Prediction-Powered Adaptive Inference with Pretrained AI Models for Contextual Bandits

Gabriel Sargent [1] **Will Wei Sun** [2] **Zhengwu Zhang** [1] **Yufeng Liu** [3]

Abstract

In adaptive experiments, statistical inference is essential for reliable decision-making and scientific discovery. Often in these settings, collecting labeled data is expensive, but decision-makers have access to large unlabeled datasets and strong pretrained AI models that can predict outcomes. Effectively leveraging these predictions in online experiments poses fundamental challenges: AI predictions may be inaccurate, and data collected under adaptive policies are inherently non-i.i.d., invalidating classical inference techniques. To address these challenges, we propose a Prediction-Powered Adaptive Inference (PPAI) estimator that integrates unlabeled data, predicted labels, and adaptively collected labeled data. We establish asymptotic normality of the PPAI estimator under mild conditions on the data-collection policy, enabling valid confidence intervals and hypothesis tests for a broad class of Z-functionals. The estimator incorporates a data-driven tuning mechanism that weights AI predictions according to their informativeness, guaranteeing that the asymptotic variance is no worse than that of the labeled-only baseline, and is strictly smaller when predictions are informative. Numerical experiments and a movie recommendation application further support the theory, illustrating efficiency gains with informative AI predictions and robust performance with inaccurate predictions.

[1]Department of Statistics and Operations Research, University of North Carolina at Chapel Hill, Chapel Hill, NC, USA [2]Department of Quantitative Methods, Purdue University, West Lafayette, IN, USA [3]Department of Statistics, University of Michigan, Ann Arbor, MI, USA. Correspondence to: Yufeng Liu <yufliu@umich.edu>, Will Wei Sun <sun244@purdue.edu>.

Proceedings of the 43^{rd} International Conference on Machine Learning, Seoul, South Korea. PMLR 306, 2026. Copyright 2026 by the author(s).

1. Introduction

Adaptive experimental designs are used in a wide variety of modern decision-making problems, like clinical trials (Varatharajah & Berry, 2022; Villar et al., 2015), mobile health (Liao et al., 2020; Yom-Tov et al., 2017), recommendation systems (Tang et al., 2014; Li et al., 2010; Zhu & Roy, 2023; McInerney et al., 2018), and education technology (Shaikh et al., 2019; Liu et al., 2014), because they enable agents to adjust decision-making strategies in real time. For instance, clinicians tailor treatments in response to observed health outcomes, recommender systems adjust recommendations based on recent engagement, and tutoring systems dynamically alter assignments based on student performance.

While the primary objective in these settings is often to optimize outcomes (e.g., minimize regret), an equally important goal is to support *statistical inference* about underlying design parameters. Inference enables scientific discovery and principled decision-making by answering questions such as "What is the average treatment effect?" or "Do outcomes differ across subpopulations?" through valid confidence intervals and hypothesis tests. Moreover, inferential insights can guide the design of future experiments and improve downstream decision-making (Shi et al., 2021; 2024; Zhang et al., 2021; 2022; Simchi-Levi & Wang, 2025).

A central bottleneck is that reliable inference typically requires substantial amounts of *labeled* data, yet collecting labels in adaptive experiments can be expensive, slow, or ethically constrained. At the same time, many applications provide abundant *unlabeled* covariates and access to externally trained AI models that can generate cheap outcome predictions. For instance, in clinical trials one may have rich covariates from electronic health records (EHRs) and pretrained models that predict patient outcomes (Wornow et al., 2025). Similarly, in recommendation systems, models trained on historical logs can predict clicks, and in online education, models can predict student performance. These predictions are not substitutes for experimental outcomes, but they constitute valuable *auxiliary information* that could, in principle, improve inferential efficiency.

This motivates a central and technically nontrivial question

of this project: *how can adaptively collected experimental data be combined with AI-generated predictions from potentially inaccurate external models to enable valid and efficient statistical inference?* Treating AI-generated predictions as ground truth can introduce large bias, while ignoring them discards potentially substantial information. Addressing adaptivity and prediction bias simultaneously therefore requires new estimator construction and new asymptotic analysis.

We consider a contextual bandit setting, in which the decision-maker sequentially observes a context $\boldsymbol{X}_t$ (e.g., covariates describing a patient), selects an action A_t (e.g., a treatment) based on a bandit policy, and receives a reward Y_t (e.g., a measure of the patient's health outcome), which depends on $\boldsymbol{X}_t$ and A_t. In addition, the decision-maker has a large pool of unlabeled contexts $\{\tilde{\boldsymbol{X}}_i\}_{i=1}^N$ and a pretrained AI model $f_a(\cdot)$ predicting the rewards of taking an action a for various contexts. The inferential goal is to estimate a parameter $\boldsymbol{\theta}_a^*$ of a working reward function for some action a and construct its confidence interval.

To address this problem, we propose a prediction-powered adaptive estimator $\hat{\boldsymbol{\theta}}_{a,T}$ whose form is explicitly designed to accommodate adaptively collected data $\{(\boldsymbol{X}_t, A_t, Y_t)\}_{t=1}^{T-1}$ (i.e. the experimental history), unlabeled contexts $\{\tilde{\boldsymbol{X}}_i\}_{i=1}^N$, and potentially inaccurate AI predictions $\{f_a(\tilde{\boldsymbol{X}}_i)\}_{i=1}^N$. At each time T, the decision-maker observes the current context $\boldsymbol{X}_T$, updates the estimate $\hat{\boldsymbol{\theta}}_{a,T}$, selects an action A_T based on the bandit policy, and receives a reward Y_T, which is then added to the labeled dataset (see Figure 1).

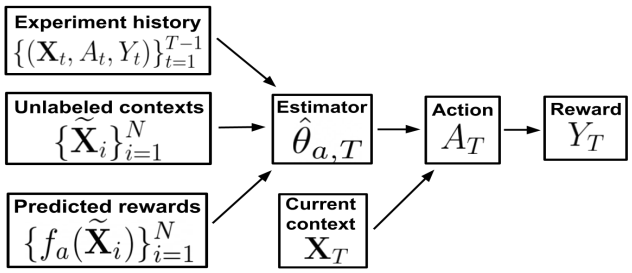

Figure 1. A high-level illustration of our online prediction-powered estimation procedure. At time T, our estimator for $\boldsymbol{\theta}_a^*$ uses adaptively collected labeled data from times $t = 1, \ldots, T-1$ and predicted rewards for selecting action a with the unlabeled contexts. The decision-maker then selects an action and receives a reward, which is then added to the labeled dataset.

On the theoretical side, we establish asymptotic normality of our estimator under mild assumptions on the bandit policy used to collect the data (Theorem 3.7). The proof relies on a martingale central limit theorem tailored to a prediction-powered estimating equation under adaptivity, a setting not covered by existing results. With this result, one can construct valid confidence intervals and hypothesis tests for a broad class of Z-functionals of the reward distribution. Additionally, our estimator adapts to the quality of the AI predictions by employing a data-driven tuning procedure that weights predictions with a parameter $\lambda \in \mathbb{R}$ based on their accuracy. We derive the optimal tuning parameter λ^* and a consistent estimate $\hat{\lambda}_T$ (Theorem 3.8). By weighting according to $\hat{\lambda}_T$, our estimator achieves the minimal asymptotic variance (Corollary 3.10). We compare our estimator to an analogous one that only uses labeled data, which we call the "labeled-only baseline." If the AI prediction model f_a is strong, our estimator leverages the predictions and unlabeled data to achieve a much smaller asymptotic variance than the labeled-only baseline. On the other hand, if f_a is poor, our estimator downweights the predictions, reverting to this baseline. Thus our estimator incorporates the predictions safely, as its asymptotic variance is never larger than the labeled-only baseline and is much smaller if the predictions are informative.

Our **contributions** can be summarized as follows.

1. (**Method**) We propose a prediction-powered adaptive inference framework for contextual bandits that integrates unlabeled contexts and black-box AI predictions while preserving valid inference under adaptive data collection and misspecification. This is not a straightforward combination of existing prediction-powered and adaptive inference methods: naïvely applying prediction-powered estimators to adaptively collected data breaks unbiasedness, while directly treating predictions as additional labels in adaptive inference also induces bias. Our approach resolves these challenges through a new estimator design tailored to both data adaptivity and prediction bias.

2. (**Theory**) We establish asymptotic normality for general Z-functionals of the reward distribution under mild regularity. In adaptive experiments, classical i.i.d. asymptotic analysis is no longer applicable, leading to invalid confidence intervals (Chen et al., 2021; Zhang et al., 2021; Guo & Xu, 2025). To remedy this, we incorporate inverse propensity weighting into the prediction-powered estimating equation to correct for the bias induced by data adaptivity (see Section 3). This construction enables us to establish asymptotic normality of the resulting estimator via a martingale central limit theorem.

3. (**Robustness and Efficiency**) We propose a data-driven tuning rule $\hat{\lambda}_T$ that adaptively weights AI predictions according to their informativeness. The resulting estimator achieves oracle-optimal weighting, guarantees no first-order efficiency loss relative to a labeled-only baseline, and yields strict efficiency gains when predictions are informative. Importantly, in many common settings, our estimation procedure can be efficiently updated in an online fashion.

1.1. Related Work

Our work connects two largely separate lines of research: prediction-powered inference and statistical inference in adaptive experiments.

Prediction-Powered Inference. Prediction-Powered Inference (PPI) leverages predictions from black-box AI models to improve statistical efficiency when only a small i.i.d. labeled dataset is available alongside a large unlabeled dataset (Angelopoulos et al., 2023). PPI++ further improves upon this framework by introducing a power-tuning procedure that adapts to prediction quality and enhances both statistical and computational efficiency (Angelopoulos et al., 2024). More recently, Xu et al. (2025) reinterpret PPI through a semi-supervised learning lens, showing that independently trained predictors do not improve inference for well-specified models and proposing an estimator that improves upon PPI++ in terms of efficiency.

Despite advances, existing PPI-type methods (Angelopoulos et al., 2023; 2024; Kluger et al., 2025; Ji et al., 2025; Xu et al., 2025; Zrnic & Candès, 2024; Wang et al., 2020) are fundamentally designed for offline settings with i.i.d. labeled data. They are therefore ill-suited for adaptive experiments, where data are collected sequentially and actions depend on historical observations. In such settings, naïvely applying PPI-type estimators fails to yield valid inference, as the underlying estimating equations are no longer unbiased and standard asymptotic guarantees break down.

Inference in Adaptive Experiments. A complementary line of work studies valid statistical inference under adaptive data collection (Hadad et al., 2021; Deshpande et al., 2018; Khamaru & Zhang, 2024; Waudby-Smith et al., 2024; Zhang et al., 2020), particularly in contextual bandit and reinforcement learning settings. In the contextual bandit literature, Chen et al. (2021) develop inference procedures for linear reward models under ϵ-greedy policies, while Zhang et al. (2021) propose a general M-estimation framework for adaptive inference. Guo & Xu (2025) extend these results to settings with model misspecification, and Han et al. (2025) study inference for bandits with matrix-valued contexts. Related work in reinforcement learning includes confidence interval construction for policy value functions in Markov decision processes (Shi et al., 2021), as well as extensions to confounded and doubly inhomogeneous environments (Shi et al., 2024; Bian et al., 2025).

However, existing adaptive inference methods rely exclusively on labeled data and do not incorporate unlabeled data or external predictive models. To the best of our knowledge, our work is the first to develop a prediction-powered adaptive inference framework that remains valid under adaptive experimental design, and to demonstrate, both theoretically and empirically, that predictive models can substantially improve inference efficiency in online adaptive experiments.

1.2. PPI++ Background

Before introducing our method, we briefly review PPI++ (Angelopoulos et al., 2023; 2024), which forms the conceptual foundation of our approach. PPI++ assumes access to i.i.d. labeled data $\{(\boldsymbol{X}_i, Y_i)\}_{i=1} \overset{i.i.d.}{\sim} \mathcal{P}$, unlabeled data $\tilde{\boldsymbol{X}}_i \overset{i.i.d.}{\sim} P_X$, and an AI model $f(\cdot)$ predicting labels. Here $\mathcal{P}_X$ is the marginal of $\mathcal{P}$. They consider the task of inferring a parameter $\boldsymbol{\theta}^* = \arg\min_{\boldsymbol{\theta}\in\mathbb{R}^d} \mathbb{E}[\ell(\boldsymbol{X}, Y; \boldsymbol{\theta})]$, for some loss function $\ell(\boldsymbol{x}, y; \boldsymbol{\theta})$, e.g., squared loss for linear regression $\ell(\boldsymbol{x}_i, y_i; \boldsymbol{\theta}) = (\boldsymbol{x}_i^\top \boldsymbol{\theta} - y_i)^2$. PPI++ augments the classical estimating equation using predictions on unlabeled data together with a correction term estimated from labeled data by minimizing a "rectified" loss function:

$$\hat{\boldsymbol{\theta}}^{\text{PP}} = \arg\min_{\theta} L_\lambda^{\text{PP}}(\boldsymbol{\theta}), \quad \text{where} \tag{1}$$

$$L_\lambda^{\text{PP}}(\boldsymbol{\theta}) = \frac{1}{n}\sum_{i=1}^{n} \ell(\boldsymbol{X}_i, Y_i; \boldsymbol{\theta}) + \lambda\left(\frac{1}{N}\sum_{i=1}^{N} \ell(\tilde{\boldsymbol{X}}_i, f(\tilde{\boldsymbol{X}}_i); \boldsymbol{\theta}) - \frac{1}{n}\sum_{i=1}^{n} \ell(\boldsymbol{X}_i, f(\boldsymbol{X}_i); \boldsymbol{\theta})\right).$$

Here, the first term of $L_\lambda^{\text{PP}}(\boldsymbol{\theta})$ is the loss of the classical M-estimator, and the second term is a correction that incorporates the unlabeled data and predictions. The λ is a weighting parameter that controls the influence of the predictions $f(\tilde{\boldsymbol{X}}_i)$. Heuristically, PPI++ sets $\lambda = 0$ when the predictions are uninformative and $\lambda = 1$ when they are very accurate. If $\lambda = 0$, we ignore predictions and unlabeled data, relying only on the labeled data; this recovers the classical estimator. On the other hand, if $\lambda = 1$, Equation (1) can be interpreted as an f-based loss plus a correction term, which is computed with the differences between the losses of the true and predicted labels on the labeled dataset. This decomposition illustrates a simple idea behind PPI++: to evaluate the bias of f, one can compare the true labels Y_t against the predictions $f(\boldsymbol{X}_t)$ on the labeled dataset.

Finally, we note that Equation (1) expresses $\hat{\boldsymbol{\theta}}^{\text{PP}}$ as an M-estimator, but by setting the derivative of $L_\lambda^{\text{PP}}(\boldsymbol{\theta})$ to zero, we can equivalently write $\hat{\boldsymbol{\theta}}^{\text{PP}}$ as a Z-estimator solving

$$\boldsymbol{G}_n^\lambda(\hat{\boldsymbol{\theta}}^{\text{PP}}) = 0, \tag{2}$$

where $\boldsymbol{g}(\boldsymbol{x}, y; \boldsymbol{\theta}) = \nabla_{\boldsymbol{\theta}} \ell(\boldsymbol{x}, y; \boldsymbol{\theta})$ and

$$\boldsymbol{G}_n^\lambda(\boldsymbol{\theta}) = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{g}(\boldsymbol{X}_i, Y_i; \boldsymbol{\theta}) + \lambda\left(\frac{1}{N}\sum_{i=1}^{N} \boldsymbol{g}(\tilde{\boldsymbol{X}}_i, f(\tilde{\boldsymbol{X}}_i); \boldsymbol{\theta}) - \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{g}(\boldsymbol{X}_i, f(\boldsymbol{X}_i); \boldsymbol{\theta})\right). \tag{3}$$

2. Problem Setting

We consider a contextual bandit setting, in which data are collected via the following procedure.

At time T,

1. The environment reveals a context $\boldsymbol{X}_T \in \mathcal{X} \subseteq \mathbb{R}^d$, drawn from a distribution $\mathcal{P}_X$.
2. The environment reveals unlabeled contexts $\{\tilde{\boldsymbol{X}}_{N_{T-1}+1}, \ldots, \tilde{\boldsymbol{X}}_{N_T}\}$ sampled i.i.d. from $\mathcal{P}_X$. Let $\mathcal{U}_T = \sigma(\tilde{\boldsymbol{X}}_1, \ldots, \tilde{\boldsymbol{X}}_{N_T})$ denote the information from all unlabeled contexts available at time T.
3. The decision-maker selects an action $A_T \in \mathcal{A}$ according to a bandit policy $\pi_T(\cdot|\boldsymbol{X}_T, \mathcal{H}_{T-1}, \mathcal{U}_T) \in \Delta(\mathcal{A})$, which depends on the current context $\boldsymbol{X}_T$, the history $\mathcal{H}_{T-1} = \sigma(\{\boldsymbol{X}_t, A_t, Y_t\}_{t=1}^{T-1})$, and the unlabeled data. Here $\Delta(\mathcal{A})$ is the set of probability distributions over the action space $\mathcal{A}$.
4. The decision-maker receives a reward $Y_T \in \mathbb{R}$. We let $\{Y_T(a) : a \in \mathcal{A}\}$ denote the potential outcomes (the rewards the decision-maker would have received for playing each action), so that $Y_T = Y_T(A_T)$.

In addition, the decision-maker has an externally trained model $f_a : \mathcal{X} \to \mathbb{R}$ for $a \in \mathcal{A}$, where $f_a(\boldsymbol{x})$ is a prediction of the reward for playing action a under context $\boldsymbol{x}$. In Appendix F, we discuss extensions to settings with possible covariate shift, where the unlabeled data may come from a distribution different from the target population.

Our inferential target is the parameter $\boldsymbol{\theta}_a^\star \in \mathbb{R}^d$ associated with a specific action $a \in \mathcal{A}$, defined through a general Z-estimation problem:

$$\mathbb{E}\big[\boldsymbol{g}(\boldsymbol{X}, Y(a); \boldsymbol{\theta}_a^\star)\big] = \mathbf{0}, \tag{4}$$

where $\boldsymbol{g}(\cdot)$ is a known score function. Unlike many prior works on inference with adaptively collected data (Shi et al., 2021; 2024; Zhang et al., 2021; Han et al., 2025), we do not assume a well-specified outcome model. In particular, the validity of our inference procedure does not rely on correctly modeling the conditional distribution of Y_T given A_T and $\boldsymbol{X}_T$. This feature makes our approach especially well-suited for adaptive experiments conducted in complex environments, where model misspecification is common. In Section 2.1, we introduce our method via a key special case of misspecified linear bandits (Chen et al., 2021). Other common choices of $\boldsymbol{g}$ can model bandits with noisy contexts, generalized linear models, and off-policy evaluation (Guo & Xu, 2025).

2.1. Warm-Up: Misspecified Linear Bandits

In the misspecified linear bandits (Chen et al., 2021), the score function in (4) is $g(\boldsymbol{x}, y; \boldsymbol{\theta}) = \boldsymbol{x}(y - \boldsymbol{x}^\top \boldsymbol{\theta})$ and hence the inferential parameter $\boldsymbol{\theta}_a^\star$ solves $\mathbb{E}\big[\boldsymbol{X}(Y(a) - \boldsymbol{X}^\top \boldsymbol{\theta}_a^\star)\big] = \mathbf{0}$. This score function corresponds to the best linear approximation of $Y(a)$ based on $\boldsymbol{X}$ for arm a. When the true dependence of $Y(a)$ on $\boldsymbol{X}$ and a is linear, this yields the true linear parameter. Otherwise, it defines the best linear projection of $Y(a)$ onto the covariates $\boldsymbol{X}$ in the least squares sense.

When ignoring the actions, the classical least squares estimator with labeled data $\{\boldsymbol{X}_t, Y_t\}_{t=1}^T$ is the OLS estimator:

$$\hat{\boldsymbol{\theta}}^{\text{OLS}} = \left[\sum_{t=1}^T \boldsymbol{X}_t \boldsymbol{X}_t^\top\right]^{-1} \sum_{t=1}^T \boldsymbol{X}_t Y_t. \tag{5}$$

For adaptive data with actions collected via a bandit policy, Chen et al. (2021) proposes a Weighted Least Squares (WLS) estimator, which uses inverse propensity weights to account for the adaptive bias:

$$\hat{\boldsymbol{\theta}}_{a,T}^{\text{WLS}} = \left[\sum_{t=1}^T \frac{1_{A_t=a}}{\pi_t(a)} \boldsymbol{X}_t \boldsymbol{X}_t^\top\right]^{-1} \sum_{t=1}^T \frac{1_{A_t=a}}{\pi_t(a)} \boldsymbol{X}_t Y_t. \tag{6}$$

Here $\pi_t(a) = \pi_t(a|\boldsymbol{X}_t, \mathcal{H}_{t-1}) := \mathbb{E}\left[1_{A_t=a}|\mathcal{H}_{t-1}, \boldsymbol{X}_t\right]$ was the probability of selecting action a at time t. For example, the ϵ-greedy policy (Lattimore & Szepesvári, 2020) plays the apparently best arm with probability $1 - \epsilon/2$ and chooses randomly with probability $\epsilon/2$. In a 2-armed linear bandit, it plays arm 1 at time t with probability

$$\pi_t(1) = (1 - \epsilon) 1_{\boldsymbol{X}_t^\top \hat{\boldsymbol{\theta}}_{1,t-1} > \boldsymbol{X}_t^\top \hat{\boldsymbol{\theta}}_{0,t-1}} + \frac{\epsilon}{2}.$$

Another example is the softmax policy (Lattimore & Szepesvári, 2020), which computes selection probabilities via exponential weighting:

$$\pi_t(1) = \frac{\exp(\boldsymbol{X}_t^\top \hat{\boldsymbol{\theta}}_{1,t-1})}{\exp(\boldsymbol{X}_t^\top \hat{\boldsymbol{\theta}}_{0,t-1}) + \exp(\boldsymbol{X}_t^\top \hat{\boldsymbol{\theta}}_{1,t-1})}.$$

Dividing by these probabilities removes the adaptive bias induced by A_t; observe that

$$\begin{aligned}
&\mathbb{E}\left[\frac{1_{A_t=a}}{\pi_t(a)} \boldsymbol{X}_t Y_t | \mathcal{H}_{t-1}\right] \\
&= \mathbb{E}\left[\mathbb{E}\left[\frac{1_{A_t=a}}{\pi_t(a)} \boldsymbol{X}_t Y_t(a) | \mathcal{H}_{t-1}, \boldsymbol{X}_t, Y_t(a)\right] | \mathcal{H}_{t-1}\right] \\
&= \mathbb{E}\left[\mathbb{E}\left[1_{A_t=a} | \mathcal{H}_{t-1}, \boldsymbol{X}_t\right] \frac{1}{\pi_t(a)} \boldsymbol{X}_t Y_t(a) | \mathcal{H}_{t-1}\right] \\
&= \mathbb{E}\left[\boldsymbol{X}_t Y_t(a) | \mathcal{H}_{t-1}\right].
\end{aligned}$$

Here the first equality uses the law of iterated expectations; the second uses that $\pi_t(a), \boldsymbol{X}_t$, and $Y_t(a)$ are measurable with respect to the inner conditioning and independence of A_t and $Y_t(a)$; and the last uses the definition of $\pi_t(a)$.

Our proposed prediction-powered adaptive inference (PPAI) framework integrates unlabeled contexts $\{\tilde{\boldsymbol{X}}_i\}_{i=1}^N$ and AI

predictions f_a while preserving valid inference under adaptive data collection and misspecification. Specifically, our PPAI estimator for the misspecified linear bandits is:

$$\hat{\boldsymbol{\theta}}_{a,T}^{\text{PPAI}} = \left[\frac{1-\lambda}{T}\sum_{t=1}^{T}\frac{1_{A_t=a}}{\pi_t(a)}\boldsymbol{X}_t\boldsymbol{X}_t^T + \frac{\lambda}{N}\sum_{i=1}^{N}\tilde{\boldsymbol{X}}_i\tilde{\boldsymbol{X}}_i^T\right]^{-1} \times\left[\frac{1}{T}\sum_{t=1}^{T}\frac{1_{A_t=a}}{\pi_t(a)}\boldsymbol{X}_t Y_t + \lambda\Big(\frac{1}{N}\sum_{i=1}^{N}\tilde{\boldsymbol{X}}_i f_a(\tilde{\boldsymbol{X}}_i) - \frac{1}{T}\sum_{t=1}^{T}\frac{1_{A_t=a}}{\pi_t(a)}\boldsymbol{X}_t f_a(\boldsymbol{X}_t)\Big)\right]. \tag{7}$$

Like the PPI++ estimator (1), our PPAI estimator uses a tuning parameter λ that weights the AI model based on its informativeness. If $\lambda = 0$, we ignore predictions and unlabeled data, relying only on the labeled data; this recovers the WLS estimator (6). On the other hand, if $\lambda = 1$, we weight the predictions highly, and the second bracketed term of (7) can be rewritten as

$$\frac{1}{N}\sum_{i=1}^{N}\tilde{\boldsymbol{X}}_i f_a(\tilde{\boldsymbol{X}}_i) + \frac{1}{T}\sum_{t=1}^{T}\frac{1_{A_t=a}}{\pi_t(a)}\boldsymbol{X}_t\left(Y_t - f_a(\boldsymbol{X}_t)\right).$$

This can be interpreted as an f_a-based prediction plus a correction term, which represents the bias of f_a.

3. Our Method under General Score Function

We now develop our PPAI estimator for the general inferential target $\boldsymbol{\theta}_a^*$ defined in (4) that solves $\mathbb{E}[\boldsymbol{g}(\boldsymbol{X}, Y(a); \boldsymbol{\theta}_a^*)] = 0$ for some known score function $\boldsymbol{g}(\cdot)$. Our PPAI estimator $\widehat{\boldsymbol{\theta}}_{a,T}$ is defined as the solution to

$$\boldsymbol{G}_{a,T}^{\lambda_{a,T}}(\boldsymbol{\theta}) := \frac{\lambda_{a,T}}{N}\sum_{i=1}^{N}\boldsymbol{g}(\tilde{\boldsymbol{X}}_i, f_a(\tilde{\boldsymbol{X}}_i); \boldsymbol{\theta}) + \frac{1}{T}\sum_{t=1}^{T}\frac{1_{\{A_t=a\}}}{\pi_t(a)}\Big(\boldsymbol{g}(\boldsymbol{X}_t, Y_t; \boldsymbol{\theta}) - \lambda_{a,T}\boldsymbol{g}(\boldsymbol{X}_t, f_a(\boldsymbol{X}_t); \boldsymbol{\theta})\Big) = o_p(1/\sqrt{T}). \tag{8}$$

Our estimator extends the PPI++ estimator in (2) to the adaptive data setting by incorporating the inverse propensity weighting. Specifically, the indicator $1_{A_t=a}$ restricts the calculations to times where arm a was played, and inverse propensity weights $1/\pi_t(a)$ cancels the adaptive bias. Like PPI++, $\{\lambda_{a,T}\}_{T\geq 1}$ is a (data-dependent) real-valued sequence that controls the weighting of the AI model over the course of the experiment. Crucially, the quantity weighted by the inverse propensity weights is a small residual $\boldsymbol{g}(\boldsymbol{X}_t, Y_t(a); \boldsymbol{\theta}) - \lambda_{a,T}\boldsymbol{g}(\boldsymbol{X}_t, f_a(\boldsymbol{X}_t); \boldsymbol{\theta})$, which mitigates the high variance introduced by small weights. We will soon exhibit a specific choice of $\{\lambda_{a,T}\}_{T\geq 1}$ that shrinks this residual by weighting AI predictions according to their informativeness (Corollary 3.10).

Before establishing the theoretical properties of PPAI, we first introduce all regularity conditions. Similarly to PPI, we assume that the unlabeled data is independent from the potential outcomes: $\{\tilde{\boldsymbol{X}}_i\}_i \perp \{\boldsymbol{X}_t, \{Y_t(a)\}_{a\in\mathcal{A}}\}_{t\geq 1}$. We also assume $\{\boldsymbol{X}_t, Y_t(a) : a \in \mathcal{A}\} \overset{i.i.d.}{\sim} \mathcal{P}$ for all $t \geq 1$. That is, the environment's draw of contexts and potential outcomes are independent across time. This is a standard assumption in contextual bandit work (Guo & Xu, 2025; Zhang et al., 2021; Chen et al., 2021). While the contexts and potential outcomes are independent, the actions are not, as they are chosen based on the bandit policy and depend on the evolving experimental history. Consequently, our labeled data $\{(\boldsymbol{X}_t, Y_t) : A_t = a\}$ is not i.i.d., unlike the standard PPI setup. We also emphasize that we do not assume a functional form of the reward model $Y_t(a)|\boldsymbol{X}_t$, so our method does not require a well-specified reward model.

We first make the following assumptions on the policy π.

Assumption 3.1 (Policy Convergence)**.** There exists a stationary policy $\bar{\pi} : \mathcal{X} \to \Delta(\mathcal{A})$ such that

$$\pi_t(a|\boldsymbol{X}_t, \mathcal{H}_{t-1}, \mathcal{U}_t) - \bar{\pi}(a|\boldsymbol{X}_t) \xrightarrow{p} 0 \quad \text{as } t \to \infty.$$

Assumption 3.2 (Minimum sampling probability)**.** For $a \in \mathcal{A}$, there exists a constant $\pi_{\min} \in (0, 1)$ such that $\pi_t(a) := \pi_t(a|\boldsymbol{X}_t, \mathcal{H}_{t-1}, \mathcal{U}_t) \geq \pi_{\min}$ almost surely for all $t \geq 1$.

Policy convergence says that the probability of selecting arm a given each context $\boldsymbol{X}_t$ stabilizes in the long run. This ensures an experiment's replicability by requiring that the policy's long-term behavior is always the same, regardless of the initial observations. This condition is satisfied by a broad class of commonly employed policies and is frequently assumed in adaptive inference works under misspecification (Chen et al., 2021; Guo & Xu, 2025). Minimum sampling probabilities are essential for inference under misspecification, as they can ensure sufficient exploration in a poorly understood environment, and they are also commonly assumed in the adaptive inference literature (Zhang et al., 2021; Chen et al., 2021; Zhang et al., 2023; Guo & Xu, 2025; Simchi-Levi & Wang, 2025; Han et al., 2025).

We also make the following causal assumption.

Assumption 3.3 (Unconfoundedness)**.** $A_t \perp \{Y_t(a)\}_{a\in\mathcal{A}}|(\mathcal{H}_{t-1}, \boldsymbol{X}_t, \mathcal{U}_t)$ for $t = 1, \ldots, T$.

Assumption 3.3 is a common unconfoundedness assumption in bandit causal framing (Zhang et al., 2021; Guo & Xu, 2025; Han et al., 2025); it says that the decision-maker acts without knowledge of the potential outcomes beyond what they have already observed. We also adopt the following

regularity conditions on the score function, which can be easily satisfied under misspecified linear bandits, bandits with noisy contexts, and score functions corresponding to other common settings (Guo & Xu, 2025).

Assumption 3.4 (Well-separated solution)**.** $\forall \epsilon > 0$, $\inf_{||\boldsymbol{\theta}-\boldsymbol{\theta}_a^*||_2>\epsilon} ||\mathbb{E}[\boldsymbol{g}(\boldsymbol{X}_t, Y_t(a); \boldsymbol{\theta})]||_2 > 0$.

Assumption 3.5 (Bounded moments)**.** There exist constants R_θ, M_2, M_3, M_4 such that, a.e. $\boldsymbol{X}_t$,

(i) $||\mathbb{E}[\boldsymbol{g}(\boldsymbol{X}_t, Y_t(a); \boldsymbol{\theta}_a^*)\boldsymbol{g}(\boldsymbol{X}_t, Y_t(a); \boldsymbol{\theta}_a^*)^\mathsf{T}|\boldsymbol{X}_t]||_2 \leq M_2$, $||\mathbb{E}[\boldsymbol{g}(\boldsymbol{X}_t, f_a(\boldsymbol{X}_t); \boldsymbol{\theta}_a^*)\boldsymbol{g}(\boldsymbol{X}_t, f_a(\boldsymbol{X}_t); \boldsymbol{\theta}_a^*)^\mathsf{T}|\boldsymbol{X}_t]||_2 \leq M_3$, $||\mathbb{E}[\boldsymbol{g}(\boldsymbol{X}_t, Y_t(a); \boldsymbol{\theta}_a^*)\boldsymbol{g}(\boldsymbol{X}_t, f_a(\boldsymbol{X}_t); \boldsymbol{\theta}_a^*)^\mathsf{T}|\boldsymbol{X}_t]||_2 \leq M_4$;

(ii) $||\boldsymbol{\theta}_a^*||_2 < R_\theta$, $\sup_{||\boldsymbol{\theta}||\leq R_\theta} \mathbb{E}[||\boldsymbol{g}(\boldsymbol{X}_t, Y_t(a); \boldsymbol{\theta})||_2^2] < \infty$;

(iii) $\mathbb{E}[||\boldsymbol{g}(\boldsymbol{X}_t, Y_t(a); \boldsymbol{\theta}_a^*)||_2^4], \mathbb{E}[||\boldsymbol{g}(\boldsymbol{X}_t, f_a(\boldsymbol{X}_t); \boldsymbol{\theta}_a^*)||_2^4] < \infty$.

Assumption 3.6 (Smoothness)**.** (i) The function $\boldsymbol{g}(\boldsymbol{x}, y; \boldsymbol{\theta})$ is twice continuously differentiable with respect to $\boldsymbol{\theta}$, with $\mathbb{E}[\nabla \boldsymbol{g}(\boldsymbol{X}_t, Y_t(a); \boldsymbol{\theta}_a^*)]$ nonsingular;

(ii) There exists a function $\phi : \mathbb{R}^{d_X} \times \mathbb{R} \to \mathbb{R}$ such that $\forall \boldsymbol{x}, y, \sup_{||\boldsymbol{\theta}||\leq R_\theta} ||\nabla \boldsymbol{g}(\boldsymbol{x}, y; \boldsymbol{\theta})||_2 \leq \phi(\boldsymbol{x}, y)$, and $\mathbb{E}[\phi(\boldsymbol{X}_t, Y_t(a))^2], \mathbb{E}[\phi(\boldsymbol{X}_t, f_a(\boldsymbol{X}_t))^2] < \infty$;

(iii) There exists a constant $\epsilon_0 > 0$ and a function $\Phi : \mathbb{R}^{d_X} \times \mathbb{R} \to \mathbb{R}$ such that $\sup_{||\boldsymbol{\theta}-\boldsymbol{\theta}^*||\leq\epsilon_0, i\in[d]} ||\nabla^2 \boldsymbol{g}^{(i)}(\boldsymbol{x}, y; \boldsymbol{\theta})||_2 \leq \Phi(\boldsymbol{x}, y)$ and $\mathbb{E}[\Phi(\boldsymbol{X}_t, Y_t(a))], \mathbb{E}[\Phi(\boldsymbol{X}_t, f_a(\boldsymbol{X}_t))] < \infty$, where $\boldsymbol{g}^{(i)}(\boldsymbol{x}, y; \boldsymbol{\theta})$ is the ith entry of $\boldsymbol{g}(\boldsymbol{x}, y; \boldsymbol{\theta})$.

Assumption 3.4 is a standard assumption in Z-estimation that requires the expected score to be bounded away from zero outside any neighborhood of the true parameter $\boldsymbol{\theta}_a^*$; this ensures that $\boldsymbol{\theta}_a^*$ is identifiable (Vaart, 1998). Assumption 3.5 imposes mild regularity on the moments of the score functions $\boldsymbol{g}(\boldsymbol{x}, y; \boldsymbol{\theta}_a^*)$ and $\boldsymbol{g}(\boldsymbol{x}, f(\boldsymbol{x}); \boldsymbol{\theta}_a^*)$ with true and predicted labels, respectively. These types of boundedness assumptions are standard for establishing asymptotic normality of estimators in adaptive experiments (Zhang et al., 2021; Chen et al., 2021; Guo & Xu, 2025). Finally, part (i) of Assumption 3.6 is required for identifiability of $\boldsymbol{\theta}_a^*$, and parts (ii) and (iii) impose smoothness on $\boldsymbol{g}$; all three are standard for Z-estimation in both classical and adaptive settings (Vaart, 1998; Zhang et al., 2023).

In what follows, we let $r := \lim_{T\to\infty} T/N$ denote the limit of the (typically small) ratio of the sizes of the labeled and unlabeled datasets, and we let $\overline{\lambda}_a$ denote the limit in probability of $\lambda_{a,T}$ whenever it exists. Also, for $\boldsymbol{v} \in \mathbb{R}^d$, we write $\boldsymbol{v}^{\otimes 2} := \boldsymbol{v}\boldsymbol{v}^\top$. For brevity, we sometimes suppress dependence on the action a in our notation, writing λ_T for $\lambda_{a,T}$, $\overline{\lambda}$ for $\overline{\lambda}_a$, f for f_a, $\widehat{\boldsymbol{\theta}}_T$ for $\widehat{\boldsymbol{\theta}}_{a,T}$, $\boldsymbol{\theta}^*$ for $\boldsymbol{\theta}_a^*$, and $\boldsymbol{G}_T^{\lambda_T}(\boldsymbol{\theta})$ for $\boldsymbol{G}_{a,T}^{\lambda_{a,T}}(\boldsymbol{\theta})$.

Theorem 3.7 (Asymptotic Normality)**.** *Suppose Assumptions 3.1-3.6 hold for an action $a \in \mathcal{A}$ and $\lambda_T \xrightarrow{p} \bar{\lambda}$. Then there exists an estimator sequence $\{\widehat{\boldsymbol{\theta}}_T\}_{T\geq 1}$ satisfying the estimating equation (8) with $||\widehat{\boldsymbol{\theta}}_T||_2 \leq R_\theta$ for all T. Moreover, for any such sequence, as $T \to \infty$,*

$$\sqrt{T}(\widehat{\boldsymbol{\theta}}_T - \boldsymbol{\theta}^*) \xrightarrow{d} \mathcal{N}(\boldsymbol{0}, \boldsymbol{\Sigma}_{\overline{\lambda}}), \tag{9}$$

where

$$\boldsymbol{\Sigma}_{\overline{\lambda}} = \boldsymbol{J}^{-1}(\overline{\lambda}^2 r\mathrm{Cov}(\boldsymbol{g}(\boldsymbol{X}_t, f(\boldsymbol{X}_t); \boldsymbol{\theta}^*)) + \boldsymbol{V})\boldsymbol{J}^{-1\top}, \tag{10}$$

$$\boldsymbol{J} = \mathbb{E}\nabla \boldsymbol{g}(\boldsymbol{X}_t, Y_t(a); \boldsymbol{\theta}^*),$$

$$\boldsymbol{V} = \mathbb{E}\left[\frac{1}{\overline{\pi}(a)}\left(\boldsymbol{g}(\boldsymbol{X}_t, Y_t(a); \boldsymbol{\theta}^*) - \overline{\lambda}\boldsymbol{g}(\boldsymbol{X}_t, f(\boldsymbol{X}_t); \boldsymbol{\theta}^*)\right)^{\otimes 2}\right]$$

$$-\overline{\lambda}^2\mathbb{E}[\boldsymbol{g}(\boldsymbol{X}_t, f(\boldsymbol{X}_t); \boldsymbol{\theta}^*)]^{\otimes 2}.$$

In Appendix C, we give a consistent estimator of (10), enabling construction of confidence intervals. Also, note that if $\overline{\lambda} = 0$, (10) reduces to

$$\Sigma_0 = \boldsymbol{J}^{-1}\mathbb{E}\left[\frac{1}{\overline{\pi}(a)}\boldsymbol{g}(\boldsymbol{X}_t, Y_t(a); \boldsymbol{\theta}^*)^{\otimes 2}\right]\boldsymbol{J}^{-1\top},$$

which coincides with the variance from Theorem 3.2 of Guo & Xu (2025).

Establishing this result is technically nontrivial for several reasons. Firstly, we explicitly allow misspecification of the reward model, which invalidates classical likelihood-based arguments. Secondly, since the labeled data are collected adaptively, the standard empirical process methods used to derive PPI asymptotics are not applicable; instead, we employ martingale limit theory.

A key step in our proof is to derive the asymptotic distribution of the scaled estimating equation $\sqrt{T}\boldsymbol{G}_T^{\lambda_T}(\boldsymbol{\theta}^*)$. This term depends on both unlabeled and labeled data, which are coupled by the weighting parameter λ_T. We analyze this expression by decomposing it into its labeled and unlabeled components and applying a martingale CLT. The unlabeled components are independent of previously observed information, while the labeled components inherit dependence through the adaptive policy, necessitating martingale methods.

The asymptotic normality result (9) holds for any fixed choice of the weighting parameter $\bar{\lambda}$, ensuring valid inference regardless of how AI predictions are weighted. However, different choices of $\bar{\lambda}$ lead to different asymptotic variances. To fully exploit the auxiliary information provided by the predictions, we now characterize the choice of $\bar{\lambda}$ that minimizes the asymptotic variance of the estimator. In particular, we derive the power-tuning parameter λ^* that minimizes the trace of $\boldsymbol{\Sigma}_{\bar{\lambda}}$.

Theorem 3.8 (Optimal $\bar{\lambda}$). *The trace of $\boldsymbol{\Sigma}_{\bar{\lambda}}$ is minimized at*

$$\lambda^* = \frac{\mathrm{tr}\left(\boldsymbol{J}^{-1}\mathbb{E}[\frac{1}{\bar{\pi}(a)}(\boldsymbol{g}_f^*\boldsymbol{g}_Y^{*\mathsf{T}} + \boldsymbol{g}_Y^*\boldsymbol{g}_f^{*\mathsf{T}})]\boldsymbol{J}^{-1\top}\right)}{2\mathrm{tr}\left(\boldsymbol{J}^{-1}(r\mathrm{Cov}(\boldsymbol{g}_f^*) + E[\frac{\boldsymbol{g}_f^{*\otimes 2}}{\bar{\pi}(a)}] - E[\boldsymbol{g}_f^*]^{\otimes 2})\boldsymbol{J}^{-1\top}\right)}, \tag{11}$$

where $\boldsymbol{g}_Y^* = \boldsymbol{g}(\boldsymbol{X}, Y(a); \boldsymbol{\theta}^*)$, $\boldsymbol{g}_f^* = \boldsymbol{g}(\boldsymbol{X}, f(\boldsymbol{X}); \boldsymbol{\theta}^*)$, and $\boldsymbol{J} = E[\nabla \boldsymbol{g}_Y^*]$.

The optimal $\bar{\lambda}$ balances two objectives: limiting variance contributed by the AI model (which favors small λ) and minimizing the IPW-weighted residual (which favors λ that aligns λg_f with g_Y). The denominator acts as a normalizing factor that penalizes high variance in $f(\boldsymbol{X})$, reflecting the first objective. The numerator of λ^* measures the (weighted) alignment of the AI predictions and true rewards, reflecting the second objective. If they are well aligned, $\lambda^* \approx 1$, and if they are poorly aligned, $\lambda^* \approx 0$.

Algorithm 1 summarizes the online procedure for updating $\hat{\boldsymbol{\theta}}_T$ to $\hat{\boldsymbol{\theta}}_{T+1}$. To avoid repeatedly summing over the labeled and unlabeled datasets, it stores intermediate sums B, C, D, E, F to compute an estimate $\hat{\lambda}$. It also stores sums $F(\boldsymbol{\theta}) = \sum_{t=1}^{T} \frac{1_{A_t=a}}{\pi_t(a)} \boldsymbol{g}(\boldsymbol{X}_t, Y_t; \boldsymbol{\theta})$, $H(\boldsymbol{\theta}) = \frac{1}{N}\sum_{i=1}^{N} \boldsymbol{g}(\tilde{\boldsymbol{X}}_i, f_a(\tilde{\boldsymbol{X}}_i); \boldsymbol{\theta})$, and $I(\boldsymbol{\theta}) = \sum_{t=1}^{T} \frac{1_{A_t=a}}{\pi_t(a)} \boldsymbol{g}(\boldsymbol{X}_t, f_a(\boldsymbol{X}_t); \boldsymbol{\theta})$ to compute the estimating function $G_{T+1}^{\hat{\lambda}}(\boldsymbol{\theta})$. We also denote $\boldsymbol{g}_{f,t} := \boldsymbol{g}(\boldsymbol{X}_t, f(\boldsymbol{X}_t); \hat{\boldsymbol{\theta}}_t)$ and $\boldsymbol{g}_{Y,t} := \boldsymbol{g}(\boldsymbol{X}_t, Y_t; \hat{\boldsymbol{\theta}}_t)$. Since $\hat{\boldsymbol{\theta}}_{T+1}$ is defined as a root of $\boldsymbol{G}_{T+1}^{\hat{\lambda}}(\boldsymbol{\theta})$, it can be computed with standard root-finding algorithms like Newton's method. Excluding root-finding, the per-step complexity of Algorithm 1 is $\mathcal{O}(d^3)$ for a single action and $\mathcal{O}(Kd^3)$ for all K actions, independent of T. For scores such as least-squares, these bounds also include root-finding. To ensure tightness, $\hat{\lambda}_T$ may be projected onto an interval $[-M, M]$ with $M > |\lambda^*|$; we suppress this projection in the notation.

Corollary 3.9. *A consistent estimate of λ^* is*

$$\hat{\lambda}_T := \frac{\mathrm{tr}(\boldsymbol{B}_T^{-1}\boldsymbol{C}_T\boldsymbol{B}_T^{-\top})}{2\mathrm{tr}(\boldsymbol{B}_T^{-1}(\frac{T}{N}(\boldsymbol{D}_T - \boldsymbol{E}_T\boldsymbol{E}_T^\top) + \boldsymbol{F}_T - \boldsymbol{E}_T\boldsymbol{E}_T^{\mathsf{T}})\boldsymbol{B}_T^{-\top})}, \tag{12}$$

projected as described above, where

$$\boldsymbol{B}_T = \frac{1}{T-1}\sum_{t=1}^{T-1} \frac{1_{A_t=a}}{\pi_t(a)} \nabla \boldsymbol{g}_{Y,t},$$

$$\boldsymbol{C}_T = \frac{1}{T-1}\sum_{t=1}^{T-1} \frac{1_{\{A_t=a\}}}{\pi_t(a)^2}\left(\boldsymbol{g}_{f,t}\boldsymbol{g}_{Y,t}^{\mathsf{T}} + \boldsymbol{g}_{Y,t}\boldsymbol{g}_{f,t}^{\mathsf{T}}\right)$$

$$\boldsymbol{D}_T = \frac{1}{T-1}\sum_{t=1}^{T-1} \frac{1_{\{A_t=a\}}}{\pi_t(a)} \boldsymbol{g}_{f,t}\boldsymbol{g}_{f,t}^{\mathsf{T}},$$

Algorithm 1 PPAI Online Update

Input: Experiment history $\{(\boldsymbol{X}_t, A_t, Y_t)\}_{t=1}^T$, unlabeled contexts $\{\tilde{\boldsymbol{X}}_i\}_{i=1}^{N(T+1)}$, predictive model f, current estimate $\hat{\boldsymbol{\theta}}$, intermediate sums $\boldsymbol{B}, \boldsymbol{C}, \boldsymbol{D}, \boldsymbol{E}, \boldsymbol{F}(\boldsymbol{\theta}), \boldsymbol{H}(\boldsymbol{\theta}), \boldsymbol{I}(\boldsymbol{\theta})$
At time $T+1$, observe $\boldsymbol{X}_{T+1}$, update
$\boldsymbol{B} \leftarrow \boldsymbol{B} + \frac{1_{A_T=a}}{\pi_T(a)} \nabla \boldsymbol{g}_{Y,T}$
$\boldsymbol{C} \leftarrow \boldsymbol{C} + \frac{1_{\{A_T=a\}}}{\pi_T(a)^2}\left(\boldsymbol{g}_{f,T}\boldsymbol{g}_{Y,T}^\top + \boldsymbol{g}_{Y,T}\boldsymbol{g}_{f,T}^\top\right)$
$\boldsymbol{D} \leftarrow \boldsymbol{D} + \frac{1_{A_T=a}}{\pi_T(a)} \boldsymbol{g}_{f,T}^{\otimes 2}$
$\boldsymbol{E} \leftarrow \boldsymbol{E} + \frac{1_{A_T=a}}{\pi_T(a)} \boldsymbol{g}_{f,T}$
$\boldsymbol{F} \leftarrow \boldsymbol{F} + \frac{1_{A_T=a}}{\pi_T(a)^2} \boldsymbol{g}_{f,T}\boldsymbol{g}_{f,T}^\top$
$\hat{\lambda} \leftarrow \frac{\mathrm{tr}(\boldsymbol{B}^{-1}\boldsymbol{C}\boldsymbol{B}^{-\top})}{2\mathrm{tr}(\boldsymbol{B}^{-1}(\frac{T}{N}(\boldsymbol{D} - \frac{1}{T}\boldsymbol{E}\boldsymbol{E}^\top) + \boldsymbol{F} - \frac{1}{T}\boldsymbol{E}\boldsymbol{E}^\top)\boldsymbol{B}^{-\top})}$
$\boldsymbol{F}(\boldsymbol{\theta}) \leftarrow F(\boldsymbol{\theta}) + \frac{1_{A_T=a}}{\pi_T(a)} \boldsymbol{g}(\boldsymbol{X}_T, Y_T; \boldsymbol{\theta})$
$\boldsymbol{H}(\boldsymbol{\theta}) \leftarrow H(\boldsymbol{\theta}) + \sum_{i=N(T)+1}^{N(T+1)} \boldsymbol{g}(\tilde{\boldsymbol{X}}_i, f(\tilde{\boldsymbol{X}}_i); \boldsymbol{\theta})$
$\boldsymbol{I}(\boldsymbol{\theta}) \leftarrow I(\boldsymbol{\theta}) + \frac{1_{A_T=a}}{\pi_T(a)} \boldsymbol{g}(\boldsymbol{X}_T, f(\boldsymbol{X}_T); \boldsymbol{\theta})$
$\boldsymbol{G}(\boldsymbol{\theta}) \leftarrow \frac{1}{T+1}F(\boldsymbol{\theta}) + \hat{\lambda}\left(\frac{1}{N(T+1)}\boldsymbol{H}(\boldsymbol{\theta}) - \frac{1}{T+1}\boldsymbol{I}(\boldsymbol{\theta})\right)$
$\hat{\boldsymbol{\theta}} = \mathrm{root}(\boldsymbol{G}(\boldsymbol{\theta}))$
return $\hat{\boldsymbol{\theta}}$

$$\boldsymbol{E}_T = \frac{1}{T-1}\sum_{t=1}^{T-1} \frac{1_{\{A_t=a\}}}{\pi_t(a)} \boldsymbol{g}_{f,t},$$

$$\boldsymbol{F}_T = \frac{1}{T-1}\sum_{t=1}^{T-1} \frac{1_{A_t=a}}{\pi_t(a)^2} \boldsymbol{g}_{f,t}\boldsymbol{g}_{f,t}^{\mathsf{T}}.$$

Having derived an estimate $\hat{\lambda}$ of the optimal λ^*, we next prove its consistency; then we establish the minimal asymptotic variance of our estimator.

Corollary 3.10. *Let $\lambda_T = \hat{\lambda}_T$ from (12). Then the asymptotic variance of $\widehat{\boldsymbol{\theta}}_T$ from Algorithm 1 has minimal trace among all variances of the form (10).*

This adaptive power-tuning enables our estimator to adapt to the quality of f_a. If f_a is accurate ($f_a(\boldsymbol{X}_t) \approx Y_t(a)$), then $\boldsymbol{g}(\boldsymbol{X}_t, Y_t(a); \boldsymbol{\theta}^*) \approx \boldsymbol{g}(\boldsymbol{X}_t, f_a(\boldsymbol{X}_t); \boldsymbol{\theta}^*)$ and $\lambda^* \approx 1$, so $\boldsymbol{V} \approx 0$. In this case, the asymptotic variance is roughly $r\boldsymbol{J}^{-1}\mathrm{Cov}(\boldsymbol{g}(\boldsymbol{X}, f(\boldsymbol{X}); \boldsymbol{\theta}^*))\boldsymbol{J}^{-\top}$. When we have abundant unlabeled data, r is small, and this variance is small. On the other hand, if f_a is not informative, then $\lambda^* = 0$, in which case the asymptotic variance is $\boldsymbol{J}^{-1}\boldsymbol{V}\boldsymbol{J}^{-\top}$ which coincides with the variance from Theorem 3.2 of Guo & Xu (2025). In summary, if f_a is accurate, our estimator leverages the predictions to reduce variance, and if f_a is totally uninformative, our estimator ignores the predictions and recovers the variance of the labeled-only baseline. In this way, our estimator incorporates the predictions safely.

4. Numerical Results

We simulate our method on a two-armed contextual bandit with reward models

$$Y_t(0) = \exp(0.2\boldsymbol{X}_t^{(1)} - 0.1\boldsymbol{X}_t^{(2)} + 0.4\boldsymbol{X}_t^{(3)}) + e_t,$$

$$Y_t(1) = \exp(0.5\boldsymbol{X}_t^{(1)} + 0.2\boldsymbol{X}_t^{(2)} - 0.1\boldsymbol{X}_t^{(3)}) + e_t,$$

for arms 0 and 1 respectively, where $e_t \sim N(0, \sigma^2)$ with $\sigma = 0.1$. Our inferential target is the least squares parameter $\boldsymbol{\theta}_0^*$ for arm 0, which satisfies $\mathbb{E}[\boldsymbol{X}(Y(0) - \boldsymbol{X}^\top \boldsymbol{\theta}_0^*)] = 0$. The contexts $\boldsymbol{X}_t \sim \mathcal{N}_3(\mathbf{0}, \boldsymbol{I})$ are drawn i.i.d. In addition, 100 i.i.d. unlabeled contexts arrive at each time step, so at time t, the decision-maker has unlabeled contexts $\tilde{\boldsymbol{X}}_1, \ldots, \tilde{\boldsymbol{X}}_{100t} \sim \mathcal{N}_3(\mathbf{0}, \boldsymbol{I})$.

We estimate $\boldsymbol{\theta}_0^*$ with our PPAI estimator $\hat{\boldsymbol{\theta}}_{0,T}$, which leverages the unlabeled dataset and a predictive model f_0. We vary the informativeness of f_0 by simulating with f_0^{poor} and f_0^{strong}, given by $f_0^{\text{poor}}(\boldsymbol{x}) = 10$ and

$$f_0^{\text{strong}}(\boldsymbol{x}) = \exp(0.21\boldsymbol{x}^{(1)} - 0.09\boldsymbol{x}^{(2)} + 0.42\boldsymbol{x}^{(3)}).$$

Note that f_0^{strong} is strongly aligned with the true reward model for arm 0 (i.e. highly informative), whereas f_0^{poor} is poorly aligned (uninformative). For comparison, we also estimate $\boldsymbol{\theta}_0^*$ with the labeled-only baseline estimator, Weighted Least Squares (WLS) (see Chen et al. (2021)).

Actions are chosen according to an ϵ-greedy policy: at each time t, with probability ϵ, we select $A_t \in \{0, 1\}$ randomly, and with probability $1-\epsilon$, we select $A_t = \text{argmax}_a \boldsymbol{X}_t^T \hat{\boldsymbol{\theta}}_{a,t}$, where $\hat{\boldsymbol{\theta}}_{0,t}$ is described above and $\hat{\boldsymbol{\theta}}_{1,t}$ is the WLS estimator for arm 1. Here we set $\epsilon = 0.3$.

Each simulation lasts $T = 5000$ time steps. The first 100 time steps are devoted to random exploration: we randomly assign 50 of the first 100 times to arm 0 and the rest to arm 1 in order to ensure that each arm has enough samples to form the initial estimators. We run 300 simulations in which $\hat{\boldsymbol{\theta}}_{0,T}$ is PPAI with f_0^{strong}, 300 simulations in which $\hat{\boldsymbol{\theta}}_{0,T}$ is PPAI with f_0^{poor}, and 300 simulations in which $\hat{\boldsymbol{\theta}}_{0,T}$ is WLS.

We approximate the true $\boldsymbol{\theta}_0^*$ with the OLS estimator of 10^8 random samples, finding $\boldsymbol{\theta}_0^* \approx (0.22, -0.11, 0.44)$. To illustrate our results, we restrict attention to the first component of $\boldsymbol{\theta}_0^*$. Figure 2 compares the sampling distributions of (the first components of) the PPAI estimator for $\boldsymbol{\theta}_0^*$ with f_0^{strong}, PPAI with f_0^{poor}, and the labeled-only baseline (WLS). Our PPAI estimators reduce variance relative to the labeled-only baseline, and this reduction is largest when PPAI uses f_0^{strong}. We also compute 90% confidence intervals for the first component of $\boldsymbol{\theta}_0^*$ halfway through $(T = 2500)$ and at the end of each experiment $(T = 5000)$. Further implementation details are discussed in Appendix D. Table 1 summarizes the average width and coverage of the confidence intervals.

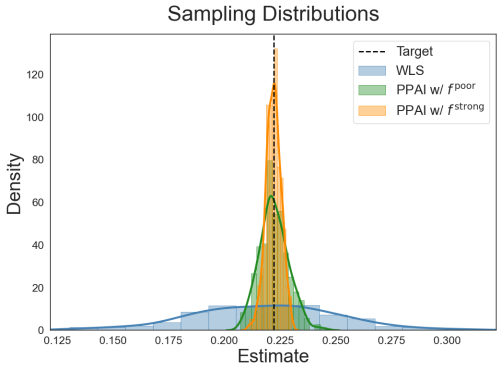

Figure 2. Sampling distributions of PPAI with a strong predictive model, PPAI with a poor predictive model, and the labeled-only baseline (WLS), using an ϵ-greedy policy with $\epsilon = 0.3$. The inferential target is approximately 0.22. PPAI with f^{strong} has the smallest variance, and the labeled-only baseline has the largest.

		$\hat{\boldsymbol{\theta}}$		
		WLS	**PPAI** (f^{poor})	**PPAI** (f^{strong})
T	**2500**	0.161 (92%)	0.030 (88%)	0.016 (91%)
	5000	0.111 (91%)	0.021 (90%)	0.011 (91%)

Table 1. Average width and coverage (in parentheses) of 90% confidence intervals for WLS, PPAI with f^{poor}, and PPAI with f^{strong}. Both PPAI variants yield narrower intervals relative to WLS, and PPAI with f^{strong} achieves a tenfold reduction in width.

Both PPAI variants yield narrower intervals relative to WLS, and PPAI with f_0^{strong} achieves a tenfold reduction in width.

Figure 3 shows the convergence of the power-tuning parameter estimates, summarizing 300 trajectories of $\hat{\lambda}_T$ for each of f^{strong} and f^{poor}. We approximate the oracle power-tuning parameters by approximating each of the expectations in (11) using Monte Carlo simulations with 10^6 samples. We find $\lambda^* \approx 0.988$ for f_0^{strong} and $\lambda^* \approx 0.116$ for f_0^{poor}.

Additional experiments are reported in Appendix D. These results show that PPAI can substantially reduce variance in low-exploration regimes with a strong AI model, that the bias-correction term is necessary for maintaining accuracy, and that adaptive tuning of $\hat{\lambda}_T$ achieves lower variance than a fixed-parameter baseline. We also vary the degree of reward misspecification and find that PPAI maintains robust performance across a range of settings.

5. Real Data Analysis

We evaluate PPAI on a movie recommendation task using *The Movies Dataset*, which contains approximately 26 million ratings from 270,000 users. Gemini is used to predict a user's movie ratings based on their individual rating history.

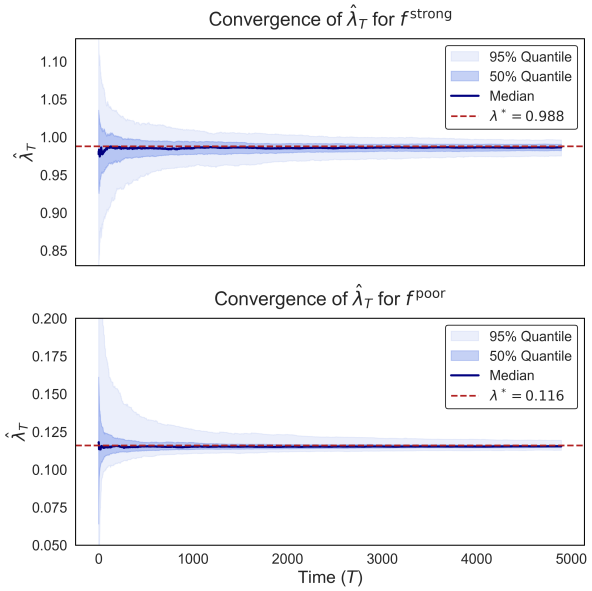

Figure 3. Quantile summary of 300 independent trajectories of $\hat{\lambda}_T$ for each predictive model, f^{strong} and f^{poor}. The oracle parameters are $\lambda^* \approx 0.988$ and $\lambda^* \approx 0.116$, for f^{strong} and f^{poor}, respectively.

5.1. Contexts and Actions

We choose five movies to serve as "representatives": *Independence Day* (1996), *The Lord of the Rings: The Return of the King* (2003), *Star Wars: Episode IV - A New Hope* (1977), *Pulp Fiction* (1994), and *Shrek* (2001). We filter our dataset to include only the 11,198 users who have rated all five of these movies.

Each user's context is defined as a standardized vector $\boldsymbol{x} \in \mathbb{R}^5$ of their ratings for these representative movies. Specifically, if $\boldsymbol{x}_{\text{raw}}$ denotes the vector of raw ratings, then $\boldsymbol{x} = (\boldsymbol{x}_{\text{raw}} - \boldsymbol{\mu})/\boldsymbol{\sigma}$, where $\boldsymbol{\mu} \in \mathbb{R}^5$ and $\boldsymbol{\sigma} \in \mathbb{R}^5$ are the entrywise means and standard deviations.

We let the action set $\mathcal{A} = \{0, 1\}$ consist of two actions, which correspond to recommending *X-Men* (2000) or *Good Will Hunting* (1997), respectively.

5.2. Reward Model

Since we do not observe the rating (i.e., reward) each user would give to every movie, we train a reward model $y : \mathcal{U} \times \mathcal{M} \to \{0.5, 1.0, 1.5, \ldots, 5.0\}$, where $\mathcal{U}$ denotes the set of 11,198 users and $\mathcal{M}$ denotes the set of movies. We treat $y(u, m)$ as the true rating that user $u \in \mathcal{U}$ would assign to movie $m \in \mathcal{M}$. Further details on the reward model are in Appendix E.1.

In our experiments, we query this model for the "true" ratings, though we use the data to infer the parameters $\boldsymbol{\beta}_0, \boldsymbol{\beta}_1$ of a simpler linear reward model $y = 1_{A_t=0}\boldsymbol{\beta}_0^\top \boldsymbol{x} + 1_{A_t=1}\boldsymbol{\beta}_1^\top \boldsymbol{x}$. These parameters represent how ratings of the two movies in the action set vary linearly with ratings of the five representative movies; for instance, the second coordinate of $\boldsymbol{\beta}_0$ characterizes the linear relationship between a user's ratings of *X-Men* (2000) and *The Lord of the Rings: Return of the King* (2003).

5.3. Experiment Setup

We randomly select 10% of the users to be in the labeled dataset and the remaining 90% to be unlabeled. We now simulate the adaptive experiment by applying an ϵ-greedy policy and computing the PPAI least squares estimator (7). To get predicted ratings, we ask Gemini to predict how a user will rate a particular movie given their rating history. For comparison, we repeat the experiment with the Weighted Least Squares estimator (6), which does not use unlabeled data or Gemini predictions. For more details, see Appendix E.2.

5.4. Results

For illustration, we focus on two estimands: the third coordinate of $\boldsymbol{\beta}_0$, which characterizes the linear relationship between a user's rating of *X-Men* (2000) and *Star Wars: Episode IV - A New Hope* (1977), and the first coordinate of $\boldsymbol{\beta}_1$, which characterizes the linear relationship between a user's rating of *Good Will Hunting* (1997) and *Independence Day* (1996). The true values, computed from the simulated ratings of the users in $\mathcal{U}$, are $\boldsymbol{\beta}_0^{(3)} \approx 0.1736$ and $\boldsymbol{\beta}_1^{(1)} \approx 0.0287$.

We construct 90% confidence intervals with both approaches; Table 2 summarizes their average length and coverage, showing that PPAI produces intervals that are approximately 5-7 times shorter than those of WLS.

	WLS	**PPAI**
$\boldsymbol{\beta}_0^{(3)}$	1.24 (83%)	0.21 (95%)
$\boldsymbol{\beta}_1^{(1)}$	1.46 (82%)	0.22 (99%)

Table 2. Average length and coverage for 90% confidence intervals ($T = 1000, N = 10,000$). For both targets, PPAI confidence intervals are 5-7 times tighter.

We also perform hypothesis tests to illustrate the inferential benefits of PPAI, finding that it delivers substantially higher testing power (see Appendix E.3).

Acknowledgments

Gabriel Sargent was partially supported by NSF RTG grant DMS-2134107 and NSF Grant SES-2217440. The research of Will Wei Sun and Yufeng Liu was partially supported by NSF Grant SES-2217440.

Impact Statement

This paper presents work whose goal is to advance the field of Machine Learning. There are many potential societal consequences of our work, none which we feel must be specifically highlighted here.

References

Angelopoulos, A. N., Bates, S., Fannjiang, C., Jordan, M. I., and Zrnic, T. Prediction-powered inference. *Science*, 382(6671):669–674, 2023. doi: 10.1126/science.adi6000. URL https://www.science.org/doi/abs/10.1126/science.adi6000.

Angelopoulos, A. N., Duchi, J. C., and Zrnic, T. PPI++: Efficient prediction-powered inference, 2024. URL https://arxiv.org/abs/2311.01453.

Bian, Z., Shi, C., Qi, Z., and Wang, L. Off-policy evaluation in doubly inhomogeneous environments. *Journal of the American Statistical Association*, 120 (550):1102–1114, 2025. doi: 10.1080/01621459.2024.2395593. URL https://doi.org/10.1080/01621459.2024.2395593.

Chen, H., Lu, W., and Song, R. Statistical inference for online decision making: In a contextual bandit setting. *Journal of the American Statistical Association*, 116(533):240–255, 2021. doi: 10.1080/01621459.2020.1770098. URL https://doi.org/10.1080/01621459.2020.1770098. PMID: 33737759.

Deshpande, Y., Mackey, L., Syrgkanis, V., and Taddy, M. Accurate inference for adaptive linear models. In Dy, J. and Krause, A. (eds.), *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 1194–1203. PMLR, 10–15 Jul 2018. URL https://proceedings.mlr.press/v80/deshpande18a.html.

Guo, Y. and Xu, Z. Statistical inference for misspecified contextual bandits, 2025. URL https://arxiv.org/abs/2509.06287.

Hadad, V., Hirshberg, D. A., Zhan, R., Wager, S., and Athey, S. Confidence intervals for policy evaluation in adaptive experiments. *Proceedings of the National Academy of Sciences*, 118(15):e2014602118, 2021. doi: 10.1073/pnas.2014602118. URL https://www.pnas.org/doi/abs/10.1073/pnas.2014602118.

Hall, P. and Heyde, C. *Martingale Limit Theory and its Application*. Academic Press, 1980.

Han, Q., Sun, W. W., and Zhang, Y. Online statistical inference in decision-making with matrix context. *The Annals of Statistics*, 53(5):1963–1986, 2025.

Ji, W., Lei, L., and Zrnic, T. Predictions as surrogates: Revisiting surrogate outcomes in the age of AI, 2025. URL https://arxiv.org/abs/2501.09731.

Khamaru, K. and Zhang, C.-H. Inference with the upper confidence bound algorithm, 2024. URL https://arxiv.org/abs/2408.04595.

Kluger, D. M., Lu, K., Zrnic, T., Wang, S., and Bates, S. Prediction-powered inference with imputed covariates and nonuniform sampling, 2025. URL https://arxiv.org/abs/2501.18577.

Lattimore, T. and Szepesvári, C. *Bandit algorithms*. Cambridge University Press, 2020.

Li, L., Chu, W., Langford, J., and Schapire, R. E. A contextual-bandit approach to personalized news article recommendation. In *Proceedings of the 19th International Conference on World Wide Web*, WWW '10, pp. 661–670, New York, NY, USA, 2010. Association for Computing Machinery. ISBN 9781605587998. doi: 10.1145/1772690.1772758. URL https://doi.org/10.1145/1772690.1772758.

Liao, P., Greenewald, K., Klasnja, P., and Murphy, S. Personalized heartsteps: A reinforcement learning algorithm for optimizing physical activity. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.*, 4(1), March 2020. doi: 10.1145/3381007. URL https://doi.org/10.1145/3381007.

Liu, Y.-E., Mandel, T., Brunskill, E., and Popovic, Z. Trading off scientific knowledge and user learning with multi-armed bandits. In *Educational Data Mining*, 2014. URL https://api.semanticscholar.org/CorpusID:4103970.

McInerney, J., Lacker, B., Hansen, S., Higley, K., Bouchard, H., Gruson, A., and Mehrotra, R. Explore, exploit, and explain: personalizing explainable recommendations with bandits. In *Proceedings of the 12th ACM Conference on Recommender Systems*, RecSys '18, pp. 31–39, New York, NY, USA, 2018. Association for Computing Machinery. ISBN 9781450359016. doi: 10.1145/3240323.3240354. URL https://doi.org/10.1145/3240323.3240354.

Shaikh, H., Modiri, A., Williams, J. J., and Rafferty, A. N. Balancing student success and inferring personalized effects in dynamic experiments. In *Proceedings of the 12th International Conference on Educational Data Mining (EDM)*, pp. 426–431, 2019.

Shi, C., Zhang, S., Lu, W., and Song, R. Statistical inference of the value function for reinforcement learning in infinite-horizon settings. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 84(3):765–793, 12 2021. ISSN 1369-7412. doi: 10.1111/rssb.12465. URL https://doi.org/10.1111/rssb.12465.

Shi, C., Zhu, J., Shen, Y., Luo, S., Zhu, H., and Song, R. Off-policy confidence interval estimation with confounded Markov decision process. *Journal of the American Statistical Association*, 119(545):273–284, 2024. doi: 10.1080/01621459.2022.2110878. URL https://doi.org/10.1080/01621459.2022.2110878.

Simchi-Levi, D. and Wang, C. Multi-armed bandit experimental design: Online decision-making and adaptive inference. *Management Science*, 71(6):4828–4846, 2025. doi: 10.1287/mnsc.2023.00492. URL https://doi.org/10.1287/mnsc.2023.00492.

Tang, L., Jiang, Y., Li, L., and Li, T. Ensemble contextual bandits for personalized recommendation. In *Proceedings of the 8th ACM Conference on Recommender Systems*, RecSys '14, pp. 73–80, New York, NY, USA, 2014. Association for Computing Machinery. ISBN 9781450326681. doi: 10.1145/2645710.2645732. URL https://doi.org/10.1145/2645710.2645732.

Vaart, A. W. v. d. *Asymptotic Statistics*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 1998.

Varatharajah, Y. and Berry, B. A contextual-bandit-based approach for informed decision-making in clinical trials. *Life*, 12(8):1277, 2022.

Villar, S. S., Wason, J., and Bowden, J. Response-adaptive randomization for multi-arm clinical trials using the forward looking Gittins index rule. *Biometrics*, 71(4):969–978, 2015.

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

Waudby-Smith, I., Wu, L., Ramdas, A., Karampatziakis, N., and Mineiro, P. Anytime-valid off-policy inference for contextual bandits. *ACM / IMS J. Data Sci.*, 1(3), May 2024. doi: 10.1145/3643693. URL https://doi.org/10.1145/3643693.

Wornow, M., Lozano, A., Dash, D., Jindal, J., Mahaffey, K. W., and Shah, N. H. Zero-shot clinical trial patient matching with LLMs. *NEJM AI*, 2(1):AIcs2400360, 2025. doi: 10.1056/AIcs2400360. URL https://ai.nejm.org/doi/full/10.1056/AIcs2400360.

Xu, Z., Witten, D., and Shojaie, A. A unified framework for semiparametrically efficient semi-supervised learning, 2025. URL https://arxiv.org/abs/2502.17741.

Yom-Tov, E., Feraru, G., Kozdoba, M., Mannor, S., Tennenholtz, M., and Hochberg, I. Encouraging physical activity in patients with diabetes: Intervention using a reinforcement learning system. *J Med Internet Res*, 19(10):e338, Oct 2017. ISSN 1438-8871. doi: 10.2196/jmir.7994. URL http://www.jmir.org/2017/10/e338/.

Zhang, K., Janson, L., and Murphy, S. Inference for batched bandits. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., and Lin, H. (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 9818–9829. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/6fd86e0ad726b778e37cf270fa0247d7-Paper.pdf.

Zhang, K., Janson, L., and Murphy, S. Statistical inference after adaptive sampling in non-Markovian environments. 2022.

Zhang, K. W., Janson, L., and Murphy, S. A. Statistical inference with M-estimators on adaptively collected data. In *Proceedings of the 35th International Conference on Neural Information Processing Systems*, NIPS '21, Red Hook, NY, USA, 2021. Curran Associates Inc. ISBN 9781713845393.

Zhang, K. W., Janson, L., and Murphy, S. A. Statistical inference after adaptive sampling for longitudinal data, 2023. URL https://arxiv.org/abs/2202.07098.

Zhu, Z. and Roy, B. V. Scalable neural contextual bandit for recommender systems, 2023. URL https://arxiv.org/abs/2306.14834.

Zrnic, T. and Candès, E. J. Cross-prediction-powered inference. *Proceedings of the National Academy of Sciences*, 121(15):e2322083121, 2024. doi: 10.1073/pnas.2322083121. URL https://www.pnas.org/doi/abs/10.1073/pnas.2322083121.

A. Proofs of Main Results

A.1. Proof of Theorem 3.7

Proof. The proof of this theorem (and its dependent lemmas) mirrors that of Theorem 3.2 from Guo & Xu (2025) but with the modifications required by the presence of unlabeled data, AI predictions, and data-driven weighting in our Z-equation.

Throughout the proof, we let $\mathcal{F}_{t-1} = \mathcal{U}_t \vee \mathcal{H}_{t-1}$ denote the information from the unlabeled data and experimental history available before the context at time t is revealed. For the remainder of this proof, we write $\widehat{\boldsymbol{\theta}}_T$ as $\widehat{\boldsymbol{\theta}}$. Let $\boldsymbol{G}_T^{\lambda_T,(i)}(\boldsymbol{\theta})$ denote the i-th entry of $\boldsymbol{G}_T^{\lambda_T}(\boldsymbol{\theta})$. By Taylor expansion, we have for each $i \in \{1,\dots,d\}$ some $\tilde{\boldsymbol{\theta}}_i$ on the line segment between $\boldsymbol{\theta}^*$ and $\widehat{\boldsymbol{\theta}}$ such that

$$-\boldsymbol{G}_T^{\lambda_T,(i)}(\boldsymbol{\theta}^*) = \boldsymbol{G}_T^{\lambda_T,(i)}(\widehat{\boldsymbol{\theta}}) - \boldsymbol{G}_T^{\lambda_T,(i)}(\boldsymbol{\theta}^*) + o_p(1/\sqrt{T})$$
$$= \langle \nabla \boldsymbol{G}_T^{\lambda_T,(i)}(\boldsymbol{\theta}^*), \widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^* \rangle + \frac{1}{2}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*)^{\mathsf{T}} \nabla^2 \boldsymbol{G}_T^{\lambda_T,(i)}(\tilde{\boldsymbol{\theta}}_i)(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*) + o_p(1/\sqrt{T}).$$

Stacking these expansions over entries $i = 1,\dots,d$ gives

$$-\boldsymbol{G}_T^{\lambda_T}(\boldsymbol{\theta}^*) = \nabla \boldsymbol{G}_T^{\lambda_T}(\boldsymbol{\theta}^*)(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*) + \frac{1}{2}\tilde{\boldsymbol{\delta}}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*) + o_p(1/\sqrt{T}),$$

where

$$\tilde{\boldsymbol{\delta}} = \begin{pmatrix} (\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*)^\top \nabla^2 \boldsymbol{G}_T^{\lambda_T,(1)}(\tilde{\boldsymbol{\theta}}_1) \\ \vdots \\ (\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*)^\top \nabla^2 \boldsymbol{G}_T^{\lambda_T,(d)}(\tilde{\boldsymbol{\theta}}_d) \end{pmatrix}.$$

Rearranging gives

$$[\nabla \boldsymbol{G}_T^{\lambda_T}(\boldsymbol{\theta}^*) + \frac{1}{2}\tilde{\boldsymbol{\delta}}]\sqrt{T}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*) = -\sqrt{T}\boldsymbol{G}_T^{\lambda_T}(\boldsymbol{\theta}^*) + o_p(1). \tag{13}$$

Since $\tilde{\boldsymbol{\theta}}_i$ is on the line segment between $\boldsymbol{\theta}^*$ and $\widehat{\boldsymbol{\theta}}$, we have by Lemmas B.1 and B.3 that for all i,

$$||\nabla^2 \boldsymbol{G}_T^{\lambda_T,(i)}(\tilde{\boldsymbol{\theta}}_i)||_{1,1} \le ||\nabla^2 \boldsymbol{G}_T^{\lambda_T}(\tilde{\boldsymbol{\theta}}_i)||_1$$
$$= ||\nabla^2 \boldsymbol{G}_T^{\lambda_T}(\tilde{\boldsymbol{\theta}}_i)||_1 1_{||\tilde{\boldsymbol{\theta}}_i - \boldsymbol{\theta}^*||_2 \le \epsilon_0} + ||\nabla^2 \boldsymbol{G}_T^{\lambda_T}(\tilde{\boldsymbol{\theta}}_i)||_1 1_{||\tilde{\boldsymbol{\theta}}_i - \boldsymbol{\theta}^*||_2 > \epsilon_0}$$
$$\le \sup_{||\boldsymbol{\theta} - \boldsymbol{\theta}^*|| \le \epsilon_0} ||\nabla^2 \boldsymbol{G}_T^{\lambda_T}(\boldsymbol{\theta})||_1 + ||\nabla^2 \boldsymbol{G}_T^{\lambda_T}(\tilde{\boldsymbol{\theta}}_i)||_1 1_{||\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*||_2 > \epsilon_0}$$
$$= \mathcal{O}_p(1),$$

where for a matrix $\boldsymbol{B} \in \mathbb{R}^{d_1 \times d_2}$, we define $||\boldsymbol{B}||_{1,1} = \sum_{i \in [d_1], j \in [d_2]} |\boldsymbol{B}_{i,j}|$. Note Lemma B.1 gives $\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^* = o_p(1)$, and Lemma B.2 ensures convergence of $\nabla \boldsymbol{G}_T^{\lambda_T}(\boldsymbol{\theta}^*)$. Together they give

$$\nabla \boldsymbol{G}_T^{\lambda_T}(\boldsymbol{\theta}^*) + \frac{1}{2}\tilde{\boldsymbol{\delta}} \xrightarrow{p} \mathbb{E}[\nabla \boldsymbol{g}(\boldsymbol{X}_t, Y_t(a); \boldsymbol{\theta}^*)].$$

Applying this result and Lemma B.4 to (13) gives the desired result. □

A.2. Proof of Theorem 3.8

Proof. Using the linearity of the trace, we can write $\mathrm{tr}(\boldsymbol{\Sigma}_{\lambda^*}) = a\lambda^2 + b\lambda + c$, where

$$a = \mathrm{tr}\left(\boldsymbol{J}^{-1}(r\mathrm{Cov}(\boldsymbol{g}_f^*) + E[\frac{1}{\bar{\pi}(a|\boldsymbol{X}_t)}\boldsymbol{g}_f^*\boldsymbol{g}_f^{*\mathsf{T}}] - E[\boldsymbol{g}_f^*]E[\boldsymbol{g}_f^{*\mathsf{T}}])\boldsymbol{J}^{-\top}\right)$$
$$b = -\mathrm{tr}\left(\boldsymbol{J}^{-1}\mathbb{E}[\frac{1}{\bar{\pi}(a|\boldsymbol{X}_t)}(\boldsymbol{g}_f^*\boldsymbol{g}_Y^{*\mathsf{T}} + \boldsymbol{g}_Y^*\boldsymbol{g}_f^{*\mathsf{T}})]\boldsymbol{J}^{-\top}\right)$$
$$c = \mathrm{tr}\left(\boldsymbol{J}^{-1}\mathbb{E}\left[\frac{\boldsymbol{g}_Y^*\boldsymbol{g}_Y^{*\mathsf{T}}}{\bar{\pi}(a)}\right]\boldsymbol{J}^{-\top}\right).$$

Since the minimizer is $\lambda^* = -\frac{b}{2a}$, the result follows.

□

A.3. Proof of Corollary 3.9

Proof. Let

$$\tilde{\lambda}_T = \frac{\mathrm{tr}(\boldsymbol{B}_T^{-1}\boldsymbol{C}_T\boldsymbol{B}_T^{-1})}{2\mathrm{tr}(\boldsymbol{B}_T^{-1}(\frac{T}{N}(\boldsymbol{D}_T - \boldsymbol{E}_T\boldsymbol{E}_T^\top) + \boldsymbol{F}_T - \boldsymbol{E}_T\boldsymbol{E}_T^\mathsf{T})\boldsymbol{B}_T^{-\top})}$$

be the unprojected ratio; then

$$\hat{\lambda}_T = \Pi_{[-M,M]}(\tilde{\lambda}_T)$$

is its projection onto $[-M, M]$, where M is large enough that $|\lambda^*| < M$. This clipping ensures that $\lambda_T = O_p(1)$, so we can apply Lemma B.1 to get consistency of $\hat{\boldsymbol{\theta}}_T$, which will be needed. It suffices to show that $\tilde{\lambda}_T \xrightarrow{p} \lambda^*$; then by continuity of $\Pi_{[-M,M]}$, the continuous mapping theorem gives

$$\hat{\lambda}_T = \Pi_{[-M,M]}(\tilde{\lambda}_T) \xrightarrow{p} \Pi_{[-M,M]}(\lambda^*) = \lambda^*,$$

as desired. Now note that for consistency of $\tilde{\lambda}_T$, by continuous mapping theorem it suffices to show that

$$\boldsymbol{B}_T \xrightarrow{p} \mathbb{E}[\nabla \boldsymbol{g}(\boldsymbol{X}, Y(a); \boldsymbol{\theta}^*)]$$

$$\boldsymbol{C}_T \xrightarrow{p} \mathbb{E}\left[\frac{1}{\bar{\pi}(a)}(\boldsymbol{g}_f^*\boldsymbol{g}_Y^{*\mathsf{T}} + \boldsymbol{g}_Y^*\boldsymbol{g}_f^{*\mathsf{T}})\right]$$

$$\boldsymbol{D}_T \xrightarrow{p} \mathbb{E}\left[\boldsymbol{g}_f^*\boldsymbol{g}_f^{*\mathsf{T}}\right]$$

$$\boldsymbol{E}_T \xrightarrow{p} \mathbb{E}\left[\boldsymbol{g}_f^*\right]$$

$$\boldsymbol{F}_T \xrightarrow{p} \mathbb{E}\left[\frac{1}{\bar{\pi}(a)}\boldsymbol{g}_f^*\boldsymbol{g}_f^{*\mathsf{T}}\right].$$

We will show convergence of $\boldsymbol{B}_T$ and $\boldsymbol{F}_T$; the arguments for the rest are similar. First note

$$\boldsymbol{B}_T = \frac{1}{T}\sum_{t=1}^{T}\frac{1_{A_t=a}}{\pi_t(a)}\nabla \boldsymbol{g}(\boldsymbol{X}_t, Y_t(a); \boldsymbol{\theta}^*) - \frac{1}{T}\sum_{t=1}^{T}\frac{1_{A_t=a}}{\pi_t(a)}\left(\nabla \boldsymbol{g}(\boldsymbol{X}_t, Y_t(a); \boldsymbol{\theta}^*) - \nabla \boldsymbol{g}(\boldsymbol{X}_t, Y_t(a); \hat{\boldsymbol{\theta}}_t)\right),$$

so for convergence of $\boldsymbol{B}_T$, it suffices to show the following two facts:

$$\frac{1}{T}\sum_{t=1}^{T}\frac{1_{A_t=a}}{\pi_t(a)}\nabla \boldsymbol{g}(\boldsymbol{X}_t, Y_t(a); \boldsymbol{\theta}^*) \xrightarrow{p} \mathbb{E}[\nabla \boldsymbol{g}(\boldsymbol{X}, Y(a); \boldsymbol{\theta}^*)] \tag{14}$$

$$\frac{1}{T}\sum_{t=1}^{T}\frac{1_{A_t=a}}{\pi_t(a)}\left(\nabla \boldsymbol{g}(\boldsymbol{X}_t, Y_t(a); \boldsymbol{\theta}^*) - \nabla \boldsymbol{g}(\boldsymbol{X}_t, Y_t(a); \hat{\boldsymbol{\theta}}_t)\right) \xrightarrow{p} 0. \tag{15}$$

For (14), it suffices to show that for any $\boldsymbol{c}, \boldsymbol{c}' \in \mathbb{R}^d$, $\boldsymbol{c}^\mathsf{T}\frac{1}{T}\sum_{t=1}^{T}\frac{1_{A_t=a}}{\pi_t(a)}\nabla \boldsymbol{g}(\boldsymbol{X}_t, Y_t(a); \boldsymbol{\theta}^*)\boldsymbol{c}' \xrightarrow{p} \boldsymbol{c}^\mathsf{T}\mathbb{E}[\nabla \boldsymbol{g}(\boldsymbol{X}, Y(a); \boldsymbol{\theta}^*)]\boldsymbol{c}'$. To that end, define $Z_t = \boldsymbol{c}^\mathsf{T}\frac{1_{A_t=a}}{\pi_t(a)}\nabla \boldsymbol{g}(\boldsymbol{X}_t, Y_t(a); \boldsymbol{\theta}^*)\boldsymbol{c}'$ and let $Z = \boldsymbol{c}^\mathsf{T}\frac{1}{\pi_{\min}}\nabla \boldsymbol{g}(\boldsymbol{X}, Y(a); \boldsymbol{\theta}^*)\boldsymbol{c}'$. Note $\mathbb{E}|Z| < \infty$ and $\mathbb{P}(|Z_t| > x) \leq \mathbb{P}(|Z| > x)$ for all x. Thus by Theorem 2.19 from Hall & Heyde (1980),

$$\frac{1}{T}\sum_{t=1}^{T}(Z_t - \mathbb{E}[Z_t|\mathcal{F}_{t-1}]) \xrightarrow{p} 0.$$

Since $\mathbb{E}[Z_t|\mathcal{F}_{t-1}] = \boldsymbol{c}^\mathsf{T}\mathbb{E}[\nabla \boldsymbol{g}(\boldsymbol{X}_t, Y_t(a); \boldsymbol{\theta}^*)]\boldsymbol{c}$, this implies

$$\frac{1}{T}\sum_{t=1}^{T}Z_t \xrightarrow{p} \boldsymbol{c}^\mathsf{T}\mathbb{E}[\nabla \boldsymbol{g}(\boldsymbol{X}_t, Y_t(a); \boldsymbol{\theta}^*)]\boldsymbol{c}',$$

establishing (14). For (15), we can modify an argument from Guo & Xu (2025): Observe for all i,

$$\left\|\frac{1}{T}\sum_{t=1}^{T}\frac{1_{A_t=a}}{\pi_t(a)}\left(\nabla \boldsymbol{g}^{(i)}(\boldsymbol{X}_t, Y_t(a); \boldsymbol{\theta}^*) - \nabla \boldsymbol{g}^{(i)}(\boldsymbol{X}_t, Y_t(a); \hat{\boldsymbol{\theta}}_t)\right)\right\|_2$$

$$\leq \frac{1}{T}\sum_{t=1}^{T}\frac{1_{A_t=a}}{\pi_t(a)}\left\|\nabla \boldsymbol{g}^{(i)}(\boldsymbol{X}_t, Y_t(a); \boldsymbol{\theta}^*) - \nabla \boldsymbol{g}^{(i)}(\boldsymbol{X}_t, Y_t(a); \hat{\boldsymbol{\theta}}_t)\right\|_2$$

$$\leq \frac{1}{\pi_{\min} T}\sum_{t=1}^{T} 1_{\|\hat{\boldsymbol{\theta}}_t - \boldsymbol{\theta}^*\|_2 \leq \epsilon_0}\left\|\int_0^1 \nabla^2 \boldsymbol{g}^{(i)}\left(\boldsymbol{X}_t, Y_t(a); \boldsymbol{\theta}^* + u(\hat{\boldsymbol{\theta}}_t - \boldsymbol{\theta}^*)\right) du \cdot (\hat{\boldsymbol{\theta}}_t - \boldsymbol{\theta}^*)\right\|_2$$
$$+ \frac{1}{\pi_{\min} T}\sum_{t=1}^{T} 1_{\|\hat{\boldsymbol{\theta}}_t - \boldsymbol{\theta}^*\|_2 > \epsilon_0}\left\|\nabla \boldsymbol{g}^{(i)}(\boldsymbol{X}_t, Y_t(a); \boldsymbol{\theta}^*) - \nabla \boldsymbol{g}^{(i)}(\boldsymbol{X}_t, Y_t(a); \hat{\boldsymbol{\theta}}_t)\right\|_2$$

$$\leq \frac{1}{\pi_{\min} T}\sum_{t=1}^{T}\Phi(\boldsymbol{X}_t, Y_t(a))||\hat{\boldsymbol{\theta}}_t - \boldsymbol{\theta}^*||_2 1_{\|\hat{\boldsymbol{\theta}}_t - \boldsymbol{\theta}^*\|_2 \leq \epsilon_0} + \frac{2}{\pi_{\min} T}\sum_{t=1}^{T}\phi(\boldsymbol{X}_t, Y_t(a)) 1_{\|\hat{\boldsymbol{\theta}}_t - \boldsymbol{\theta}^*\|_2 > \epsilon_0}$$

$$= o_p(1).$$

To see why this is $o_p(1)$, note first that the summands are $o_p(1)$ by consistency of $\hat{\boldsymbol{\theta}}_t$. Also, the summands are bounded by i.i.d. integrable functions $\epsilon_0\Phi(\boldsymbol{X}_t, Y_t(a))$ and $\phi(\boldsymbol{X}_t, Y_t(a))$, respectively, and are therefore uniformly integrable. Hence the expectations of the summands converge to 0. Cesaro's lemma and Markov's inequality then imply that both summands are $o_p(1)$. This implies (15), establishing convergence of $\boldsymbol{B}_T$. Similarly, to show convergence of $\boldsymbol{F}_T$, it suffices to show

$$\frac{1}{T}\sum_{t=1}^{T}\frac{1_{A_t=a}}{\pi_t(a)^2}\boldsymbol{g}^*_{f,t}\boldsymbol{g}^{*\mathsf{T}}_{f,t} \xrightarrow{p} \mathbb{E}\left[\frac{1}{\bar{\pi}(a)}\boldsymbol{g}^*_f\boldsymbol{g}^{*\mathsf{T}}_f\right] \tag{16}$$

$$\frac{1}{T}\sum_{t=1}^{T}\frac{1_{A_t=a}}{\pi_t(a)^2}\left(\boldsymbol{g}^*_{f,t}\boldsymbol{g}^{*\mathsf{T}}_{f,t} - \hat{\boldsymbol{g}}_{f,t}\hat{\boldsymbol{g}}^{\mathsf{T}}_{f,t}\right) \xrightarrow{p} 0. \tag{17}$$

Note for (16), it suffices to show that $\frac{1}{T}\sum_{t=1}^{T}\boldsymbol{c}^{\mathsf{T}}\frac{1_{A_t=a}}{\pi_t(a)^2}\boldsymbol{g}^*_{f,t}\boldsymbol{g}^{*\mathsf{T}}_{f,t}\boldsymbol{c}' \xrightarrow{p} \boldsymbol{c}^{\mathsf{T}}\mathbb{E}\left[\frac{1}{\bar{\pi}(a)}\boldsymbol{g}^*_f\boldsymbol{g}^{*\mathsf{T}}_f\right]\boldsymbol{c}'$ for any $\boldsymbol{c}, \boldsymbol{c}' \in \mathbb{R}^d$. If $Z_t = \boldsymbol{c}^{\mathsf{T}}\frac{1_{A_t=a}}{\pi_t(a)^2}\boldsymbol{g}^*_{f,t}\boldsymbol{g}^{*\mathsf{T}}_{f,t}\boldsymbol{c}'$, then we need to show

$$\frac{1}{T}\sum_{t=1}^{T}(Z_t - \mathbb{E}[Z_t|\mathcal{F}_{t-1}]) \xrightarrow{p} 0 \tag{18}$$

$$\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}[Z_t|\mathcal{F}_{t-1}] = \frac{1}{T}\sum_{t=1}^{T}\boldsymbol{c}^{\top}\mathbb{E}\left[\frac{1}{\pi_t(a)}\boldsymbol{g}^*_{f,t}\boldsymbol{g}^{*\top}_{f,t}|\mathcal{F}_{t-1}\right]\boldsymbol{c}' \xrightarrow{p} \boldsymbol{c}^{\top}\mathbb{E}\left[\frac{1}{\bar{\pi}(a)}\boldsymbol{g}^*_f\boldsymbol{g}^{*\top}_f\right]\boldsymbol{c}' \tag{19}$$

For (18), let $Z = \frac{||\boldsymbol{c}||_2||\boldsymbol{c}'||_2}{\pi^2_{\min}}||\boldsymbol{g}^*_{f,t}||_2^2$. Note $\mathbb{P}(|Z_t| > x) \leq \mathbb{P}(|Z| > x)$ for all x, and $E|Z| < \infty$ by Assumption 3.5. Thus by Theorem 2.19 from Hall & Heyde (1980),

$$\frac{1}{T}\sum_{t=1}^{T}(Z_t - \mathbb{E}[Z_t|\mathcal{F}_{t-1}]) \xrightarrow{p} 0,$$

establishing (18). For (19), note

$$\mathbb{E}\left|\frac{1}{T}\sum_{t=1}^{T}\boldsymbol{c}^{\top}\mathbb{E}\left[\frac{1}{\pi_t(a)}\boldsymbol{g}^*_{f,t}\boldsymbol{g}^{*\top}_{f,t}|\mathcal{F}_{t-1}\right]\boldsymbol{c}' - \boldsymbol{c}^{\top}\mathbb{E}\left[\frac{1}{\bar{\pi}(a)}\boldsymbol{g}^*_f\boldsymbol{g}^{*\top}_f\right]\boldsymbol{c}'\right|$$

$$\leq \frac{||\boldsymbol{c}||_2||\boldsymbol{c}'||_2\sqrt{||\mathbb{E}||\boldsymbol{g}^*_f||_2^4}}{\pi^2_{\min}T}\sum_{t=1}^{T}\sqrt{\mathbb{E}\left[|\bar{\pi}(a) - \pi_t(a)|^2\right]}$$

$$\xrightarrow{p} 0$$

by Cauchy-Schwarz, policy convergence, and Cesaro's lemma. This implies (19), thus establishing (16). For (17), we can again modify an argument from Guo & Xu (2025): Note

$$\left\| \frac{1}{T} \sum_{t=1}^{T} \frac{1_{A_t=a}}{\pi_t(a)^2} \left(\boldsymbol{g}^*_{f,t} \boldsymbol{g}^{*\mathsf{T}}_{f,t} - \hat{\boldsymbol{g}}_{f,t} \hat{\boldsymbol{g}}^{\mathsf{T}}_{f,t} \right) \right\|_2$$

$$\leq \frac{1}{T} \cdot \frac{1}{\pi^2_{\min}} \sum_{t=1}^{T} \left\| \boldsymbol{g}^*_{f,t} \boldsymbol{g}^{*\mathsf{T}}_{f,t} - \hat{\boldsymbol{g}}_{f,t} \hat{\boldsymbol{g}}^{\mathsf{T}}_{f,t} \right\|_2$$

$$\leq \frac{1}{\pi^2_{\min} T} \sum_{t=1}^{T} \left(||\widehat{\boldsymbol{g}}_{f,t} \widehat{\boldsymbol{g}}^{\mathsf{T}}_{f,t} - \widehat{\boldsymbol{g}}_{f,t} \boldsymbol{g}^{*\mathsf{T}}_{f,t}||_2 + ||\widehat{\boldsymbol{g}}_{f,t} \boldsymbol{g}^{*\mathsf{T}}_{f,t} - \boldsymbol{g}^*_{f,t} \boldsymbol{g}^{*\mathsf{T}}_{f,t}||_2 \right)$$

$$\leq \frac{1}{\pi^2_{\min} T} \sum_{t=1}^{T} ||\widehat{\boldsymbol{g}}_{f,t} - \boldsymbol{g}^*_{f,t}||_2 \cdot (||\widehat{\boldsymbol{g}}_{f,t}||_2 + ||\boldsymbol{g}^*_{f,t}||_2)$$

$$\leq \frac{1}{\pi^2_{\min} T} \sum_{t=1}^{T} ||\widehat{\boldsymbol{g}}_{f,t} - \boldsymbol{g}^*_{f,t}||_2 \cdot (||\widehat{\boldsymbol{g}}_{f,t} - \boldsymbol{g}^*_{f,t}||_2 + 2||\boldsymbol{g}^*_{f,t}||_2)$$

Here,

$$||\widehat{\boldsymbol{g}}_{f,t} - \boldsymbol{g}^*_{f,t}||_2 = \left\| \int_0^1 \nabla \boldsymbol{g} \left(\boldsymbol{X}_t, f_a(\boldsymbol{X}_t); \boldsymbol{\theta}^* + u(\widehat{\boldsymbol{\theta}}_t - \boldsymbol{\theta}^*) \right) du \cdot (\widehat{\boldsymbol{\theta}}_t - \boldsymbol{\theta}^*) \right\|_2$$

$$\leq \phi(\boldsymbol{X}_t, f_a(\boldsymbol{X}_t)) ||\widehat{\boldsymbol{\theta}}_t - \boldsymbol{\theta}^*||_2$$

by Assumption 3.6. Thus,

$$\left\| \frac{1}{T} \sum_{t=1}^{T} \frac{1_{A_t=a}}{\pi_t(a)^2} \left(\boldsymbol{g}^*_{f,t} \boldsymbol{g}^{*\mathsf{T}}_{f,t} - \hat{\boldsymbol{g}}_{f,t} \hat{\boldsymbol{g}}^{\mathsf{T}}_{f,t} \right) \right\|_2$$

$$\leq \frac{1}{\pi^2_{\min} T} \sum_{t=1}^{T} \phi(\boldsymbol{X}_t, f_a(\boldsymbol{X}_t))^2 ||\widehat{\boldsymbol{\theta}}_t - \boldsymbol{\theta}^*||_2^2 + \frac{2}{\pi^2_{\min} T} \sum_{t=1}^{T} \phi(\boldsymbol{X}_t, f_a(\boldsymbol{X}_t)) ||\boldsymbol{g}(\boldsymbol{X}_t, f_a(\boldsymbol{X}_t); \boldsymbol{\theta}^*)||_2 ||\widehat{\boldsymbol{\theta}}_t - \boldsymbol{\theta}^*||_2$$

$$= o_p(1),$$

where convergence in probability follows from Cesaro's lemma and Markov's inequality, as above. This establishes (17), completing the proof of convergence of $\boldsymbol{F}_T$.

□

A.4. Proof of Corollary 3.10

Proof. By Theorem 3.9, $\hat{\lambda}_T \xrightarrow{p} \lambda^*$. Thus if $\lambda_T = \hat{\lambda}_T$, then by Theorem 3.7, the asymptotic variance of $\widehat{\boldsymbol{\theta}}_T$ is $\boldsymbol{\Sigma}_{\lambda^*}$, which has minimal trace by definition of λ^*. □

B. Lemmas

Lemma B.1 (Existence, boundedness, consistency of $\widehat{\boldsymbol{\theta}}_T$)**.** *Under Assumptions 3.2-3.6, there exists a sequence of estimators $\{\widehat{\boldsymbol{\theta}}_T\}_{T \geq 1}$ such that (8) holds, and $||\widehat{\boldsymbol{\theta}}_T||_2 \leq R_\theta$ for all T. Moreover, for any such sequence, as $T \to \infty$, $\widehat{\boldsymbol{\theta}}_T \xrightarrow{p} \boldsymbol{\theta}^*$.*

Proof. We first show that for R_θ from Assumption 3.5, as $T \to \infty$,

$$\sup_{||\boldsymbol{\theta}||_2 \leq R_\theta} ||\boldsymbol{G}_T^{\lambda_T}(\boldsymbol{\theta}) - \mathbb{E}[\boldsymbol{g}(\boldsymbol{X}, Y(a); \boldsymbol{\theta})]||_2 \xrightarrow{p} 0. \tag{20}$$

It suffices to show the coordinate-wise result for each $i = 1, \ldots, d$:

$$\sup_{||\boldsymbol{\theta}||_2 \leq R_\theta} |\boldsymbol{G}_T^{\lambda_T, (i)}(\boldsymbol{\theta}) - \mathbb{E}[\boldsymbol{g}^{(i)}(\boldsymbol{X}, Y(a); \boldsymbol{\theta})]| \xrightarrow{p} 0. \tag{21}$$

Fix $\epsilon > 0$ and let $\Theta_\epsilon = \{\boldsymbol{\theta}_j : j = 1, \ldots, N_\epsilon\}$ be an ϵ-net of the set $\Theta := \overline{\mathcal{B}(\mathbf{0}, R_\theta)}$ with finite cardinality N_ϵ. By Assumption 3.6, we have

$$\begin{aligned}|\boldsymbol{g}^{(i)}(\boldsymbol{X}_t, Y_t(a); \boldsymbol{\theta}) - \boldsymbol{g}^{(i)}(\boldsymbol{X}_t, Y_t(a); \boldsymbol{\theta}_j)| &\leq ||\boldsymbol{g}(\boldsymbol{X}_t, Y_t(a); \boldsymbol{\theta}) - \boldsymbol{g}(\boldsymbol{X}_t, Y_t(a); \boldsymbol{\theta}_j)||_2 \\ &\leq \left|\left|\int_0^1 \nabla \boldsymbol{g}(\boldsymbol{X}_t, Y_t(a); \boldsymbol{\theta}_j + u(\boldsymbol{\theta} - \boldsymbol{\theta}_j))du \cdot (\boldsymbol{\theta} - \boldsymbol{\theta}_j)\right|\right|_2 \\ &\leq \phi(\boldsymbol{X}_t, Y_t(a)) \cdot ||\boldsymbol{\theta} - \boldsymbol{\theta}_j||_2 \\ &\leq \epsilon\phi(\boldsymbol{X}_t, Y_t(a)).\end{aligned}$$

(Here $\boldsymbol{g}^{(i)}(\boldsymbol{X}_t, Y_t(a); \boldsymbol{\theta})$ denotes the ith entry of $\boldsymbol{g}(\boldsymbol{X}_t, Y_t(a); \boldsymbol{\theta})$.) Thus,

$$l_j(\boldsymbol{X}_t, Y_t(a)) \leq \boldsymbol{g}^{(i)}(\boldsymbol{X}_t, Y_t(a); \boldsymbol{\theta}) \leq u_j(\boldsymbol{X}_t, Y_t(a)) \tag{22}$$

where we define $l_j(\boldsymbol{X}_t, Y_t(a)) = \boldsymbol{g}^{(i)}(\boldsymbol{X}_t, Y_t(a); \boldsymbol{\theta}_j) - \epsilon\phi(\boldsymbol{X}_t, Y_t(a))$ and $u_j(\boldsymbol{X}_t, Y_t(a)) = \boldsymbol{g}^{(i)}(\boldsymbol{X}_t, Y_t(a); \boldsymbol{\theta}_j) + \epsilon\phi(\boldsymbol{X}_t, Y_t(a))$. Thus,

$$\begin{aligned}&\sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} |\boldsymbol{G}_T^{\lambda_T,(i)}(\boldsymbol{\theta}) - \mathbb{E}[\boldsymbol{g}^{(i)}(\boldsymbol{X}, Y(a); \boldsymbol{\theta})]| \\ = &\sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \left| \frac{1}{T}\sum_{t=1}^T \frac{1_{\{A_t=a\}}}{\pi_t(a)} \boldsymbol{g}^{(i)}(\boldsymbol{X}_t, Y_t; \boldsymbol{\theta}) + \frac{1}{N}\sum_{i=1}^N \lambda_T \boldsymbol{g}^{(i)}(\tilde{\boldsymbol{X}}_i, f(\tilde{\boldsymbol{X}}_i), \boldsymbol{\theta}) - \frac{1}{T}\sum_{t=1}^T \frac{1_{\{A_t=a\}}}{\pi_t(a)} \lambda_T \boldsymbol{g}^{(i)}(\boldsymbol{X}_t, f(\boldsymbol{X}_t), \boldsymbol{\theta}) \right. \\ &\left. -\mathbb{E}\left[\frac{1_{\{A_t=a\}}}{\pi_t(a)} \boldsymbol{g}^{(i)}(\boldsymbol{X}_t, Y_t; \boldsymbol{\theta}) \mid \mathcal{F}_{t-1}\right]\right| \\ &\leq \sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} |A_T| + |\lambda_T| \sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} |B_N| + |\lambda_T| \sup_{\theta\in\Theta} |C_T|,\end{aligned} \tag{23}$$

where

$$A_T = \frac{1}{T}\sum_{t=1}^T \left(\frac{1_{A_t=a}}{\pi_t(a)} \boldsymbol{g}^{(i)}(\boldsymbol{X}_t, Y_t; \boldsymbol{\theta}) - \mathbb{E}\left[\frac{1_{A_t=a}}{\pi_t(a)} \boldsymbol{g}^{(i)}(\boldsymbol{X}_t, Y_t; \boldsymbol{\theta}) | \mathcal{F}_{t-1}\right] \right)$$

$$B_N = \frac{1}{N}\sum_{i=1}^N \boldsymbol{g}^{(i)}(\tilde{\boldsymbol{X}}_i, f(\tilde{\boldsymbol{X}}_i); \theta) - \mathbb{E}[\boldsymbol{g}^{(i)}(\boldsymbol{X}, f(\boldsymbol{X}); \boldsymbol{\theta})]$$

$$C_T = \frac{1}{T}\sum_{t=1}^T \left(\mathbb{E}\left[\frac{1_{A_t=a}}{\pi_t(a)} \boldsymbol{g}^{(i)}(\boldsymbol{X}_t, f(\boldsymbol{X}_t); \boldsymbol{\theta}) | \mathcal{F}_{t-1}\right] - \frac{1_{A_t=a}}{\pi_t(a)} \boldsymbol{g}^{(i)}(\boldsymbol{X}_t, f(\boldsymbol{X}_t); \boldsymbol{\theta}) \right).$$

Since $\lambda_T = \mathcal{O}_p(1)$, it suffices to show that $\sup_{\theta\in\Theta} |A_T|, \sup_{\theta\in\Theta} |B_N|, \sup_{\theta\in\Theta} |C_T| \xrightarrow{p} 0$. The proof for A_T is provided in Lemma B.1 of Guo & Xu (2025), but we reproduce it here for completeness. Note

$$\begin{aligned}\sup_{\theta\in\Theta} A_T &\leq \max_{k\in[N_\epsilon]} \frac{1}{T}\sum_{t=1}^T \left\{ \frac{1_{A_t=a}}{\pi_t(a)} u_k(\boldsymbol{X}_t, Y_t(a)) - \mathbb{E}\left[\frac{1_{A_t=a}}{\pi_t(a)} l_k(\boldsymbol{X}_t, Y_t(a)) | \mathcal{F}_{t-1}\right] \right\} \\ &\leq \Delta_1 + \Delta_2,\end{aligned} \tag{24}$$

where

$$\Delta_1 = \max_{k\in[N_\epsilon]} \frac{1}{T}\sum_{t=1}^T \left\{ \frac{1_{A_t=a}}{\pi_t(a)} u_k(\boldsymbol{X}_t, Y_t(a)) - \mathbb{E}\left[\frac{1_{A_t=a}}{\pi_t(a)} u_k(\boldsymbol{X}_t, Y_t(a)) | \mathcal{F}_{t-1}\right] \right\}$$

$$\Delta_2 = \max_{k\in[N_\epsilon]} \frac{1}{T}\sum_{t=1}^T \mathbb{E}\left[\frac{1_{A_t=a}}{\pi_t(a)} (u_k(\boldsymbol{X}_t, Y_t(a)) - l_k(\boldsymbol{X}_t, Y_t(a))) | \mathcal{F}_{t-1}\right].$$

We first study Δ_1. Observe

$$|\Delta_1| \leq \sum_{k\in[N_\epsilon]} \left| \frac{1}{T}\sum_{t=1}^T \left\{ \frac{1_{A_t=a}}{\pi_t(a)} u_k(\boldsymbol{X}_t, Y_t(a)) - \mathbb{E}\left[\frac{1_{A_t=a}}{\pi_t(a)} u_k(\boldsymbol{X}_t, Y_t(a)) | \mathcal{F}_{t-1}\right] \right\} \right|.$$

For all $\epsilon' > 0$,

$$\mathbb{P}\left(\left|\frac{1}{T}\sum_{t=1}^{T}\left\{\frac{1_{A_t=a}}{\pi_t(a)}u_k(\boldsymbol{X}_t, Y_t(a)) - \mathbb{E}\left[\frac{1_{A_t=a}}{\pi_t(a)}u_k(\boldsymbol{X}_t, Y_t(a))|\mathcal{F}_{t-1}\right]\right\}\right| > \epsilon'\right)$$

$$\leq \frac{1}{T^2\epsilon'^2}\mathbb{E}\left(\sum_{t=1}^{T}\left\{\frac{1_{A_t=a}}{\pi_t(a)}u_k(\boldsymbol{X}_t, Y_t(a)) - \mathbb{E}\left[\frac{1_{A_t=a}}{\pi_t(a)}u_k(\boldsymbol{X}_t, Y_t(a))|\mathcal{F}_{t-1}\right]\right\}\right)^2$$

$$= \frac{1}{T^2\epsilon'^2}\sum_{t=1}^{T}\mathbb{E}\left(\frac{1_{A_t=a}}{\pi_t(a)}u_k(\boldsymbol{X}_t, Y_t(a)) - \mathbb{E}\left[\frac{1_{A_t=a}}{\pi_t(a)}u_k(\boldsymbol{X}_t, Y_t(a))|\mathcal{F}_{t-1}\right]\right)^2$$

$$\leq \frac{1}{T^2\epsilon'^2}\sum_{t=1}^{T}\mathbb{E}\left(\frac{1_{A_t=a}}{\pi_t(a)}u_k(\boldsymbol{X}_t, Y_t(a))\right)^2$$

$$\leq \frac{1}{T^2\epsilon'^2\pi_{\min}^2}\sum_{t=1}^{T}\mathbb{E}u_k^2(\boldsymbol{X}_t, Y_t(a))$$

$$= \frac{1}{T^2\epsilon'^2\pi_{\min}^2}\sum_{t=1}^{T}\mathbb{E}(\boldsymbol{g}^{(i)}(\boldsymbol{X}_t, Y_t(a); \boldsymbol{\theta}_j) + \epsilon\phi(\boldsymbol{X}_t, Y_t(a)))^2$$

$$\leq \frac{1}{T^2\epsilon'^2\pi_{\min}^2}\cdot T\mathbb{E}[2(\boldsymbol{g}^{(i)}(\boldsymbol{X}_1, Y_1(a); \boldsymbol{\theta}_j))^2 + 2\epsilon^2\phi^2(\boldsymbol{X}_1, Y_1(a))]$$

$$\leq \frac{2M_2' + 2\epsilon^2 M_\phi}{\pi_{\min}^2 T\epsilon'^2} \to 0,$$

where $M_2' = \sup_{||\boldsymbol{\theta}||_2 \leq R_\theta}\mathbb{E}||\boldsymbol{g}(\boldsymbol{X}_t, Y_t(a); \boldsymbol{\theta})||_2^2$, $M_\phi = \mathbb{E}\phi(\boldsymbol{X}_t, Y_t(a))^2$. The first inequality follows from Chebyshev, and the first equality uses the following fact: if $u_{j,t} = \frac{1_{A_t=a}}{\pi_t(a)}u_k(\boldsymbol{X}_t, Y_t(a))$. Then for $t_1 < t_2$,

$$\mathbb{E}[(u_{j,t_1} - \mathbb{E}[u_{j,t_1}|\mathcal{F}_{t_1-1}])(u_{j,t_2} - \mathbb{E}[u_{j,t_2}|\mathcal{F}_{t_2-1}])]$$

$$= \mathbb{E}[\mathbb{E}[(u_{j,t_1} - \mathbb{E}[u_{j,t_1}|\mathcal{F}_{t_1-1}])(u_{j,t_2} - \mathbb{E}[u_{j,t_2}|\mathcal{F}_{t_2-1}])|\mathcal{F}_{t_2-1}]$$

$$= \mathbb{E}[(u_{j,t_1} - \mathbb{E}[u_{j,t_1}|\mathcal{F}_{t_1-1}]) \cdot \mathbb{E}[u_{j,t_2} - \mathbb{E}[u_{j,t_2}|\mathcal{F}_{t_2-1}]|\mathcal{F}_{t_2-1}]$$

$$= \mathbb{E}[(u_{j,t_1} - \mathbb{E}[u_{j,t_1}|\mathcal{F}_{t_1-1}]) \cdot 0] = 0.$$

The penultimate inequality uses that $(a+b)^2 \leq 2(a^2+b^2)$ for any $a, b \in \mathbb{R}$, and the final inequality uses Assumptions 3.5 and 3.6, which implies $M_2' < \infty$ and $M_\phi < \infty$. This establishes that $\Delta_1 \xrightarrow{p} 0$. We now study Δ_2. Note

$$\Delta_2 \leq \max_{j\in[N_\epsilon]}\frac{1}{T}\sum_{t=1}^{T}\frac{1}{\pi_{\min}}\mathbb{E}\left[u_j(\boldsymbol{X}_t, Y_t(a)) - l_j(\boldsymbol{X}_t, Y_t(a))|\mathcal{F}_{t-1}\right]$$

$$\leq \max_{j\in[N_\epsilon]}\frac{1}{T}\sum_{t=1}^{T}\frac{1}{\pi_{\min}}\mathbb{E}\left[2\epsilon\phi(\boldsymbol{X}_t, Y_t(a))|\mathcal{F}_{t-1}\right]$$

$$= \frac{2\epsilon}{\pi_{\min}}\mathbb{E}\phi(\boldsymbol{X}_1, Y_1(a))$$

$$\leq \frac{2\epsilon}{\pi_{\min}}\sqrt{\mathbb{E}\phi(\boldsymbol{X}_1, Y_1(a))^2}$$

$$\leq \frac{2\epsilon\sqrt{M_\phi}}{\pi_{\min}}.$$

Plugging in the analysis of Δ_1 and Δ_2 into (24) gives that for all $\epsilon, \epsilon' > 0$,

$$\mathbb{P}\left(\sup_{\theta\in\Theta}A_T > \frac{2\epsilon\sqrt{M_\phi}}{\pi_{\min}} + \epsilon'\right) \to 0.$$

This implies that for all $\delta > 0$,

$$\mathbb{P}\left(\sup_{\theta\in\Theta} A_T > \delta\right) \to 0.$$

By a similar argument,

$$\mathbb{P}\left(\sup_{\theta\in\Theta} -A_T > \delta\right) \to 0,$$

so

$$\sup_{\theta\in\Theta} |A_T| \xrightarrow{p} 0.$$

Similar reasoning shows $\sup_{\theta\in\Theta} |B_N| \xrightarrow{p} 0$ and $\sup_{\theta\in\Theta} |C_T| \xrightarrow{p} 0$. Thus by (23),

$$\sup_{\boldsymbol{\theta}\in\Theta} |\boldsymbol{G}_T^{\lambda_T,(i)}(\boldsymbol{\theta}) - \mathbb{E}[\boldsymbol{g}^{(i)}(\boldsymbol{X}, Y(a); \boldsymbol{\theta})]| \xrightarrow{p} 0.$$

So (21) holds and therefore (20) holds.

We now focus on the main results. Define

$$\widehat{\boldsymbol{\theta}}_T = \underset{||\boldsymbol{\theta}||_2 \le R_\theta}{\operatorname{argmin}} ||\boldsymbol{G}_T^{\lambda_T}(\boldsymbol{\theta})||_2$$

and

$$\tilde{\boldsymbol{\theta}}_T = \boldsymbol{\theta}^* - [\nabla \mathbb{E}\boldsymbol{g}(\boldsymbol{X}_t, Y_t(a); \boldsymbol{\theta}^*)]^{-1} \boldsymbol{G}_T^{\lambda_T}(\boldsymbol{\theta}^*).$$

Note Lemma B.4 implies

$$\tilde{\boldsymbol{\theta}}_T - \boldsymbol{\theta}^* = O_p(1/\sqrt{T}) \tag{25}$$

Moreover, Taylor expansion gives

$$\begin{aligned}
\boldsymbol{G}_T^{\lambda_T}(\tilde{\boldsymbol{\theta}}_T) &= \boldsymbol{G}_T^{\lambda_T}(\boldsymbol{\theta}^*) + \nabla \boldsymbol{G}_T^{\lambda_T}(\boldsymbol{\theta}^*)(\tilde{\boldsymbol{\theta}}_T - \boldsymbol{\theta}^*) + o_p(||\tilde{\boldsymbol{\theta}}_T - \boldsymbol{\theta}^*||_2) \\
&= \boldsymbol{G}_T^{\lambda_T}(\boldsymbol{\theta}^*) - \nabla \boldsymbol{G}_T^{\lambda_T}(\boldsymbol{\theta}^*)[\nabla \mathbb{E}\boldsymbol{g}(\boldsymbol{X}_t, Y_t(a); \boldsymbol{\theta}^*)]^{-1} \boldsymbol{G}_T^{\lambda_T}(\boldsymbol{\theta}^*) + o_p(1/\sqrt{T}). \\
&= \boldsymbol{G}_T^{\lambda_T}(\boldsymbol{\theta}^*) - (\boldsymbol{I} + o_p(1)) \boldsymbol{G}_T^{\lambda_T}(\boldsymbol{\theta}^*) + o_p(1/\sqrt{T}) \\
&= o_p(1/\sqrt{T}).
\end{aligned} \tag{26}$$

The second equality follows from the definition of $\tilde{\boldsymbol{\theta}}_T$ and (25). The third follows from Lemma B.2 and the fact that, by Assumption 3.6, $\nabla \mathbb{E}\boldsymbol{g}(\boldsymbol{X}_t, Y_t(a); \boldsymbol{\theta}^*) = \frac{\partial}{\partial \boldsymbol{\theta}} \mathbb{E}\boldsymbol{g}(\boldsymbol{X}_t, Y_t(a); \boldsymbol{\theta})_{\boldsymbol{\theta}=\boldsymbol{\theta}^*} = \mathbb{E}\frac{\partial}{\partial \boldsymbol{\theta}} \boldsymbol{g}(\boldsymbol{X}_t, Y_t(a); \boldsymbol{\theta})_{\boldsymbol{\theta}=\boldsymbol{\theta}^*} = \mathbb{E}\nabla \boldsymbol{g}(\boldsymbol{X}_t, Y_t(a); \boldsymbol{\theta}^*)$ is nonsingular. The last equality follows from the fact that $\sqrt{T}\boldsymbol{G}_T^{\lambda_T}(\boldsymbol{\theta}^*) = O_p(1)$.

Combining (25) (which implies $\mathbb{P}(||\tilde{\boldsymbol{\theta}}_T||_2 > R_\theta) \to 0$) and (26), we have

$$\begin{aligned}
||\boldsymbol{G}_T^{\lambda_T}(\widehat{\boldsymbol{\theta}}_T)||_2 &= ||\boldsymbol{G}_T^{\lambda_T}(\widehat{\boldsymbol{\theta}}_T)||_2 1_{\{||\tilde{\boldsymbol{\theta}}_T||_2 \le R_\theta\}} + ||\boldsymbol{G}_T^{\lambda_T}(\widehat{\boldsymbol{\theta}}_T)||_2 1_{\{||\tilde{\boldsymbol{\theta}}_T||_2 > R_\theta\}} \\
&\le ||\boldsymbol{G}_T^{\lambda_T}(\tilde{\boldsymbol{\theta}}_T)||_2 + o_p(1/\sqrt{T}) \\
&= o_p(1/\sqrt{T}).
\end{aligned}$$

Thus this choice of $\hat{\boldsymbol{\theta}}_T$ satisfies (8). This completes the existence proof.

Lastly, we show consistency. Suppose $\widehat{\boldsymbol{\theta}}_T$ satisfies $||\widehat{\boldsymbol{\theta}}_T|| \le R_\theta$ and (8). By Assumption 3.4, for any $\epsilon > 0$,

$$\delta_\epsilon := \inf_{||\boldsymbol{\theta}-\boldsymbol{\theta}^*||_2 > \epsilon} ||\mathbb{E}\boldsymbol{g}(\boldsymbol{X}_t, Y_t(a); \boldsymbol{\theta})||_2 > 0.$$

Combining this with (20) gives

$$\inf_{\boldsymbol{\theta}\in\boldsymbol{\Theta}, ||\boldsymbol{\theta}-\boldsymbol{\theta}^*||_2 > \epsilon} ||\boldsymbol{G}_T^{\lambda_T}(\boldsymbol{\theta})||_2 \ge \inf_{||\boldsymbol{\theta}-\boldsymbol{\theta}^*||_2 > \epsilon} ||\mathbb{E}[\boldsymbol{g}(\boldsymbol{X}, Y(a); \boldsymbol{\theta})]||_2 - \sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} ||\boldsymbol{G}_T^{\lambda_T}(\boldsymbol{\theta}) - \mathbb{E}[\boldsymbol{g}(\boldsymbol{X}, Y(a); \boldsymbol{\theta})]||_2 \ge \delta_\epsilon - o_p(1)$$

where $\Theta = \overline{\mathcal{B}(\mathbf{0}, R_\theta)}$. This implies

$$\lim_{T\to\infty} \mathbb{P}\left(\inf_{\boldsymbol{\theta}\in\boldsymbol{\Theta}, ||\boldsymbol{\theta}-\boldsymbol{\theta}^*||_2>\epsilon} ||\boldsymbol{G}_T^{\lambda_T}(\boldsymbol{\theta})||_2 \leq \frac{1}{2}\delta_\epsilon\right) = 0. \tag{27}$$

We also have, for any $\epsilon' > 0$,

$$\lim_{T\to\infty} \mathbb{P}\left(||\boldsymbol{G}_T^{\lambda_T}(\widehat{\boldsymbol{\theta}}_T)||_2 > \epsilon'/\sqrt{T}\right) = 0.$$

Therefore

$$\lim_{T\to\infty} \mathbb{P}\left(||\boldsymbol{G}_T^{\lambda_T}(\widehat{\boldsymbol{\theta}}_T)||_2 > \frac{1}{2}\delta_\epsilon\right) = 0. \tag{28}$$

Combining (27) and (28) gives

$$\begin{aligned}
&\mathbb{P}(||\widehat{\boldsymbol{\theta}}_T - \boldsymbol{\theta}^*||_2 > \epsilon) \\
&= \mathbb{P}\left(||\widehat{\boldsymbol{\theta}}_T - \boldsymbol{\theta}^*||_2 > \epsilon, ||\boldsymbol{G}_T^{\lambda_T}(\widehat{\boldsymbol{\theta}}_T)||_2 \leq \frac{1}{2}\delta_\epsilon\right) + \mathbb{P}\left(||\widehat{\boldsymbol{\theta}}_T - \boldsymbol{\theta}^*||_2 > \epsilon, ||\boldsymbol{G}_T^{\lambda_T}(\widehat{\boldsymbol{\theta}}_T)||_2 > \frac{1}{2}\delta_\epsilon\right) \\
&\leq \mathbb{P}\left(\inf_{\boldsymbol{\theta}\in\boldsymbol{\Theta}, ||\boldsymbol{\theta}-\boldsymbol{\theta}^*||_2>\epsilon} ||\boldsymbol{G}_T^{\lambda_T}(\boldsymbol{\theta})||_2 \leq \frac{1}{2}\delta_\epsilon\right) + \mathbb{P}\left(||\boldsymbol{G}_T^{\lambda_T}(\widehat{\boldsymbol{\theta}}_T)||_2 > \frac{1}{2}\delta_\epsilon\right) \\
&\to 0
\end{aligned}$$

as $T \to \infty$. This establishes consistency of $\widehat{\boldsymbol{\theta}}_T$. □

Lemma B.2. *Under the assumptions of Theorem 3.7, as* $T \to \infty, \nabla \boldsymbol{G}_T^{\lambda_T}(\boldsymbol{\theta}^*) \xrightarrow{p} \mathbb{E}\nabla \boldsymbol{g}(\boldsymbol{X}_t, Y_t(a); \boldsymbol{\theta}^*)$.

Proof. It suffices to show that $\boldsymbol{c}^\mathsf{T}\nabla \boldsymbol{G}_T^{\lambda_T}(\boldsymbol{\theta}^*)\boldsymbol{c}' \xrightarrow{p} \boldsymbol{c}^\mathsf{T}\mathbb{E}\nabla \boldsymbol{g}(\boldsymbol{X}_t, Y_t(a); \boldsymbol{\theta}^*)\boldsymbol{c}'$ for any nonrandom vectors $\boldsymbol{c}, \boldsymbol{c}' \in \mathbb{R}^d$. To that end, fix $\boldsymbol{c}, \boldsymbol{c}' \in \mathbb{R}^d$ and note

$$\boldsymbol{c}^\mathsf{T}\nabla \boldsymbol{G}_T^{\lambda_T}(\boldsymbol{\theta}^*)\boldsymbol{c}' = A_T + \lambda_T B_T - \lambda_T C_T, \tag{29}$$

where

$$A_T = \frac{1}{T}\sum_{t=1}^{T} \frac{1_{\{A_t=a\}}}{\pi_t(a)} \boldsymbol{c}^\mathsf{T}\nabla \boldsymbol{g}(\boldsymbol{X}_t, Y_t; \boldsymbol{\theta}^*)\boldsymbol{c}'$$

$$B_T = \frac{1}{N}\sum_{i=1}^{N} \boldsymbol{c}^\mathsf{T}\nabla \boldsymbol{g}(\tilde{\boldsymbol{X}}_i, f(\tilde{\boldsymbol{X}}_i); \boldsymbol{\theta}^*)\boldsymbol{c}'$$

$$C_T = \frac{1}{T}\sum_{t=1}^{T} \frac{1_{\{A_t=a\}}}{\pi_t(a)} \boldsymbol{c}^\mathsf{T}\nabla \boldsymbol{g}(\boldsymbol{X}_t, f(\boldsymbol{X}_t); \boldsymbol{\theta}^*)\boldsymbol{c}'.$$

We first show that $A_T \xrightarrow{p} \boldsymbol{c}^\mathsf{T}\mathbb{E}\nabla \boldsymbol{g}(\boldsymbol{X}_t, Y_t(a); \boldsymbol{\theta}^*)\boldsymbol{c}'$. Write $Z_t = \frac{1_{\{A_t=a\}}}{\pi_t(a)} \boldsymbol{c}^\mathsf{T}\nabla \boldsymbol{g}(\boldsymbol{X}_t, Y_t; \boldsymbol{\theta}^*)\boldsymbol{c}'$ so that $A_T = \sum_{t=1}^{T} Z_t$. Let $Z = \frac{1}{\pi_{\min}} \boldsymbol{c}^\mathsf{T}\nabla \boldsymbol{g}(\boldsymbol{X}, Y(a); \boldsymbol{\theta}^*)\boldsymbol{c}'$, and note for all x, $\mathbb{P}(|Z_t| \geq x) \leq \mathbb{P}(|Z| \geq x)$ and $\mathbb{E}|Z| < \infty$. To see integrability of Z, note by our assumptions,

$$||\nabla \boldsymbol{g}(\boldsymbol{X}_t, Y_t(a); \boldsymbol{\theta}^*)||_2 \leq \sup_{||\boldsymbol{\theta}|| \leq R_\theta} ||\nabla \boldsymbol{g}(\boldsymbol{X}_t, Y_t(a); \boldsymbol{\theta})||_2 \leq \phi(\boldsymbol{X}_t, Y_t(a)). \tag{30}$$

Our assumptions also imply that $\mathbb{E}[\phi(\boldsymbol{X}_t, Y_t(a))] < \infty$, so taking the expectation of (30) gives $\mathbb{E}||\nabla \boldsymbol{g}(\boldsymbol{X}_t, Y_t(a); \boldsymbol{\theta}^*||_2 < \infty$. Thus,

$$\mathbb{E}|Z| \leq \frac{1}{\pi_{\min}} ||c||_2 ||c'||_2 E||\nabla \boldsymbol{g}(\boldsymbol{X}_t, Y_t(a); \boldsymbol{\theta}^*||_2 < \infty.$$

We can now apply Theorem 2.19 from Hall & Heyde (1980) to get

$$\frac{1}{T}\sum_{t=1}^{T} (Z_t - \mathbb{E}[Z_t|\mathcal{F}_{t-1}]) \xrightarrow{p} 0.$$

Since $\mathbb{E}[Z_t|\mathcal{F}_{t-1}] = \boldsymbol{c}^\mathsf{T}\mathbb{E}[\nabla \boldsymbol{g}(\boldsymbol{X}_t, Y_t(a); \boldsymbol{\theta}^*)]\boldsymbol{c}'$, it follows that

$$A_T \xrightarrow{p} \boldsymbol{c}^\mathsf{T}\mathbb{E}\nabla[\boldsymbol{g}(\boldsymbol{X}, Y(a); \boldsymbol{\theta}^*)]\boldsymbol{c}'.$$

A similar argument shows that $C_T \xrightarrow{p} \boldsymbol{c}^\mathsf{T}\mathbb{E}\nabla \boldsymbol{g}(\boldsymbol{X}, f(\boldsymbol{X}); \boldsymbol{\theta}^*)\boldsymbol{c}'$. And by the Law of Large Numbers, $B_T \xrightarrow{p} \boldsymbol{c}^\mathsf{T}\mathbb{E}\nabla \boldsymbol{g}(\boldsymbol{X}, f(\boldsymbol{X}); \boldsymbol{\theta}^*)\boldsymbol{c}'$ as well. Taking the limit of (29), the result follows.

□

Lemma B.3. *Under the assumptions of Theorem 3.7,* $\sup_{||\boldsymbol{\theta}-\boldsymbol{\theta}^*||_2 \le \epsilon_0} ||\nabla^2 \boldsymbol{G}_T^{\lambda_T}(\boldsymbol{\theta})||_1 = \mathcal{O}_p(1)$, *where the tensor norm is defined as* $||\boldsymbol{B}||_1 = \sum_{i\in[d_1], j\in[d_2], k\in[d_3]} |\boldsymbol{B}_{i,j,k}|$ *for* $\boldsymbol{B} \in \mathbb{R}^{d_1\times d_2\times d_3}$.

Proof. Note that

$$\nabla^2 \boldsymbol{G}_T^{\lambda_T}(\boldsymbol{\theta}) = \frac{1}{T}\sum_{t=1}^{T} \frac{1_{\{A_t=a\}}}{\pi_t(a)} \nabla^2 \boldsymbol{g}(\boldsymbol{X}_t, Y_t; \boldsymbol{\theta}) + \frac{1}{N}\sum_{i=1}^{N} \lambda_T \nabla^2 \boldsymbol{g}(\tilde{\boldsymbol{X}}_i, f(\tilde{\boldsymbol{X}}_i), \boldsymbol{\theta}) - \frac{1}{T}\sum_{t=1}^{T} \frac{1_{\{A_t=a\}}}{\pi_t(a)} \lambda_T \nabla^2 \boldsymbol{g}(\boldsymbol{X}_t, f(\boldsymbol{X}_t), \boldsymbol{\theta})).$$

By triangle inequality,

$$||\nabla^2 \boldsymbol{G}_T^{\lambda}(\boldsymbol{\theta})||_1 \le \frac{1}{\pi_{\min}} A_T + |\lambda_T| B_T + \frac{1}{\pi_{\min}} |\lambda_T| C_T$$

where

$$A_T(\boldsymbol{\theta}) = \frac{1}{T}\sum_{t=1}^{T} ||\nabla^2 \boldsymbol{g}(\boldsymbol{X}_t, Y_t(a); \boldsymbol{\theta})||_1$$

$$B_T(\boldsymbol{\theta}) = \frac{1}{N}\sum_{i=1}^{N} ||\nabla^2 \boldsymbol{g}(\tilde{\boldsymbol{X}}_i, f(\tilde{\boldsymbol{X}}_i); \boldsymbol{\theta})||_1$$

$$C_T(\boldsymbol{\theta}) = \frac{1}{T}\sum_{t=1}^{T} ||\nabla^2 \boldsymbol{g}(\boldsymbol{X}_t, f(\boldsymbol{X}_t); \boldsymbol{\theta})||_1.$$

Note λ_T is tight, so to establish tightness of $\sup_{||\boldsymbol{\theta}-\boldsymbol{\theta}^*||_2 \le \epsilon_0} ||\nabla^2 \boldsymbol{G}_T^{\lambda}(\boldsymbol{\theta})||_1$, it suffices to show tightness of $\sup_{||\boldsymbol{\theta}-\boldsymbol{\theta}^*||_2 \le \epsilon_0} A_T(\theta)$, $\sup_{||\boldsymbol{\theta}-\boldsymbol{\theta}^*||_2 \le \epsilon_0} B_T(\theta)$, and $\sup_{||\boldsymbol{\theta}-\boldsymbol{\theta}^*||_2 \le \epsilon_0} C_T(\theta)$. Assumption 3.6 implies that for all $\boldsymbol{\theta} \in \overline{\mathcal{B}(\boldsymbol{\theta}^*, \epsilon_0)}$,

$$||A_T(\boldsymbol{\theta})||_1 \le \frac{1}{T}\sum_{t=1}^{T} d^{5/2} \sup_{i\in[d]} ||\nabla^2 \boldsymbol{g}^{(i)}(\boldsymbol{X}_t, Y_t(a); \boldsymbol{\theta})||_2 \le \frac{d^{5/2}}{T}\sum_{t=1}^{T} \Phi(\boldsymbol{X}_t, Y_t(a)).$$

Thus

$$\sup_{||\boldsymbol{\theta}-\boldsymbol{\theta}^*||_2 \le \epsilon_0} ||A_T(\boldsymbol{\theta})||_1 \le d^{5/2} \cdot \frac{1}{T}\sum_{t=1}^{T} \Phi(\boldsymbol{X}_t, Y_t(a)) = \mathcal{O}_p(1).$$

A similar argument shows

$$\sup_{||\boldsymbol{\theta}-\boldsymbol{\theta}^*||_2 \le \epsilon_0} B_T(\boldsymbol{\theta}) = \mathcal{O}_p(1)$$

$$\sup_{||\boldsymbol{\theta}-\boldsymbol{\theta}^*||_2 \le \epsilon_0} C_T(\boldsymbol{\theta}) = \mathcal{O}_p(1),$$

as desired.

□

Lemma B.4. *Under the assumptions of Theorem 3.7,* $\sqrt{T}\boldsymbol{G}_T^{\lambda_T}(\boldsymbol{\theta}^*) \xrightarrow{d} \mathcal{N}(\boldsymbol{0}, \overline{\lambda}^2 r\mathrm{Cov}(\boldsymbol{g}(\boldsymbol{X}, f(\boldsymbol{X}); \boldsymbol{\theta}^*)) + \boldsymbol{V})$ *as* $T \to \infty$.

Proof. Note that $\sqrt{T}\boldsymbol{G}_T^{\lambda_T}(\boldsymbol{\theta}^*) - \sqrt{T}\boldsymbol{G}_T^{\overline{\lambda}}(\boldsymbol{\theta}^*) = o_p(1)$, so it suffices to show that

$$\sqrt{T}\boldsymbol{G}_T^{\overline{\lambda}}(\boldsymbol{\theta}^*) \xrightarrow{d} \mathcal{N}(\mathbf{0}, \overline{\lambda}^2 r\mathrm{Cov}(\boldsymbol{g}(\boldsymbol{X}, f(\boldsymbol{X}); \boldsymbol{\theta}^*)) + \boldsymbol{V}). \tag{31}$$

To that end, note that the scaled estimating equation $\sqrt{T}\boldsymbol{G}_T^{\overline{\lambda}}(\boldsymbol{\theta}^*)$ can be expressed as a sum of terms of the form

$$\bar{\lambda}\frac{\sqrt{T}}{N}\left(\boldsymbol{g}(\tilde{\boldsymbol{X}}_i, f(\tilde{\boldsymbol{X}}_i); \boldsymbol{\theta}^*) - \mathbb{E}\left[\boldsymbol{g}(\boldsymbol{X}, f(\boldsymbol{X}); \boldsymbol{\theta}^*)\right]\right)$$

$$\frac{1}{\sqrt{T}}\left(\frac{1_{A_t=a}}{\pi_t(a)}\left(\boldsymbol{g}(\boldsymbol{X}_t, Y_t(a); \boldsymbol{\theta}^*) - \bar{\lambda}\boldsymbol{g}(\boldsymbol{X}_t, f(\boldsymbol{X}_t); \boldsymbol{\theta}^*)\right) + \bar{\lambda}\mathbb{E}[\boldsymbol{g}(\boldsymbol{X}, f(\boldsymbol{X}); \boldsymbol{\theta}^*)]\right),$$

summed in the order in which they appear in the experiment. Let N_t be the number of unlabeled data points available at time t, with $N_0 = 0$. We now recursively define a σ-field summarizing the information available from the labeled and unlabeled data. For each row $T \geq 1$, we define $\mathcal{G}_{T,k}$ to be the information in the first k data points (labeled or unlabeled). Let $\mathcal{G}_{T,0}$ be the trivial σ-algebra. For each $t = 1, \ldots, T$ define

$$\mathcal{G}_{T,i+t-1} = \mathcal{G}_{T,i+t-2} \vee \sigma(\tilde{\boldsymbol{X}}_i), \qquad i = N_{t-1} + 1, \ldots, N_t.$$

Then define

$$\mathcal{G}_{T,N_t+t} = \mathcal{G}_{T,N_t+t-1} \vee \sigma(\boldsymbol{X}_t, A_t, Y_t).$$

In particular, the information available before the tth labeled data point is simply $\mathcal{G}_{T,N_t+t-1} = \mathcal{F}_{t-1}$.

By Cramer-Wold, it suffices to show that for any $\boldsymbol{c} \in \mathbb{R}^d$, $\boldsymbol{c}^\mathsf{T}\sqrt{T}\boldsymbol{G}_T^{\overline{\lambda}}(\boldsymbol{\theta}^*) \xrightarrow{d} \mathcal{N}(\mathbf{0}, \boldsymbol{c}^\mathsf{T}(\overline{\lambda}^2 r\mathrm{Cov}(\boldsymbol{g}(\boldsymbol{X}, f(\boldsymbol{X}); \boldsymbol{\theta}^*)) + \boldsymbol{V})\boldsymbol{c})$. We now verify the conditions of Theorem 3.2 from Hall & Heyde (1980). We first need to show that the summands form a martingale difference sequence, i.e. the conditional expectation of the kth increment given $\mathcal{G}_{T,k-1}$ equals 0. If the kth increment corresponds to an unlabeled term, then since $\tilde{\boldsymbol{X}}_i$ is independent of previously observed data, its conditional expectation equals its unconditional expectation, giving

$$\mathbb{E}\left[\boldsymbol{c}^\top\bar{\lambda}\frac{\sqrt{T}}{N}\left(\boldsymbol{g}(\tilde{\boldsymbol{X}}_i, f(\tilde{\boldsymbol{X}}_i); \boldsymbol{\theta}^*) - \mathbb{E}\left[\boldsymbol{g}(\boldsymbol{X}, f(\boldsymbol{X}); \boldsymbol{\theta}^*)\right]\right)\right] = 0,$$

as desired. Otherwise, if the summand corresponds to the tth labeled data point, then its conditional expectation is

$$\mathbb{E}\left[\boldsymbol{c}^\mathsf{T}\frac{1}{\sqrt{T}}\left(\frac{1_{\{A_t=a\}}}{\pi_t(a)}\left[\boldsymbol{g}(\boldsymbol{X}_t, Y_t(a); \boldsymbol{\theta}^*) - \overline{\lambda}\boldsymbol{g}(\boldsymbol{X}_t, f(\boldsymbol{X}_t); \boldsymbol{\theta}^*)\right] + \overline{\lambda}\mathbb{E}[\boldsymbol{g}(\boldsymbol{X}, f(\boldsymbol{X}); \boldsymbol{\theta}^*)]|\mathcal{F}_{t-1}\right]\right)$$

$$= \frac{1}{\sqrt{T}}\boldsymbol{c}^\mathsf{T}\left(\mathbb{E}\left[\frac{1_{\{A_t=a\}}}{\pi_t(a)}\left[\boldsymbol{g}(\boldsymbol{X}_t, Y_t(a); \boldsymbol{\theta}^*) - \overline{\lambda}\boldsymbol{g}(\boldsymbol{X}_t, f(\boldsymbol{X}_t); \boldsymbol{\theta}^*)\right]|\mathcal{F}_{t-1}\right] + \overline{\lambda}\mathbb{E}[\boldsymbol{g}(\boldsymbol{X}, f(\boldsymbol{X}); \boldsymbol{\theta}^*)]\right)$$

$$= \frac{1}{\sqrt{T}}\boldsymbol{c}^\mathsf{T}\left(\mathbb{E}\left[\mathbb{E}\left[\frac{1_{\{A_t=a\}}}{\pi_t(a)}\left[\boldsymbol{g}(\boldsymbol{X}_t, Y_t(a); \boldsymbol{\theta}^*) - \overline{\lambda}\boldsymbol{g}(\boldsymbol{X}_t, f(\boldsymbol{X}_t); \boldsymbol{\theta}^*)\right]|\mathcal{F}_{t-1}, \boldsymbol{X}_t, Y_t(a)\right]|\mathcal{F}_{t-1}\right] + \overline{\lambda}\mathbb{E}[\boldsymbol{g}(\boldsymbol{X}, f(\boldsymbol{X}); \boldsymbol{\theta}^*)]\right)$$

$$= \frac{1}{\sqrt{T}}\boldsymbol{c}^\mathsf{T}\left(\mathbb{E}\left[\frac{1}{\pi_t(a)}\mathbb{E}\left[1_{\{A_t=a\}}|\mathcal{F}_{t-1}, \boldsymbol{X}_t\right]\left[\boldsymbol{g}(\boldsymbol{X}_t, Y_t(a); \boldsymbol{\theta}^*) - \overline{\lambda}\boldsymbol{g}(\boldsymbol{X}_t, f(\boldsymbol{X}_t); \boldsymbol{\theta}^*)\right]|\mathcal{F}_{t-1}\right] + \overline{\lambda}\mathbb{E}[\boldsymbol{g}(\boldsymbol{X}, f(\boldsymbol{X}); \boldsymbol{\theta}^*)]\right)$$

$$= \frac{1}{\sqrt{T}}\boldsymbol{c}^\mathsf{T}\left(\mathbb{E}\left[\boldsymbol{g}(\boldsymbol{X}_t, Y_t(a); \boldsymbol{\theta}^*) - \overline{\lambda}\boldsymbol{g}(\boldsymbol{X}_t, f(\boldsymbol{X}_t); \boldsymbol{\theta}^*)|\mathcal{F}_{t-1}\right] + \overline{\lambda}\mathbb{E}[\boldsymbol{g}(\boldsymbol{X}, f(\boldsymbol{X}); \boldsymbol{\theta}^*)]\right)$$

$$= \frac{1}{\sqrt{T}}\boldsymbol{c}^\mathsf{T}\left(\mathbb{E}[\boldsymbol{g}(\boldsymbol{X}_t, Y_t(a); \boldsymbol{\theta}^*)] - \overline{\lambda}\mathbb{E}[\boldsymbol{g}(\boldsymbol{X}, f(\boldsymbol{X}); \boldsymbol{\theta}^*)] + \overline{\lambda}\mathbb{E}[\boldsymbol{g}(\boldsymbol{X}, f(\boldsymbol{X}); \boldsymbol{\theta}^*)]\right)$$

$$= \frac{1}{\sqrt{T}}\boldsymbol{c}^\mathsf{T}\mathbb{E}[\boldsymbol{g}(\boldsymbol{X}_t, Y_t(a); \boldsymbol{\theta}^*)]$$

$$= 0.$$

The second equality uses the tower property, the third uses that $\boldsymbol{g}(\boldsymbol{X}_t, Y_t(a); \boldsymbol{\theta}^*) - \overline{\lambda}\boldsymbol{g}(\boldsymbol{X}_t, f(\boldsymbol{X}_t); \boldsymbol{\theta}^*)$ is measurable with respect to $\sigma(\boldsymbol{X}_t, Y_t(a))$, the fourth uses unconfoundedness and that $\pi_t(a) := \mathbb{E}[1_{\{A_t=a\}}|\mathcal{F}_{t-1}, \boldsymbol{X}_t]$, and the last uses that $(\boldsymbol{X}_t, Y_t(a))$ is independent of the observed information $\mathcal{F}_{t-1}$.

Secondly, we must show that the sum of conditional variances of the kth increment given $\mathcal{G}_{T,k-1}$ converges in probability to $\boldsymbol{c}^\top(\bar{\lambda}^2 r\mathrm{Cov}(\boldsymbol{g}(\boldsymbol{X}, f(\boldsymbol{X}); \boldsymbol{\theta}^*)) + \boldsymbol{V})\boldsymbol{c}$. Again by independence, the conditional variance of an unlabeled increment will equal the unconditional variance:

$$\mathrm{Var}\big(\boldsymbol{c}^\top \bar{\lambda}\frac{\sqrt{T}}{N}\boldsymbol{g}(\tilde{\boldsymbol{X}}_i, f(\tilde{\boldsymbol{X}}_i); \boldsymbol{\theta}^*)\big)$$

$$= \bar{\lambda}^2 \frac{T}{N^2}\boldsymbol{c}^\top \mathrm{Cov}(\boldsymbol{g}(\boldsymbol{X}, f(\boldsymbol{X}); \boldsymbol{\theta}^*))\boldsymbol{c}.$$

Thus the sum of conditional variances of all N unlabeled terms is

$$\boldsymbol{c}^\top \bar{\lambda}^2 \frac{T}{N}\mathrm{Cov}(\boldsymbol{g}(\boldsymbol{X}, f(\boldsymbol{X}); \boldsymbol{\theta}^*))\boldsymbol{c} \xrightarrow{p} \boldsymbol{c}^\top \bar{\lambda}^2 r\mathrm{Cov}(\boldsymbol{g}(\boldsymbol{X}, f(\boldsymbol{X}); \boldsymbol{\theta}^*))\boldsymbol{c}. \tag{32}$$

We now compute the sum of conditional variances of labeled increments:

$$\sum_{t=1}^{T} \mathrm{Var}\left(\boldsymbol{c}^\mathsf{T}\frac{1}{\sqrt{T}}\left[\frac{1_{\{A_t=a\}}}{\pi_t(a)}\big[\boldsymbol{g}(\boldsymbol{X}_t, Y_t(a); \boldsymbol{\theta}^*) - \bar{\lambda}\boldsymbol{g}(\boldsymbol{X}_t, f(\boldsymbol{X}_t); \boldsymbol{\theta}^*)\big] + \bar{\lambda}\mathbb{E}[\boldsymbol{g}(\boldsymbol{X}, f(\boldsymbol{X}); \boldsymbol{\theta}^*)]\right]|\mathcal{F}_{t-1}\right).$$

Since the expression inside the variance has mean zero, we can replace the variance with the expectation of the square. Expanding the square gives

$$\frac{1}{T}\sum_{t=1}^{T}\Bigg(\mathbb{E}\left[\frac{1_{\{A_t=a\}}}{\pi_t(a)^2}(\boldsymbol{c}^\mathsf{T}(\boldsymbol{g}(\boldsymbol{X}_t, Y_t; \boldsymbol{\theta}^*) - \bar{\lambda}\boldsymbol{g}(\boldsymbol{X}_t, f(\boldsymbol{X}_t); \boldsymbol{\theta}^*)))^2|\mathcal{F}_{t-1}\right]$$
$$+2\bar{\lambda}\boldsymbol{c}^\mathsf{T}\mathbb{E}[\boldsymbol{g}(\boldsymbol{X}, f(\boldsymbol{X}); \boldsymbol{\theta}^*)]\boldsymbol{c}^\mathsf{T}\mathbb{E}\left[\frac{1_{\{A_t=a\}}}{\pi_t(a)}(\boldsymbol{g}(\boldsymbol{X}_t, Y_t; \boldsymbol{\theta}^*) - \bar{\lambda}\boldsymbol{g}(\boldsymbol{X}_t, f(\boldsymbol{X}_t); \boldsymbol{\theta}^*)|\mathcal{F}_{t-1}\right] + (\bar{\lambda}\boldsymbol{c}^\mathsf{T}\mathbb{E}[\boldsymbol{g}(\boldsymbol{X}, f(\boldsymbol{X}); \boldsymbol{\theta}^*)])^2\Bigg)$$

$$= \frac{1}{T}\sum_{t=1}^{T}\left(\mathbb{E}\left[\frac{1_{\{A_t=a\}}}{\pi_t(a)^2}(\boldsymbol{c}^\mathsf{T}(\boldsymbol{g}(\boldsymbol{X}_t, Y_t; \boldsymbol{\theta}^*) - \bar{\lambda}\boldsymbol{g}(\boldsymbol{X}_t, f(\boldsymbol{X}_t); \boldsymbol{\theta}^*)))^2|\mathcal{F}_{t-1}\right] - \bar{\lambda}^2\boldsymbol{c}^\mathsf{T}\mathbb{E}[g(\boldsymbol{X}, f(\boldsymbol{X}); \boldsymbol{\theta}^*)]\mathbb{E}[g(\boldsymbol{X}, f(\boldsymbol{X}); \boldsymbol{\theta}^*)]^\top\boldsymbol{c}\right)$$

$$= \frac{1}{T}\sum_{t=1}^{T}\mathbb{E}[\boldsymbol{c}^T\boldsymbol{I}_t\boldsymbol{c}|\mathcal{F}_{t-1}] - \bar{\lambda}^2\boldsymbol{c}^T\mathbb{E}[\boldsymbol{g}(\boldsymbol{X}, f(\boldsymbol{X}); \boldsymbol{\theta}^*)]^{\otimes 2}\boldsymbol{c}, \tag{33}$$

where $\boldsymbol{I}_t = \frac{1}{\pi_t(a)}\mathbb{E}_{Y_t(a)}[\big(\boldsymbol{g}(\boldsymbol{X}_t, Y_t(a); \boldsymbol{\theta}^*) - \bar{\lambda}\boldsymbol{g}(\boldsymbol{X}_t, f(\boldsymbol{X}_t); \boldsymbol{\theta}^*)\big)^{\otimes 2}]$. We now study the limit of this. Let $\bar{\boldsymbol{I}}_t = \frac{1}{\bar{\pi}(a|\boldsymbol{X}_t)}\mathbb{E}_{Y_t(a)}[\big(\boldsymbol{g}(\boldsymbol{X}_t, Y_t(a); \boldsymbol{\theta}^*) - \bar{\lambda}\boldsymbol{g}(\boldsymbol{X}_t, f(\boldsymbol{X}_t); \boldsymbol{\theta}^*)\big)^{\otimes 2}]$. Observe

$$|\boldsymbol{c}^T\boldsymbol{I}_t\boldsymbol{c} - \boldsymbol{c}^T\bar{\boldsymbol{I}}_t\boldsymbol{c}|$$

$$= |\bar{\pi}(a|\boldsymbol{X}_t) - \pi_t(a)|\frac{1}{\bar{\pi}(a|\boldsymbol{X}_t)\pi_t(a)}\boldsymbol{c}^T\mathbb{E}[\big(\boldsymbol{g}(\boldsymbol{X}_t, Y_t(a); \boldsymbol{\theta}^*) - \bar{\lambda}\boldsymbol{g}(\boldsymbol{X}_t, f(\boldsymbol{X}_t); \boldsymbol{\theta}^*)\big)^{\otimes 2}|\boldsymbol{X}_t]\boldsymbol{c}$$

$$\leq M|\bar{\pi}(a|\boldsymbol{X}_t) - \pi_t(a)|$$

for some M, since we have bounded moments by Assumption 3.5 and minimum sampling probabilities by Assumption 3.2. Note the random variables $\{\bar{\pi}(a|\boldsymbol{X}_t) - \pi_t(a)\}$ are uniformly integrable. Thus $\bar{\pi}(a|\boldsymbol{X}_t) - \pi_t(a) \xrightarrow{p} 0$ can be strengthened to convergence in L^1 :

$$\lim_{t\to\infty}\mathbb{E}|\bar{\pi}(a|\boldsymbol{X}_t) - \pi_t(a)| = 0.$$

Now define $U_{t,c} = \boldsymbol{c}^T\boldsymbol{I}_t\boldsymbol{c} - \boldsymbol{c}^T\bar{\boldsymbol{I}}_t\boldsymbol{c}$. Combining the above facts gives

$$\lim_{t\to\infty}\mathbb{E}|U_{t,c}| = 0.$$

Also, note that

$$\mathbb{E}|\mathbb{E}|U_{t,c}|\mathcal{F}_{t-1}| \leq \mathbb{E}\mathbb{E}[|U_{t,c}|\mathcal{F}_{t-1}] = \mathbb{E}|U_{t,c}|.$$

Thus

$$\lim_{t\to\infty}\mathbb{E}|\mathbb{E}[U_{t,c}|\mathcal{F}_{t-1}] = 0.$$

Note that if a sequence of random variables converges to 0 in L_1, then its running average does as well. Therefore

$$\frac{1}{t}\sum_{\tau\le t}\mathbb{E}[U_{\tau,\boldsymbol{c}}|\mathcal{F}_{\tau-1}] \xrightarrow{L_1} 0.$$

Plugging in the definition of $U_{\tau,\boldsymbol{c}}$ gives

$$\frac{1}{T}\sum_{t\le T}\mathbb{E}[\boldsymbol{c}^T\boldsymbol{I}_t\boldsymbol{c}|\mathcal{F}_{t-1}] - \frac{1}{T}\sum_{t\le T}\mathbb{E}[\boldsymbol{c}^T\bar{\boldsymbol{I}}_t\boldsymbol{c}|\mathcal{F}_{t-1}] \xrightarrow{L_1} 0.$$

Note also

$$\frac{1}{T}\sum_{t\le T}\mathbb{E}[\boldsymbol{c}^T\bar{\boldsymbol{I}}_t\boldsymbol{c}|\mathcal{F}_{t-1}] = \frac{1}{T}\sum_{t\le T}\boldsymbol{c}^T\mathbb{E}[\bar{\boldsymbol{I}}_t]\boldsymbol{c} = \boldsymbol{c}^T\mathbb{E}[\bar{\boldsymbol{I}}_t]\boldsymbol{c} = \boldsymbol{c}^T\mathbb{E}\left[\frac{1}{\bar{\pi}(a|\boldsymbol{X}_t)}(\boldsymbol{g}(\boldsymbol{X}_t,Y_t(a);\boldsymbol{\theta}^*) - \bar{\lambda}\boldsymbol{g}(\boldsymbol{X}_t,f(\boldsymbol{X}_t);\boldsymbol{\theta}^*))^{\otimes 2}\right]\boldsymbol{c}.$$

These imply that the following convergence holds in L_1 and probability:

$$\frac{1}{T}\sum_{t\le T}\mathbb{E}[\boldsymbol{c}^T\boldsymbol{I}_t\boldsymbol{c}|\mathcal{F}_{t-1}] \to \boldsymbol{c}^T\mathbb{E}\left[\frac{1}{\bar{\pi}(a|\boldsymbol{X}_t)}(\boldsymbol{g}(\boldsymbol{X}_t,Y_t(a);\boldsymbol{\theta}^*) - \bar{\lambda}\boldsymbol{g}(\boldsymbol{X}_t,f(\boldsymbol{X}_t);\boldsymbol{\theta}^*))^{\otimes 2}\right]\boldsymbol{c}.$$

Therefore the limit in probability of (33) is

$$\boldsymbol{c}^T\left(\mathbb{E}\left[\frac{1}{\bar{\pi}(a)}\big(g(\boldsymbol{X}_t,Y_t(a);\boldsymbol{\theta}^*) - \bar{\lambda}\boldsymbol{g}(\boldsymbol{X}_t,f(\boldsymbol{X}_t);\boldsymbol{\theta}^*)\big)^{\otimes 2}\right] - \bar{\lambda}^2\mathbb{E}[g(\boldsymbol{X},f(\boldsymbol{X});\boldsymbol{\theta}^*)]^{\otimes 2}\right)\boldsymbol{c}$$

$$= \boldsymbol{c}^{\mathsf{T}}\boldsymbol{V}\boldsymbol{c}.$$

This shows that

$$\sum_{t=1}^{T}\mathrm{Var}\left(\boldsymbol{c}^{\mathsf{T}}\frac{1}{\sqrt{T}}\left[\frac{1_{\{A_t=a\}}}{\pi_t(a)}\left[\boldsymbol{g}(\boldsymbol{X}_t,Y_t(a);\boldsymbol{\theta}^*) - \bar{\lambda}\boldsymbol{g}(\boldsymbol{X}_t,f(\boldsymbol{X}_t);\boldsymbol{\theta}^*)\right] + \bar{\lambda}\mathbb{E}[\boldsymbol{g}(\boldsymbol{X},f(\boldsymbol{X});\boldsymbol{\theta}^*)]\right]|\mathcal{F}_{t-1}\right) \xrightarrow{p} \boldsymbol{c}^{\mathsf{T}}\boldsymbol{V}\boldsymbol{c}. \quad (34)$$

Together, (32) and (34) show that the sum of conditional variances of the kth increment given $\mathcal{G}_{T,k-1}$ converges in probability to $\boldsymbol{c}^\top(\bar{\lambda}^2 r\mathrm{Cov}(\boldsymbol{g}(\boldsymbol{X},f(\boldsymbol{X});\boldsymbol{\theta}^*)) + \boldsymbol{V})\boldsymbol{c}$, as desired.

Lastly, we must verify the conditional Lindeberg condition, which requires controlling the tail second moments of the summands. Specifically, this requires, for all $\delta > 0$

$$\sum_{i=1}^{N}\mathbb{E}\left[\left(\boldsymbol{c}^{\mathsf{T}}\boldsymbol{W}_i\right)^2 1_{|\boldsymbol{c}^{\mathsf{T}}\boldsymbol{W}_i|>\delta}\right] \xrightarrow{p} 0 \quad (35)$$

$$\sum_{t=1}^{T}\mathbb{E}\left[\left(\boldsymbol{c}^{\mathsf{T}}\boldsymbol{Z}_t\right)^2 1_{|\boldsymbol{c}^{\mathsf{T}}\boldsymbol{Z}_t|>\delta}|\mathcal{F}_{t-1}\right] \xrightarrow{p} 0 \quad (36)$$

where $\boldsymbol{W}_i = \bar{\lambda}\frac{\sqrt{T}}{N}\left(\boldsymbol{g}(\tilde{\boldsymbol{X}}_i,f(\tilde{\boldsymbol{X}}_i);\boldsymbol{\theta}^*) - \mathbb{E}\left[\boldsymbol{g}(\boldsymbol{X},f(\boldsymbol{X});\boldsymbol{\theta}^*)\right]\right)$ is an unlabeled term and $\boldsymbol{Z}_t = \frac{1}{\sqrt{T}}\left(\frac{1_{A_t=a}}{\pi_t(a)}\left[\boldsymbol{g}(\boldsymbol{X}_t,Y_t(a);\boldsymbol{\theta}^*) - \bar{\lambda}\boldsymbol{g}(\boldsymbol{X}_t,f(\boldsymbol{X}_t);\boldsymbol{\theta}^*)\right] + \bar{\lambda}\mathbb{E}[\boldsymbol{g}(\boldsymbol{X},f(\boldsymbol{X});\boldsymbol{\theta}^*)]\right)$ is a labeled term. Note that (35) involves unconditional expectations, again since the unlabeled data is independent of the previous information. To see (35), observe

$$\sum_{i=1}^{N}\mathbb{E}\left[(\boldsymbol{c}^\top\boldsymbol{W}_i)^2 1_{|\boldsymbol{c}^\top\boldsymbol{W}_i|>\delta}\right]$$

$$\le \frac{1}{\delta^2}\sum_{i=1}^{N}\mathbb{E}[(\boldsymbol{c}^\top\boldsymbol{W}_i)^4]$$

$$= \frac{\bar{\lambda}^4T^2}{\delta^2N^3}\mathbb{E}\left[(\boldsymbol{c}^\top(\boldsymbol{g}(\boldsymbol{X},f(\boldsymbol{X});\boldsymbol{\theta}^*) - \mathbb{E}[\boldsymbol{g}(\boldsymbol{X},f(\boldsymbol{X});\boldsymbol{\theta}^*)])^4\right]$$

$$= \frac{1}{N} \cdot \left(\frac{T}{N}\right)^2 \frac{\bar{\lambda}^4}{\delta^2} \mathbb{E}\left[(\boldsymbol{c}^\top(\boldsymbol{g}(\boldsymbol{X}, f(\boldsymbol{X}); \boldsymbol{\theta}^*) - \mathbb{E}[\boldsymbol{g}(\boldsymbol{X}, f(\boldsymbol{X}); \boldsymbol{\theta}^*)])^4\right]$$

$$\to 0$$

by Assumption 3.5 and since $T/N \to r$.

For (36), observe

$$\sum_{t=1}^{T} \mathbb{E}\left[\left(\boldsymbol{c}^\mathsf{T} \boldsymbol{Z}_t\right)^2 1_{|\boldsymbol{c}^\mathsf{T}\boldsymbol{Z}_t| > \delta} | \mathcal{F}_{t-1}\right]$$

$$\leq \frac{1}{\delta^2} \sum_{t=1}^{T} \mathbb{E}[(\boldsymbol{c}^\mathsf{T} \boldsymbol{Z}_t)^4 | \mathcal{F}_{t-1}]$$

$$\leq \frac{27||\boldsymbol{c}||_2^4}{T\delta^2} \mathbb{E}\left[\frac{||\boldsymbol{g}(\boldsymbol{X}_t, Y_t(a); \boldsymbol{\theta}^*)||_2^4}{\pi_{\min}^4} + \frac{\bar{\lambda}^4 ||\boldsymbol{g}(\boldsymbol{X}_t, f(\boldsymbol{X}_t); \boldsymbol{\theta}^*)||_2^4}{\pi_{\min}^4} + \bar{\lambda}^4 \mathbb{E}||\boldsymbol{g}(\boldsymbol{X}_t, f(\boldsymbol{X}_t); \boldsymbol{\theta}^*)||_2^4\right]$$

$$\xrightarrow{p} 0,$$

using Cauchy-Schwarz, Assumption 3.5, and $(|x| + |y| + |z|)^4 \leq 27(x^4 + y^4 + z^4)$. This establishes (36).

Thus by the martingale central limit theorem, $\sqrt{T}\boldsymbol{G}_T^{\hat{\lambda}}(\boldsymbol{\theta}^*) \xrightarrow{d} \mathcal{N}(\boldsymbol{0}, \bar{\lambda}^2 r \mathrm{Cov}(\boldsymbol{g}(\boldsymbol{X}, f(\boldsymbol{X}); \boldsymbol{\theta}^*)) + \boldsymbol{V})$, as desired. □

C. Consistent Estimate of Asymptotic Variance

Theorem C.1. *A consistent estimate of $\boldsymbol{\Sigma}_{\bar{\lambda}}$ from Theorem 3.7 is*

$$\hat{\boldsymbol{\Sigma}} = \boldsymbol{B}_T^{-1} \left(\lambda_T \frac{T}{N}(\boldsymbol{D}_T - \boldsymbol{E}_T^{\otimes 2}) + \boldsymbol{H}_T - \lambda_T^2 \boldsymbol{E}_T^{\otimes 2}\right) \boldsymbol{B}_T^{-\top}, \tag{37}$$

where $\lambda_T, \boldsymbol{B}_T, \boldsymbol{D}_T$, and $\boldsymbol{E}_T$ are defined as in Theorem 3.8 and $\boldsymbol{H}_T = \frac{1}{T-1}\sum_{t=1}^{T-1} \frac{1_{A_t=a}}{\pi_t(a)^2}(\boldsymbol{g}_{Y,t} - \lambda_T \boldsymbol{g}_{f,t})^{\otimes 2}$.

Proof. By assumption, $\lambda_T \xrightarrow{p} \bar{\lambda}$, and $T/N \to r$. Also, from the proof of Theorem 3.8, we have

$$\boldsymbol{B}_T \xrightarrow{p} \mathbb{E}[\nabla \boldsymbol{g}(\boldsymbol{X}, Y(a); \boldsymbol{\theta}^*)]$$

$$\boldsymbol{D}_T \xrightarrow{p} \mathbb{E}[\boldsymbol{g}_f^* \boldsymbol{g}_f^{*\top}]$$

$$\boldsymbol{E}_T \xrightarrow{p} \mathbb{E}[\boldsymbol{g}_f^*].$$

So by continuous mapping theorem, it suffices to show

$$\boldsymbol{H}_T \xrightarrow{p} \mathbb{E}\left[\frac{1}{\bar{\pi}(a)}(\boldsymbol{g}_Y^* - \bar{\lambda}\boldsymbol{g}_f^*)^{\otimes 2}\right] \tag{38}$$

To that end, note

$$\boldsymbol{H}_T = \frac{1}{T-1}\sum_{t=1}^{T-1} \frac{1_{A_t=a}}{\pi_t(a)^2} \boldsymbol{g}_{Y,t}^{\otimes 2} + \lambda_T^2 \boldsymbol{F}_T - \lambda_T \boldsymbol{C}_T, \tag{39}$$

where $\boldsymbol{F}_T$ and $\boldsymbol{C}_T$ are defined in the proof of Theorem 3.8. Again from that proof, we have

$$\boldsymbol{F}_T \xrightarrow{p} \mathbb{E}\left[\frac{1}{\bar{\pi}(a)} \boldsymbol{g}_f^* \boldsymbol{g}_f^{*\mathsf{T}}\right]$$

$$\boldsymbol{C}_T \xrightarrow{p} E\left[\frac{1}{\bar{\pi}(a)}(\boldsymbol{g}_f^* \boldsymbol{g}_Y^{*\mathsf{T}} + \boldsymbol{g}_Y^* \boldsymbol{g}_f^{*\mathsf{T}})\right]$$

and by the same argument,

$$\frac{1}{T-1}\sum_{t=1}^{T-1}\frac{1_{A_t=a}}{\pi_t(a)^2}\boldsymbol{g}_{Y,t}^{\otimes 2}\xrightarrow{p}\mathbb{E}\left[\frac{\boldsymbol{g}_Y^{*\otimes 2}}{\bar{\pi}(a)}\right].$$

Thus by continuous mapping theorem applied to (39),

$$\boldsymbol{H}_T\xrightarrow{p}\mathbb{E}\left[\frac{\boldsymbol{g}_{Y,t}^{*\otimes 2}}{\bar{\pi}(a)}\right]+\bar{\lambda}^2\mathbb{E}\left[\frac{1}{\bar{\pi}(a)}\boldsymbol{g}_f^*\boldsymbol{g}_f^{*\mathsf{T}}\right]-\bar{\lambda}E\left[\frac{1}{\bar{\pi}(a)}(\boldsymbol{g}_f^*\boldsymbol{g}_Y^{*\mathsf{T}}+\boldsymbol{g}_Y^*\boldsymbol{g}_f^{*\mathsf{T}})\right]$$

$$=\mathbb{E}\left[\frac{1}{\bar{\pi}(a)}(\boldsymbol{g}_Y^*-\bar{\lambda}\boldsymbol{g}_f^*)^{\otimes 2}\right].$$

establishing (38) and completing the proof.

□

D. Further Simulations

D.1. Low-Exploration Regimes

To show strong predictions can reduce variance due to small inverse propensity weights, we reran the experiment of Section 4 with smaller ϵ and a stronger predictive model $f^{\text{excellent}}(\boldsymbol{x})=\exp(0.201\boldsymbol{x}^{(1)}-0.097\boldsymbol{x}^{(2)}+0.398\boldsymbol{x}^{(3)})$. Table 3 shows the average confidence interval width of 90% confidence intervals with coverage in parentheses. In this low-exploration regime, PPAI confidence intervals are 10-16 times tighter than those of WLS and maintain near-nominal coverage.

ϵ	**WLS**	**PPAI ($f^{\text{excellent}}$)**
0.01	0.47 (89%)	0.03 (88%)
0.05	0.19 (90%)	0.02 (89%)

Table 3. Confidence interval width and coverage for WLS and PPAI under small choices of ϵ. PPAI intervals are substantially tighter and achieve near-nominal coverage.

D.2. Ablation Study: Importance of the Correction Term

We conduct an ablation study to evaluate the impact of the PPAI correction term, by comparing standard PPAI against an estimator that relies entirely on the AI predictions and unlabeled data:

$$\hat{\boldsymbol{\theta}}_{a,T}^{\text{NC}}=\left[\sum_{i=1}^{N}\tilde{\boldsymbol{X}}_i\tilde{\boldsymbol{X}}_i^T\right]^{-1}\times\sum_{i=1}^{N}\tilde{\boldsymbol{X}}_i f_a(\tilde{\boldsymbol{X}}_i).\tag{40}$$

As shown in Table 4, the inclusion of the correction term significantly reduces the mean absolute error across 300 trials. As expected, the error reduction is largest when the AI model is poor. This demonstrates that the correction term is vital for maintaining accuracy, especially for poorly calibrated AI models.

	PPAI w/o correction	**PPAI**
f^{poor}	0.222	0.005
f^{strong}	0.013	0.003

Table 4. Mean absolute error over 300 trials comparing PPAI with and without the correction term.

D.3. Ablation Study: Adaptive Lambda Selection

To assess the effectiveness of the data-driven parameter $\hat{\lambda}_T$, we compare our procedure against an approach where λ is fixed at $\lambda=0.5$. The results in Table 5 show that using $\hat{\lambda}_T$ drastically reduces the confidence interval width while maintaining or improving coverage.

	PPAI with $\lambda = 0.5$	**PPAI with $\hat{\lambda}_T$**
f^{poor}	0.355 (88%)	0.021 (89%)
f^{strong}	0.054 (93%)	0.011 (93%)

Table 5. CI width (coverage) comparing fixed λ against adaptive $\hat{\lambda}_T$ selection.

D.4. Misspecification Analysis

We evaluate the robustness of PPAI under varying levels of reward model misspecification. We define the true reward as $Y = \boldsymbol{X}^\top \beta + c \cdot h(\boldsymbol{X})$, where $\beta = [0.3, 0.1, 0.2]^\top$ and $h(\boldsymbol{x}) = \exp(0.2\boldsymbol{x}^{(1)} - 0.1\boldsymbol{x}^{(2)} + 0.3\boldsymbol{x}^{(3)})$. We test across $c \in \{0, 0.5, 1\}$, comparing a strong predictive model $f^{\text{strong}}(\boldsymbol{x}) = \boldsymbol{x}^\top \beta + c \cdot \exp(0.29\boldsymbol{x}^{(1)} + 0.11\boldsymbol{x}^{(2)} + 0.19\boldsymbol{x}^{(3)})$ and a poor model $f^{\text{poor}}(\boldsymbol{x}) = 1$. The remaining experimental setup from Section 4 is unchanged. As shown in Table 6, PPAI CIs consistently outperform standard WLS as the misspecification constant c increases, maintaining near-nominal coverage with significantly narrower widths.

c	**WLS**	**PPAI (f^{poor})**	**PPAI (f^{strong})**
0.0	0.009 (91%)	0.009 (91%)	0.009 (90%)
0.5	0.054 (91%)	0.031 (89%)	0.010 (88%)
1.0	0.107 (91%)	0.063 (89%)	0.011 (92%)

Table 6. Width and coverage of 90% confidence intervals under increasing reward misspecification.

D.5. Confidence Interval Implementation Details

To construct the PPAI confidence intervals summarized in Table 1, we use the plug-in estimate of the PPAI asymptotic variance Σ_{λ^*} described in Appendix C. We construct $1 - \alpha$ confidence intervals for the first component of $\boldsymbol{\theta}_0^*$ via

$$\hat{\boldsymbol{\theta}}_0^{(0)} \pm z_{1-\alpha/2} \sqrt{\frac{\hat{\boldsymbol{\Sigma}}_{0,0}}{T}}, \tag{41}$$

where z_q denotes the qth quantile of the standard normal distribution. Because our theoretical guarantees are asymptotic, finite-sample coverage may deviate from the nominal level in challenging settings, such as with small sample sizes, high-variance rewards, or out-of-scale predictions.

E. Movie Recommendation Experiment Details

E.1. Movie Recommendation Reward Model

To construct the reward model, we trained a random forest regressor to predict the rating for a user-movie pair (u, m) based on an embedding of the movie's summary and on the following five features:

1. (User Mean) The average rating provided by user u.
2. (Movie Mean) The average rating received by movie m.
3. (Interaction: User $\times$ Movie) The product of the user mean and the movie mean.
4. (Genre Affinity) The user's historical average rating for the specific genres associated with movie m.
5. (Interaction: Genre $\times$ Movie) The product of the genre affinity and the movie mean.

We then clip the output of the model to $[0.5, 5]$. The model was initially trained on 80% of the data. On the remaining 20% test set, the model achieved a Mean Absolute Error of 0.597, indicating that simulated ratings typically deviate by approximately half of a star from observed ratings. Following this validation, we retrained the model on the full dataset to serve as our simulation environment.

E.2. Experiment Steps

We simulate our method as follows: At time step $t = 1, \ldots, 1000$,

1. We sample the context $\boldsymbol{X}_t$ of the tth user.
2. We apply an ϵ-greedy policy with $\epsilon = 0.1$ to choose the movie $m(A_t)$ to show the user. With probability ϵ, we select $A_t \in \{0, 1\}$ randomly; with probability $1 - \epsilon$, we select $A_t = \arg\max_a \hat{\boldsymbol{\beta}}_{\boldsymbol{a}}^\top \boldsymbol{X}_t$.
3. We query the reward model to compute the rating $Y_t = y(\boldsymbol{X}_t, m(A_t))$.
4. We draw 10 unlabeled user contexts $\tilde{\boldsymbol{X}}_{10(t-1)+1}, \ldots, \tilde{\boldsymbol{X}}_{10t}$ and compute predicted ratings for both movies $m(0), m(1)$ by querying Gemini.
5. With the labeled data $(\boldsymbol{X}_\tau, A_\tau, Y_\tau)_{\tau=1}^t$ and unlabeled data $(\tilde{\boldsymbol{X}}_i, a, f(\tilde{\boldsymbol{X}}_i))_{i=1}^{10t}$ for $a = 0, 1$, we compute the PPAI least squares estimator (7).

To obtain predicted ratings from Gemini, we use prompts of the following form: "Predict how many stars (1-5) a viewer would give X-Men (2000) given that they gave these ratings: Independence Day (1996): 4.5, The Lord of the Rings: The Return of the King (2003): 3.0, Star Wars: Episode IV - A New Hope (1977): 4.0, Pulp Fiction (1994): 2.5, Shrek (2001): 5.0. State your predicted rating within brackets, e.g. [4.5]."

E.3. Hypothesis Tests

We apply hypothesis tests to demonstrate how PPAI can enable discoveries. First, since $\boldsymbol{\beta}_0^{(3)} \approx 0.17$ is well separated from zero, it suggests a non-zero relationship between ratings of *Star Wars* and *X-Men*. We therefore test $H_0 : \boldsymbol{\beta}_0^{(3)} = 0$ against $H_1 : \boldsymbol{\beta}_0^{(3)} \neq 0$. With PPAI, we (correctly) reject in 89% of trials, whereas with WLS, we reject in just 21% of trials; this shows our method gives greater power against H_0.

To test estimates of $\boldsymbol{\beta}_1^{(1)}$, we first note the proximity of $\boldsymbol{\beta}_1^{(1)} \approx 0.03$ to zero demonstrates a negligible relationship between ratings of *Good Will Hunting* and *Independence Day*. We therefore employ a negligibility test: the null hypothesis is $H_0 : \boldsymbol{\beta}_1^{(1)} \notin [-\delta, \delta]$ ($\boldsymbol{\beta}_1^{(1)}$ is not negligible), for some $\delta > 0$, and the alternative is $H_1 : \boldsymbol{\beta}_1^{(1)} \in [-\delta, \delta]$ ($\boldsymbol{\beta}_1^{(1)}$ is negligible). The procedure is to reject when the 90% confidence interval is fully contained in $[-\delta, \delta]$. The choice of $\delta > 0$ is somewhat arbitrary and depends on domain knowledge, but to demonstrate our method, we set $\delta = 0.2$. For this δ, our method rejects (correctly concludes negligibility) in 87% of trials, whereas WLS *never* rejects, as its confidence intervals are consistently too large to fit in $[-\delta, \delta]$.

F. Extension to Covariate Shift

In practice, the unlabeled covariates may come from a different distribution than the labeled covariates. In this section, we briefly discuss how our framework can be extended to accommodate such covariate shift.

Specifically, suppose

$$\boldsymbol{X}_t \overset{\text{i.i.d.}}{\sim} \mathcal{P}_X^\ell, \qquad \tilde{\boldsymbol{X}}_i \overset{\text{i.i.d.}}{\sim} \mathcal{P}_X^u,$$

where $\mathcal{P}_X^\ell$ and $\mathcal{P}_X^u$ denote the labeled and unlabeled covariate distributions, with corresponding densities p_ℓ and p_u, respectively. We assume that $\mathcal{P}_X^\ell$ is absolutely continuous with respect to $\mathcal{P}_X^u$, and define the density ratio

$$w(\boldsymbol{x}) := \frac{p_\ell(\boldsymbol{x})}{p_u(\boldsymbol{x})}.$$

We further assume that w is bounded almost surely.

We first consider the case where w is known. The key idea is to reweight the unlabeled data so that expectations under the unlabeled covariate distribution are converted into expectations under the labeled covariate distribution. Indeed,

$$\mathbb{E}_{\mathcal{P}_X^u}\Big[w(\tilde{\boldsymbol{X}})\, \boldsymbol{g}(\tilde{\boldsymbol{X}}, f_a(\tilde{\boldsymbol{X}}); \boldsymbol{\theta}^*)\Big] = \int \frac{p_\ell(\tilde{\boldsymbol{x}})}{p_u(\tilde{\boldsymbol{x}})} \boldsymbol{g}(\tilde{\boldsymbol{x}}, f_a(\tilde{\boldsymbol{x}}); \boldsymbol{\theta}^*)\, p_u(\tilde{\boldsymbol{x}})\, d\tilde{\boldsymbol{x}},$$

$$= \int \boldsymbol{g}(\tilde{\boldsymbol{x}}, f_a(\tilde{\boldsymbol{x}}); \boldsymbol{\theta}^*)\, p_\ell(\tilde{\boldsymbol{x}})\, d\tilde{\boldsymbol{x}} = \mathbb{E}_{\mathcal{P}_X^\ell}\left[\boldsymbol{g}(\tilde{\boldsymbol{X}}, f_a(\tilde{\boldsymbol{X}}); \boldsymbol{\theta}^*)\right].$$

Accordingly, when w is known, we modify the estimating equation by reweighting the unlabeled term:

$$\boldsymbol{G}_T^{\lambda_T, w}(\boldsymbol{\theta}) = \frac{1}{T}\sum_{t=1}^{T} \frac{\mathbf{1}\{A_t = a\}}{\pi_t(a)}\, \boldsymbol{g}(\boldsymbol{X}_t, Y_t; \boldsymbol{\theta}) + \lambda_T \left[\frac{1}{N}\sum_{i=1}^{N} w(\tilde{\boldsymbol{X}}_i)\, \boldsymbol{g}(\tilde{\boldsymbol{X}}_i, f_a(\tilde{\boldsymbol{X}}_i); \boldsymbol{\theta}) - \frac{1}{T}\sum_{t=1}^{T} \frac{\mathbf{1}\{A_t = a\}}{\pi_t(a)}\, \boldsymbol{g}(\boldsymbol{X}_t, f_a(\boldsymbol{X}_t); \boldsymbol{\theta})\right].$$

In practice, the density ratio w is typically unknown and must be estimated. Suppose $\hat{w}$ is an estimator constructed from an auxiliary sample independent of $\{(\boldsymbol{X}_t, A_t, Y_t)\}_{t=1}^T$ and $\{\tilde{\boldsymbol{X}}_i\}_{i=1}^N$. One simple way to obtain such an auxiliary sample is to reserve a small fraction of the labeled and unlabeled data for this purpose. When $\hat{w}$ is a reasonable estimator for w, say $\|\hat{w} - w\|_\infty = o_p(T^{-1/2})$, then we may replace w by $\hat{w}$ in the estimating equation:

$$\boldsymbol{G}_T^{\lambda_T, \hat{w}}(\boldsymbol{\theta}) = \frac{1}{T}\sum_{t=1}^{T} \frac{\mathbf{1}\{A_t = a\}}{\pi_t(a)}\, \boldsymbol{g}(\boldsymbol{X}_t, Y_t; \boldsymbol{\theta}) + \lambda_T \left[\frac{1}{N}\sum_{i=1}^{N} \hat{w}(\tilde{\boldsymbol{X}}_i)\, \boldsymbol{g}(\tilde{\boldsymbol{X}}_i, f_a(\tilde{\boldsymbol{X}}_i); \boldsymbol{\theta}) - \frac{1}{T}\sum_{t=1}^{T} \frac{\mathbf{1}\{A_t = a\}}{\pi_t(a)}\, \boldsymbol{g}(\boldsymbol{X}_t, f_a(\boldsymbol{X}_t); \boldsymbol{\theta})\right].$$

A full theoretical treatment of this covariate-shift extension is left for future work.