Batch-Adaptive Causal Annotations

Anonymous ACL submission

Abstract

Estimating the causal effects of interventions is crucial to policy and decision-making, yet outcome data are often missing or subject to non-standard measurement error. While ground-truth outcomes can sometimes be obtained through costly data annotation or follow-up, budget constraints typically allow only a fraction of the dataset to be labeled. We address this challenge by optimizing *which data points should be sampled for outcome information* in order to improve efficiency in average treatment effect estimation with missing outcomes. We derive a closed-form solution for the optimal sampling probability in batches. We optimize the asymptotic variance of a doubly-robust estimator for causal inference with missing outcomes, and show the resulting asymptotic convergence to the optimal variance. Motivated by a collaboration with a street outreach provider generating millions of case notes, we also extend this framework to costly annotations of unstructured data, such as text or images, common in healthcare and social services. Across simulated and real-world datasets, including one of outreach interventions in homelessness services, our approach achieves substantially lower mean-squared error and recovers the AIPW estimate with fewer labels than existing baselines. In practice, we show that our method can match confidence intervals obtained with 361 random samples using only 90 optimized samples—saving 75% of the labeling budget.

1 INTRODUCTION

Estimating causal effects is challenging to begin with, but a common challenge arises when outcome data is missing but potentially observable via active querying. In this paper, we study observational causal inference with missing outcomes, where we can obtain information about ground-truth outcomes at a cost, via expert data annotation or follow-up. Recent machine learning tools can label outcome information from noisy text or image observations, but naively using biased or potentially erroneous label *predictions* as stand-ins for ground truth can invalidate statistical inference and confidence intervals. Small ground-truth annotation budgets allow valid estimation on a subsample, but introduces high variance. We build on doubly-robust causal inference with missing outcomes to determine where to sample additional outcome annotations to minimize the asymptotic variance of downstream treatment effect estimation.

Our methodology is motivated by a collaboration with a nonprofit to evaluate the impact of street outreach on housing outcomes, where rich information about outcomes of outreach are embedded in case notes written by outreach workers. Street outreach is an intensive intervention; caseworkers canvass for and build relationships with homeless clients and write case notes after each interaction. These notes are a noisy view on the ground truth of what happens during the open-ended process of outreach. Was a client progressing towards housing or their goals, or were they facing other barriers? In our experience, outreach workers can extract structured ground-truth information from the unstructured text of case notes. They can provide context and recognize important milestones. Yet under-resourced outreach workers cannot label millions of case notes. While modern natural language processing tools can facilitate annotation at scale, they are often inaccurate. *Given an annotation budget constraint, how can we strategically assign expert labels while leveraging weaker ML-predicted annotations to optimize causal effect estimation?* In this paper, we develop general methodology for optimizing data annotation and we demonstrate its effectiveness empirically, including on ground-truthed housing outcome data.

This problem is not unique to the social work domain and can generally apply to cases of measurement error with misaligned modalities (such as text or images), where we can query the ground truth for

some portion of the data at a cost. In some settings, we can query other data sources for ground-truth labels directly, while in other settings, outcomes may be recorded in complex information such as text or images. However, due to dimensionality issues, these cannot be directly substituted for ground-truth outcomes Y. Throughout the paper, we refer to these as "complex embedded outcomes", or $\tilde{Y}$. Weaker imputation of this auxiliary information is feasible at scale, but second-best due to inaccuracies. For example, when an outcome variable, wages, is only observed from self-reported working individuals, surveyors could conduct follow-up interviews with participants to obtain wage data, but this can be expensive. Noisy measures from the same dataset (such as last year's wages) or transporting prediction models from national wage databases can be predictive. Such trade-offs between expert annotation and scalable, weaker imputation are pervasive in data-intensive machine learning, for example as in the recent "LLM-as-a-judge" framework (Zheng et al., 2023).

This study makes the following contributions: we propose a two-stage batch-adaptive algorithm for efficient average treatment effect (ATE) estimation from complex embedded outcomes. We derive the expert labeling probability that minimizes the asymptotic variance of an orthogonal estimator (Bia et al., 2021). We design a two-stage adaptive annotation procedure. The first stage estimates nuisance functions for the asymptotic variance on the fully observed data. We use the estimates and functions from the first stage to estimate the optimal labeling probabilities in the second stage. The final proposed estimator combines the model-annotated labels and the expert labels in a doubly robust estimator for the average treatment effect (ATE). We show that this two-stage design achieves the optimal asymptotic variance with weaker double-machine learning requirements on nuisance function estimates. We leverage our closed-form characterizations to provide insights on how to improve downstream treatment-effect estimation. We validate and show improvements upon random sampling on semi-synthetic and real-world datasets from retail and street outreach.

2 RELATED WORK

Our work is closest to adaptive designs that optimize (asymptotic) variance objectives for causal estimands. Prior work studies treatment allocation under such objectives (Hahn et al., 2011; Li and Owen, 2024; Cook et al., 2024), while we instead allocate *outcome annotation* probability under a labeling budget in a missing-outcomes setting. We also relate to work on design-based supervised learning and data annotation with auxiliary signals (e.g., text) (Egami et al., 2023), and discuss why prediction-focused active learning is generally misaligned with minimizing ATE uncertainty. Additional discussion and broader taxonomy are in the supplement/appendix.

3 PROBLEM SETUP

We study causal inference with missing outcomes, where a simpler ground truth outcome $Y \in \mathbb{R}$ can be revealed via annotation of a more complex observation thereof (e.g., text or images), denoted $\tilde{Y}$. We also discuss extensions to a setting where we can use $\tilde{Y}$ to enhance nuisance function estimation.

In both cases, we assume the ground-truth data-generating process follows that of standard causal inference. The ground-truth data $(X, Z, Y(Z))$ includes covariates $X \in \mathcal{X}$, a binary treatment $Z \in \{0, 1\}$, and potential outcomes $Y(Z)$ in the Neyman-Rubin potential outcome framework. We only observe $Y(Z)$ for the historically-assigned Z and assume the usual stable unit value treatment assumption (SUTVA). *If* all ground-truth outcomes were observed, estimation would reduce to the standard causal setting; the key challenge is missingness. Let $R \in \{0, 1\}$ denote the presence ($R = 1$) or absence ($R = 0$) of the outcome Y. The *observed* dataset is (X, Z, R, RY), i.e. with missing outcomes. Causal identification relies on the following assumptions:

Assumption 1 (Treatment ignorability (Hernan and Robins, 2025))**.** $Y(Z) \perp\!\!\!\perp Z \mid X$.

Assumption 2 (R-ignorability (Rubin, 1976; Bia et al., 2021))**.** $R \perp\!\!\!\perp Y(Z) \mid Z, X$.

Assumption 1, or unconfoundedness, posits that the observed covariates are fully informative of treatment. It is generally untestable but robust estimation is possible in its absence, e.g. via sensitivity analysis and partial identification (Zhao et al., 2019; Kallus and Zhou, 2021). On the other hand, Assumption 2 *is true by design*, since we choose what datapoints are annotated for ground-truth labels based on (Z, X) alone.

Though completely random sampling enables doubly-robust causal inference, we ask: how can we optimize our choice of annotated datapoints to improve the *variance* of downstream estimation?

We assume a fixed annotation budget $B \in [0,1]$ that determines the fraction of the dataset that can be annotated. We define the propensity score and annotation (outcome observation) probability as follows:

$$e_z(X) := P(Z = z|X) \quad \text{(propensity score)}$$
$$\pi(Z,X) := P(R = 1|Z,X) \quad \text{(annotation probability)}$$

We assume positivity/overlap; that we observe treatment and outcome with nonzero probability.

Assumption 3 (Treatment and annotation positivity (Hernan and Robins, 2025))**.** $\epsilon < \pi(z,X) \leq 1-\epsilon, z \in \{0,1\}$ and $\epsilon < e_1(X) < 1-\epsilon$, with $\epsilon > 0$.

Assumptions 1 to 3 are standard in causal inference and we point the reader to textbook references for further discussion (Hernan and Robins, 2025; Imbens, 2004; Kennedy, 2020).

We define the outcome model, which is identified on the $R = 1$ data by Assumption 2, and the conditional variance:

$$\mu_z(X) := \mathbb{E}[Y \mid Z = z, X] \underset{asn.2}{=} \mathbb{E}[Y \mid Z = z, R = 1, X]$$
$$\sigma_z^2(X) := \mathbb{E}[(Y - \mu_z(X))^2 \mid Z = z, X].$$

Batch allocation setup. We consider a two-batch adaptive protocol, where n i.i.d. observations are randomly split into two batches. We consider a proportional asymptotic where the size of first batch, n_1, is a fixed proportion $\kappa \in (0,1)$ of n.

Assumption 4 (Proportional asymptotic (Hahn et al., 2011; Li and Owen, 2024))**.** $\lim_{n\to\infty} \frac{n_1}{n} = \kappa$.

In the first batch, we randomly assign annotations according to a small but asymptotically nontrivial fraction of the budget. Outcomes are realized and observed, and the nuisance models $(\hat{\mu}_z(x), \hat{e}_z(x), \hat{\sigma}_z^2(x))$ are trained on the observed data. In the second batch, we solve for optimal annotation probabilities π^* and sample data so that the mixture distribution over outcome observations achieves π^*. We combine the results from both batches and use the data for ATE estimation.

Extension to missing outcomes with context. We provide an extension of our missing outcomes framework to settings where complex-embedded outcomes might be used not only for data annotation but also to enhance outcome model predictions. Though our method assumes ground-truth outcomes could be revealed for each datapoint, for example via follow-up surveys, in practice this is most likely relevant in *data annotation* settings. Expert data annotation only works when there is some data to annotate: we denote this noisy observation $\tilde{Y}$, which could be text or images. Given that a noisy observation $\tilde{Y}$ is available, a natural question is, when can $\tilde{Y}$ be included to further improve outcome prediction? We need an additional assumption: an exclusion restriction that the direct causal effect of treatment passes through the ground truth Y alone. Similar assumptions are in the measurement error literature (Shu and Yi, 2019). For example, in a medical setting, treatment may shrink a tumor (changing Y), which is recorded in clinical notes or imaging data $\tilde{Y}$. But the treatment does not directly effect *how* text or images are *recorded*. This prevents collider bias, and is testable after the first batch of data.

Assumption 5 (Complex embedded outcomes: exclusion restriction)**.** $Z \perp \tilde{Y} \mid X, Y(Z)$

In this setting, under Assumption 5, we can allow the outcome model to depend on the complex embedded $\tilde{Y}$, and denote $\mu_Z(X,\tilde{Y}) := \mathbb{E}[Y|Z,X,\tilde{Y}]$. Note we only need Assumption 5 if using $\tilde{Y}$ to improve outcome modeling of $\mu_z(X,\tilde{Y})$. Otherwise, if there is any doubt about Assumption 5, simply revert to the original case and do not include $\tilde{Y}$ in outcome prediction.

There are several ways of incorporating the context into the outcome model. We denote an ML prediction based on $\tilde{Y}$ (with X covariates and treatment information) as $f_z(X,\tilde{Y})$; for example zero-shot prediction using an LLM. If using black-box ML or LLM predictions, we recommend ensembling with $\mathbb{E}[Y|Z,X]$ or estimating $\mathbb{E}[Y \mid Z, R = 1, f_z(X,\tilde{Y})]$ to calibrate LLM predictions, in order to satisfy statistical consistency conditions. (Egami et al. (2023) also suggests this).

4 METHOD

We outline our method, starting with a recap of the augmented inverse-propensity weighting (AIPW) estimator for causal inference with missing outcomes. Then we optimize its asymptotic variance, characterize the optimal π^*, and give a feasible estimation procedure.

Recap: Optimal asymptotic variance for the ATE with missing outcomes. We seek to estimate the average treatment effect (ATE) on ground-truth outcomes Y. Define

$$\tau = \mathbb{E}[Y(1) - Y(0)].$$

Bia et al. (2021) derives a double-machine learning estimator for ATE estimation with missing outcomes. For each $z \in \{0, 1\}$,

$$\mathbb{E}[Y(z)] = \mathbb{E}[\psi_z],$$
$$\psi_z = \frac{\mathbb{1}[Z = z]\, R\,(Y - \mu_z(X))}{e_z(X)\,\pi(z, X)} + \mu_z(X),$$

and the ATE is $\tau_{AIPW} = \mathbb{E}[\psi_1 - \psi_0]$. The outcome model $\mu_z(X)$ is estimated on data with observed outcomes. Under SUTVA and Assumption 2, $\mathbb{E}[Y(z)|X] = \mathbb{E}[Y|Z = z, X] = \mathbb{E}[Y|Z = z, R = 1, X]$.

We optimize the semiparametric efficient asymptotic variance with missing outcomes. We express the asymptotic variance of (Bia et al., 2021) in terms of μ_z, e_z, π:

Proposition 1. *The asymptotic variance (AVar) is:*

$$\text{AVar} = \text{Var}[\mu_1(X) - \mu_0(X)] + \sum_{z\in\{0,1\}} \mathbb{E}\left[\frac{\sigma_z^2(X)}{e_z(X)\,\pi(z, X)}\right]$$

The first term is independent of π; we focus on optimizing the second term with respect to π.

Remark 1. In the setting with complex embedded outcomes where the outcome predictions $\mu_z(X, \tilde{Y})$ predict based on $\tilde{Y}$ information, this only changes the outcome model for evaluating the AIPW estimator. Since we optimize annotation probabilities varying in X alone, the optimization objective and solution remain the same in the limit, marginalizing over $\tilde{Y}$.

Characterizing the optimal $\pi^*(z, x)$. We first characterize the population optimal sampling probabilities $\pi^*(z, x)$, assuming the nuisance functions are known. We optimize the asymptotic variance over π under a global sampling budget $B \in [0, 1]$ over all annotations. $\pi^*(z, x)$ solves

$$\min_{0<\pi(z,x)\le 1, \forall z,x} \sum_{z\in\{0,1\}} \mathbb{E}\left[\frac{\sigma_z^2(X)}{e_z(X)\pi(z, X)}\right] \quad \text{s.t. } \mathbb{E}[\pi(Z, X)] \le B \tag{1}$$

Note that in the global budget constraint, $\mathbb{E}[\pi(Z, X)] = \mathbb{E}[\pi(1, X)\mathbb{1}[Z = 1] + \pi(0, X)\mathbb{1}[Z = 0]]$. We can characterize the solution as follows.

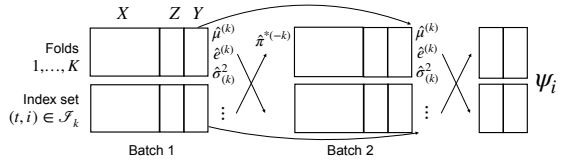

Figure 1: Illustration of cross-fitting (K folds within batches), used by our batched allocation procedure.

Theorem 1. *The optimal annotation probabilities are:*

$$\pi^*(z, X) = \frac{\sqrt{\sigma_z^2(X)}}{e_z(X)} B \times \left(\mathbb{E}\left[\sqrt{\sigma_1^2(X)} + \sqrt{\sigma_0^2(X)}\right]\right)^{-1}$$

Note that sampling probabilities increase in the conditional variance/uncertainty of the model, $\sigma^2(X)$, and the inverse propensity score. Characterizing the closed-form solution is useful for our analysis later on. Proof details are in the supplement.

Feasible two-batch adaptive design and estimator. Our characterizations above assume knowledge of true $\sigma_z^2(x)$ and propensity scores $e_z(x)$. Since these need to be estimated, we leverage the double machine learning (DML) framework and conduct a feasible two-batch adaptive design (Chernozhukov et al., 2018; Bia et al., 2021). Standard cross-fitting (Chernozhukov et al., 2018) splits the data, estimates nuisance functions on one fold, and evaluates the estimator on a datapoint leveraging nuisance functions from another fold of data. We leverage a variant (Li and Owen, 2024) that introduces folds within each batch of data. We summarize the cross-fitting approach here; details are in the supplement. First, we split the observations in each batch $t = 1, 2$ into K folds (e.g. $K = 5$). Let $\mathcal{I}_k$ denote the set of batch and observation indices (t, i) assigned to fold k and batch t. Then within each fold, we estimate nuisance models on observations in batch 1. We use cross-fitting to optimize the sampling probabilities, i.e., $\pi^{*,(-k)}$ optimizes asymptotic variance with out-of-fold nuisances $e^{(-k)}$. Finally we adaptively assign annotation probabilities in batch 2. This ensures independence, meaning that the nuisance models in batch 2, fold k rely only on observations from the *previous* batch 1, in fold k.

Algorithm 1 includes this procedure, with additional implementation details in the supplement.

Algorithm 1 Batch Adaptive Causal Estimation (With Complex Embedded Outcomes)

Input: Data $\mathcal{D} = \{(X_i, Z_i)\}_{i=1}^n$, sampling budget $B \in [0, 1]$

Step 1: Partition $\mathcal{D}$ into 2 batches and K folds $\mathcal{D}_1^{(k)}, \mathcal{D}_2^{(k)}$ for $k = 1, \ldots, K$

Step 2: On Batch 1, sample $R_1 \sim \text{Bern}(B)$. Estimate nuisances within each k-fold $\hat{\mu}_z^{(k)}(X)$ (or $\hat{\mu}_z^{(k)}(X, \tilde{Y})$), $\hat{\sigma}_z^{2(k)}(X)$, and $\hat{e}_z^{(k)}(X)$.

Step 3: On Batch 2, folds $k = 1, \ldots, K$, obtain π^* by optimizing eq. (1), plugging in nuisance estimates. Solve for $\hat{\pi}_2^{(k)}(X_i) = \frac{1}{1-\kappa}(\pi^*(X_i) - \kappa\pi_1)$

Step 4: On Batch 2, sample $R_2 \sim \text{Bern}(\hat{\pi}_2^{(k)}(X_i))$ and obtain outcomes.

Step 5: Pool data across batches and estimate ATE with AIPW estimator in eq. (2) (or eq. (RZ-plug-in.), or balancing weights) and out of fold nuisances.

Therefore the cross-fitted feasible estimator takes the form $\hat{\tau}_{AIPW} = \frac{1}{n}\sum_{t=1}^{2}\sum_{k=1}^{K}\sum_{(t,i)\in\mathcal{I}_k}\hat{\psi}_{1,i} - \hat{\psi}_{0,i}$ where

$$\hat{\psi}_{z,i} = \frac{\mathbb{1}[Z_i = z]R_i}{\hat{e}_z^{(-k)}(X_i)\,\hat{\pi}^{(-k)}(z, X_i)}\left(Y_i - \hat{\mu}_z^{(-k)}(X_i)\right) + \hat{\mu}_z^{(-k)}(X_i). \tag{2}$$

5 ANALYSIS

In this section, we provide a central limit theorem for the setting where annotation probabilities are assigned adaptively and nuisance parameters must be estimated. We provide some insights to improve estimation as well as an extension to settings with continuous treatments.

Denote $\|\cdot\|_2 = (\mathbb{E}[(\cdot)^2])^{1/2}$. The following Assumptions 6 to 8 are all standard in the double machine learning literature (Chernozhukov et al., 2018; Wager, 2024; Athey and Wager, 2021; Uehara et al., 2020; Bia et al., 2021). Assumption 9 is specific to our batch adaptive sampling design and can also be found in (Li and Owen, 2024).

Assumption 6 (Consistent estimation and boundedness)**.** Assume bounded second moments of outcomes and errors, $\|Y(z)\|_2 \le C_1, \|\mu_z(X)\|_2 \le C_2, \|(Y - \mu_z(X))\|_2^2 \le 4B_{\sigma^2}, \forall z$; and consistent estimation $\mathbb{E}[(\mu_z(X) - \hat{\mu}_z(X))^2] \le K_\mu n^{-r_\mu}$ for some constants $C_1, C_2, B_{\sigma^2}, K_\mu, r_\mu \ge 0$.

Assumption 7 (Product error rates)**.** For nuisance functions, assume the products of their mean-square convergence rates vanish faster than $n^{-1/2}$: (i) $\sqrt{n}\,\|\hat{\mu}_z(X) - \mu_z(X)\|_2 \times \|\hat{\pi}(z, X) - \pi(z, X)\|_2 \xrightarrow{p} 0$; (ii) $\sqrt{n}\,\|\hat{\mu}_z(X) - \mu_z(X)\|_2 \times \|\hat{e}_z(X) - e_z(X)\|_2 \xrightarrow{p} 0$.

Assumption 8 (VC dimension for nuisance estimation)**.** The nuisance estimation of e_z and σ_z^2 occurs over function classes with finite VC-dimension.

Assumption 9 (Sufficiently weak dependence across batches (Li and Owen, 2024))**.**

$$g_i := \mathbb{E}\left[\hat{\psi}_i(R; \hat{\eta}) - \psi_i(R; \eta)\,\middle|\,\mathcal{I}^{(-k)}, X_i\right].$$

$$\sqrt{\frac{1}{n_{t,k}}\sum_{i:(t,i)\in\mathcal{I}_k}\|g_i\|^2} = o_p(n^{-\frac{1}{4}}),$$

where $\hat{\eta}$ collects the estimated nuisance functions (propensity, annotation probability, and outcome model) and η denotes their population counterparts.

Theorem 2. *Given Assumptions 1 to 3, suppose that we construct the feasible estimator $\hat{\tau}_{AIPW}$ (Equation (2)) using CSBAE-style cross-fitting with estimators satisfying Assumptions 6 to 9. Then*

$$\sqrt{n}(\hat{\tau}_{AIPW} - \tau) \Rightarrow \mathcal{N}(0, V_{AIPW}),$$

where τ is the ATE and V_{AIPW} is

$$\sum_{z\in 0,1}\mathbb{E}\left[\frac{\sigma_z^2(X)}{e_z(X)\pi^*(z, X)}\right] + \text{Var}\left[\mu_1(X) - \mu_0(X)\right].$$

Theorem 2 shows that the batch adaptive design and feasible estimator has an asymptotic variance equal to the variance of the true ATE under missing outcomes and the optimal π^*. Therefore, our procedure gives asymptotically valid level-α confidence intervals for τ of minimum width. The proof of Theorem 2 proceeds in two steps: (i) the feasible AIPW estimator converges to the oracle-nuisance AIPW estimator, and (ii) the oracle estimator under feasible nuisances converges to the same limit under oracle nuisances (details in the supplement).

Insights and improvements

1) When is our method much better than uniform sampling? Prior works of (Egami et al., 2023; Zrnic and Candès, 2024), though they do

not study treatment effect estimation, obtain valid inference with uniform sampling (i.e. with the budget probability). When do optimized data annotation probabilities improve upon uniform sampling? To answer this, we analyze the relative efficiency (RelEff) which compares the asymptotic variance (AVar) under optimized or uniform sampling, for the same budget.

Corollary 1 (Relative efficiency)**.**

$$\begin{aligned}\text{RelEff} &= \frac{\text{AVar}(\pi^*)}{\text{AVar}(B)} \\ &= \frac{A}{D},\end{aligned}$$

where

$$\begin{aligned}V_z(X) &:= \frac{\sigma_z^2(X)}{e_z(X)}, \\ S(X) &:= \sqrt{\sigma_1^2(X)} + \sqrt{\sigma_0^2(X)}, \\ A &:= \frac{1}{B}\left(\mathbb{E}\left[S(X)\right]\right)^2 + \text{Var}[\tau(X)], \\ D &:= \frac{1}{B}\mathbb{E}\left[V_1(X)\right] + \frac{1}{B}\mathbb{E}\left[V_0(X)\right] \\ &\quad + \text{Var}[\tau(X)].\end{aligned}$$

By construction, $\text{RelEff} \leq 1$; the smaller it is, the larger the improvement from our method. Our method's improvement increases if the budget is smaller ($B \downarrow$) or if there are imbalanced propensities where $e_1(X)$ close to 0 or 1. Improvements shrink for large budgets or when treatment variances are similar.

2) Direct estimation of $(e\pi^*)^{-1}$ mitigates estimation stability. It is well known that estimating propensities and then inverting estimates can be unstable in practice. This problem is doubly-so for causal inference with missing outcomes. We find many papers on adaptive treatment allocation note this challenge and mix their optimized allocation probabilities with uniform in the experimental sections (Dimakopoulou et al., 2021; Zrnic and Candès, 2024; Cook et al., 2024); just as many papers in causal inference clip the weights in practice (Wang et al., 2017). Our closed-form solution reveals that estimating propensity scores for the *final* ATE estimation on the full dataset is *fundamentally unnecessary*, though it is needed to estimate π^*. At π^*, observe that[1] $(e_z(x)\pi^*(z,x))^{-1} \propto \sqrt{\sigma_z^2}(x)^{-1}$ and is *independent of the propensity score* $e_z(x)$. Therefore estimating the optimal inverse propensity function directly can exploit its *lower* statistical complexity. In causal inference and covariate shift, many methods (such as balancing weights) avoid the plug-in approach for inverse propensity methods in favor of direct estimation of the inverse propensity score (Tsuboi et al., 2009; Zubizarreta, 2015; Imai and Ratkovic, 2014; Kallus, 2018a,b; Cohn et al., 2023; Bruns-Smith et al., 2025). We recommend estimation on the final dataset with such approaches or other types of direct estimation. For example, even estimation of $P(Z = z, R = 1 \mid X)$ directly helps:

$$p_z(X) := P(Z = z, R = 1 \mid X).$$

$$\psi_z(e, \pi^*) = \tfrac{\mathbb{1}[Z=z,R=1]}{p_z(X)}(Y - \mu_z(X)) + \mu_z(X). \quad (RZ\text{-plug-in.})$$

3) Insights extend beyond binary treatments. Our analysis extends to other static estimands (e.g., continuous treatments with kernel localization); we include details in the appendix.

6 EXPERIMENTS

We evaluate our batch adaptive allocation protocol on synthetic and real-world datasets. We show that our method enables consistent and efficient ATE estimation even under limited labeling budgets, ultimately helping resource-constrained organizations obtain reliable estimates from their data.

Baselines. Across all experimental setups, we compare against completely random sampling for AIPW, and evaluate MSE relative to an oracle full-data skyline (infeasible in practice). We compare MSE to the skyline of the standard AIPW estimator with fully observed outcomes, that is when the budget equals 1 or $R = 1$ for all data points. In our setting, completely random sampling for AIPW is the strong baseline, because AIPW is optimal causal estimation for the ATE. Other more complicated methods target other objectives. **Active learning baselines.** We also evaluate standard pool-based active learning (AL) heuristics (e.g., uncertainty sampling and variance-reduction criteria) adapted to query outcomes. Empirically, they underperform for ATE inference: AL optimizes prediction error for μ_z rather than the AIPW variance contribution $\sigma_z^2(X)/(e_z(X)\pi(z,X))$, and aggressive AL policies can induce near-deterministic π that inflates inverse-weight variance. Full AL details and results are in the supplement.

[1] This depends on some joint properties of κ, p_1, whether it is feasible to find second-stage batch sampling probabilities π_2 so that $\kappa p_1 + (1-\kappa)\pi_2(x) = \pi^*(x)$

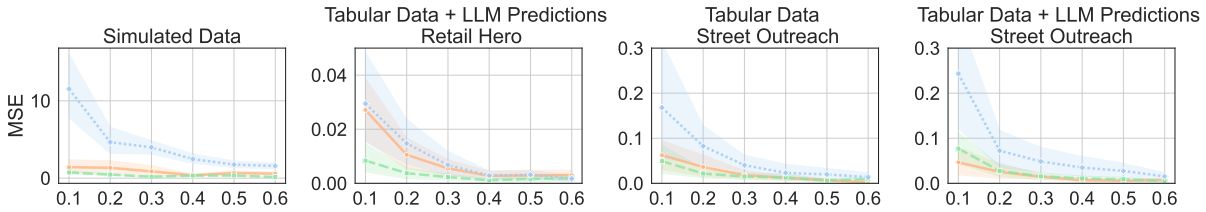

Figure 2: **Main results (MSE).** Mean squared error averaged over trials across labeling budgets for synthetic data (leftmost), Retail Hero (center left), and Street Outreach (center right and rightmost). Confidence interval width plots and additional experimental details are in the supplement.

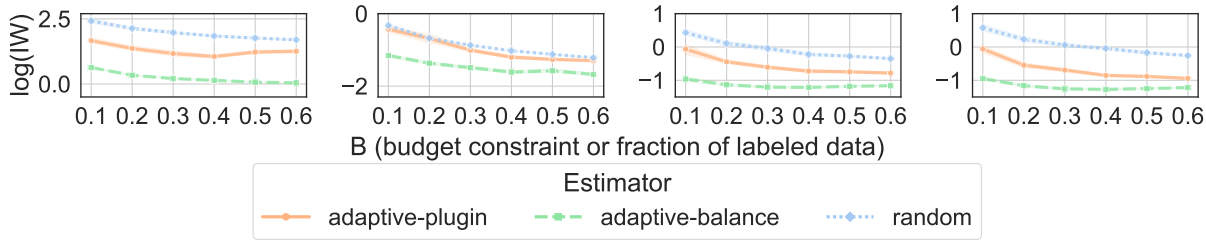

Figure 3: **Main results (CI width).** Average 95% confidence interval width (log scale) across labeling budgets for the same three settings as Figure 2.

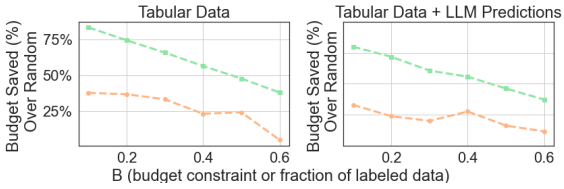

Figure 4: **Budget saved (RetailHero).** Percentage of annotation budget saved by adaptive sampling to match the confidence-interval width of uniform sampling.

Datasets and setup. We consider (i) a controlled synthetic setting to validate the theory, and (ii) two real-world motivated datasets with text-embedded outcome signals: RetailHero, where the treatment is an SMS ad and $\tilde{Y}$ are LLM-generated social-media posts tied to outcomes, and Street Outreach, where $\tilde{Y}$ are LLM summaries of casenotes and Y is housing placement. In each case, we allocate a small initial batch of labels uniformly, estimate nuisance components, compute $\hat{\pi}^*$, and then collect the remaining labels according to the second-stage allocation.

In addition to reporting MSE (Fig. 2), we track the width of valid 95% confidence intervals as a direct proxy for inferential precision under a fixed annotation budget. This aligns with the workshop setting: the goal is not only accurate point estimates, but also reliable uncertainty quantification under limited labels.

Synthetic Data. We include a simulation study in the supplement to validate the method in a fully controlled setting.

Results summary. We report additional plots and implementation details in the supplement. Across synthetic and two real-world datasets (RetailHero with text-embedded outcomes; Street Outreach casenotes), batch-adaptive annotation improves over uniform random sampling for AIPW, especially at small budgets where uncertainty is most acute. Figure 2 shows large MSE reductions; we also observe materially tighter confidence intervals and substantial annotation savings to match a target interval width.

RetailHero and Street Outreach. RetailHero is a semi-synthetic setting with tabular covariates, randomized SMS treatment, and outcomes with text-embedded signals; Street Outreach uses baseline tabular covariates and LLM-generated summaries of casenotes, with outcomes reflecting housing placement. In both cases, we estimate μ_z with flexible learners and use the first batch to estimate the components of π^*; the second batch concentrates labels where the asymptotic variance contribution is largest. This yields the biggest gains at low budgets, where random sampling suffers most from heavy-tailed inverse-weight terms.

Practical takeaways. The gains are most pronounced in regions with poor overlap (small $e_z(X)$) or high conditional outcome variance $\sigma_z^2(X)$, where the AIPW influence function up-

weights observations. Our closed-form π^* naturally re-allocates labeling effort toward these high-variance regions, improving precision without changing the estimand. In practice, we recommend pairing batch-adaptive sampling with direct estimation of $(e\pi^*)^{-1}$ (e.g., balancing weights) to avoid unstable plug-in inverse weights; this stabilizes finite-sample behavior while preserving the asymptotic efficiency guarantees.

What the figure shows. Across datasets, the gap between adaptive and uniform sampling is largest at small budgets ($B \in [0.1, 0.3]$): the adaptive policy reduces MSE quickly by prioritizing labels in regions that dominate the asymptotic variance term $\mathbb{E}[\sigma_z^2(X)/(e_z(X)\pi(z, X))]$. As the budget increases, both methods approach the fully-observed AIPW skyline and the marginal benefit of re-allocation decreases, consistent with the RelEff characterization.

In the real-world settings, these gains translate into practical budget savings: for a target interval width, adaptive annotation achieves comparable precision with substantially fewer expert labels than uniform sampling (see supplement for budget-saved curves and confidence-interval results). This is particularly important when labels require human review of unstructured text (casenotes) or costly follow-up.

Conclusion, limitations, and future work. We have introduced a batch-adaptive causal annotation procedure for efficient data labeling. Limitations include assuming that annotations reveal ground truth, since annotators might disagree. Our theory also requires LLM statistical consistency - we suggest using them in ensembled predictions. In future work, we plan to explore other causal estimators.

References

Susan Athey, Raj Chetty, Guido W Imbens, and Hyunseung Kang. 2019. The surrogate index: Combining short-term proxies to estimate long-term treatment effects more rapidly and precisely. Technical report, National Bureau of Economic Research.

Susan Athey and Stefan Wager. 2021. Policy learning with observational data. *Econometrica*, 89(1):pp. 133–161.

Michela Bia, Martin Huber, and Lukäs Laffërs. 2021. Double machine learning for sample selection models. *arXiv prepint arXiv:2012.00745*.

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

Wenbin Cai, Ya Zhang, and Jun Zhou. 2013. Maximizing expected model change for active learning in regression. In *2013 IEEE 13th International Conference on Data Mining*, pages 51–60.

Kamalika Chaudhuri, Prateek Jain, and Nagarajan Natarajan. 2017. Active heteroscedastic regression. In *International Conference on Machine Learning*, pages 694–702. PMLR.

Kamalika Chaudhuri, Sham M Kakade, Praneeth Netrapalli, and Sujay Sanghavi. 2015. Convergence rates of active learning for maximum likelihood estimation. In *Advances in Neural Information Processing Systems*, volume 28. Curran Associates, Inc.

Jacob M Chen, Rohit Bhattacharya, and Katherine A Keith. 2024. Proximal causal inference with text data. *arXiv preprint arXiv:2401.06687*.

Victor Chernozhukov, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen, Whitney Newey, and James Robins. 2018. Double/debiased machine learning for treatment and structural parameters.

Victor Chernozhukov, Whitney Newey, Vıctor M Quintas-Martınez, and Vasilis Syrgkanis. 2022. Riesznet and forestriesz: Automatic debiased machine learning with neural nets and random forests. In *International Conference on Machine Learning*, pages 3901–3914. PMLR.

David A. Cohn, Zoubin Ghahramani, and Michael I. Jordan. 1996. Active learning with statistical models. *J. Artif. Int. Res.*, 4(1):129–145.

Eric R Cohn, Eli Ben-Michael, Avi Feller, and José R Zubizarreta. 2023. Balancing weights for causal inference. In *Handbook of Matching and Weighting Adjustments for Causal Inference*, pages 293–312. Chapman and Hall/CRC.

Kyle Colangelo and Ying-Ying Lee. 2020. Double debiased machine learning nonparametric inference with continuous treatments. *arXiv preprint arXiv:2004.03036*.

Thomas Cook, Alan Mishler, and Aaditya Ramdas. 2024. Semiparametric efficient inference in adaptive experiments. In *Causal Learning and Reasoning*, pages 1033–1064. PMLR.

Yifan Cui, Hongming Pu, Xu Shi, Wang Miao, and Eric Tchetgen Tchetgen. 2024. Semiparametric proximal causal inference. *Journal of the American Statistical Association*, 119(546):1348–1359.

Nikita Dhawan, Leonardo Cotta, Karen Ullrich, Rahul Krishnan, and Chris J Maddison. 2023. End-to-end causal effect estimation from unstructured natural language data. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

Maria Dimakopoulou, Zhimei Ren, and Zhengyuan Zhou. 2021. Online multi-armed bandits with adaptive inference. *Advances in Neural Information Processing Systems*, 34:1939–1951.

Naoki Egami, Christian J Fong, Justin Grimmer, Margaret E Roberts, and Brandon M Stewart. 2022. How to make causal inferences using texts. *Science Advances*, 8(42):eabg2652.

Naoki Egami, Musashi Hinck, Brandon Stewart, and Hanying Wei. 2023. Using imperfect surrogates for downstream inference: Design-based supervised learning for social science applications of large language models. In *Advances in Neural Information Processing Systems*, volume 36, pages 68589–68601. Curran Associates, Inc.

Claudio Gentile, Zhilei Wang, and Tong Zhang. 2024. Fast rates in pool-based batch active learning. *J. Mach. Learn. Res.*, 25(1).

Jinyong Hahn, Keisuke Hirano, and Dean Karlan. 2011. Adaptive experimental design using the propensity score. *Journal of Business & Economic Statistics*, 29(1):96–108.

M.A. Hernan and J.M. Robins. 2025. *Causal Inference: What If*. Chapman & Hall/CRC Monographs on Statistics & Applied Probab. CRC Press.

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

Guido W. Imbens. 2004. Nonparametric estimation of average treatment effects under exogeneity: A review. *The Review of Economics and Statistics*, 86(1):4–29.

Andrew Jesson, Panagiotis Tigas, Joost van Amersfoort, Andreas Kirsch, Uri Shalit, and Yarin Gal. 2021. Causal-bald: Deep bayesian active learning of outcomes to infer treatment-effects from observational data. *Advances in Neural Information Processing Systems*, 34:30465–30478.

Zhijing Jin, Julius von Kügelgen, Jingwei Ni, Tejas Vaidhya, Ayush Kaushal, Mrinmaya Sachan, and Bernhard Schoelkopf. 2021. Causal direction of data collection matters: Implications of causal and anticausal learning for nlp. *arXiv preprint arXiv:2110.03618*.

Nathan Kallus. 2018a. Balanced policy evaluation and learning. *Advances in neural information processing systems*, 31.

Nathan Kallus. 2018b. Optimal a priori balance in the design of controlled experiments. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 80(1):85–112.

Nathan Kallus and Xiaojie Mao. 2024. On the role of surrogates in the efficient estimation of treatment effects with limited outcome data. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, page qkae099.

Nathan Kallus and Angela Zhou. 2018. Policy evaluation and optimization with continuous treatments. In *International conference on artificial intelligence and statistics*, pages 1243–1251. PMLR.

Nathan Kallus and Angela Zhou. 2021. Minimax-optimal policy learning under unobserved confounding. *Management Science*, 67(5):2870–2890.

Edward H. Kennedy. 2020. Efficient nonparametric causal inference with missing exposure information. *The International Journal of Biostatistics*.

Tor Lattimore and Csaba Szepesvári. 2020. *Bandit algorithms*. Cambridge University Press.

Harrison H Li and Art B Owen. 2024. Double machine learning and design in batch adaptive experiments. *Journal of Causal Inference*, 12(1):20230068.

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

Bernhard Schölkopf, Dominik Janzing, Jonas Peters, Eleni Sgouritsa, Kun Zhang, and Joris Mooij. 2012. On causal and anticausal learning. *arXiv preprint arXiv:1206.6471*.

Burr Settles. 2009. Active learning literature survey.

Di Shu and Grace Y Yi. 2019. Causal inference with measurement error in outcomes: Bias analysis and estimation methods. *Statistical Methods in Medical Research*, 28(7):2049–2068. PMID: 29241426.

Dhanya Sridhar and David M Blei. 2022. Causal inference from text: A commentary. *Science Advances*, 8(42):eade6585.

Iiris Sundin, Peter Schulam, Eero Siivola, Aki Vehtari, Suchi Saria, and Samuel Kaski. 2019. Active learning for decision-making from imbalanced observational data. In *International conference on machine learning*, pages 6046–6055. PMLR.

Eric J Tchetgen Tchetgen, Andrew Ying, Yifan Cui, Xu Shi, and Wang Miao. 2024. An introduction to proximal causal inference. *Statistical Science*, 39(3):375–390.

Yuta Tsuboi, Hisashi Kashima, Shohei Hido, Steffen Bickel, and Masashi Sugiyama. 2009. Direct density ratio estimation for large-scale covariate shift adaptation. *Journal of Information Processing*, 17:138–155.

Masatoshi Uehara, Masahiro Kato, and Shota Yasui. 2020. Off-policy evaluation and learning for external validity under a covariate shift. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS '20, Red Hook, NY, USA. Curran Associates Inc.

Victor Veitch, Dhanya Sridhar, and David Blei. 2020. Adapting text embeddings for causal inference. In *Conference on Uncertainty in Artificial Intelligence*, pages 919–928. PMLR.

Roman Vershynin. 2018. *High-dimensional probability: An introduction with applications in data science*, volume 47. Cambridge university press.

Stefan Wager. 2024. Causal inference: A statistical learning approach.

Yu-Xiang Wang, Alekh Agarwal, and Miroslav Dudık. 2017. Optimal and adaptive off-policy evaluation in contextual bandits. In *International Conference on Machine Learning*, pages 3589–3597. PMLR.

Dongrui Wu, Chin-Teng Lin, and Jian Huang. 2019. Active learning for regression using greedy sampling. *Information Sciences*, 474:90–105.

Shu Yang and Peng Ding. 2020. Combining multiple observational data sources to estimate causal effects. *Journal of the American Statistical Association*.

Jinglong Zhao. 2023. Adaptive neyman allocation. *arXiv preprint arXiv:2309.08808*.

Qingyuan Zhao, Dylan S Small, and Bhaswar B Bhattacharya. 2019. Sensitivity analysis for inverse probability weighting estimators via the percentile bootstrap. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 81(4):735–761.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, and 1 others. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in Neural Information Processing Systems*, 36:46595–46623.

Yinglun Zhu and Robert Nowak. 2022. Active learning with neural networks: Insights from nonparametric statistics. *Advances in Neural Information Processing Systems*, 35:142–155.

Tijana Zrnic and Emmanuel J Candès. 2024. Active statistical inference. *arXiv preprint arXiv:2403.03208*.

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

	Predictive error objective (MSE of $E[Y(z) \mid X]$)	Decision objective	Inference (Optimize asymptotic variance for ATE)
Choose treatments	Experimental design	Bandits for simple regret, best-arm identification (Lattimore and Szepesvári, 2020)	(Hahn et al., 2011; Li and Owen, 2024; Cook et al., 2024; Zhao, 2023)
Annotate outcomes	Active learning; regression-based for CATE (Jesson et al., 2021).	Best-arm identification. For CATE, (Sundin et al., 2019). Else n/a.	Our work

Table 1: Taxonomy of adaptive data collection methods for causal inference by what is sampled (treatments vs. outcomes) and by target objective (prediction, decision, inference).

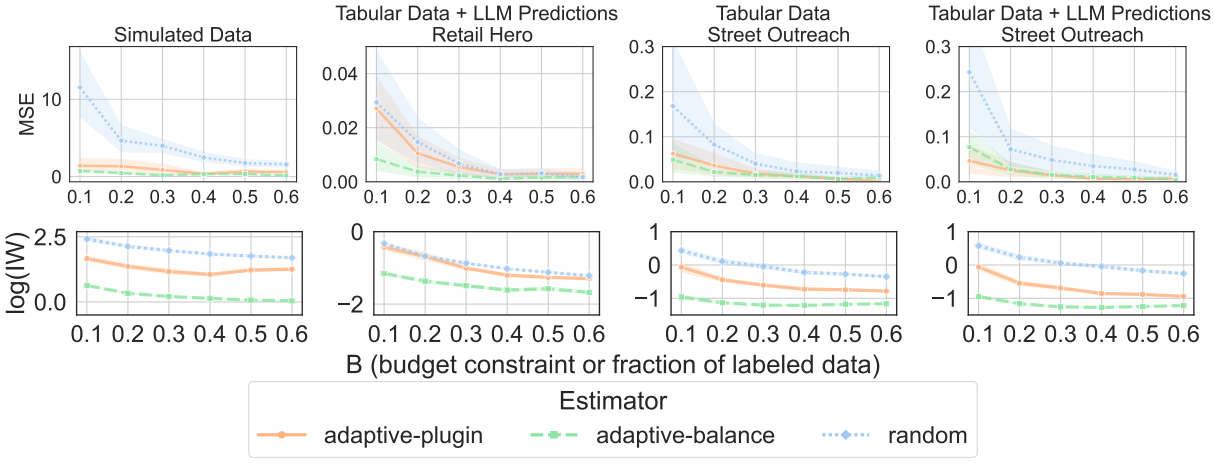

Figure 5: **Full main-results figure.** Mean squared error (top) and 95% confidence interval width on the log scale (bottom) averaged over trials across labeling budgets for synthetic data (leftmost), Retail Hero (center left), and Street Outreach (center right and rightmost).

Supplementary Figures and Tables

A NOTATION

B ADDITIONAL DISCUSSION ON RELATED WORK

Additional discussion on surrogate estimation In much of the surrogate literature, surrogates measure an outcome that is impossible to measure at the time of analysis. The canonical example in (Athey et al., 2019) studies the long-term intervention effects of job training on lifetime earnings, by using only short-term outcomes (surrogates) such as yearly earnings. In this regime, the ground truth cannot be obtained at the time of analysis. In this paper, we focus a different regime where obtaining the ground truth from expert data annotators is feasible but budget-binding.

We leverage the fact that we can design sampling probabilities of outcome observations (ground-truth annotations) or patterns of missingness for doubly-robust estimation, aligning with some methods in the surrogate outcomes and data combination literature (Yang and Ding, 2020; Kallus and Mao, 2024). But we treat the underlying setting as a single unconfounded dataset with missingness. The different setting of proximal causal inference (Tchetgen Tchetgen et al., 2024; Cui et al., 2024) seeks proxy outcomes/treatments that are informative of unobserved confounders; we assume unconfoundedness holds. Recently, (Chen et al., 2024) study the "design-based supervised learning" perspective of (Egami et al., 2023) specifically for proxies for unobserved confounding.

Additional discussion on more adaptive allocation methods beyond batch. We outline how our approach is a good fit for our motivating data annotation setting. Full-adaptivity is less relevant in our setting with ground-truth annotation from human experts, due to distributed-computing-type issues with random times of annotation completion. But standard tools such as the martingale CLT can be applied to extend our theoretical results to full adaptivity. Additionally, many recent works primarily focus on

Y_i	Ground truth outcomes, observed when label is provided by experts
$\tilde{Y}_i$	Complex embedded outcomes, such as raw text
X_i	Covariates included in estimation
Z_i	Treatment assignment indicator
R_i	Missingness indicator, indicates whether i is expertly labeled
$e_z(X_i)$	Propensity score, probability of being assigned treatment $Z = z$
$\pi(Z_i, X_i)$	Annotation probability, probability of sampling unit i for expert annotation
$f_z(X_i, \tilde{Y}_i)$	Estimated function of covariates and complex embedded outcomes, e.g. zero-shot LLM prediction from raw text
$\hat{\mu}_z(X_i, f(\tilde{Y}_i))$	Estimated model predicting Y as function of $(X_i, f(\tilde{Y}_i))$

the different problem of treatment allocation for ATE estimation. In-sample regret is less relevant for our setting of data annotation, which is a pure-exploration problem.

Optimizing asymptotic variance of the ATE vs. active learning. An extensive literature in machine learning studies where to sample data to improve machine learning predictors, in the subfield of active learning. The biggest difference is that we target functional estimation, aka improving estimation and inference on the average treatment effect, rather than improving estimation of the black-box nuisance predictors, so our approach is complementary to other approaches for active learning. Approaches for active learning with nonparametric regression include Zhu and Nowak (2022); Chaudhuri et al. (2017). Active learning generally requires additional structural conditions, such as margin or low-noise conditions, in order to show improvements. Our work highlights optimality leveraging the structure of our final treatment effect inferential goal.

Other works on causal inference and active learning for heterogeneous treatment effect estimation Some papers combine active learning and causal inference, but they primarily focus on estimating the conditional average treatment effect, or CATE = $E[Y(1) - Y(0) \mid X]$. Most of these papers consider estimation via the difference of two regression functions, i.e. CATE estimators that look like $\mu_1(X) - \mu_0(X)$, and therefore focus on active learning for regression methods in general, with a twist of learning the two treated/control regression functions. (Jesson et al., 2021) adapts Bayesian active learning for deep models, but modifies them to avoid sampling in non-overlap regions. (Sundin et al., 2019) focuses on sampling *counterfactual* outcome information with a best-arm identification objective (type-S error, to identify the correct sign of treatment effect). While these earlier papers also aim to reveal outcome information when treatment is already assigned, they primarily focus on reducing regression estimation error of an *inefficient/non-doubly-robust* estimator for the CATE. We instead focus on estimating the ATE, and optimizing the asymptotic variance of *semiparametrically efficient* estimation of the averaged ATE functional.

Relationship to causal inference and NLP There is a large and rapidly growing literature on causal inference with text data (Egami et al., 2022; Sridhar and Blei, 2022; Veitch et al., 2020). Throughout, we have deliberately used the terminology of measurement error to characterize our approach: that text measures outcomes of interest. (Dhawan et al., 2023) also adopt this stance towards text and note that it differs from prior works on causal inference and NLP, which focuses on questions of substantive interest related to the text itself.

Although we can define a potential outcome $\tilde{Y}(Z)$, we are generally uninterested in causal inference in the ambient high-dimensional space of $\tilde{Y}(Z)$ itself - corresponding to, in our examples, the effect of the presence of a tumor on the pixel image, the effect of street outreach on the linguistic characteristics of casenotes written for documentation, etc — $\tilde{Y}(Z)$ is relevant to causal estimation insofar as it is informative of latent outcomes $Y(Z)$.

This is consistent with viewing certain types of NLP tasks as “anti-causal learning" (Schölkopf et al., 2012), wherein outcomes cause measurements thereof, in analogy to anti-causal learning in supervised classification where a label of “cat" or “dog" causes the classification covariates (e.g. image) (Jin et al., 2021). Analogously, we view the underlying ground-truth outcomes Y as causing the measurement

thereof, $\tilde{Y}$.

C ALGORITHM

Algorithm 2 (Full Algorithm) Batch Adaptive Causal Estimation With Complex Embedded Outcomes

Input: Data $\mathcal{D} = \{(X_i, Z_i, Y_i, \tilde{Y}_i)\}_{i=1}^n$, sampling budget B_z for $z \in \{0, 1\}$
Output: ATE estimator $\hat{\tau}_{AIPW}$
Partition $\mathcal{D}$ into 2 batches and K folds $\mathcal{D}_1^{(k)}, \mathcal{D}_2^{(k)}$ for $k = 1, \ldots, K$
Batch 1:
for $k = 1, \ldots, K$ **do**
  On $\mathcal{D}_1^{(k)}$: Sample $R_1 \sim \text{Bern}(\pi_1(Z, X))$, where $\pi_1(z, x) = B_z$.
  Estimate nuisance models: Where $R = 1$, estimate $\hat{\mu}_z^{(k)}$ by regressing Y on X (or $X, \tilde{Y}$), and $\hat{\sigma}_z^{2(k)}$ by regressing $(Y - \hat{\mu}_z)^2$ on X. Estimate $\hat{e}_z^{(k)}$ by regressing Z on X.
end for
Batch 2:
for $k = 1, \ldots, K$ **do**
  On $\mathcal{D}_2^{(k)}$: Obtain π^* by optimizing eq. (1), plugging in $\hat{\mu}_z^{(-k)}$, $\hat{\sigma}_z^{2(-k)}$, and $\hat{e}_z^{(-k)}$.
  Solve for $\hat{\pi}_2^{(k)}(X_i) = \frac{1}{1-\kappa}(\pi^*(X_i) - \kappa\pi_1)$
  Sample $R_2 \sim \text{Bern}(\hat{\pi}_2^{(k)}(X_i)$
end for
Obtain $\mathcal{D}^{(k)}$ for $k = 1, \ldots, K$ by pooling across batches $\mathcal{D}_1^{(k)}$ and $\mathcal{D}_2^{(k)}$
On $\mathcal{D}^{(k)}$, re-estimate $\hat{\mu}_z^{(k)}$, $\hat{\sigma}_z^{2(k)}$,and $\hat{e}_z^{(k)}$ on observed outcomes RY for $k = 1, \ldots, K$
On $\mathcal{D}^{(k)}$, run optimization procedure to get $\pi^{*(-k)}$ with out of fold nuisances $\hat{\mu}_z^{(-k)}$, $\hat{\sigma}_z^{2(-k)}$, and $\hat{e}_z^{(-k)}$.
On full data $\mathcal{D}$, estimate ATE by using AIPW estimator in eq. (2) and out of fold nuisances $\pi^{*(-k)}$, $\hat{\mu}_z^{(-k)}$, $\hat{\sigma}_z^{2(-k)}$, and $\hat{e}_z^{(-k)}$

D ADDITIONAL RESULTS

D.1 Treatment$-z$-specific budgets B_z

We also consider a setting with different a priori fixed budgets within each treatment group, where

$$\text{sampling budget proportion } B_z \in [0, 1]$$

is the max percentage of the treated group $Z = z$ that can be annotated. Given that we are trying to choose the π that minimizes this variance bound, we only need to focus on the terms that depend on π and can drop the rest. Supposing oracle knowledge of propensities and outcome models, the optimization problem, for each $z \in \{0, 1\}$ is:

$$\min_{0<\pi(z,x)\leq 1, \forall z,x} \left\{ \mathbb{E}\left[\frac{\sigma_z^2(X)}{e_z(X)\pi(z, X)} \right] : \mathbb{E}\left[\pi(z, X) \mid Z = z\right] \leq B_z, z \in \{0, 1\} \right\} \qquad \text{(z-budget)}$$

Theorem 3. *The solution to the within-z-budget problem is:*

$$\pi^*(z, X) = \frac{\sqrt{\sigma_z^2(X)/e_z^2(X)}}{\mathbb{E}\left[\sqrt{\sigma_z^2(X)/e_z^2(X)} \mid Z = z\right]} \cdot B_z$$

.

D.2 Extension to continuous treatments

In the continuous setting, consider estimation of a counterfactual mean:

$$\mathbb{E}[Y(z)].$$

(We can extend to contrasts for different values of treatment, in analogy to the ATE). Let $(Y_i, X_i, Z_i)_{i=1}^n$ be an i.i.d. sample from $Q = (Y, X, Z) \in \mathcal{Q} = \mathcal{Y} \times \mathcal{X} \times \mathcal{Z}_0 \subseteq \mathcal{R}^{1 \times d_x \times 1}$, i.e. consider a univariate continuous treatment $Z \in \mathcal{Z}_0$. This can extend to the case of multiple continuous treatments d_Z but for ease of mathematical computation, we start with the one-dimensional continuous treatment setting. We derive the form of the asymptotic variance as well as the bias term for an estimator for continuous treatments with missing outcomes.

We introduce an estimator for continuous treatments with missing outcomes that is a direct extension of (Kallus and Zhou, 2018; Colangelo and Lee, 2020), while building on the Riesz representer characterization of (**?**)'s automatic double machine learning estimator for continuous treatment effects. We introduce what we call the "partial" Riesz representer, $\alpha(z, X) = \frac{1}{P(Z=z|X)}$ which is the inverse generalized propensity score or the balancing function for treatment alone. (We term it "partial" since we are optimizing over the $\pi(z, x)$ missingness probabilities in the denominator). We introduce the partial Riesz representer following our earlier insight as to the improved finite-sample performance of using balancing weights estimators on the final collected data. We also introduce $\bar{\alpha}$ to account for mispecification of the nuisance function. Under the correct specification of this nuisance function, $\bar{\alpha} = \alpha$.

The following estimator for continuous treatments with missing outcomes is a direct extension of (Kallus and Zhou, 2018; Colangelo and Lee, 2020), that replaces the indicator function $\mathbb{I}[Z = z]$ with a local kernel function smoother localizing around z, $K_h(Z - z)$:

$$\mathbb{E}[Y(z)] = E[\psi_z(\alpha, \mu)]$$

where,

$$\psi_z(\alpha, \mu) = \mu(z, X_i) + \frac{K_h(Z_i - z)\mathbb{I}[R=1]\alpha(z, X_i)}{\pi(z, X_i)} (Y_i - \mu(z, X_i)). \tag{3}$$

and

$$\alpha(z, x) = \frac{1}{f_{Z|X}(z|x)}.$$

Here $f_{Z|X}(z|x)$ is defined as conditional probability density of treatment given covariates and later we will use $f_{ZX}(z, x)$ to refer to the joint distribution between treatments and covariates.

Following our analysis in the binary treatment setting, we derive the asymptotic variance of this estimator. In the continuous treatment setting, the asymptotic variance does incur bias and we derive the expressions of both the variance and bias terms in the following proposition.

Proposition 2. *The asymptotic variance (AVar) for the continuous treatment setting is:*

$$AVar = V_z + B_z,$$

where $V_z \equiv h^{-1}\mathbb{E}\left[\frac{\bar{\alpha}^2(z,x)}{\pi(z,x)} f_{Z|X}(z \mid x)\mathbb{E}\left[(Y - \mu(z, x))^2 \mid Z = z, X = x\right]\right]\xi_k$ *and* $B_z \equiv h^4\left(\left[2\frac{d}{dz}\bar{\mu}(z, X)\frac{d}{dz}f_{Z|X}(z|x) + f_{Z|X}(z|x)\frac{d^2}{dz}\bar{\mu}(z, X) + (\bar{\mu}(z, X) - \mu(z, X))\frac{d^2}{dz}f_{Z|X}(z|x)\right]\kappa\right)^2.$

Most notably, we see that the bias term does not depend on $\pi(z, x)$. Therefore, we can focus our optimization on V_z with respect to $\pi(z, x)$.

For this optimization procedure, we consider the same assumptions required as in (Colangelo and Lee, 2020), standard in kernel density estimation analysis such as sufficient smoothness of the underlying function and kernel function, and rate conditions $h \to 0, nh \to \infty, nh^4 \to C \in [0, \infty)$. Suppose that $\alpha(z, X)$ is well-specified. Let $\sigma^2(z, x) = \mathbb{E}\left[(Y - \mu(z, X))^2 \mid Z = z, X = x\right]$. We need to optimize the expression for variance that explicitly has the integration over K_h. The objective function arises from the asymptotic variance expression in (Colangelo and Lee, 2020, Thm. 3); it follows readily from following

their proof of Thm. 3 with our analysis of the asymptotic variance as in Proposition 1. The proof of the optimal solution follows our analysis in Theorem 1 with a few slightly different expressions. The optimization problem can be written as follows:

$$\pi^*(z,x) \in \arg\min_{\pi(z,x)} \int_{\mathcal{X}} \int_{Z_0} \frac{K_h^2(s-z)\bar{\alpha}^2(s,x)}{\pi(s,x)} \sigma^2(s,x) f_{ZX}(s,x) ds dx$$

The closed-form solution form solution for the optimal annotation probability for the continuous treatments case is:

$$\pi^*(z,X) = \frac{K_h(Z-z)\bar{\alpha}(z,X)\sqrt{\sigma^2(z,X)}}{\mathbb{E}\left[K_h(Z-z)\sqrt{\bar{\alpha}^2(Z,X)\sigma^2(Z,X)}\right]} B.$$

E PROOFS

E.1 Optimal annotation probability analysis

Proof of Proposition 1 . We simplify the expression for the asymptotic variance of the ATE with missing outcomes to isolate the components affected by the data annotation probability.

First the variance of the ATE defined in terms of the efficient influence function ψ_z for $z \in \{0,1\}$ is

$$\begin{aligned}
\mathrm{Var}[\psi_1 - \psi_0] &= \mathrm{Var}\left[\frac{\mathbb{1}[Z=1]\cdot R\cdot[Y-\mu_1(X)]}{e_1(X)\cdot\pi(1,X)} + \mu_1(X) - \frac{\mathbb{1}[Z=0]\cdot R\cdot[Y-\mu_0(X)]}{e_0(X)\cdot\pi(0,X)} + \mu_0(X)\right] \\
&= \underbrace{\mathrm{Var}\left[\frac{\mathbb{1}[Z=1]\cdot R\cdot[Y-\mu_1(X)]}{e_1(X)\cdot\pi(1,X)} + \mu_1(X)\right]}_{V_1} + \underbrace{\mathrm{Var}\left[\frac{\mathbb{1}[Z=0]\cdot R\cdot[Y-\mu_0(X)]}{e_0(X)\cdot\pi(0,X)} + \mu_0(X)\right]}_{V_2} \\
&\quad \underbrace{- 2\mathrm{Cov}\left[\frac{\mathbb{1}[Z=1]\cdot R\cdot[Y-\mu_1(X)]}{e_1(X)\cdot\pi(1,X)} + \mu_1(X), \frac{\mathbb{1}[Z=0]\cdot R\cdot[Y-\mu_0(X)]}{e_0(X)\cdot\pi(0,X)} + \mu_0(X)\right]}_{V_3}
\end{aligned}$$

For V_3:

$$\begin{aligned}
&2\mathrm{Cov}\left[\frac{\mathbb{1}[Z=1]\cdot R\cdot[Y-\mu_1(X)]}{e_1(X)\cdot\pi(1,X)} + \mu_1(X), \frac{\mathbb{1}[Z=0]\cdot R\cdot[Y-\mu_0(X)]}{e_0(X)\cdot\pi(0,X)} + \mu_0(X)\right] \\
=&2\left[\mathbb{E}\left[\frac{\mathbb{1}[Z=1]\cdot R}{e_1(X)\cdot\pi(1,X)}[\underbrace{\mathbb{E}[Y|Z=1,R=1,X] - \mu_1(X)}_{=0}]\right]\right] \\
&+\left[\mathbb{E}\left[\mu_1(X)\cdot\frac{\mathbb{1}[Z=0]\cdot R}{e_0(X)\cdot\pi(0,X)}[\underbrace{\mathbb{E}[Y|Z=0,R=1,X] - \mu_0(X)}_{=0}] + \mu_0(X)\right]\right] \\
&-\mathbb{E}\left[\frac{\mathbb{1}[Z=1]\cdot R}{e_1(X)\cdot\pi(1,X)}[\underbrace{\mathbb{E}[Y|Z=1,R=1,X] - \mu_1(X)}_{=0}] + \mu_1(X)\right] \\
&\times\mathbb{E}\left[\frac{\mathbb{1}[Z=0]\cdot R}{e_0(X)\cdot\pi(0,X)}[\underbrace{\mathbb{E}[Y|Z=0,R=1,X] - \mu_0(X)}_{=0}] + \mu_0(X)\right]\Bigg] \\
=&2\Big[\mathbb{E}[\mu_1(X)\cdot\mu_0(X)] - \mathbb{E}[\mu_1(X)\mu_0(X)]\Big]
\end{aligned}$$

For V_1**:**

$$\mathrm{Var}\left[\frac{\mathbb{1}[Z=1]\cdot R\cdot[Y-\mu_1(X)]}{e_1(X)\cdot\pi(1,X)} + \mu_1(X)\right]$$

$$= \text{Var}\left[\frac{\mathbb{1}[Z=1]\cdot R\cdot[Y-\mu_1(X)]}{e_1(X)\cdot\pi(1,X)}\right] + \text{Var}[\mu_1(X)] + 2\underbrace{\text{Cov}\left[\frac{\mathbb{1}[Z=1]\cdot R\cdot[Y-\mu_1(X)]}{e_1(X)\cdot\pi(1,X)}, \mu_1(X)\right]}_{=0}$$

$$= \mathbb{E}\left[\left[\frac{\mathbb{1}[Z=1]\cdot R\cdot[Y-\mu_1(X)]}{e_1(X)\cdot\pi(1,X)}\right]^2\right] - \left[\frac{\mathbb{1}[Z=1]\cdot R\cdot}{e_1(X)\cdot\pi(1,X)}[\underbrace{\mathbb{E}[Y|Z=1,R=1,X]-\mu_1(X)}_{=0}]\right]^2$$

$$+ \mathbb{E}\left[\mu_1(X)^2\right] - \mathbb{E}\left[\mu_1(X)\right]^2$$

$$= \mathbb{E}\left[\left[\frac{\mathbb{1}[Z=1]^2\cdot R^2}{e_1^2(X)\cdot\pi^2(1,X)}\cdot[Y-\mu_1(X)]^2\right]\right] + \mathbb{E}\left[\mu_1(X)^2\right] - \mathbb{E}\left[\mu_1(X)\right]^2$$

$$= \mathbb{E}\left[\frac{\mathbb{1}[Z=1]\cdot R}{e_1^2(X)\cdot\pi^2(1,X)}\cdot[Y-\mu_1(X)]^2\right] + \mathbb{E}\left[\mu_1(X)^2\right] - \mathbb{E}\left[\mu_1(X)\right]^2$$

$$= \mathbb{E}\left[\frac{1}{e_1(X)\cdot\pi(1,X)}\cdot[Y-\mu_1(X)]^2\right] + \mathbb{E}\left[\mu_1(X)^2\right] - \mathbb{E}\left[\mu_1(X)\right]^2$$

Lastly, $V_1 = V_2$. So the full variance term is

$$\begin{aligned}\text{Var}[\psi_1-\psi_0] &= \mathbb{E}\left[\frac{1}{e_1(X)\cdot\pi(1,X)}\cdot[Y-\mu_1(X)]^2\right] + \mathbb{E}\left[\frac{1}{e_0(X)\cdot\pi(0,X)}\cdot[Y-\mu_0(X)]^2\right] \\ &\quad + \mathbb{E}\left[(\mu_1(X)-\mu_0(X))^2\right] - \mathbb{E}\left[\mu_1(X)-\mu_0(X)\right]^2 \\ &= \mathbb{E}\left[\frac{1}{e_1(X)\cdot\pi(1,X)}\cdot[Y-\mu_1(X)]^2\right] + \mathbb{E}\left[\frac{1}{e_0(X)\cdot\pi(0,X)}\cdot[Y-\mu_0(X)]^2\right] \\ &\quad + \text{Var}\left[\mu_1(X)-\mu_0(X)\right]\end{aligned}$$

Rewriting the bound from Hahn (1998), we get

$$\begin{aligned}V &\geq \mathbb{E}\left[\frac{1}{e_1(X)\cdot\pi(1,X)}\cdot[Y-\mu_1(X)]^2\right] + \mathbb{E}\left[\frac{1}{e_0(X)\cdot\pi(0,X)}\cdot[Y-\mu_0(X)]^2\right] \\ &\quad + \text{Var}\left[\mu_1(X)-\mu_0(X)\right]\end{aligned}$$

□

Proof of Theorem 3 . Finding the optimal π can be separated into sub-problems for each treatment $z \in \{0,1\}$, since the objective and dual variables are separable across z. We first look at a solution for $\pi(z,X)$ for a given z:

$$\begin{aligned}\min_{\pi(z,x)} \mathbb{E}\left[\frac{\sigma_z^2(X)}{e_z(X)\pi(z,X)}\right] \\ \text{s.t. } \mathbb{E}\left[\pi(z,X) \mid Z=z\right] \leq B_z, \\ 0 < \pi(z,x) \leq 1,\ \forall x\end{aligned} \tag{z-budget}$$

We define the Lagrangian of the optimization problem and introduce dual variables λ for the budget constraint and η and ν for the the constraint that $0 < \pi(z,X) \leq 1$:

$$\mathcal{L} = \mathbb{E}\left[\frac{(Y-\mu_z(X))^2}{e_z(X)\pi(z,X)}\right] + \lambda_z(\mathbb{E}\left[\pi(z,X) \mid Z=z\right] - B_z) + \sum_{x\in\mathcal{X}}(\nu_x^z(\pi(z,x)-1) - \eta_x^z\pi(z,x))$$

Define the conditional outcome variance $\sigma^2(X) = \mathbb{E}\left[(Y-\mu(z,1,X))^2|X\right]$. Note that by iterated expectations,

$$\mathcal{L} = \mathbb{E}\left[\frac{\sigma_z^2(X)}{e_z(X)\pi(z,X)}\right] + \lambda_z(\mathbb{E}\left[\pi(z,X) \mid Z=z\right] - B_z) + \sum_{x\in\mathcal{X}}(\nu_x^z(\pi(z,x)-1) - \eta_x^z\pi(z,x))$$

We can find the optimal solution by setting the derivative equal to 0. Since $p(X = x \mid Z = z) = \frac{e_z(x)p(x)}{p(Z=z)}$

$$\frac{\partial \mathcal{L}}{\partial \pi(z,X)} = -\frac{\sigma^2(X)}{e_z(X)(\pi^2(z,X))}p(x) + \lambda_z \frac{e_z(x)p(x)}{p(Z=z)} + \nu_x - \eta_x = 0, \text{ where } p(x) > 0$$

$$= -\frac{\sigma^2(X)}{e_z^2(X)\pi^2(z,X)} + \frac{\lambda_z}{p(Z=z)} + \frac{(\nu_x^z - \eta_x^z)}{p(x)e_z(x)} = 0$$

Therefore

$$\pi(z,x) = \sqrt{\frac{\sigma^2(x)}{e_z^2(x)(\frac{\lambda_z}{p(Z=z)} + \frac{(\nu_x^z - \eta_x^z)}{p(x)e_z(x)})}}$$

Next we give a choice of λ that results in an interior solution with $0 \leq \pi(z,x) \leq 1$, so that ν_x^z, η_x^z can be set to 0 without loss of generality to satisfy complementary slackness.

We posit a closed form solution

$$\pi^*(z,X) = \frac{\sqrt{\sigma_z^2(X)/e_z^2(X)}}{\mathbb{E}\left[\sqrt{\sigma_z^2(X)/e_z^2(X)} \mid Z = z\right]} \cdot B_z$$

.

Note that this solution is self-normalized to satisfy the budget constraint such that

$$\mathbb{E}\left[\pi^*(z,X)\mathbb{I}[Z=z]\right] = \mathbb{E}\left[\frac{\sqrt{\sigma^2(X)/e_z^2(X)}}{\mathbb{E}\left[\sqrt{\sigma_z^2(X)/e_z^2(X)} \mid Z = z\right]} B_z \mid Z = z\right] = B_z$$

This solution corresponds to a choice of $\lambda_z^* = p(Z=z)\mathbb{E}\left[\sqrt{\sigma^2(X)/e_z^2(X)}|Z=z\right]^2/B_z^2$ in the prior parametrized expression.

$$\pi_\lambda(z,X) = \pi^*(z,X)$$

$$\sqrt{\frac{\sigma_z^2(X)}{e_z^2(X)\frac{\lambda}{p(Z=z)}}} = \frac{\sqrt{\sigma_z^2(X)/e_z^2(X)}}{\mathbb{E}\left[\sqrt{\sigma_z^2(X)/e_z^2(X)} \mid Z = z\right]} \cdot B_z$$

We can check that the KKT conditions are satisfied at $\pi^*(z,X)$ and λ^*. We note that since $\pi^*(z,X)$ is an interior solution then w.l.o.g we can fix $\nu_x, \eta_x = 0$ to satisfy complementary slackness.

It remains to check that $\frac{\partial \mathcal{L}}{\partial \pi^*(z,X)} = 0$, we have that:

$$\frac{\partial \mathcal{L}}{\partial \pi(z,X)} = -\frac{\sigma_z^2(X)}{e_z(X)} \cdot \frac{e_z^2(X)\mathbb{E}\left[\sqrt{\sigma_z^2(X)/e_z(X)} \mid Z = z\right]^2}{\sigma_z^2(X) \cdot B_z^2} + \frac{\mathbb{E}\left[\sqrt{\sigma^2(X)/e_z(X)} \mid Z = z\right]^2 \sigma_z^2(X)e_z(X)}{\sigma_z^2(X) \cdot B_z^2} + 0 = 0.$$

Thus we have shown that $\pi^*(z,X)$ is optimal.

□

Proof of Theorem 1 . Proceed as in the proof of Theorem 3.

The Lagrangian of the optimization problem (with a single global budget constraint) is:

$$\mathcal{L} = \sum_{z \in \{0,1\}} \mathbb{E}\left[\frac{(Y - \mu_z(X))^2}{e_z(X)\pi(z,X)}\right] + \sum_{x \in \mathcal{X}} (\nu_x^z(\pi(z,x) - 1) - \eta_x^z \pi(z,x))$$

$$+ \lambda(\mathbb{E}\left[\pi(1,X)\mathbb{I}[Z=1] + \pi(0,X)\mathbb{I}[Z=0]\right] - B)$$

Again by iterated expectations,

$$\mathcal{L} = \mathbb{E}\left[\frac{\sigma_z^2(X)}{e_z(X)\pi(z,X)}\right] + \lambda(\mathbb{E}\left[\pi(1,X)e_1(X) + \pi(0,X)e_0(X)\right] - B_z) + \sum_{x\in\mathcal{X}}(\nu_x^z(\pi(z,x)-1) - \eta_x^z\pi(z,x))$$

We can find the optimal solution by setting the derivative equal to 0.

$$\begin{aligned}\frac{\partial\mathcal{L}}{\partial\pi(z,X)} &= -\frac{\sigma^2(X)}{e_z(X)(\pi^2(z,X))}p(x) + \lambda p(x)e_z(x) + \nu_x^z - \eta_x^z = 0, \text{ where } p(x) > 0 \\ &= -\frac{\sigma^2(X)}{e_z^2(X)\pi^2(z,X)} + \lambda + \frac{(\nu_x^z - \eta_x^z)}{p(x)e_z(x)} = 0\end{aligned}$$

Therefore we obtain a similar expression parametrized in λ, but this parameter is the same across both groups under a global budget.

$$\pi(z,x) = \sqrt{\frac{\sigma^2(x)}{e_z^2(x)(\lambda + \frac{(\nu_x^z - \eta_x^z)}{p(x)e_z(x)})}}$$

We can similarly give a closed-form expression for a different choice of λ yielding an interior solution, so that we can set $\nu_x^z, \eta_x^z = 0$ without loss of generality.

$$\lambda = \frac{\mathbb{E}\left[\mathbb{I}[Z=1]\sqrt{\sigma_1^2(X)/e_1^2(X)} + \mathbb{I}[Z=0]\sqrt{\sigma_0^2(X)/e_0^2(X)}\right]^2}{B^2}$$

Notice that this satisfies the normalization requirement that $\mathbb{E}[\pi^\lambda(1,X)\mathbb{I}[Z=1] + \pi^\lambda(0,X)\mathbb{I}[Z=0]] \leq B$, and similarly note that the partial derivatives with respect to $\pi(z,x)$ are 0. □

Proof of Proposition 2 . We simplify the expression for the asymptotic variance of the ATE with missing outcomes and continuous treatments. We derive the variance and the bias terms and isolate the components affected by the data annotation probability. Again, here $f_{Z|X}(z|x)$ is defined as conditional probability density of treatment given covariates and later we will use $f_{ZX}(z,x)$ to refer to the joint distribution between treatments and covariates. And the "partial" Riesz representer is $\alpha(z,x) = \frac{1}{f_{Z|X}(z,x)}$ and we introduce $\bar{\alpha}$ to account for mispecification.

$$\begin{aligned}\text{Var}[\psi_z] &= \text{Var}\left[\mu(z,X) + \frac{K_h(Z-z)\alpha(z,X)R}{\pi(z,X)}(Y - \mu(z,X))\right] \\ &= \text{Var}\left[\frac{K_h(Z-z)\alpha(z,X)R}{\pi(z,X)}(Y - \mu(z,X))\right] + \text{Var}[\mu(z,X)] \\ &\quad \underbrace{+ 2\text{Cov}\left[\frac{K_h(Z-z)\alpha(z,X)R}{\pi(z,X)}(Y - \mu(z,X)), \mu(z,X)\right]}_{=0}\end{aligned}$$

We focus on the first term as it is the part that depends on $\pi(z,x)$:

$$V = V\left[\mathbb{E}\left[\frac{K_h(Z-z)\alpha(z,X)R}{\pi(z,X)}(Y - \mu(z,X))\right]\right] + \mathbb{E}\left[V\left[\frac{K_h(Z-z)\alpha(z,X)R}{\pi(z,X)}(Y - \mu(z,X))\right]\right]$$

(Law of total variance)

$$+\mathbb{E}\left[\left(\mathbb{E}\left[\frac{K_h(Z-z)\alpha(z,X)R}{\pi(z,X)}(Y-\mu(z,X))\right]\right)^2\right]-\left(\mathbb{E}\left[\mathbb{E}\left[\frac{K_h(Z-z)\alpha(z,X)R}{\pi(z,X)}(Y-\mu(z,X))\right]\right]\right)^2$$

$$+\mathbb{E}\left[\mathbb{E}\left[\left(\frac{K_h(Z-z)\alpha(z,X)R}{\pi(z,X)}(Y-\mu(z,X))\right)^2\right]\right]-\mathbb{E}\left[\left(\mathbb{E}\left[\frac{K_h(Z-z)\alpha(z,X)R}{\pi(z,X)}(Y-\mu(z,X))\right]\right)^2\right]$$

$$=\underbrace{\mathbb{E}\left[\mathbb{E}\left[\left(\frac{K_h(Z-z)\alpha(z,X)R}{\pi(z,X)}(Y-\mu(z,X))\right)^2\right]\right]}_{V_z}-\underbrace{\left(\mathbb{E}\left[\mathbb{E}\left[\frac{K_h(Z-z)\alpha(z,X)R}{\pi(z,X)}(Y-\mu(z,X))\right]\right]\right)^2}_{B_z}$$

(canceled out first and fourth term of expansion)

For V_z:

$$\mathbb{E}\left[\mathbb{E}\left[\left(\frac{K_h(Z-z)\bar{\alpha}(z,X)R}{\pi(z,X)}(Y-\mu(z,X))\right)^2\right]\right]=\mathbb{E}\left[\mathbb{E}\left[\frac{K_h^2(Z-z)\bar{\alpha}^2(z,X)R^2}{\pi^2(z,X)}(Y-\mu(z,X))^2\right]\right]$$

$$=\int_{\mathcal{X}}\int_{Z_0}\frac{K_h^2(s-z)\bar{\alpha}^2(s,x)R^2}{\pi^2(s,x)}\mathbb{E}\left[(Y-\mu(s,x))^2|Z=s,X=x\right]f_{ZX}(s,x)dsdx$$

$$=h^{-1}\int_{\mathcal{X}}\int_{\mathcal{Q}}\frac{k^2(u)\bar{\alpha}^2(s,x)R^2}{\pi^2(s,x)}\mathbb{E}\left[(Y-\mu(s,x))^2\mid Z=z+uh,X=x\right]f_{ZX}(z+uh,x)dudx$$

(change of variables: u=s-z/h,s=z+uh)

$$=h^{-1}\int_{\mathcal{X}}\int_{\mathcal{Q}}\left(\mathbb{E}\left[(Y-\mu_(z,x))^2\mid Z=z,X=x\right]+uh\frac{d}{dz}\mathbb{E}\left[(Y-\mu(z,x))^2\mid Z=z,X=x\right]\Big|_{z=\bar{z}}\right)$$

$$\times\left(f_{ZX}(z,x)+uh\frac{d}{dz}f_{ZX}(z,x)\Big|_{z=z'}\right)k^2(u)\frac{\bar{\alpha}^2(z,x)R^2}{\pi^2(z,x)}dudx$$

(taylor expansion and mean value theorem for $\bar{z}, z'$ between $z, z+uh$)

$$=h^{-1}\int_{\mathcal{X}}\frac{\bar{\alpha}^2(z,x)R^2}{\pi^2(z,x)}\left[\int_{\mathcal{R}}k^2(u)\mathbb{E}\left[(Y-\mu(z,x))^2\mid Z=z,X=x\right]f_{ZX}(z,x)+o(h^2)du\right]dx$$

$$=h^{-1}\int_{\mathcal{X}}\frac{\bar{\alpha}^2(z,x)R^2}{\pi^2(z,x)}\xi_k\mathbb{E}\left[(Y-\mu(z,x))^2\mid Z=z,X=x\right]f_{ZX}(z,x)dx+o(h^2)\qquad(\xi_k\equiv\textstyle\int k^2(u))$$

$$=h^{-1}\mathbb{E}\left[\frac{\bar{\alpha}^2(z,x)R^2}{\pi^2(z,x)}f_{Z|X}(z\mid x)\mathbb{E}\left[(Y-\mu(z,x))^2\mid Z=z,X=x\right]\right]\xi_k$$

$$=h^{-1}\mathbb{E}\left[\frac{\bar{\alpha}^2(z,x)}{\pi(z,x)}f_{Z|X}(z\mid x)\mathbb{E}\left[(Y-\mu(z,x))^2\mid Z=z,X=x\right]\right]\xi_k$$

$(\mathbb{E}\left[R^2|X\right]=\mathbb{E}\left[R|X\right]=\pi(z,x))$

For B_z:

$$\left(\mathbb{E}\left[\mathbb{E}\left[\frac{K_h(Z-z)\bar{\alpha}(z,X)R}{\pi(z,X)}(Y-\mu(z,X)\right]\right]\right)^2=\left(\mathbb{E}\left[\frac{\bar{\alpha}(z,X)R}{\pi(z,X)}\mathbb{E}\left[K_h(Z-z)(Y-\mu(z,X))\right]\right]\right)^2$$

$$=\left(\mathbb{E}\left[\bar{\alpha}(z,X)\underbrace{\mathbb{E}\left[K_h(Z-z)(Y-\mu(z,X))\mid Z=z,R=1,X\right]}\right]\right)^2\qquad(\mathbb{E}\left[R|X\right]=\pi(z,x))$$

$$=\int_{\mathcal{Z}_0}K_h(s-z)(\bar{\mu}(z,X)-\mu(z,X))f_{Z|X}(s|x)ds\qquad(\bar{\mu}(z,X)=\mathbb{E}\left[Y\mid Z=z,R=1,X\right])$$

$$=\int_{\mathcal{Q}}k(u)(\bar{\mu}(z+uh,X)-\mu(z,X))f_{Z|X}(z+uh|x)du\qquad\text{(change of variables)}$$

$$
\begin{aligned}
&= \int_Q \left((\bar{\mu}(z,X) - \mu(z,X)) + uh\frac{d}{dz}\bar{\mu}(z,X) + \frac{u^2h^2}{2}\frac{d^2}{dz}\bar{\mu}(z,X) \right) \\
&\times \left(f_{Z|X}(z|x) + uh\frac{d}{dz}f_{Z|X}(z|x) + \frac{u^2h^2}{2}\frac{d^2}{dz}f_{Z|X}(z|x) \right) && \text{(taylor expansion)} \\
&\times k(u)du + O(h^3) \\
&= \int_Q (\bar{\mu}(z,X) - \mu(z,X)) f_{Z|X}(z|x)k(u)\,du \\
&\quad + h\Bigg[\underbrace{(\bar{\mu}(z,X) - \mu(z,X))uk(u)\frac{d}{dz}f_{Z|X}(z|x)}_{\int uk(u)\,du=0} + \underbrace{f_{Z|X}(z|x)uk(u)\frac{d}{dz}\bar{\mu}_z(X)}_{\int uk(u)\,du=0} \Bigg] \\
&\quad + h^2\Bigg[\frac{1}{2}(\bar{\mu}(z,X) - \mu(z,X))u^2k(u)\frac{d^2}{dz}f_{Z|X}(z|x) + \frac{1}{2}u^2k(u)f_{Z|X}(z|x)\frac{d^2}{dz}\bar{\mu}(z,X) \\
&\qquad + u^2k(u)\frac{d}{dz}\bar{\mu}(z,X)\frac{d}{dz}f_{Z|X}(z|x) \Bigg] + O(h^3) \\
&= (\bar{\mu}(z,X) - \mu_z(X))f_{Z|X}(z|x) + h^2\Bigg[\frac{d}{dz}\bar{\mu}(z,X)\frac{d}{dz}f_{Z|X}(z|x) + \frac{1}{2}f_{Z|X}(z|x)\frac{d^2}{dz}\bar{\mu}(z,X) \\
&\qquad + \frac{1}{2}(\bar{\mu}(z,X) - \mu_z(X))\frac{d^2}{dz}f_{z|X}(z|x) \Bigg] \\
&\times \int_{-\infty}^{\infty} u^2k(u)du + O(h^3) \\
&= \underbrace{\mathbb{E}\left[(\bar{\mu}(z,X) - \mu(z,X)) f_{Z|X}(z|x)\bar{\alpha}(z,x) \right]^2}_{=0} + h^4 \Bigg(\Bigg[2\frac{d}{dz}\bar{\mu}(z,X)\frac{d}{dz}f_{Z|X}(z|x) + f_{Z|X}(z|x)\frac{d^2}{dz}\bar{\mu}(z,X) \\
&\qquad + (\bar{\mu}(z,X) - \mu(z,X))\frac{d^2}{dz}f_{Z|X}(z|x) \Bigg]\kappa \Bigg)^2 && (\kappa \equiv \textstyle\int u^2k(u)du)
\end{aligned}
$$

□

Proof of **??** . The objective function arises from the asymptotic variance expression in (Colangelo and Lee, 2020, Thm. 3); it follows readily from following their proof of Thm. 3 with our analysis of the asymptotic variance as in Proposition 1. The proof of the optimal solution follows our analysis in Theorem 1 with a few slightly different expressions, discussed as follows.

The Lagrangian can be written as follows:

$$
\begin{aligned}
\mathcal{L} = &\int_{\mathcal{X}}\int_{Z_0} \frac{K_h^2(s-z)\bar{\alpha}^2(s,x)}{\pi(s,x)}\sigma^2(s,x)f_{ZX}(s,x)\,ds\,dx \\
&+ \lambda\int\int(\pi(s,x) - B)f_{zx}(s,x)\,ds\,dx \\
&+ \nu\int\int(\pi(s,x) - 1)f_{ZX}(s,x)\,ds\,dx + \eta\int\int(-\pi(s,x))f_{ZX}(s,x)\,ds\,dx
\end{aligned}
$$

We can take the pointwise derivative w.r.t. $\pi(s,x)$ to obtain the FOC

$$
\frac{\partial\mathcal{L}}{\partial\pi(s,x)} = \frac{-K_h^2(s-z)\bar{\alpha}^2(s,x)\sigma^2(s,x)}{\pi^2(s,x)}f_{ZX}(s,x) + (\lambda + \nu - \eta)f_{ZX}(s,x) = 0
$$

Solving the FOC, we obtain

$$(\lambda+\nu-\eta)f_{ZX}(s,x)=\frac{K_h^2(s-z)\bar{\alpha}^2(s,x)\sigma^2(s,x)}{\pi^2(s,x)}f_{ZX})(s,x)$$

$$\sqrt{\pi^2(s,x)}=\sqrt{\frac{K_h^2(s-z)\bar{\alpha}^2(s,x)\sigma^2(s,x)}{\lambda+\nu-\eta}}$$

$$\pi^*(s,x)=\sqrt{\frac{K_h^2(s-z)\bar{\alpha}^2(s,x)\sigma^2(s,x)}{\lambda+\nu-\eta}}$$

We can solve for λ^* and set ν and η to be zero:

$$\mathbb{E}\left[\sqrt{\frac{K_h^2(Z-z)\bar{\alpha}^2(Z,X)\sigma^2(Z,X)}{\lambda}}\right]=B$$

$$\lambda^*=\frac{\mathbb{E}\left[K_h(Z-z)\sqrt{\bar{\alpha}^2(Z,X)\sigma^2(Z,X)}\right]^2}{B^2}$$

Then plug back into our optimal $\pi^*(Z,X)$,

$$\pi^*(Z,X)=\pi_\lambda(Z,X)=\sqrt{\frac{K_h^2(Z-z)\bar{\alpha}^2(Z,X)\sigma^2(Z,X)}{\frac{\mathbb{E}\left[K_h(Z-z)\sqrt{\bar{\alpha}^2(Z,X)\sigma^2(Z,X)}\right]^2}{B^2}}}=\frac{\sqrt{K_h^2(Z-z)\bar{\alpha}^2(Z,X)\sigma^2(Z,X)}}{\mathbb{E}\left[K_h(Z-z)\sqrt{\bar{\alpha}^2(Z,X)\sigma^2(Z,X)}\right]}B$$

We can check that $\frac{\partial\mathcal{L}}{\partial\pi^*}=0$

$$-\frac{K_h^2(Z-z)\bar{\alpha}^2(Z,X)\sigma^2(Z,X)}{\pi^2(Z,X)}+(\lambda+\nu-\eta)=0$$

$$-\frac{K_h^2(Z-z)\bar{\alpha}^2(Z,X)\sigma^2(Z,X)}{\frac{K_h^2(Z-z)\bar{\alpha}^2(Z,X)\sigma^2(Z,X)}{\mathbb{E}\left[K_h(Z-z)\sqrt{\bar{\alpha}^2(Z,X)\sigma^2(Z,X)}\right]^2}B^2}+\frac{\mathbb{E}\left[K_h(Z-z)\sqrt{\bar{\alpha}^2(Z,X)\sigma^2(Z,X)}\right]^2}{B^2}=0$$

$$-\frac{\mathbb{E}\left[K_h(Z-z)\sqrt{\bar{\alpha}^2(Z,X)\sigma^2(Z,X)}\right]^2}{B^2}+\frac{\mathbb{E}\left[K_h(Z-z)\sqrt{\bar{\alpha}^2(Z,X)\sigma^2(Z,X)}\right]^2}{B^2}=0$$

□

E.2 Estimation analysis

Proof of Theorem 2 . **Proof sketch.**

The proof proceeds in two steps. The first establishes that the feasible AIPW estimator converges to the AIPW estimator with oracle nuisances. It follows from standard analysis with cross-fitting, in particular the variant used across batches.

Preliminaries In the analysis, we write the score function as a function of R in addition to other nuisance functions:

$$\psi_{z,i}(R_i,e,\pi,\mu)=\frac{\mathbb{I}[Z_i=z]R_i(Y_i-\mu_z(X_i))}{e_z(X_i)\pi(z,X_i)}+\mu_z(X_i)$$

The AIPW estimator can be rewritten as a sum over estimators within batch-t, fold-k, $\hat{\tau}^{(t,k)}_{AIPW}$, as follows:

$$\hat{\tau}_{AIPW} = \sum_{t=1}^{2}\sum_{k=1}^{K}\frac{n_{t,k}}{n}\sum_{(t,i)\in\mathcal{I}_k}\frac{1}{n_{t,k}}\{\hat{\psi}_{1,i}(R,\hat{e},\hat{\pi},\hat{\mu}) - \hat{\psi}_{0,i}(R,\hat{e},\hat{\pi},\hat{\mu})\} = \sum_{t=1}^{2}\sum_{k=1}^{K}\frac{n_{t,k}}{n}\hat{\tau}^{(t,k)}_{AIPW}$$

We introduce an intermediate quantity. The realized treatments are sampled with probability $\hat{\pi}(X_i)$, $R_i \sim Bern(\hat{\pi}(Z_i, X_i))$. In the asymptotic framework, we study treatments sampled from a mixture distribution over the two batches, $\tilde{R}_i \sim Bern(\pi^*(Z_i, X_i))$.

$$\tilde{\tau}_{AIPW} = \sum_{t=1}^{2}\sum_{k=1}^{K}\frac{n_{t,k}}{n}\sum_{(t,i)\in\mathcal{I}_k}\frac{1}{n_{t,k}}\{\hat{\psi}_{1,i}(\tilde{R},\hat{e},\hat{\pi},\hat{\mu}) - \hat{\psi}_{0,i}(R,\hat{e},\hat{\pi},\hat{\mu})\}$$

We also denote the AIPW estimator with oracle nuisances, $\hat{\tau}^{*}_{AIPW}$, as

$$\hat{\tau}^{*}_{AIPW} = \sum_{t=1}^{2}\sum_{k=1}^{K}\frac{n_{t,k}}{n}\sum_{(t,i)\in\mathcal{I}_k}\frac{1}{n_{t,k}}\{\psi_{1,i}(\tilde{R}_i,e,\pi,\mu) - \psi_{0,i}(\tilde{R}_i,e,\pi,\mu)\}$$

We study convergence within a batch$-t$, fold$-k$ subset; the decompositions above give that convergence also holds for the original estimators.

The first step studies the limiting mixture distribution propensity arising from the two-batch process and shows that the use of the double-machine learning estimator (AIPW), under the weaker product error assumptions, gives that the oracle estimator is asymptotically equivalent to the oracle estimator where missingness follows the limiting mixture missingness probability. The latter of these is a sample average of iid terms and follows a standard central limit theorem. Recalling that $\tilde{R}_i = \mathbb{I}[U_i \geq \pi^*(X_i)]$, we wish to show:

$$\sum_{z}\mathbb{E}_n[\psi_{z,i}(R,\hat{e},\hat{\pi},\hat{\mu})] - \mathbb{E}_n[\psi_{z,i}(\tilde{R},e,\pi,\mu)] = o_p(n^{-\frac{1}{2}}).$$

Next we show that the estimator with feasible nuisance estimators converges to the estimator with oracle knowledge of the nuisance functions

$$\sqrt{n}(\tilde{\tau}^{(t,k)}_{AIPW} - \hat{\tau}^{*,(t,k)}_{AIPW}) \to_p 0.$$

The result follows by the standard limit theorem applied to the estimator with oracle nuisance functions.

Step 1

Let $\tilde{R}_i = \mathbb{I}[U_i \geq \pi^*(Z_i, X_i)]$. Restricting attention to a single treatment value $z \in \{0, 1\}$, we want to show that:

$$\sum_{t=1}^{2}\sum_{k=1}^{K}\frac{n_{t,k}}{n}\sum_{(t,i)\in\mathcal{I}_k}\frac{1}{n_{t,k}}\left\{\hat{\psi}_{1,i}(\tilde{R},\hat{e},\hat{\pi},\hat{\mu}) - \hat{\psi}_{1,i}(R,\hat{e},\hat{\pi},\hat{\mu})\right\}$$

$$= \sum_{t=1}^{2}\sum_{k=1}^{K}\frac{n_{t,k}}{n}\sum_{(t,i)\in\mathcal{I}_k}\frac{1}{n_{t,k}}\left\{\frac{\mathbb{I}[Z_i=z]\tilde{R}_i(Y_i-\hat{\mu}_z(X_i))}{\hat{e}_z(X_i)\hat{\pi}(z,X_i)} - \frac{\mathbb{I}[Z_i=z]R_i(Y_i-\hat{\mu}_z(X_i))}{\hat{e}_z(X_i)\hat{\pi}(z,X_i)}\right\} = o_p(n^{-1/2}).$$

Without loss of generality we further consider one summand on batch-t, fold-k data, the same argument will apply to the other summands and the final estimator.

Note that by consistency of potential outcomes, for any data point we have that

$$\frac{\mathbb{I}[Z_i=z]\tilde{R}_i(Y_i-\hat{\mu}_z(X_i))}{\hat{e}_z(X_i)\hat{\pi}(z,X_i)} - \frac{\mathbb{I}[Z_i=z]R_i(Y_i-\hat{\mu}_z(X_i))}{\hat{e}_z(X_i)\hat{\pi}(z,X_i)} = \frac{\mathbb{I}[Z_i=z](\tilde{R}_i-R_i)(Y_i(z)-\hat{\mu}_z(X_i))}{\hat{e}_z(X_i)\hat{\pi}(z,X_i)}$$

For each batch $t = 1, \ldots, T$ and fold $k = 1, \ldots, K$, according to the CSBAE crossfitting procedure, we observe that conditional on $\mathcal{I}_{(-k)}$ for a given batch and the observed covariates, the summands (namely

$R_i = \mathbb{I}[U_i \leq \hat{\pi}^{(-k)}(X_i)]$) are independent mean-zero. The final estimator will consist of the sum over batches and folds. We start by looking at the estimator over one batch t and one fold k and the rest follows for the other batches and folds.

$$\frac{1}{n_{t,k}} \sum_{(t,i)\in\mathcal{I}_k} \frac{\mathbb{I}[Z_i = z](\tilde{R}_i - R_i)(Y_i(z) - \hat{\mu}_z(X_i))}{\hat{e}_z(X_i)\hat{\pi}(z, X_i)}$$

$$= \frac{1}{n_{t,k}} \sum_{(t,i)\in\mathcal{I}_k} \frac{\mathbb{I}[Z_i = z]\left((\tilde{R}_i - \pi^*(z, X_i)) + (\pi^*(z, X_i) - \hat{\pi}(z, X_i)) + (\hat{\pi}(z, X_i) - R_i)\right)(Y_i(z) - \hat{\mu}_z(X_i))}{\hat{e}_z(X_i)\hat{\pi}(z, X_i)}$$

$$\leq \nu_e\gamma_{\sigma^2} \frac{1}{n_{t,k}} \sum_{(t,i)\in\mathcal{I}_k} \mathbb{I}[Z_i = z]\left((\tilde{R}_i - \pi^*(z, X_i)) + (\pi^*(X_i) - \hat{\pi}(z, X_i)) + (\hat{\pi}(z, X_i) - R_i)\right)(Y_i(z) - \hat{\mu}_z(X_i))$$

Applying Cauchy-Schwarz to each of these terms, we obtain product error rate terms. For the second term, we obtain that

$$\begin{aligned}
&\nu_e\gamma_{\sigma^2} \frac{1}{n_{t,k}} \sum_{(t,i)\in\mathcal{I}_k^z} (\pi^*(X_i) - \hat{\pi}(X_i))(Y_i(z) - \hat{\mu}_z(X_i)) \\
&\leq \nu_e\gamma_{\sigma^2} \sqrt{\frac{1}{n_{t,k}} \sum_{(t,i)\in\mathcal{I}_k^z} (\pi^*(X_i) - \hat{\pi}(X_i))^2} \sqrt{\frac{1}{n_{t,k}} \sum_{(t,i)\in\mathcal{I}_k^z} (Y_i(z) - \hat{\mu}_z(X_i))^2} \\
&= \nu_e\gamma_{\sigma^2} \left\|\pi^*(X_i) - \hat{\pi}(X_i)\right\|_{2,n} \left\|Y_i(z) - \hat{\mu}_z(X_i)\right\|_{2,n} \\
&= o_p(n^{-\frac{1}{2}}) \qquad \text{(Assumption 7)}
\end{aligned}$$

Analogously, we conclude that the first and third terms are $o_p(n^{-\frac{1}{2}})$, applying Cauchy-Schwarz to each of them in turn.

Step 2 (feasible estimator converges to oracle)

If we look at one term for one treatment and datapoint in the above (the rest follows for the others), we obtain the following decomposition into error and product-error terms:

$$\frac{Z_i\tilde{R}_i(Y_i - \hat{\mu}_1(X_i))}{\hat{e}_1(X_i)\hat{\pi}(1, X_i)} - \frac{Z_i\tilde{R}_i(Y_i - \mu_1(X_i))}{e_1(X_i)\pi(1, X_i)} + (\hat{\mu}_1(X_i) - \mu_1(X_i))$$

$$= (\mu_1(X_i) - \hat{\mu}_1(X_i))\left(\frac{Z_i\tilde{R}_i}{e_1(X_i)\pi(1, X_i)} - 1\right) + Z_i\tilde{R}_i(Y_i - \hat{\mu}_1(X_i))(\frac{1}{\hat{e}_1(X_i)\hat{\pi}(1, X_i)} - \frac{1}{e_1(X_i)\pi(1, X_i)})$$

(by $\pm\frac{Z_i\tilde{R}_i(Y_i - \hat{\mu}_1(X_i))}{e_1(X_i)\pi(1, X_i)}$)

$$= (\mu_1(X_i) - \hat{\mu}_1(X_i))\left(\frac{Z_i\tilde{R}_i}{e_1(X_i)\pi(1, X_i)} - 1\right) + Z_i\tilde{R}_i(Y_i - \mu_1(X_i))(\frac{1}{\hat{e}_1(X_i)\hat{\pi}(1, X_i)} - \frac{1}{e_1(X_i)\pi(1, X_i)})$$

$$+ Z_i\tilde{R}_i(\mu_1(X_i) - \hat{\mu}_1(X_i))(\frac{1}{\hat{e}_1(X_i)\hat{\pi}(1, X_i)} - \frac{1}{e_1(X_i)\pi(1, X_i)})$$

(by $\pm Z_i\tilde{R}_i\mu_1(X_i)(\frac{1}{\hat{e}_1(X_i)\hat{\pi}(1,X_i)} - \frac{1}{e_1(X_i)\pi(1,X_i)})$)

$$= (\mu_1(X_i) - \hat{\mu}_1(X_i))\left(\frac{Z_i\tilde{R}_i}{e_1(X_i)\pi(1, X_i)} - 1\right)$$

$$+ Z_i\tilde{R}_i(Y_i - \mu_1(X_i))\left(\hat{\pi}(1, X_i)^{-1}(\hat{e}_1(X_i)^{-1} - e_1(X_i)^{-1}) + e_1(X_i)^{-1}(\hat{\pi}(1, X_i)^{-1} - \pi(1, X_i)^{-1})\right)$$

$$+ Z_i\tilde{R}_i(\mu_1(X_i) - \hat{\mu}_1(X_i))\left(\hat{\pi}(1, X_i)^{-1}(\hat{e}_1(X_i)^{-1} - e_1(X_i)^{-1}) + e_1(X_i)^{-1}(\hat{\pi}(1, X_i)^{-1} - \pi(1, X_i)^{-1})\right)$$

(by $\pm\frac{1}{e\hat{\pi}}$)

We want to show that

$$\sqrt{n_{t,k}}(\hat{\tau}^{(t,k)}_{AIPW} - \hat{\tau}^{*,(t,k)}_{AIPW}) \to_p 0$$

Now that we have written out this expansion for one datapoints, we can write out this expansion within a batch-t, fold-k subset, and write out the cross-fitting terms for reference:

$$\begin{aligned}
&\sqrt{n_{t,k}}\left(\hat{\tau}^{(t,k)}_{AIPW} - \hat{\tau}^{*,(t,k)}_{AIPW}\right) \\
&= \frac{1}{\sqrt{n_{t,k}}} \sum_{i:(t,i)\in\mathcal{I}_k} (\mu_1(X_i) - \hat{\mu}_1^{(-k)}(1, X_i)) \left(\frac{Z_i\tilde{R}_i}{e_1(X_i)\pi(1,X_i)} - 1\right) \\
&\quad + \frac{1}{\sqrt{n_{t,k}}} \sum_{i:(t,i)\in\mathcal{I}_k} Z_i\tilde{R}_i(Y_i - \mu_1(X_i))\times \\
&\qquad\qquad \left(\hat{\pi}^{(-k)}(1,X_i)^{-1}(\hat{e}_1^{(-k)}(X_i)^{-1} - e_1(X_i)^{-1}) + e_1(X_i)^{-1}(\hat{\pi}^{(-k)}(1,X_i)^{-1} - \pi(1,X_i)^{-1})\right) \\
&\quad + \frac{1}{\sqrt{n_{t,k}}} \sum_{i:(t,i)\in\mathcal{I}_k} Z_i\tilde{R}_i(\mu_1(X_i) - \hat{\mu}_1^{(-k)}(1, X_i))\times \\
&\qquad\qquad \left(\hat{\pi}^{(-k)}(1,X_i)^{-1}(\hat{e}_1^{(-k)}(X_i)^{-1} - e_1(X_i)^{-1}) + e_1(X_i)^{-1}(\hat{\pi}^{(-k)}(1,X_i)^{-1} - \pi(1,X_i)^{-1})\right)
\end{aligned}$$

Bound for third term:

$$\begin{aligned}
&\frac{1}{\sqrt{n_{t,k}}} \sum_{i:(t,i)\in\mathcal{I}_k} Z_i\tilde{R}_i(\mu_1(X_i) - \hat{\mu}_1^{(-k)}(X_i))(\hat{\pi}^{(-k)}(1,X_i)^{-1}(\hat{e}_1^{(-k)}(X_i)^{-1} - e_1(X_i)^{-1}) \\
&\qquad\qquad + e_1(X_i)^{-1}(\hat{\pi}^{(-k)}(1,X_i)^{-1} - \pi(1,X_i)^{-1}) \\
&= \frac{1}{\sqrt{n_{t,k}}} \sum_{i:(t,i)\in\mathcal{I}_k} Z_i\tilde{R}_i\hat{\pi}^{(-k)}(1,X_i)^{-1}(\mu_1(X_i) - \hat{\mu}_1^{(-k)}(X_i))(\hat{e}_1^{(-k)}(X_i)^{-1} - e_1(X_i)^{-1}) \\
&\qquad\qquad + Z_i\tilde{R}_i e_1(X_i)^{-1}(\mu_1(X_i) - \hat{\mu}_1^{(-k)}(X_i))(\hat{\pi}^{(-k)}(1,X_i)^{-1} - \pi(1,X_i)^{-1}) \\
&\leq (\lambda_\pi + \nu_e)\frac{1}{\sqrt{n_{t,k}}} \sum_{i:(t,i)\in\mathcal{I}_k} (\mu_1(X_i) - \hat{\mu}_1^{(-k)}(X_i))(\hat{e}_1^{(-k)}(X_i)^{-1} - e_1(X_i)^{-1}) \\
&\qquad\qquad + (\mu_1(X_i) - \hat{\mu}_1^{(-k)}(X_i))(\hat{\pi}^{(-k)}(1,X_i)^{-1} - \pi(1,X_i)^{-1}) \\
&\leq (\lambda_\pi + \nu_e)\delta_n n^{-1/2}
\end{aligned}$$

where the last inequality makes use of product error rate assumptions 5-6 and nuisance function convergence rates from Lemma 4. Thus, we find that this term is $o_p(1/\sqrt{n})$

Bound for the first term:

The key to bounding the first term is that cross-fitting allows us to treat this term as the average of independent mean-zero random variables. We will bound it with Chebyshev's inequality, which requires a bound on the second moment on the summands in the first term.

$$\begin{aligned}
&\mathbb{E}\left[\frac{1}{\sqrt{n_{t,k}}} \sum_{i:(t,i)\in\mathcal{I}_k} \left((\mu_1(X_i) - \hat{\mu}_1^{(-k)}(X_i))\left(\frac{Z_i\tilde{R}_i}{e_1(X_i)\pi(1,X_i)} - 1\right)\right)^2 \mid \mathcal{I}_{(-k)}, \{X_i\}\right] \\
&\quad = \mathrm{Var}\left[\frac{1}{\sqrt{n_{t,k}}} \sum_{i:(t,i)\in\mathcal{I}_k} (\mu_1(X_i) - \hat{\mu}_1^{(-k)}(X_i))\left(\frac{Z_i\tilde{R}_i}{e_1(X_i)\pi(1,X_i)} - 1\right) \mid \mathcal{I}_{(-k)}, \{X_i\}\right]
\end{aligned}$$

$$= \frac{1}{n_{t,k}} \sum_{i:(t,i)\in\mathcal{I}_k} \mathbb{E}\left[(\mu_1(X_i) - \hat{\mu}_1^{(-k)}(X_i))^2 \left(\frac{Z_i\tilde{R}_i}{e_1(X_i)\pi(1,X_i)} - 1\right)^2 \mid \mathcal{I}_{(-k)}, \{X_i\}\right]$$

(expectation of $(\frac{Z_i\tilde{R}_i}{e_1(X_i)\pi(1,X_i)} - 1)^2$)

$$= \frac{1}{n_{t,k}} \sum_{i:(t,i)\in\mathcal{I}_k} \frac{1 - e_1(X_i)\pi(z,X_i)}{e_1(X_i)\pi(1,X_i)} (\mu_1(X_i) - \hat{\mu}_1^{(-k)}(X_i))^2$$

$$\leq \frac{1-\nu_e\lambda_\pi}{\nu_e\lambda_\pi} \frac{1}{n_{t,k}} \sum_{i:(t,i)\in\mathcal{I}_k} ((\mu_1(X_i) - \hat{\mu}_1^{(-k)}(X_i))^2 = o_p(\frac{1}{n^{1+2r_\mu}})$$

where for the third equality, we use the fact that

$$\mathbb{E}[(\frac{Z_i\tilde{R}_i}{e_1(X_i)\pi(1,X_i)} - 1)^2 \mid \mathcal{I}_{(-k)}, \{X_i\}] = \mathbb{E}[(\frac{Z_i^2R_i^2}{e_1^2(X_i)\pi^2(1,X_i)} - 2\frac{Z_i\tilde{R}_i}{e_1(X_i)\pi(1,X_i)} + 1 \mid \mathcal{I}_{(-k)}, \{X_i\}] = \frac{1}{e_1(X_i)\pi(1,X_i)} - 1$$

Since $r_\mu \geq 0$, we can conclude by Chebyshev's inequality that the first term is $o_p(n^{-1/2})$.

Bound for the second term: We bound the second term following a similar argument as above.

$$\mathbb{E}\left[\frac{1}{\sqrt{n_{t,k}}} \sum_{i:(t,i)\in\mathcal{I}_k} \left(Z_i\tilde{R}_i(Y_i - \mu_1(X_i)) \left(\hat{\pi}^{(-k)}(1,X_i)^{-1}(\hat{e}_1^{(-k)}(X_i)^{-1} - e_1(X_i)^{-1})\right)^2 \mid \mathcal{I}_{(-k)}, \{X_i\}\right]$$

$$+ \mathbb{E}\left[\frac{1}{\sqrt{n_{t,k}}} \sum_{i:(t,i)\in\mathcal{I}_k} \left(Z_i\tilde{R}_i(Y_i - \mu_1(X_i)) \left(e_1(X_i)^{-1}(\hat{\pi}^{(-k)}(1,X_i)^{-1} - \pi(1,X_i)^{-1})\right)\right)^2 \mid \mathcal{I}_{(-k)}, \{X_i\}\right]$$

$$= \mathrm{Var}\left[\frac{1}{\sqrt{n_{t,k}}} \sum_{i:(t,i)\in\mathcal{I}_k} \left(Z_i\tilde{R}_i(Y_i - \mu_1(X_i)) \left(\hat{\pi}^{(-k)}(1,X_i)^{-1}(\hat{e}_1^{(-k)}(X_i)^{-1} - e_1(X_i)^{-1})\right) \mid \mathcal{I}_{(-k)}, \{X_i\}\right]$$

$$+ \mathrm{Var}\left[\frac{1}{\sqrt{n_{t,k}}} \sum_{i:(t,i)\in\mathcal{I}_k} \left(Z_i\tilde{R}_i(Y_i - \mu_1(X_i)) \left(e_1(X_i)^{-1}(\hat{\pi}^{(-k)}(1,X_i)^{-1} - \pi(1,X_i)^{-1})\right) \mid \mathcal{I}_{(-k)}, \{X_i\}\right]$$

$$= \frac{1}{n_{t,k}} \sum_{i:(t,i)\in\mathcal{I}_k} \mathbb{E}\left[\left(\hat{\pi}^{(-k)}(1,X_i)^{-1}(\hat{e}_1^{(-k)}(X_i)^{-1} - e_1(X_i)^{-1})\right)^2 \frac{Z_i^2R_i^2}{(\hat{\pi}^{(-k)}(1,X_i))^2}(Y_i - \mu_1(X_i))^2 \mid \mathcal{I}_{(-k)}, \{X_i\}\right]$$

$$+ \frac{1}{n_{t,k}} \sum_{i:(t,i)\in\mathcal{I}_k} \mathbb{E}\left[\left(e_1(X_i)^{-1}(\hat{\pi}^{(-k)}(1,X_i)^{-1} - \pi(1,X_i)^{-1})\right)^2 \frac{Z_i^2R_i^2}{(\hat{\pi}^{(-k)}(1,X_i))^2}(Y_i - \mu_1(X_i))^2 \mid \mathcal{I}_{(-k)}, \{X_i\}\right]$$

$$= \frac{1}{n_{t,k}} \sum_{i:(t,i)\in\mathcal{I}_k} \frac{e_1(X_i)\pi(z,X_i)}{(\hat{\pi}^{(-k)}(1,X_i))^2}\mathbb{E}[\sigma^2(X_i) \mid \mathcal{I}_{(-k)}, \{X_i\}](\hat{e}_1^{(-k)}(X_i)^{-1} - e_1(X_i)^{-1})^2$$

$$+ \frac{e_1(X_i)(\pi^{(-k)}(z,X_i))}{e_1(X_i)}\mathbb{E}[\sigma^2(X_i) \mid \mathcal{I}_{(-k)}, \{X_i\}](\hat{\pi}^{(-k)}(1,X_i)^{-1} - \pi(1,X_i)^{-1})^2$$

$$\leq \frac{1}{n_{t,k}} \sum_{i:(t,i)\in\mathcal{I}_k} \frac{\nu_e^2\lambda_\pi^2}{(\hat{\pi}^{(-k)}(1,X_i))^2} B_{\sigma^2}(\hat{e}_1^{(-k)}(X_i)^{-1} - e_1(X_i)^{-1}))^2 + \frac{\nu_e^2\lambda_\pi^2}{\nu_e^2} B_{\sigma^2}(\hat{\pi}^{(-k)}(1,X_i)^{-1} - \pi(1,X_i)^{-1})^2$$

$$= o_p(\frac{1}{n^{1+2r_e+2r_\pi}})$$

where the last inequality is because $\sigma^2(X)$ is bounded above, $\sigma^2(X) \leq B_{\sigma^2}$, by Lemma 4. Thus, by similar argument to the first term, since this term is a sum of zero-mean random variables and since $r_\pi, r_e \geq 0$, we can apply Chebyshev's inequality and get that this term is also $o_p(1/\sqrt{n})$. This holds for both treatments. Therefore,

$$\sqrt{n_{t,k}}(\hat{\tau}_{AIPW}^{(t,k)} - \hat{\tau}_{AIPW}^{*,(t,k)}) \to_p 0.$$

Putting these results from Step 1 and Step 2 together, along with the fact that $\frac{n_{t,k}}{n} \to \frac{1}{K}$, gives the theorem. □

F ADDITIONAL LEMMAS

F.1 Results appearing in other works, stated for completeness.

Lemma 1 (Conditional convergence implies unconditional convergence, from (Chernozhukov et al., 2018))**.** *Lemma 6.1. (Conditional Convergence implies unconditional) Let $\{X_m\}$ and $\{Y_m\}$ be sequences of random vectors. (a) If, for $\epsilon_m \to 0, \Pr(\|X_m\| > \epsilon_m \mid Y_m) \to_{\Pr} 0$, then $\Pr(\|X_m\| > \epsilon_m) \to 0$. In particular, this occurs if $E[\|X_m\|^q / \epsilon_m^q \mid Y_m] \to_{Pr} 0$ for some $q \geq 1$, by Markov's inequality. (b) Let $\{A_m\}$ be a sequence of positive constants. If $\|X_m\| = O_P(A_m)$ conditional on Y_m, namely, that for any $\ell_m \to \infty$, $\Pr(\|X_m\| > \ell_m A_m \mid Y_m) \to_{Pr} 0$, then $\|X_m\| = O_P(A_m)$ unconditionally, namely, that for any $\ell_m \to \infty$, $\Pr(\|X_m\| > \ell_m A_m) \to 0$.*

Lemma 2 (Chebyshev's inequality)**.** *Let X be a random variable with mean μ and variance σ^2. Then, for any $t > 0$, we have*

$$P(|X - \mu| \geq t) \leq \frac{\sigma^2}{t^2}$$

Lemma 3 (Theorem 8.3.23 (Empirical processes via VC dimension), (Vershynin, 2018))**.** *Let $\mathcal{F}$ be a class of Boolean functions on a probability space (Ω, Σ, μ) with finite VC dimension $\mathrm{vc}(\mathcal{F}) \geq 1$. Let $X, X_1, X_2, \ldots, X_n$ be independent random points in Ω distributed according to the law μ. Then*

$$\mathbb{E} \sup_{f \in \mathcal{F}} \left| \frac{1}{n} \sum_{i=1}^{n} f(X_i) - \mathbb{E} f(X) \right| \leq C \sqrt{\frac{\mathrm{vc}(\mathcal{F})}{n}}$$

F.2 Lemmas

Lemma 4 (Convergence of $\hat{\pi}$)**.** *Assume that with high probability, for some large constant K, $\|\hat{e}(X) - e(X)\|_2 \leq K n^{-r_e}$, $\left\|\hat{\sigma}^2(X) - \sigma^2(X)\right\|_2 \leq K n^{-r_\sigma}$. Assume Assumption 8. Assume that $\sigma^2(X) > 0$ so that its inverse is bounded $1/\sigma^2(X) \leq \gamma_\sigma$. Recall that Theorem 1 gives that*

$$\pi^*(z, X) = \sqrt{\frac{\sigma_z^2(X)}{e_z^2(X)}} B \left(\mathbb{E}\left[\mathbb{I}[Z=1] \sqrt{\frac{\sigma_1^2(X)}{e_1^2(X)}} + \mathbb{I}[Z=0] \sqrt{\frac{\sigma_0^2(X)}{e_0^2(X)}} \right] \right)^{-1}$$

Define $\hat{\pi}^*(z, x)$ to be a plug-in version of the above (with $\hat{\sigma}^2, \hat{e}$, and $\mathbb{E}_n[\cdot]$). Then

$$\|\hat{\pi}^*(z, X) - \pi^*(z, X)\|_2 = o_p(n^{-\min(r_e, r_\sigma, 1/2)}).$$

Proof. Let $a = \frac{\sigma_z^2(X)}{e_z^2(X)}$, $b = \mathbb{E}\left[\mathbb{I}[Z=1] \sqrt{\frac{\sigma_1^2(X)}{e_1^2(X)}} + \mathbb{I}[Z=0] \sqrt{\frac{\sigma_0^2(X)}{e_0^2(X)}} \right]$.

Let $c = \frac{\hat{\sigma}_z^2(X)}{\hat{e}_z^2(X)}$, $d = \mathbb{E}_n\left[\mathbb{I}[Z=1] \sqrt{\frac{\hat{\sigma}_1^2(X)}{\hat{e}_1^2(X)}} + \mathbb{I}[Z=0] \sqrt{\frac{\hat{\sigma}_0^2(X)}{\hat{e}_0^2(X)}} \right]$.

Then $\|\pi^*(z, X) - \hat{\pi}^*(z, X)\|_2 = \|a/b - c/d\|_2$.

Positivity of $\sigma_z^2(X)$ gives the elementary equality that $\frac{a}{b} - \frac{c}{d} = \left(\frac{a-b}{b}\right) + \left(\frac{d-c}{d}\right)$.

Therefore, by triangle inequality and boundedness,

$$\begin{aligned} \|\pi^*(z, X) - \hat{\pi}^*(z, X)\|_2 &\leq \gamma_\sigma \left\| \sqrt{\sigma^2(X)/e^2(X)} - \sqrt{\hat{\sigma}^2(X)/\hat{e}^2(X)} \right\|_2 \\ &+ \gamma_\sigma \left| \mathbb{E}_n\left[\mathbb{I}[Z=1] \sqrt{\frac{\hat{\sigma}_1^2(X)}{\hat{e}_1^2(X)}} + \mathbb{I}[Z=0] \sqrt{\frac{\hat{\sigma}_0^2(X)}{\hat{e}_0^2(X)}} \right] - \mathbb{E}\left[\mathbb{I}[Z=1] \sqrt{\frac{\sigma_1^2(X)}{e_1^2(X)}} + \mathbb{I}[Z=0] \sqrt{\frac{\sigma_0^2(X)}{e_0^2(X)}} \right] \right| \end{aligned} \tag{4}$$

Next we show that for $z \in \{0, 1\}$,

$$\left\| \sqrt{\hat{\sigma}_z^2(X)/\hat{e}_z^2(X)} - \sqrt{\sigma_z^2(X)/e_z^2(X)} \right\|_2 \leq \nu_e B_{\sigma^2} \left(\left\| \sqrt{\hat{\sigma}_z^2(X)} - \sqrt{\sigma_z^2(X)} \right\|_2 + \|e_z(X) - \hat{e}_z(X)\|_2 \right) \tag{5}$$

In the below, we drop the z argument.

By the triangle inequality, boundedness of $1/\hat{e}(X) \leq \nu_e$, and of $\sigma^2(X) \leq B_{\sigma^2}$:

$$\begin{aligned}
&\left\| \sqrt{\hat{\sigma}^2(X)/\hat{e}^2(X)} - \sqrt{\sigma^2(X)/e^2(X)} \right\|_2 \\
&= \left\| \sqrt{\hat{\sigma}^2(X)/\hat{e}^2(X)} \pm \sqrt{\sigma^2(X)/\hat{e}^2(X)} - \sqrt{\sigma^2(X)/e^2(X)} \right\|_2 \\
&\leq \nu_e \left\| \sqrt{\hat{\sigma}^2(X)} - \sqrt{\sigma^2(X)} \right\|_2 + B_{\sigma^2} \left\| \frac{1}{e(X)} - \frac{1}{\hat{e}(X)} \right\|_2
\end{aligned}$$

For the second term:

$$B_{\sigma^2} \left\| \frac{1}{e(X)} - \frac{1}{\hat{e}(X)} \right\|_2 \leq B_{\sigma^2} \left\| \frac{1}{e(X)} - \frac{1}{\hat{e}(X)} \right\|_2 \leq B_{\sigma^2} \nu_e \left\| e(X) - \hat{e}(X) \right\|_2$$

since $1/e(X)$ is Lipschitz on the assumed bounded domain (overlap assumption).

For the first term:

$$\nu \left\| \sqrt{\hat{\sigma}^2(X)} - \sqrt{\sigma^2(X)} \right\|_2 \leq \nu_e B_{\sigma^2} \left\| \hat{\sigma}^2(X) - \sigma^2(X) \right\|_2$$

since $\sigma^2(X)$ is bounded away from 0, then $\sqrt{\sigma^2(X)}$ is Lipschitz.

This proves Equation (5), which bounds the first term of Equation (4). For the second term, denote for brevity

$$\hat{\beta}(\sigma, e) = \mathbb{E}_n \left[\mathbb{I}[Z=1] \sqrt{\frac{\sigma_1^2(X)}{e_1^2(X)}} + \mathbb{I}[Z=0] \sqrt{\frac{\sigma_0^2(X)}{e_0^2(X)}} \right],$$

and $\beta(\sigma, e)$ to be the above with $\mathbb{E}\left[\cdot\right]$ instead of $\mathbb{E}_n\left[\cdot\right]$. Then the second term of Equation (4) is $\hat{\beta}(\hat{\sigma}, \hat{e}) - \beta(\sigma, e)$, and decomposing further, that

$$\hat{\beta}(\hat{\sigma}, \hat{e}) - \beta(\sigma, e) = \hat{\beta}(\hat{\sigma}, \hat{e}) - \hat{\beta}(\sigma, e) + \hat{\beta}(\sigma, e) - \beta(\sigma, e).$$

Note that by Cauchy-Schwarz inequality, and Lemma 3 (chaining with VC-dimension),

$$\hat{\beta}(\hat{\sigma}, \hat{e}) - \hat{\beta}(\sigma, e) \leq 2\nu_e B_{\sigma^2} \left(\left\| \sqrt{\hat{\sigma}_z^2(X)} - \sqrt{\sigma_z^2(X)} \right\|_2 + \left\| e_z(X) - \hat{e}_z(X) \right\|_2 \right) + 2C \sqrt{\frac{\mathrm{vc}(\mathcal{F}_{\sqrt{\frac{\sigma^2}{e}}})}{n}}$$

And another application of Lemma 3 gives that

$$\hat{\beta}(\sigma, e) - \beta(\sigma, e) = (\mathbb{E}_n - \mathbb{E}) \left[\mathbb{I}[Z=1] \sqrt{\frac{\sigma_1^2(X)}{e_1^2(X)}} + \mathbb{I}[Z=0] \sqrt{\frac{\sigma_0^2(X)}{e_0^2(X)}} \right] \leq 2C \sqrt{\frac{\mathrm{vc}(\mathcal{F}_{\sqrt{\frac{\sigma^2}{e}}})}{n}}.$$

Combining the above bounds with Equation (4), we conclude that $\|\pi^*(z, X) - \hat{\pi}^*(z, X)\|_2 = o_p(n^{-\min(r_e, r_\sigma, 1/2)})$. □

G ADDITIONAL EXPERIMENTS, DETAILS, AND DISCUSSION

G.1 Additional details

All experiments using our full algorithm 2 were conducted on a 2021 13-inch MacBook Pro equipped with a 2.3 GHz Quad-Core Intel Core i7 processor and 32 GB of memory. This setup was used to train standard nuisance models using machine learning, evaluated our algorithm, and conduct the analysis tasks reported in this paper. The average compute time for the experiments on real world data with 20 trials was less than 30 minutes, while the simulated data with 100 trials took less than 60 minutes. Additionally, for all experiments, we allocate 55% of the data to batch 1 and 45% to batch 2.

We run the ML nuisance models, logistic regression, random forest and support vectors machines, using popular Python packages (i.e. sklearn and scipy). We use logistic regression to estimate the propensity scores. For the outcome and variance models, we use random forest with the following hyperparameters:

- max_depth: None
- min_samples_leaf: 4
- min_samples_split: 10
- n_estimators: 100
- random_state: 42

We also use support vector machines for the outcome models incorporating LLM predictions, and we use the following hyperparameters:

- kernel: 'rbf'
- C: 1

We chose these hyperparameters by doing a grid search over hyperparameters and chose the ones that performed the best. We ensemble predictions from the best performing random forest model trained on X and the best performing SVM model trained on X and $f(X, \tilde{Y})$ for our outcome model $\mu_z(X, \tilde{Y})$.

We run LLM calls on Together.AI since they provide enterprise-secure deployments of local models, which is required for sensitive data. Because we need to use local LLMs for the real-world street outreach data, we also use the same local LLMs for the other experiments. We use "Llama-3.3-70B-Instruct-Turbo" for all experiments using LLMs. (Larger models provide effectively similar performance).

To solve our optimization problem, we used the python package CVXPY and we specifically used the Splitting Conic Solver (SCS) solver.

Once the experiments are run, we display the means and 95% confidence interval bands, obtained through bootstrapping, in each of our figures.

G.2 Synthetic Data

Before running our batch adaptive algorithm, we split the data into a validation set (35% of data), which we use to estimate the oracle ATE. Then we use the remainder (65%) of the data to run our algorithm, which splits that data into the two batches in the way we described previously.

Data Generating Process. We generate a dataset $\mathcal{D} = \{X, Z, Y, Y(1), Y(0)\}$, of size 1000 and where the true ATE $\tau = \mathbb{E}[Y(1)] - \mathbb{E}[Y(0)] = 3$. We sample each covariate $X \in \mathbb{R}^5$ from a standard normal distribution, $X \sim \mathcal{N}(0, I_5)$. Treatment Z is drawn with logistic probability $\gamma_z(X) = (1 + e^{-X_2 + X_3 + 0.5})$. We define $\sigma_z^2(X)$ as follows:

$$\sigma_1^2(X) := \max[1.3 + 0.4\sin(X_1), 0]$$
$$\sigma_0^2(X) := \max[3.5 + 0.3\cos(X_3), 0].$$

Finally, the outcome models are defined as:

$$Y(0) = 5 + X_1 - 2X_2 + \epsilon_0$$
$$Y(1) = Y(0) + \theta_0 + \epsilon_1,$$

where $\epsilon_0 \sim \mathcal{N}(0, \sigma_0^2(X))$ and $\epsilon_1 \sim \mathcal{N}(0, \sigma_1^2(X))$. The observed outcomes are $Y = Z \cdot Y(1) + (1 - Z) \cdot Y(0)$.

Results. We see the greatest advantage with our adaptive estimation for budgets between 0.1 and 0.4. While for larger budgets, even as the MSE for both estimators converge, the interval width for the adaptive estimator is still relatively small. Adaptive annotation with a larger budget introduces additional variation in inverse annotation probabilities, as compared to uniform sampling, which is equivalent to full-information estimation at a marginally smaller budget. This regime of improvement for small budgets is nonetheless practically relevant and consistent with other works.

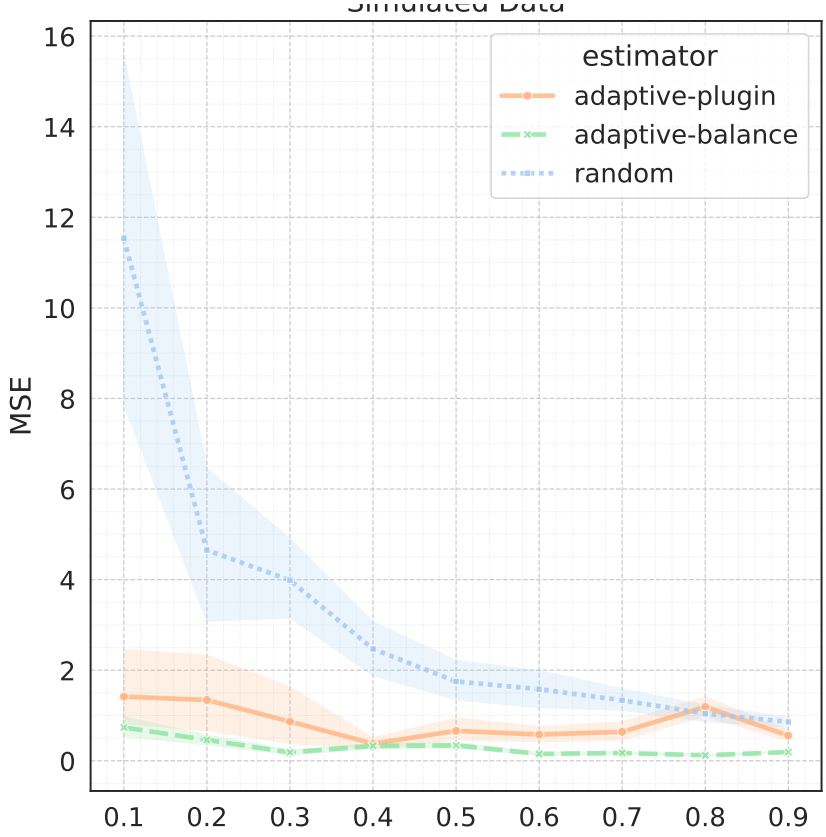

Figure 6: Mean squared error between estimated ATE and true ATE averaged over 100 trials across varying budgets.

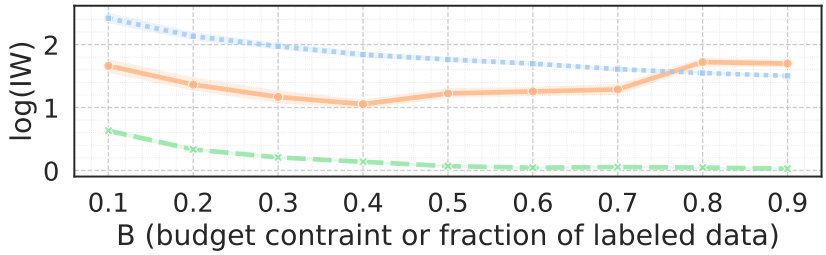

Figure 7: Average confidence interval width averaged over 100 trials across varying budgets.

To stabilize the estimation of the inverse annotation probabilities, we use the plug-in estimator following eq. (RZ-plug-in.) and the ForestReisz method to estimate the balancing weights (Chernozhukov et al., 2022).This approach provides an automatic machine learning debiasing procedure to learn the Reisz representer, or unique weights that automatically balances functions between treated and control groups using a random forest model.

G.3 Real-world Dataset Details

We provide further details about the treatment, covariates and outcomes for each dataset. Table 2 and table 3 describe the variables in the retail hero and outreach datasets, respectively. We refer the reader to (Dhawan et al., 2023) for further details about the dataset. For the outreach data, we constructed the binary treatment variable by binning the frequency of outreach engagements for each client within the first 6 months of the treatment period. We checked for overlap in propensity scores and decided to use treatments in the middle of the distribution as they had the most overlap. Additionally, by corollary 1, our method does well even when the propensity scores do not have good overlap.

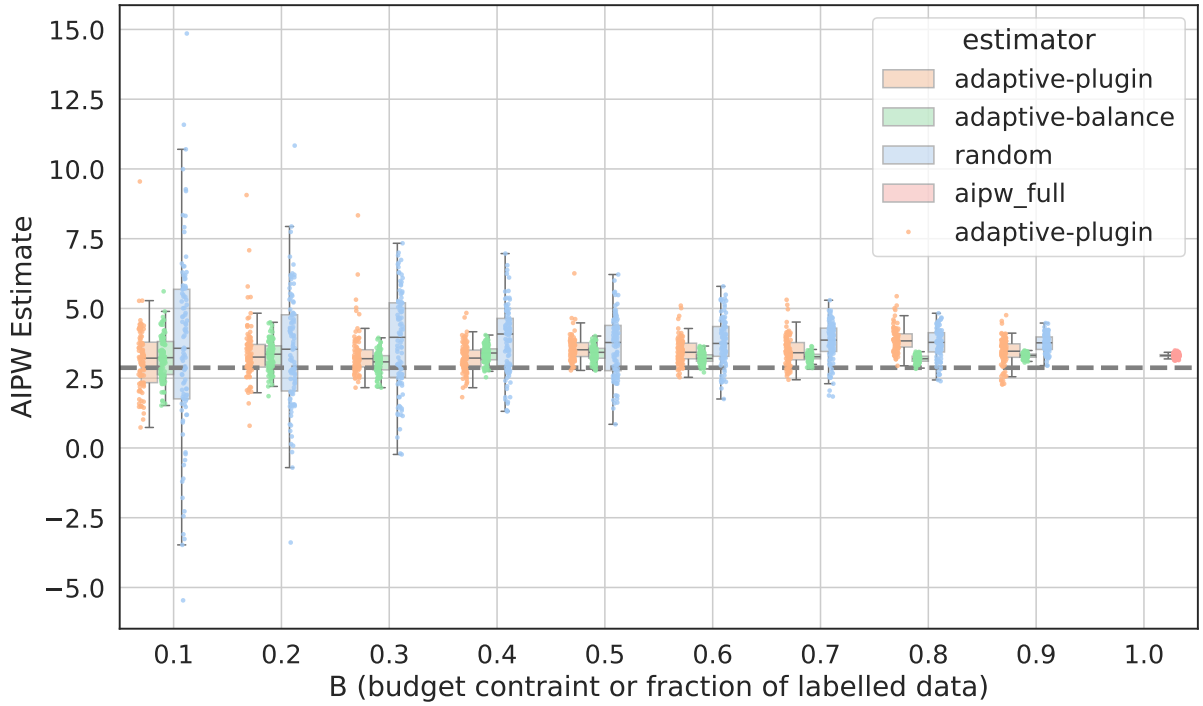

Figure 8: Boxplots of ATE estimates compared to skyline $\hat{\tau}_{AIPW}$ when the labeling budget is the entire dataset in red and the grey dotted line is τ.

Variable	Description	Discrete Category
Outcome		
Purchase	whether a customer purchased a product	[Yes,No]
Treatment		
SMS communication	whether a text was sent to encourage customer to continue shopping	[Yes, No]
Covariates		
avg. purchase	avg. purchase value per transaction	[1-263, 264-396, 397-611, $>$ 612]
avg. product quantity	avg. number of products bought	[$\leq$ 7, $>$ 7]
avg. points received	avg. number of points received	[$\leq$ 5, $>$ 5]
num transactions	total number of transactions so far	[$\leq$ 8, 9 - 15, 16 - 27, $>$ 28]
age	age of user	[$\leq$ 45, $>$ 45]

Table 2: Covariate, treatment, and outcome descriptions and discrete category definitions for RetailHero dataset.

G.4 Additional Context on Street Outreach

In New York City alone, approximately $\$80,000,000$ per year is invested in homeless street outreach to an unclear effect. It is a time-consuming process, and it is unclear how the impacts of such intensive individualized outreach might compare to other proposed approaches, such as those focusing on placing entire networks of individuals together. While the nonprofit reports key metrics such as number of completed placements in housing services, these can be somewhat rare due to length of outreach, delays in waiting for housing, matching issues, etc; moreover, much of a successful placement is out of the control of outreach due to highly limited housing capacities. Measuring the impacts of street outreach on intermediate outcomes such as accessing benefits and services, completing required appointments and interviews, can better reflect the immediate impacts of street outreach.

G.5 Robustness Check on Street Outreach Data

To further demonstrate the utility of our approach, we run experiments on the Street Outreach data with $\tilde{Y}$. To recap, our setup consists of covariates X, which includes client characteristics at baseline and

Variable	Description	Discrete Category
Outcome		
Placement	The greatest housing placement attained by the client between 2019–2021	[3:permanent housing, 2: shelter/transitional housing, 1: other (e.g., hospital), 0: streets]
Treatment		
Street outreach	Binned frequency of outreach within the first three months of 2019	[More outreach (3–15), Less outreach (1–2)]
Covariates		
DateFirstSeen	Ordinal date when the client was first seen by the outreach team	NA
Program	Outreach or service program the client belonged to	[Brooklyn Library, Grand Central Partnership, Hospital to Home, K-Mart Alley, Macy's, MetLife, Penn Post Office, Pyramid Park, S2H Bronx, S2H Brooklyn, S2H Manhattan, S2H Queens, Starbucks, Superblock, Vornado, Williamsburg Stabilization Bed]
BelievedChronic	Perceived by outreach workers as chronically homeless individual	[Yes, No]
Gender	Perceived or disclosed gender of client	[Female, Male, Transgender]
Race	Perceived or disclosed race of client	[American Indian/Alaskan Native, Asian, Black/African American, Native Hawaiian/Pacific Islander, White/Caucasian]
Ethnicity	Perceived or disclosed ethnicity of client	[Hispanic/Latino, Non-hispanic/latino]
Age	Perceived or disclosed age range of client	[$<$ 30 years old, 30–50 years old, $>$ 50 years old]
Was311Call	Whether outreach workers were responding to a 311 city call	[Yes, No]
Was911Call	Whether 911 was called to the scene	[Yes, No]
Removal958	Whether outreach workers were responding to removal hotline call	[Yes, No]
Housing application	Whether any mention of the housing application was found in casenotes	[Yes, No]
Service refusal	Whether outreach worker documented that a client refused their services in casenotes	[Yes, No]
Important documents	Whether there was mention of any important documents (i.e. social security card, drivers license, etc,) in casenotes	[Yes, No]
Benefits	Whether there was any mention of social service benefits in the casenotes (i.e. foodstamps, SSI)	[Yes, No]
num contacts	number of engagements with an outreach worker prior to 2019	NA
max Placement	maximum housing placement reached before 2019	[3:permanent housing, 2: shelter/transitional housing, 1: other (e.g., hospital), 0: streets]

Table 3: Covariates, treatment, and outcome descriptions and discrete category definitions for the Street Outreach dataset.

LLM-generated summaries of case notes recorded before the treatment period. In the main text, we used LLMs to summarize casenotes prior to outreach during the interventional period, and used them in zero-shot prediction of later placement outcomes. Here we also incorporate LLM-generated summaries of case notes recorded post-treatment. These represent $\tilde{Y}$ in our framework.

In Figure 10 and Figure 11, we see that our results and analysis are preserved, and qualitatively similar. Our adaptive approach still shows improvements over uniform random sampling. The MSE is tripled when going from our adaptive estimators to random sampling in the tabular data. The MSE is five times higher when going from adaptive to random sampling in the setting where we have added LLM predictions using post-treatment summaries $\tilde{Y}$ only and it is nearly doubled when using both pre- and post-treatment summaries.

In this experimental setup, we find that tabular estimation with ground-truth validated codes overall

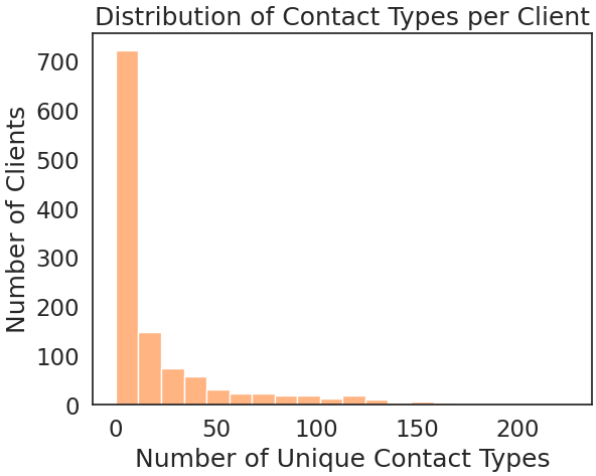

Figure 9: Distribution of street outreach engagements for client population.

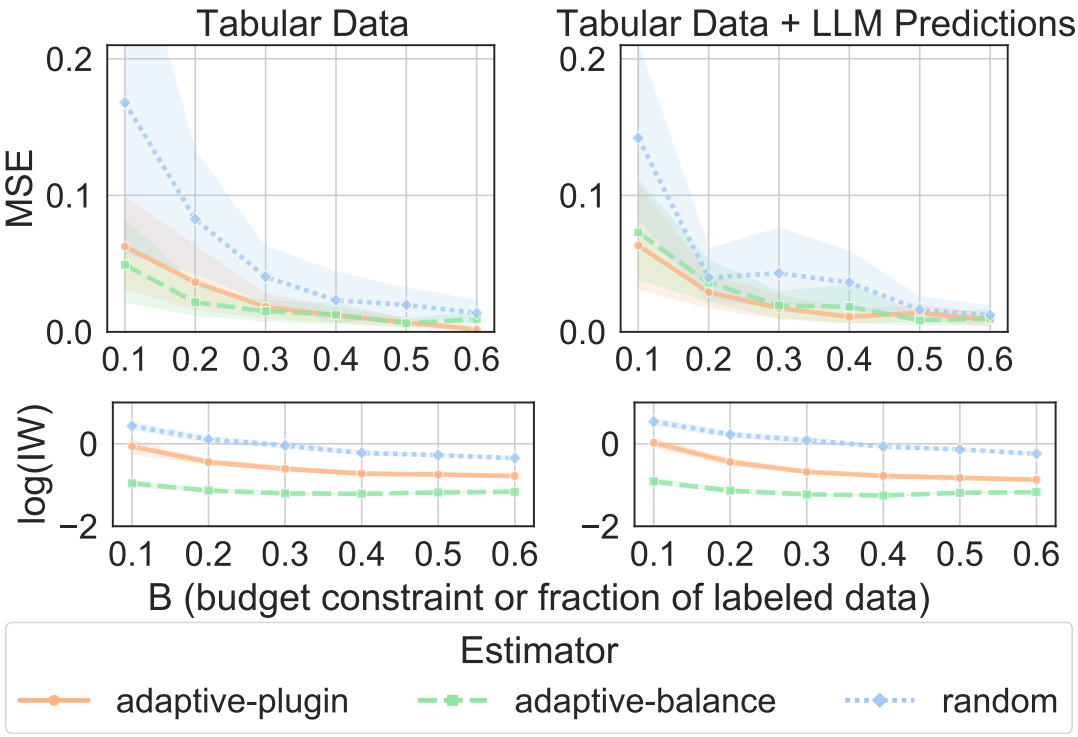

Figure 10: **Street Outreach Data with pretreatment summaries (no $\tilde{Y}$).** Mean squared error and 95% confidence interval width averaged over 20 trials across budget percentages of the data. This plot makes use of tabular data and the best-performing random forest outcome model (left) and text-encoded outcomes using LLMs (right).

performs comparably as using more advanced LLM estimation. In this setup, we use placement outcomes as the measure of interest, in part because it is (nearly) fully recorded in our dataset, and hence we can consider it as having access to the "ground-truth" outcome in our methodological setup. On the other hand, we also expect that casenotes are weakly informative of placement, as compared with other outcomes we might seek to extract from casenotes (but do not have the ground-truth for). Nonetheless, this validates the usefulness of the method, and we leave further empirical developments for future work.

G.6 Budget Saved Plots

We compute the amount of budget saved due to our batch adaptive sampling approach. We find the sample size required to achieve the same confidence interval width with batch adaptive annotations using

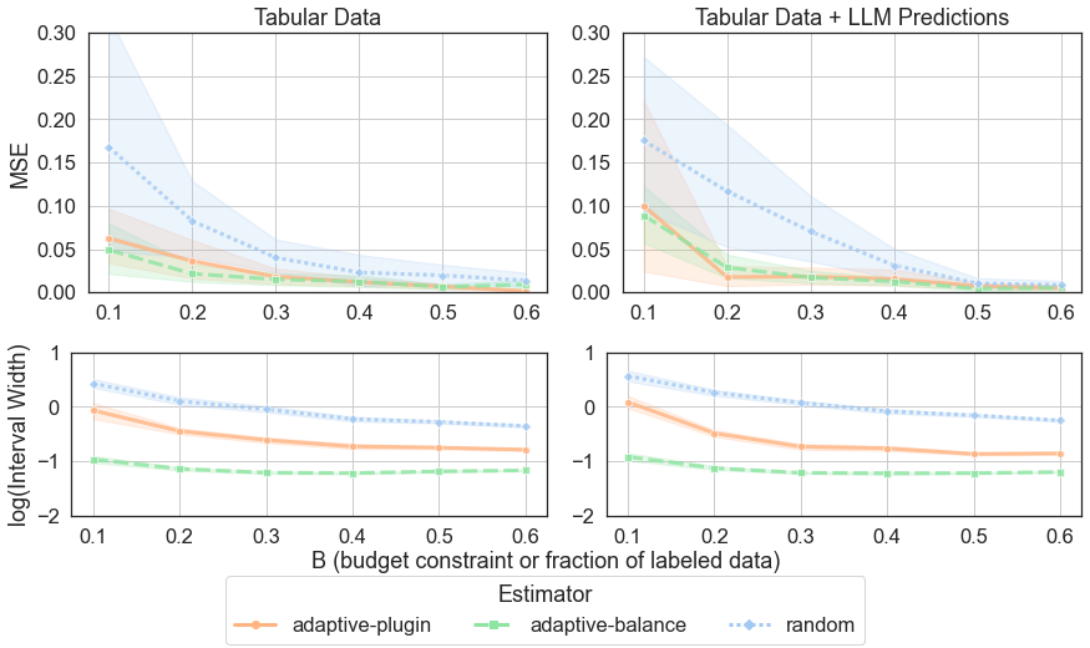

Figure 11: **Street outreach data with pre- and post-treatment summaries.** Mean squared error and 95% confidence interval width averaged over 20 trials across budget percentages of the data. This plot makes use of tabular data and the best-performing random forest outcome model (left) and text-encoded outcomes using LLMs (right).

balancing weights (green) and RZ-plug-in (orange) compared to uniform random sampling.

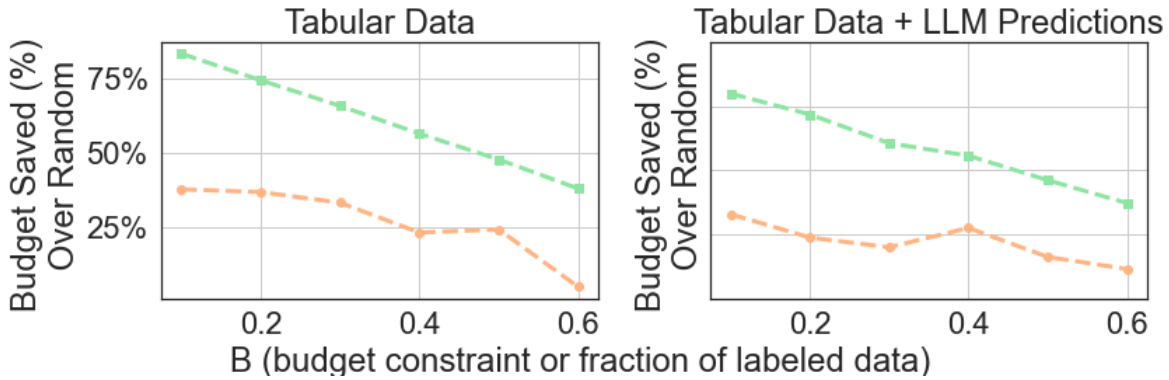

Figure 12: **RetailHero Data.** Budget saved due to batch adaptive annotation. The reduction in annotation sample size needed to achieve the same confidence interval width with batch adaptive annotation on tabular data (left) and on tabular data + complex embedded outcomes (right) compared to random sampling.

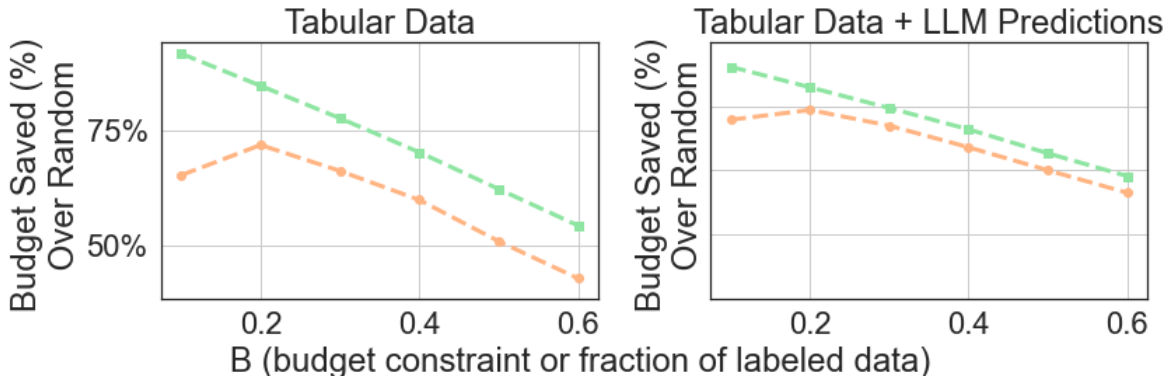

Figure 13: **Street Outreach Data.** Budget saved due to batch adaptive annotation. The reduction in annotation sample size needed to achieve the same confidence interval width with batch adaptive annotation on tabular data (left) and on tabular data + complex embedded outcomes (right) compared to random sampling.

G.7 Active Learning Baselines

Active learning is not a strong baseline and we argue this on theoretical and empirical fronts. Active learning for regression can't improve statistical rates of convergence, while the doubly-robust AIPW estimator in causal inference can, so using AIPW is optimal. Additionally, using pool-based active learning algorithms in AIPW blows up variance due to near-deterministic annotation probabilities. Active learning models only target μ_z, but the outcome model contributes $\frac{\sigma_z^2(x)}{e_z(x)\pi(z,x)}$ to the causal Avar, and our optimal annotation correctly balances the effect of all factors, but active learning only considers the first.

In summary, active learning does something *completely different for prediction error, suboptimal for causal inference*.

Empirically, we run active learning algorithms to learn μ in AIPW and find that it *totally fails* for these reasons; if these objectives line up, it can do well, but in general, the prediction and causal error objectives are different.

Theoretical comparison to active learning. As a reminder, we optimize:

$$AVar_{ATE} = Var[CATE(X)] + \sum_{z\in\{0,1\}} E[\frac{\sigma_z^2(X)}{e_z(X)\pi(z,X)}]$$

(The first term is the variance of $CATE = E[Y(1) - Y(0)|X]$; it is never observed.)

To go more in detail on our experiments 1) we compare to theoretical results in batch *pool-based active learning*, Chaudhuri et al. (2015) and Gentile et al. (2024) (henceforth GWZ), which show that active learning doesn't improve convergence rates for regression, only multiplicative constants. Instead, the AIPW estimator is optimal for causal estimation: if the outcome and propensity scores can only achieve $n^{-1/4}$ convergence, the AIPW estimator is $O(n^{-1/2})$-rate convergent, so AIPW can speed up outcome model convergence rates. Therefore using the AIPW estimator is best, and random sampling + AIPW is a stronger baseline than active learning.

To emphasize the different objectives, consider a simple example with two regions:

- Region 1 (Poor Overlap), $X > 0$: Propensity score $e(X) = 0.01$; outcome noise $\sigma_1(X), \sigma_0(X)$=1.
- Region 2 (High Prediction Uncertainty), $X < 0$: Propensity score $e(X) = 0.5$; outcome noise $\sigma_1(X), \sigma_0(X) = 10$ and the outcome model is complex.

Our method compares the ATE variance contribution in either region:

- Region 1: $\frac{\sqrt{1}}{0.01} = 100$
- Region 2: $\frac{\sqrt{100}}{0.5} = 20$

and samples in Region 1, where the causal variance is five times higher. Uncertainty-based active learning samples in Region 2, to the detriment of causal variance.

Active Learning Empirical Evaluations. We evaluate our method against 2-3 active learning baselines for each experiment from two popular and well-established python packages (scikit-activeML and modAL). Different active learning algorithms are appropriate for different outcome models, so we choose the sampling strategy based on our modeling task, and we use pool-based active learning matching our two-batch approach. (Note our approach is *model-agnostic*, while active learning methods are not). For the classification tasks on our two real-world datasets (RetailHero/Street Outreach), we use UncertaintySampling with margin sampling and least confident sampling as query strategies, which both choose x with highest uncertainty measure based on classification probabilities $P(\hat{Y} = 1 \mid x)$ (Settles, 2009). For the regression tasks, we use Expected Model Variance Reduction (Cohn et al., 1996), Expected Model Change Maximization (Cai et al., 2013), and Improved Greedy Sampling (Wu et al., 2019); these choose x that maximizes greatest future variance reduction, maximally change the current model via the loss gradient, and diversity in feature and output space, respectively.

We run each approach over 50 trials and take the average MSE. Across the board, we see that our approach does better than the popular active learning strategies that are not optimized for causal estimation.

Result Tables

Estimator	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
active-evar	**0.313**	17.3	85.1	579	1.31e+03	3.87e+03	1.27e+04	5.03e+04	8.93e+05
active-greedy	6.13	79.9	369	852	1.99e+03	5.06e+03	1.33e+04	5.09e+04	2.95e+05
active-mvar	10.6	94.3	314	883	2.17e+03	5.70e+03	1.21e+04	3.87e+04	2.99e+05
adaptive-balance	0.471	**0.227**	**0.276**	0.236	**0.265**	**0.246**	**0.198**	**0.176**	**0.203**
adaptive-plugin	1.7	1.17	0.831	**0.196**	0.83	0.449	0.507	0.93	0.481
random	8.99	4.56	2.19	1.54	1.7	1.61	1.46	0.956	0.987

Table 4: Averaged MSEs for Synthetic Data.

Estimator	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
active-margin	3.53e+03	0.047	0.087	12.5	8.38e+03	2.25e+06	1.49e+06	6.53e+05	1.43e+07
active-uncertain	16.1	38.9	70.4	75.9	115	112	168	250	402
adaptive-balance	**0.004**	0.002	0.002	**0.001**	**0.001**	0.001	**0**	**0**	**0**
adaptive-plugin	**0.004**	**0.001**	**0.001**	**0.001**	**0.001**	**0**	**0**	**0**	**0**
random	0.027	0.012	0.009	0.006	0.005	0.003	0.001	0.001	**0**

Table 5: Averaged MSEs for RetailHero Data.

Estimator	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
active-margin	**0.009**	28.5	4.47	0.501	0.449	0.044	0.099	0.412	0.209
active-uncertain	0.017	**0.009**	0.018	0.008	0.017	0.018	0.025	0.023	0.024
adaptive-balance	0.046	0.031	**0.013**	**0.006**	**0.005**	**0.003**	**0.004**	**0.003**	0.002
adaptive-plugin	0.045	0.025	0.027	0.012	0.006	0.004	**0.004**	0.006	**0.001**
random	0.113	0.061	0.037	0.045	0.014	0.012	0.011	**0.003**	**0.001**

Table 6: Averaged MSEs for Street Outreach Data.

Gentile et al. (2024) chooses a point x maximizing a diversity measure, D(x,S) that quantifies model uncertainty and is directly influenced by the observation noise, $\sigma_z^2(X)$. For general function approximation, they introduce a maximal disagreement measure over the regression function class $\mathcal{F}$ $\sup_{f,g\in\mathcal{F}} \frac{(f(x)-g(x))^2}{\sum_{z\in S}(f(z)-g(z))^2+1}$, where S is the set of already sampled points. If $\sigma^2(x)$ is large for some x, their disagreement measure is also large. Their diversity measure finds points where it is possible for two functions, f, g, to have similar predictions on the already-labeled data S (a small denominator) but different predictions for a new point x (a large numerator). When observation noise $\sigma^2(x)$ is larger, many different functions can be considered "plausible" fits and can agree on S but disagree elsewhere, leading to a high diversity score. In contrast, low noise tightly constrains all plausible functions, resulting in low disagreement.

G.8 LLM Prompts

Retail Hero Prompt

You are a user who used a website for online purchases in the past one year and want to share your background and experience with the purchases on social media.
Attributes: The following are attributes that you have, along with their descriptions.
$\{features\}$
Personality Traits: The following dictionary describes your personality with levels (High or Low) of the Big Five personality traits.
$\{traits\}$
Your Instructions:
Write a social media post in first-person, accurately describing the information provided. Write this post in the tone and style of someone with the given personality traits, without simply listing them. Only return the post that you can broadcast on social media and nothing more.
—
$\{post\}$
—

Street Outreach Casenote Summaries Prompt

Objective: Your task is to summarize a trajectory of case notes of a client in street homelessness outreach, focusing on client interactions, the challenges they are facing, goals they are working towards, and progress towards housing placement. These are all from the same client. This summary is designed to help caseworkers and organizations assess client history at a glance, remind of prior personal information and important challenges mentioned (like veteran status or other information that is relevant for eligibility for housing, medical issues, and status of their support network), allocate resources effectively, and improve support for individuals experiencing chronic homelessness.
Context: $\{task\_context\}$
The summary should be a concise overview of the client's situation, highlighting key points from the case notes. It should not include any personal opinions or assumptions about the client's future or potential outcomes. The goal is to provide a clear and informative summary that can be used by caseworkers and organizations to better understand the client's history and current status.
Here are the case notes for batch $\{batch\_num\}$ of $\{total\_batches\}$:
— START NOTES —
$\{notes\}$
— END NOTES —
Based *only* on the notes provided above for this batch, generate a comprehensive summary focusing on key events, decisions, and progress during this specific period. The target length is approximately $\{target\_length\}$ words. Ensure the summary strictly reflects the content of these notes.

Street Outreach Classification

You are an expert analyst specializing in predicting long-term housing stability for individuals experiencing homelessness. Your task is to analyze client data, including demographic information, historical interactions, and case note summaries, to predict the **most stable housing placement level** the client is likely to achieve and maintain over the **next two years**.
Input Data:
You will be provided with the following information for each client:
Prediction Task:
Based *only* on the provided attributes and the case notes summary, predict the single most stable housing placement level the client is likely to maintain over the next two years.
Housing Placement Levels (Prediction Output):
Your prediction must be an integer between 0 and 3:

- **0**: No stable placement (remains on the street or in emergency shelters).
- **1**: Transitional Housing (temporary placement with support, aiming for longer-term housing).
- **2**: Rapid Re-housing (time-limited rental assistance and services).
- **3**: Permanent Supportive Housing (long-term housing with ongoing support services).

Reasoning Guidance (Internal Thought Process - Do Not Output This):

- Consider factors that promote stability: housing application progress, possession of documents, benefit acquisition, engagement with services (unless contacts are excessive without progress), prior successful placements (even if temporary), positive recent developments in the case notes.
- Consider factors that hinder stability: chronic homelessness indicators, frequent service refusals, mental health crises (Removal958), lack of documents/income, lack of prior placements, patterns of instability noted in the summary.
- Weigh the structured data against the nuances presented in the case note summary. The summary provides vital context.

Client Information:
Prediction:
Provide *only* the predicted number (0, 1, 2, or 3) as the output. Do not include any other text, explanation, or formatting.
Examples: $\{examples\}$