 224×224 multispectral satellite images (seven Landsat bands plus a nightlights band) for 19,669 DHS survey clusters in 23 African countries, labeled with a real-valued asset-wealth index. We use WILDS's country fold A, whose ten out-of-distribution countries we pool into a single target population: the OOD test group (Angola, Côte d'Ivoire, Ethiopia, Mali, Rwanda) and the OOD validation group (Benin, Burkina Faso, Guinea, Sierra Leone, Tanzania). Doing so doubles the target population and the number of countries the correction can key on. WILDS's own CodaLab download host was unreachable, so we load the `1aurent/PovertyMap` Hugging Face mirror, which also ships the survey metadata (urban/rural, nightlight summaries, country, year) as ordinary columns. The loader identifies the fold by matching the mirror's OOD country sets against WILDS's five fold definitions.

Source predictor: ResNet18 (He et al., 2015) with its first convolution widened to the eight multispectral bands, trained from scratch on fold A's training countries, following WILDS's ERM baseline (mean squared error; Adam with learning rate 10^{-3}; step decay of 0.96 per epoch; batch size 64; up to 200 epochs; random flips). It reaches correlations of 0.902 and 0.868 on the training countries and in-distribution validation, against 0.830 on the pooled held-out countries.

Calibration features: the source prediction, the urban/rural indicator, the survey's two nightlight summaries (`nl_mean` and `nl_center`), and one-hot indicators for the ten target countries (14 features total). The mirror's binarized wealth column is never used, as it is derived from the target.

E.3.3 Model selection

Both WILDS source predictors keep the checkpoint that scores best on the *in-distribution* validation split rather than the out-of-distribution one WILDS's own protocol uses, because our setting assumes no target-domain labels at training time; this is also what makes pooling both held-out groups into a single target leak-free (`--selection-split val` recovers WILDS's protocol). One consequence is that our Camelyon17 source reaches 0.673 on the OOD test hospital against WILDS's reported ERM baseline of 70.3%, a benchmark Koh et al. (2020) single out for its 6.4% seed-to-seed standard deviation there.

E.4 PROTEIN LANGUAGE MODELS

ProteinGym. The BLAT_ECOLX_Firnberg_2014 deep-mutational-scanning assay from the ProteinGym benchmark (Notin et al., 2023), measuring the fitness effect of amino-acid substitutions in a beta-lactamase enzyme. It also includes a protein-model-ranking target and an annotator-quality sweep.

Source predictor: ESM-2 8M (Lin et al., 2022) zero-shot masked-marginal scores, with no training.

Candidate models: three ESM-2 checkpoints (Lin et al., 2022) (8M, 35M, and 150M parameters), which are ranked in the protein-model-ranking target and swapped in as the annotator in the annotator-quality sweep.

Calibration features: one-hot mutation identity (position and substituted amino acid) with the ESM-2 score (47 features total, for the 286-residue BLAT assay with all 20 canonical amino acids observed).

LoRA calibration: fine-tunes the same ESM-2 8M backbone that produces the source score, on the raw mutated sequence, with the source score as a second linear term.

AutoEval-ProteinGym. A replication of AutoEval's (Boyeau et al., 2024) released SPG1 experiment (the STRSG assay). This replication follows AutoEval's own convention of 90% nominal coverage, rather than the 95% used elsewhere in this paper.

Source predictor: a precomputed VESPA (Marquet et al., 2021) annotator score.

Candidate models: the seven zero-shot models AutoEval compares, using AutoEval's released scores: UniRep (Alley et al., 2019), CARP (640M) (Yang et al., 2024), ESM-1b (Rives et al., 2020), ESM-1v (five-model ensemble) (Meier et al., 2021), ESM-2 (15B) (Lin et al., 2022), RITA (XL) (Hesslow et al., 2022), and ProGen2 (XLarge) (Nijkamp et al., 2023). The annotator-quality sweep uses each of these as the annotator in turn, plus VESPA.

Calibration features: per-site biochemical-property deltas (hydrophobicity, volume, polarity, charge, aromaticity) and epistasis proxies with the VESPA score (37 features total).

LoRA calibration: fine-tunes a separately loaded ESM-2 8M encoder on the raw mutant sequence, with the VESPA score as an auxiliary input – source and LoRA backbone are different models here.

E.5 LLM ARENAS

Chatbot Arena. Human pairwise preferences from Chatbot Arena (Chiang et al., 2024), paired with precomputed public LLM-judge scores, restricted to the most frequently compared models.

Source predictor: the Meta-Llama-3-8B-Instruct judge's precomputed pairwise-preference probability (human agreement 62.9%).

Calibration features: a small metadata layer only – source logit, prompt/response character and word counts, response-length contrast, and one-hot model-identity dummies for both compared models (47 features total among the 20 most-compared models); a frozen-sentence-embedding feature mode also exists in code but was not used for the reported run.

AutoEval MT-Bench Arena. A replication of AutoEval's (Boyeau et al., 2024) released MT-Bench (Zheng et al., 2023) experiment, matching human ratings to GPT-4's pairwise judgments across six chat systems. This replication follows AutoEval's own convention of 90% nominal coverage, rather than the 95% used elsewhere in this paper.

Source predictor: GPT-4's own precomputed pairwise judgment, as released.

Calibration features: its Bradley–Terry design matrix and a small metadata layer (source logit, response-length contrast, and turn; 8 features total among the six chat systems) combined with frozen MiniLM-L6-v2 sentence embeddings of the prompt and a response-embedding contrast ($2 \times 384 = 768$ dimensions), for 776 features total. Question-level fixed effects are omitted in both arena applications since the smallest gold-sample sizes leave too few observations per calibration fold.

LoRA calibration (both arenas): both replace their (unfine-tunable) source entirely with `distilgpt2` (Sanh et al., 2019), which supplies its own zero-shot pairwise-preference score and is LoRA-fine-tuned as the correction model on the same formatted prompt/response text. In both cases this also makes the sparse-vs-LoRA comparison an apples-to-apples test of calibration mechanism: each LoRA-comparison's sparse arm is given `distilgpt2`'s own frozen, pooled hidden state over the identical text, so the only difference between the sparse and LoRA arms is the calibration mechanism itself, not which encoder supplies the features.

E.6 Hyperparameters and Implementation

To promote reproducibility, the source code is provided in the supplementary material, along with the SLURM script files used to launch the jobs.

Sparse-linear calibration. Continuous-outcome applications (Census mean/OLS, Capital Bikeshare, PovertyMap, ProteinGym), and Camelyon17, whose binary label enters a mean-estimation residual correction, select their calibrator's ℓ_1 penalty by 5-fold cross-validation over scikit-learn's automatic log-spaced grid (`LassoCV`). AutoEval-ProteinGym instead fixes the penalty ($\alpha = 0.01$, no cross-validation) given the cost of repeated re-fits over its much larger unlabeled pool. Binary/preference-outcome applications (Census healthcare, Chatbot Arena, AutoEval MT-Bench Arena) calibrate with a fixed, smoothed-ℓ_1-penalized offset-logistic correction ($\alpha = 10^{-3}$ for Census healthcare and Chatbot Arena, $\alpha = 0.01$ for AutoEval MT-Bench Arena) rather than cross-validation, which we found unnecessary for stability at this scale and considerably slower under the offset-logistic objective.

Joint power-tuning ridge. The joint estimator's multi-source weight solve is stabilized with a small, fixed ridge term added to its Hessian before inversion: a multiplier of 10^{-8} (relative to $\mathrm{trace}(\Sigma_H)/M$) for the mean estimator, and 10^{-6} for the logistic and OLS joint estimators. This constant is fixed across every experiment purely for numerical conditioning of the weight solve – it is never tuned from data or exposed as a command-line option.

Confidence intervals. All cross-fitted estimators use Wald intervals $\hat{\mu} \pm z_{1-\alpha/2}\widehat{\mathrm{SE}}$. For TC–Cross-PPI(++), $\widehat{\mathrm{SE}}^2 = \widehat{\mathrm{Var}}(Y - \hat{\lambda}\hat{f}_t^{\mathrm{oof}})/n + \hat{\lambda}^2\,\widehat{\mathrm{Var}}(\bar{f}_t)/N$, where $\hat{f}_t^{\mathrm{oof}}$ is the out-of-fold calibrated prediction on the gold sample and $\bar{f}_t$ the fold-averaged prediction on the unlabeled sample. Joint TC–Cross-PPI++ is the same with $\hat{\lambda}\hat{f}_t$ replaced by the weighted combination $\hat{\lambda}_s f_s + \hat{\lambda}_\Delta \hat{\Delta}$ of the source prediction and the correction. The weights are estimated from the same out-of-fold quantities, with no additional split, and are treated as fixed in $\widehat{\mathrm{SE}}$.

LoRA hyperparameters. Table 3 gives the adapter configuration for each of the four LoRA experiments; all four additionally include the source score as an auxiliary linear input to the correction and pool the backbone's final hidden states by mean pooling.

Compute Resources: Non-LoRA experiments, and those that don't require training a source model, are run on up to 8 Intel Xeon 8268s. Experiments that score or fine-tune a neural network (i.e., Came-

Table 3: LoRA adapter hyperparameters by application. Target modules follow each backbone's attention-layer naming (`query`/`value` for ESM-2, `c_attn` for GPT-2-family models). Both arena applications share `distilgpt2` as backbone.

Application	r	α	Dropout	Target modules	Epochs	LR	Batch (train/predict)	Max length
ProteinGym (BLAT)	1	16	0.05	query, value	100	10^{-3}	256 / 256	—
AutoEval-ProteinGym	2	16	0.05	query, value	100	10^{-3}	512 / 512	—
Chatbot Arena	2	16	0.05	c_attn	100	10^{-3}	16 / 64	1024
AutoEval MT-Bench Arena	2	16	0.05	c_attn	100	10^{-3}	16 / 64	1024

lyon17's and PovertyMap's source training, ProteinGym, AutoEval-ProteinGym LoRA, the Chatbot Arena and AutoEval MT-Bench Arena LoRA variants, and ASR accent transfer), additionally use a single V100, or A100 for LoRA experiments.

F Additional Results

Figures 3- 5 report the empirical ESS, interval widths, and coverage across all datasets, including ProteinGym and Census Income (OLS), which are discussed only in this appendix.

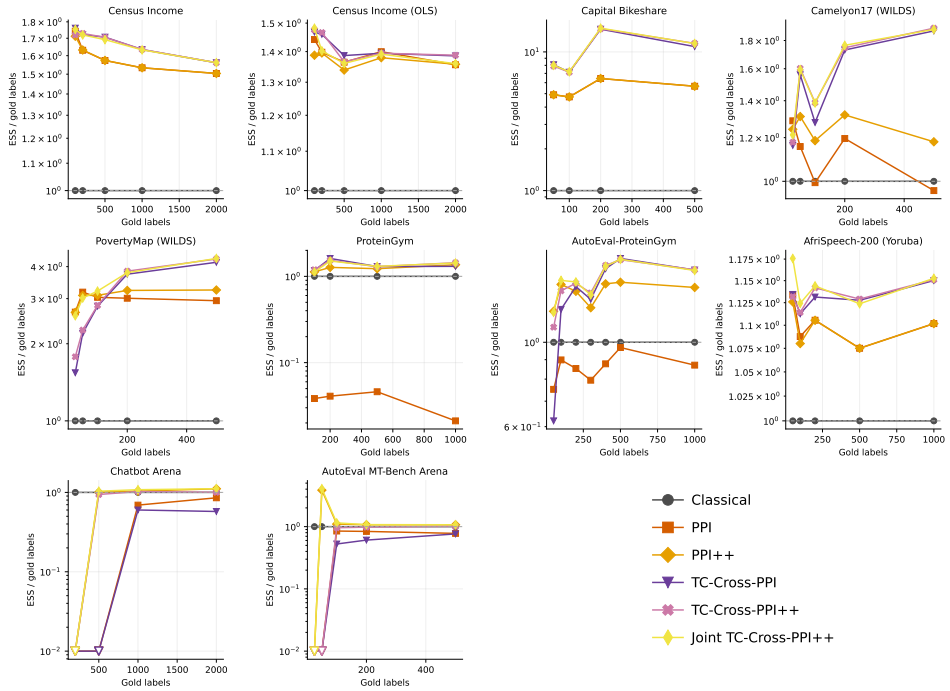


Figure 3: Empirical ESS divided by the number of gold labels (n), on a logarithmic y-axis. The dashed line is classical inference. Values above one indicate MSE improvement over the gold-only estimator. Values below 0.01 are drawn at 0.01 as open downward triangles; they occur only for the two Bradley–Terry arena applications at their smallest label budgets, where quasi-separation makes the non-classical estimators unstable.

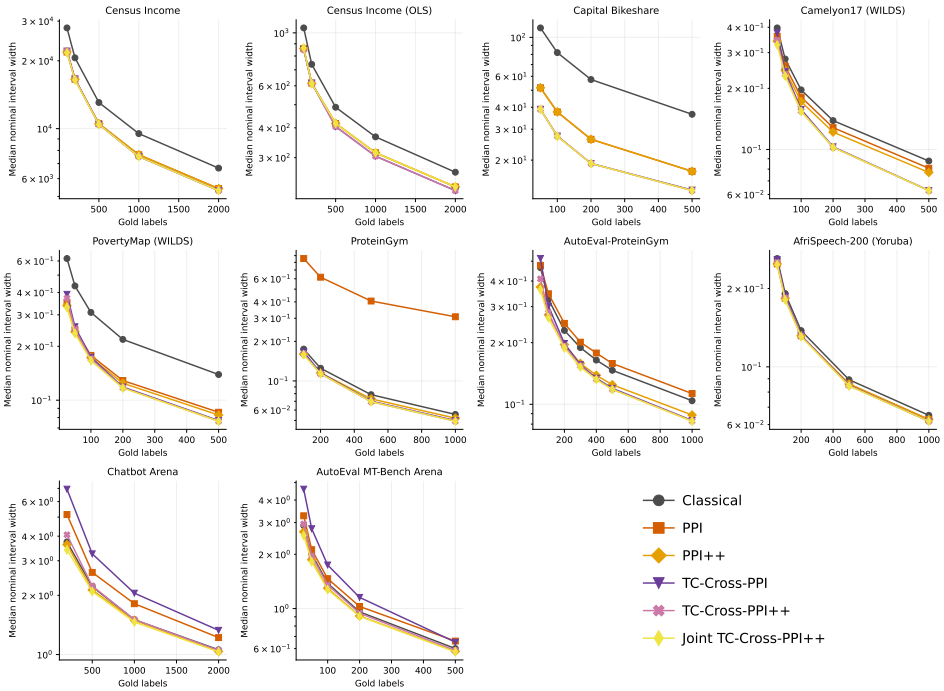


Figure 4: Median (over trials) width of nominal 95% confidence intervals (90% for the two AutoEval replications), on a logarithmic y-axis. Widths are in each estimand's natural units (e.g., dollars for Census Income, hourly rentals for Capital Bikeshare), so they should be compared across methods within a panel rather than across panels.

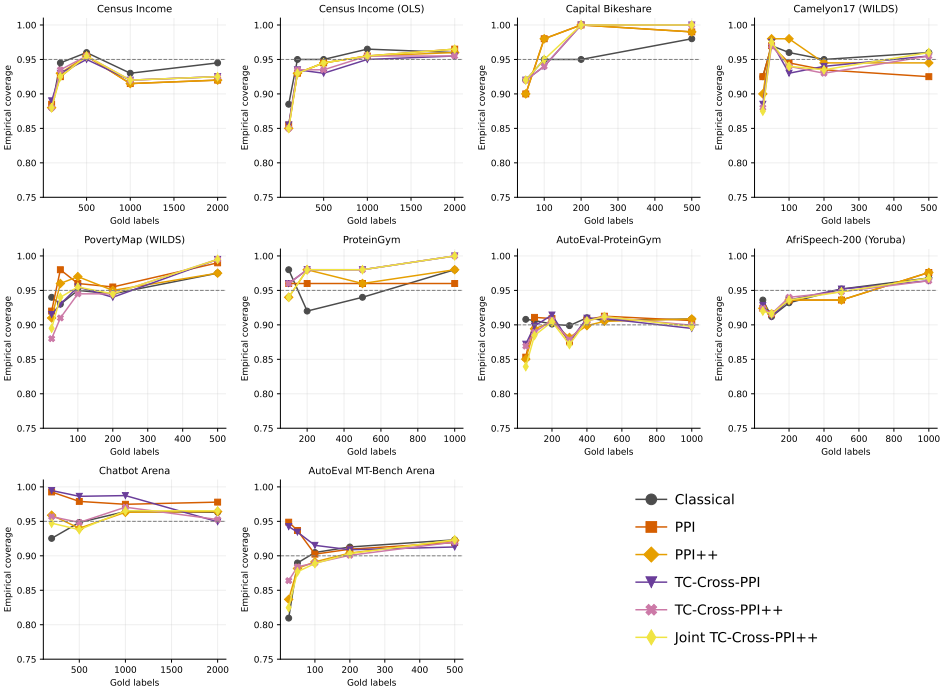


Figure 5: Empirical coverage of nominal confidence intervals. The dashed line marks the nominal level: 95% for most panels, and 90% for the two AutoEval replications (AutoEval-ProteinGym and AutoEval MT-Bench Arena), following AutoEval's convention (Appendix E). Coverage substantially below the line indicates anti-conservative inference; coverage near one can instead reflect conservative or separation-driven intervals and should be interpreted jointly with interval width.

F.1 Results with Additional Datasets

Census OLS. On the Census dataset, we additionally estimate the OLS coefficient using the same covariates (Appendix E; Figures 3 and 4). PPI++ has relative MSE of 0.717 at $n = 200$ and 0.747 at $n = 500$, and TC–Cross-PPI++ improves on it at both, with 0.682 and 0.732. Joint TC–Cross-PPI++ is more conservative here, at 0.713 and 0.735, tracking PPI++ more closely than the stronger cross-fitted correction. Its power weight can only shrink toward whichever direction it trusts least.

ProteinGym. We analyze the `BLAT_ECOLX_Firnberg_2014` assay with ESM-2 8M (`esm2_t6_8M_UR50D`) as the source predictor (Figures 3 and 4). The raw source score is poorly aligned with experimental fitness, and vanilla PPI pins λ to one, so it is highly inefficient here, with relative MSE above $20\times$ classical at every n. Power tuning fixes this. PPI++ reaches relative MSE of 0.79–0.82 across $n = 200$–500, and the cross-fitted methods improve further still, to 0.65–0.77, or ESS gains of up to $\sim 53\%$. For model ranking we target $\mu_m = \mathbb{E}[Y s_m(X)]$, where s_m is model m's standardized zero-shot score. With three candidate ESM models, TC–Cross-PPI++ reaches mean rank correlation 1.000 at $n = 100$, against 0.917 for PPI++ and 0.250 for vanilla PPI.

F.2 Results looking at Annotator Quality

Protein annotator quality. To study sensitivity to synthetic-label quality, we repeat the ProteinGym analysis with ESM-2 8M, 35M, and 150M as the annotator, at a fixed $n = 500$. Their correlations with experimental fitness are 0.396, 0.522, and 0.626. As annotator quality increases, PPI++ MSE decreases from 2.89×10^{-4} to 2.74×10^{-4} to 2.18×10^{-4}. TC–Cross-PPI++ reduces MSE further, to 2.33×10^{-4}, 2.18×10^{-4}, and 1.76×10^{-4}, and Joint TC–Cross-PPI++ tracks it closely (Figure 6, left). Transfer calibration again has its largest proportional benefit for the weakest annotator. Coverage for TC–Cross-PPI++ is 0.976, 0.972, and 0.980 across the three annotators.

AutoEval-ProteinGym annotator quality. We repeat the SPG1 replication using all eight candidate/annotator scores (UniRep, CARP, ESM-1b, ESM-1v, ESM-2, RITA, ProGen2, and VESPA) as the source, at a fixed $n = 500$. Annotator-fitness correlation ranges from -0.05 for UniRep to 0.49 for VESPA, and both MSE and ESS gain improve monotonically with it (Figure 6, right). PPI++'s ESS gain rises from $+0.3\%$ to $+36.8\%$ over that range, and TC–Cross-PPI++'s rises from $+36.8\%$ to $+70.2\%$, staying ahead of PPI++ at every annotator. As with Census OLS, Joint TC–Cross-PPI++ again tracks closer to PPI++ than to the stronger TC–Cross-PPI++ correction. Coverage stays within 0.88–0.92 across every method and annotator, against a nominal 90%.

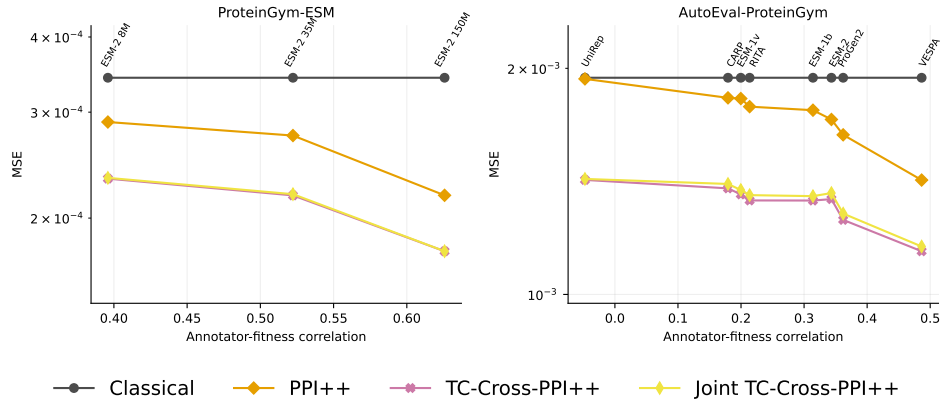


Figure 6: Sensitivity to annotator quality at $n = 500$ (MSE on a logarithmic y-axis in both panels). Left: ProteinGym-ESM MSE by method versus annotator–fitness correlation (3 annotators). Right: AutoEval-ProteinGym MSE by method versus annotator–fitness correlation (8 annotators).

F.3 VARIANCE RESULTS

For mean estimation, adding a constant to the predictor doesn't change PPI's variance, since the rectifier already corrects for such a shift. What transfer calibration adds is a lower residual variance, $\mathrm{Var}(Y-f)$, which Table 4 compares for the source and calibrated predictors. Calibration helps most when the source's errors depend on the covariates. On Capital Bikeshare, the 2011-to-2012 shift varies with the calendar and weather covariates, and calibration roughly halves the residual variance. The reduction also grows with n, since more gold labels give a better-fit correction. When little of the source's error is predictable from the calibration features, as on Census Income, calibration changes little and the cross-fitted methods perform about as well as PPI++. With very few labels, the correction can instead increase the residual variance (e.g., PovertyMap at $n = 25$). Joint power tuning matters most here, since it can down-weight the correction while keeping the source predictor, and it essentially matches PPI++ where TC–Cross-PPI falls behind (Figure 1).

Table 4: Residual variance of the transfer-calibrated predictor relative to the source predictor, $\widehat{\mathrm{Var}}(Y - \hat{f}_t)/\widehat{\mathrm{Var}}(Y - f_s)$, on the gold sample and averaged over trials, at the smallest and largest label budgets. Values below one indicate calibration reduced the residual variance.

Application	Smallest n	Largest n
Census Income	1.03 ($n = 100$)	0.96 ($n = 2000$)
Capital Bikeshare	0.61 ($n = 50$)	0.41 ($n = 500$)
Camelyon17 (WILDS)	1.07 ($n = 25$)	0.60 ($n = 500$)
PovertyMap (WILDS)	1.97 ($n = 25$)	0.79 ($n = 500$)
ProteinGym	0.04 ($n = 100$)	0.03 ($n = 1000$)
AfriSpeech-200 (Yoruba)	1.07 ($n = 50$)	0.97 ($n = 1000$)

F.4 RESULTS WITH LORA

As an alternative to sparse-linear calibration, we also fine-tune a LoRA adapter as the correction model. On BLAT this adapts the ESM-2 8M backbone that produces the source score; on AutoEval-ProteinGym the source (VESPA) is not a fine-tunable network, so the adapter is trained on a separate ESM-2 8M encoder with the VESPA score as an input; on AutoEval MT-Bench Arena, distilgpt2 replaces the GPT-4 judge as both the source and the LoRA backbone (Appendix E). The results are mixed and assay-dependent (Figure 7, Table 5). On BLAT, sparse TC–Cross-PPI++ is slightly preferable, with relative MSE of 0.854 against 0.904 at $n = 1000$. On SPG1, LoRA calibration is noisier at small n but eventually overtakes sparse calibration, giving the strongest result in this comparison at $n = 500$ (0.223 against 0.423 relative MSE, an ESS gain of 348%). On AutoEval MT-Bench Arena, distilgpt2 supplies both the source score and the LoRA backbone, and neither calibration style meaningfully improves on classical inference across the sweep from $n = 100$ to 750. Sparse TC–Cross-PPI++ stays within 1–4$\%$ of classical throughout, while its LoRA counterpart is noisier at small n (1.300 at $n = 100$) before converging to the same near-parity result (1.010 at $n = 750$). We read this as evidence that LoRA calibration can need more gold labels to stabilize than sparse-linear calibration, not that either approach dominates.

Table 5: Sparse-linear versus LoRA calibration at one representative n per application (Figure 7 shows the full ProteinGym/AutoEval-ProteinGym sweep over n; the AutoEval MT-Bench Arena sweep over $n = 100$–750 is summarized in the text). Each cell reports (MSE ratio, CI width ratio), both normalized by classical inference. Bold marks the better of the two calibrated methods, independently for each of the two ratios.

Application	n	Method	(MSE, CI width) ratios
ProteinGym (BLAT)	1000	Classical	(1.000, 1.000)
		TC-Cross-PPI++	(**0.854**, **0.886**)
		TC-Cross-PPI++-LoRA	(0.904, 0.937)
AutoEval-ProteinGym	500	Classical	(1.000, 1.000)
		TC-Cross-PPI++	(0.423, 0.822)
		TC-Cross-PPI++-LoRA	(**0.223**, **0.746**)
AutoEval MT-Bench Arena	750	Classical	(1.000, 1.000)
		TC-Cross-PPI++	(1.010, 0.998)
		TC-Cross-PPI++-LoRA	(**1.010**, **0.994**)

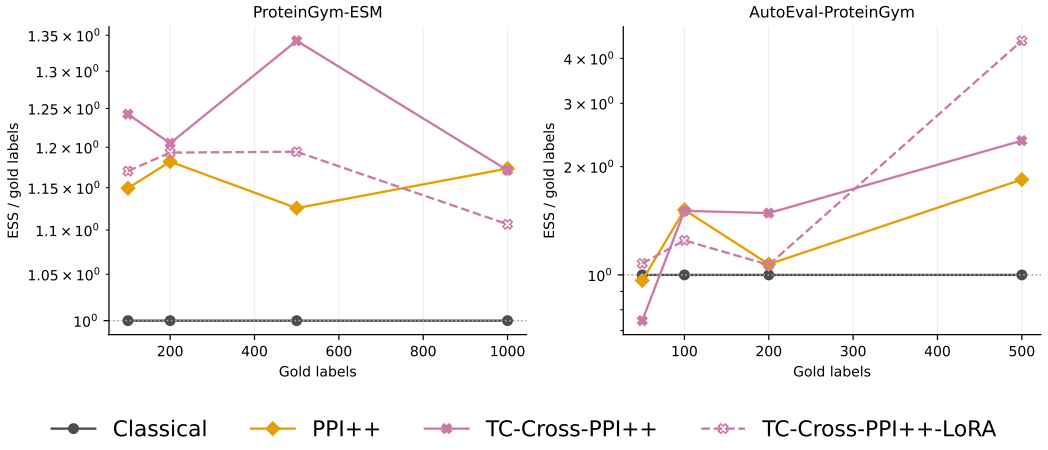


Figure 7: Sparse-linear versus LoRA calibration for ProteinGym-ESM (left) and AutoEval-ProteinGym (right), both using ESM-2 8M as the LoRA backbone. ESS/gold-labels is shown on a logarithmic y-axis; the dotted line at one is classical inference.

F.5 Results with Other PPI Variants

Comparison with Cross-PPI. Cross-prediction-powered inference (Cross-PPI) (Zrnic & Candès, 2024) also cross-fits a predictor on the gold labels, but it trains that predictor from scratch rather than adapting a pretrained source. This comparison tells us how much of our gain comes from the source predictor and how much from cross-fitting alone. For Census Income and Capital Bikeshare, whose sources are tree ensembles, we train Cross-PPI with the same model class on the target covariates (Figure 8). On Capital Bikeshare, Cross-PPI reaches a relative MSE of only 0.75–0.91, compared with 0.16–0.21 for PPI and 0.07–0.14 for TC–Cross-PPI, with coverage between 0.91 and 0.98. On Census Income it does worse than classical inference up to $n = 500$ (relative MSE 1.85 at $n = 100$) and reaches only 0.87 at $n = 2000$, while PPI, PPI++ and the transfer-calibrated methods stay between 0.57 and 0.67. A model fit to up to a few thousand gold labels from the target domain could not compete with the source model that was trained on far more data. Our approach uses both the well-trained source predictor and the scarce gold labels on learning a correction.

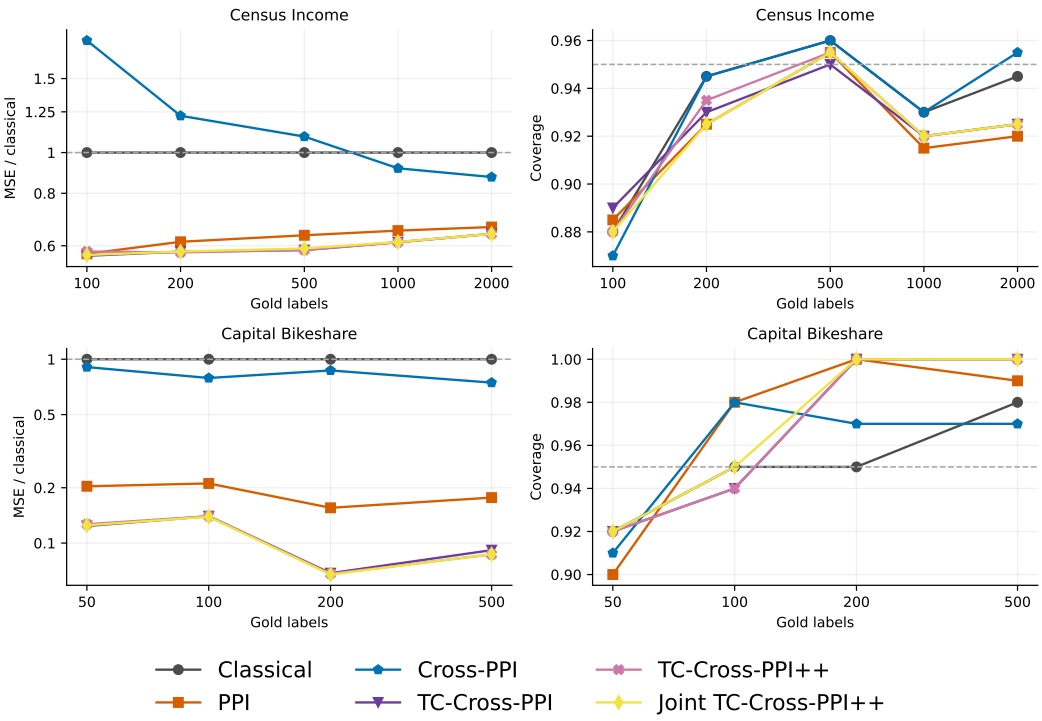


Figure 8: Cross-PPI, trained from scratch with the source's model class, compared with PPI and the transfer-calibrated methods on Census Income (top, 200 repetitions) and Capital Bikeshare (bottom, 100 repetitions). Left: MSE relative to classical inference on a logarithmic y-axis; the dashed line at one is classical inference. Right: coverage, with the dashed line at the nominal 95%. Both panels use a logarithmic x-axis.

Comparison with Stratified PPI. Stratified prediction-powered inference (StratPPI) (Fisch et al., 2024) partitions the input space into strata $\mathcal{A}_k$ with known weights w_k and draws n_k gold labels from each stratum. It then minimizes the stratified rectified loss

$$\sum_k w_k \Big[\tfrac{1}{n_k} \sum_{i \in \mathcal{L}_k} \ell_\theta(X_i, Y_i) + \lambda_k \big(\tfrac{1}{N_k} \sum_{j \in \mathcal{U}_k} \ell_\theta(X_j, f(X_j)) - \tfrac{1}{n_k} \sum_{i \in \mathcal{L}_k} \ell_\theta(X_i, f(X_i)) \big) \Big],$$

with one power-tuning weight $\lambda_k \in [0, 1]$ per stratum. Each λ_k is tuned by the PPI++ criterion restricted to its stratum, so the synthetic labels are trusted only where they track the gold labels. The strata are formed from the autorater's own output. This makes StratPPI a useful baseline for us, since it adapts to judge heterogeneity without learning any correction, while transfer calibration corrects the judge as a function of the covariates.

We evaluate StratPPI on the Arena datasets with proportional allocation, $n_k \propto w_k$, using one stratum per value of GPT-4's five-valued preference on AutoEval MT-Bench Arena and 10 equal-mass bins of the judge probability on Chatbot Arena. Figure 9 reports the median MSE across trials, relative to classical inference, since at small n the Bradley–Terry fit occasionally diverges for every method except classical. StratPPI's coverage is 0.83–0.93 on AutoEval, against a nominal 0.90, and 0.92–0.97 on Chatbot Arena, against a nominal 0.95. Its median-MSE ratio is 0.82–1.01 on AutoEval, against 0.70–0.96 for PPI++, and 0.89–1.30 on Chatbot Arena at $n \geq 500$, against 0.72–0.99. The per-stratum λ_k average 0.25–0.39, close to PPI++'s single weight.

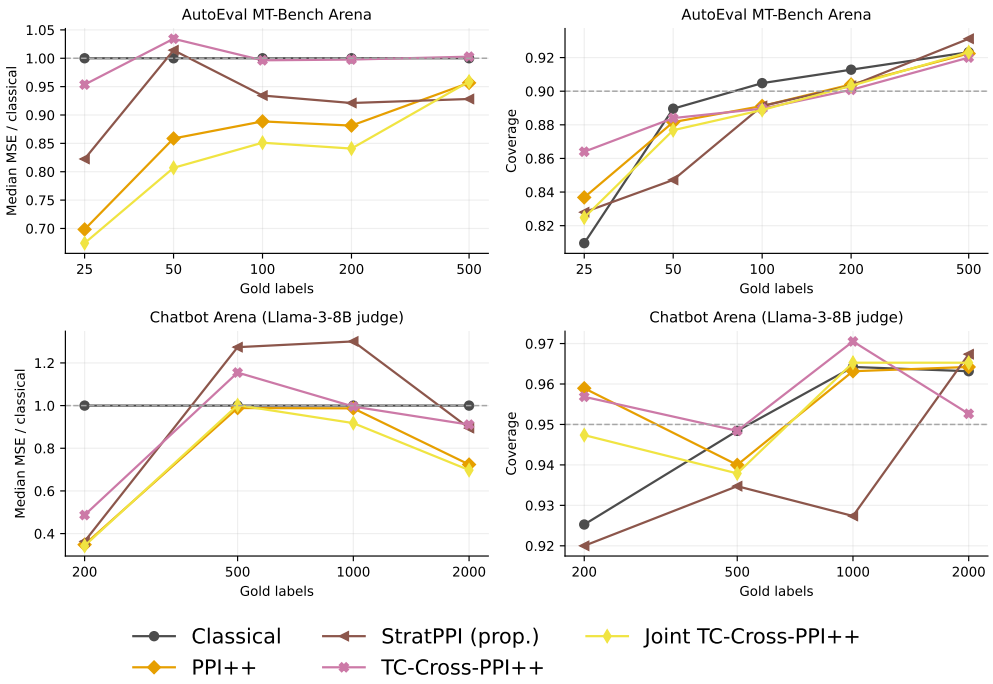


Figure 9: StratPPI (proportional allocation, strata from the judge's own output) compared with PPI++ and the transfer-calibrated methods on AutoEval MT-Bench Arena (top, 250 repetitions) and Chatbot Arena with the Llama-3-8B judge (bottom, 50 repetitions). Left: median BT-coefficient MSE relative to the classical median; the dashed line at one is classical inference. Right: average coordinate-wise coverage, with the dashed line at nominal coverage (90% for AutoEval, following its convention, and 95% for Chatbot Arena). Both panels use a logarithmic x-axis.