Batch-Adaptive Annotations for Causal Inference with Complex-Embedded Outcomes

Ezinne Nwankwo
Department of Computer Science
University of California, Berkeley
ezinne_nwankwo@berkeley.edu

Lauri Goldkind
Department of Social Work
Fordham University
lgoldkind@fordham.edu

Angela Zhou
Department of Data Science and Operations
University of Southern California
zhoua@usc.edu

Abstract

Estimating the causal effects of an intervention on outcomes is crucial to policy and decision-making. But often, information about outcomes can be missing or subject to non-standard measurement error. It may be possible to reveal ground-truth outcome information at a cost, for example via data annotation or follow-up; but budget constraints entail that only a fraction of the dataset can be labeled. In this setting, we optimize *which data points should be sampled for outcome information* and therefore efficient average treatment effect estimation with missing data. We do so by allocating data annotation in batches. We extend to settings where outcomes may be recorded in unstructured data that can be annotated at a cost, such as text or images, for example, in healthcare or social services. Our motivating application is a collaboration with a street outreach provider with millions of case notes, where it is possible to expertly label some, but not all, ground-truth outcomes. We demonstrate how expert labels and noisy imputed labels can be combined into a doubly robust causal estimator. We run experiments on simulated data and two real-world datasets, including one on street outreach interventions in homelessness services, to show the versatility of our proposed method.

1 Introduction

Evaluating causal effects of a treatment or policy intervention is a challenging problem in its own right, but an added layer of complexity comes when there is missing data. In this paper, we consider a setting of observational causal inference with missing outcomes, where it is possible to obtain information about ground-truth outcomes at a cost, via expert annotation or follow-up. Recent tools in machine learning can label outcomes, but for inferential goals, this can lead to error-prone and biased outputs. With a small budget, one can obtain valid causal effects on a small subsample without using additional contextual information or imputation, but this can be 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 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.

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Workshop: MLxOR: Mathematical Foundations and Operational Integration of Machine Learning for Uncertainty-Aware Decision-Making.

These notes are a noisy view on the ground-truth of what happens during the open-ended process of outreach. Was a client progressing on their housing application or their goals, or were they facing other barriers? In our experience, outreach workers can extract structured information, from the unstructured text of case notes. They can provide context and recognize important milestones. But it is simply impossible for under-resourced expert outreach workers to 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 collect ground truth data, such as by assigning expert annotation, while leveraging additional data sources or weaker annotation to optimize causal effect estimation?* In this paper, we develop general methodology for optimizing data annotation and we validate our methodological innovations using outcomes with plausible ground-truth information on housing placement.

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 it is possible to query the ground truth directly 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. Weaker imputation of auxiliary information is feasible at scale, but second-best due to inaccuracies.

Related work. Please see the appendix for the related work. Our model is closest to optimizing a validation set for causal inference with missing outcomes, which can be broadly useful for causal inference with non-standard measurement error. The most related work is that of [12, 29], which leverages the fact that sampling probabilities for data annotation are known to obtain doubly-robust estimation via causal inference. These works generally address non-causal estimands such as mean estimation and M-estimation (therefore without discussion of treatment effect estimation).

2 Problem setup

Our problem setting is causal inference with missing outcomes. We discuss extensions to a setting where outcomes are measured in a high-dimensional contextual variable $\tilde{Y}$, such as images or text.

In both cases, we assume the ground-truth data-generating process follows that of standard causal inference. A data instance $(X, Z, Y(Z))$, 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 realized treatment assignment Z and assume the usual stable unit value treatment assumption (SUTVA). *If* the ground-truth data were observed, we would have a standard causal inference task at hand, so the key challenge is its *missingness*. We let $R \in \{0, 1\}$ denote the presence ($R = 1$) or absence ($R = 0$) of the outcome Y. Therefore, our observational dataset for estimation is (X, Z, R, RY), i.e. with missing outcomes. For causal identification, we generally proceed under the following assumptions:

Assumption 1 (Treatment ignorability [16, 17, 19])**.** $Y(Z) \perp\!\!\!\perp Z \mid X$.

Assumption 2 (R-ignorability [21, 2])**.** $R \perp\!\!\!\perp Y(Z) \mid Z, X$

Assumption 2 is true by design as long as the full corpus of datapoints needing annotation is available from the outset, since we choose what datapoints are annotated for ground-truth labels based on (Z, X) alone.

Although one approach is completely random sampling, we are particularly concerned with *how can we select datapoints for expert annotation for optimal estimation*? We assume the budget is limited for data annotation, but we have control over the missingness mechanism, i.e. assigning data for expert annotation. Define the propensity score and annotation (outcome observation) probability:

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

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

Assumption 3 (Treatment and annotation positivity [16, 17, 19])**.** $\epsilon < \pi(z, X) \leq 1, z \in \{0, 1\}$ and $1/\nu < e_1(X) < 1 - 1/\nu, \nu > 0$

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

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

Batch allocation setup. We consider a two-batch adaptive protocol, where n iid observations are randomly split into two batches. We consider a proportional asymptotic regime where the budget and size of first batch n_1 are fixed proportions $\kappa \in (0, 1)$ of the dataset size.

Assumption 4 (Proportional asymptotic [15, 20])**.** $\lim_{n\to\infty} \frac{n_1}{n} = \kappa$.

In the first batch, we randomly assign annotations according to a small but asymptotically nontrivial fraction of the budgets. In the first batch, 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. We solve for optimal annotation probabilities $\pi*$ and sample data in the second batch so that the mixture distribution over outcome observations achieves $\pi*$. We combine the results from both batches and use the data for ATE estimation, which we describe in the next section.

3 Method

This section outlines our proposed methodology.

Recap: Optimal asymptotic variance for the ATE with missing outcomes. Our target parameter of interest is the ATE of a binary treatment vector Z on an outcome Y.

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

Bia et al. [2] derives a double-machine learning estimator for ATE estimation with missing outcomes:

$$\mathbb{E}[Y(z)] = \mathbb{E}[\psi_z], \text{where } \psi_z = \frac{\mathbb{I}[Z = z]R(Y - \mu_z(X))}{e_z(X)\pi(z, X)} + \mu_z(X), \text{ and } \tau_{AIPW} = \mathbb{E}[\psi_1 - \psi_0].$$

The outcome model $\mu_z(X)$ is estimated on data with observed outcomes, since SUTVA and assumption 2 give that $\mathbb{E}[Y(z)|X] = \mathbb{E}[Y|Z = z, X] = \mathbb{E}[Y|Z = z, R = 1, X]$.

The focus of our work is to optimize the semiparametric efficient asymptotic variance (proven in [2]), which is closely related to the ATE of [14].

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

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

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

Remark 1. We state the results for the base model, though they extend directly for the case with contexts. With contexts, by marginalizing over $\tilde{Y}$, the analogous expressions use the estimators $\hat{\mu}_z(X, \tilde{Y})$ instead of $\hat{\mu}_z(X)$ whereas $\hat{\sigma}_z^2(X)$ stays the same (sampling probabilities depend only on (Z, X) and just correspondingly marginalizes over $\tilde{Y}$, $\hat{\sigma}_z^2(X) = \mathbb{E}[(Y - \hat{\mu}_z(X, \tilde{Y}))^2 \mid Z = z, X = x]$. In the setting with noisy measurements $\tilde{Y}$, under the exclusion restriction **??**, the mean potential outcome is identified by regression adjustment: $\mathbb{E}[Y(z)] = \mathbb{E}[\mathbb{E}[Y|Z = z, R = 1, X, \tilde{Y}]] = \mathbb{E}[\mathbb{E}[Y|Z = z, R = 1, X]]$.

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 sampling budget. We consider a global budget constraint $B \in [0, 1]$ over all annotations. The setting is meaningful when the budget binds, $B \ll 1$, which is still practically relevant.

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

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

Theorem 1. *The solution to the global budget problem is:*

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

For the proof, see Appendix G

4 Analysis

See the appendix for an extension to settings with continuous treatments. Denote $\|\cdot\|_2 = (\mathbb{E}[(\cdot)^2])^{1/2}$.

Theorem 2. *Given Z- and R-ignorability, bounded second moments, $o_p(n^{-1/2})$ product error rates, nuisance function estimates have a bounded VC dimension, and sufficiently weak dependence, across batches, suppose that we construct the feasible estimator $\hat{\tau}_{AIPW}$ using the CSBAE crossfitting procedure in with estimators satisfying consistency and product error rates. Then*

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

where $V_{AIPW} = \sum_{z \in 0,1} \mathbb{E}\left[\frac{\sigma_z^2(X)}{e_z(X)\pi^*(z,X)} \right] + \mathrm{Var}\left[\mu_1(X) - \mu_0(X) \right]$. *Here τ is the ATE.*

For the proof, see Appendix G. The main result from 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^*. This implies that our procedure successfully minimizes the asymptotic variance bound. With this, we can also quantify the uncertainty of our treatment effect estimates by producing level-α confidence intervals for τ that achieve coverage with $1 - \alpha$ probability.

Experiments. See the appendix for more experiments. We demonstrate our method on street outreach casenote data collected by a partnering nonprofit providing homelessness services. The covariate data X consists of baseline characteristics on each client as tabular data (left, Figure 1), such as the number of previous outreach engagements, and (right, Figure 1) LLM generated summaries of case notes recorded before treatment. More information on the data can be found in Appendix I. We seek to estimate the causal effect of street outreach on housing placement. We use housing placement as an illustrative example because it is well-recorded ground truth data in our dataset. However, it could also be plausibly missing, in which case nonprofits have to decide how to expend their limited resources to obtain more information (i.e., caseworker follow-up calls or analyzing more recent casenotes $\tilde{Y}$). In Figure 1 we see that overall our adaptive approach shows improvements over uniform random sampling. In Figure 10, we see that we can save between $43 - 75\%$ of the budget using the plugin-in estimator on tabular data alone and by incorporating LLM predictions, and between $53 - 91\%$ using the balance estimator over the random sampling baseline.

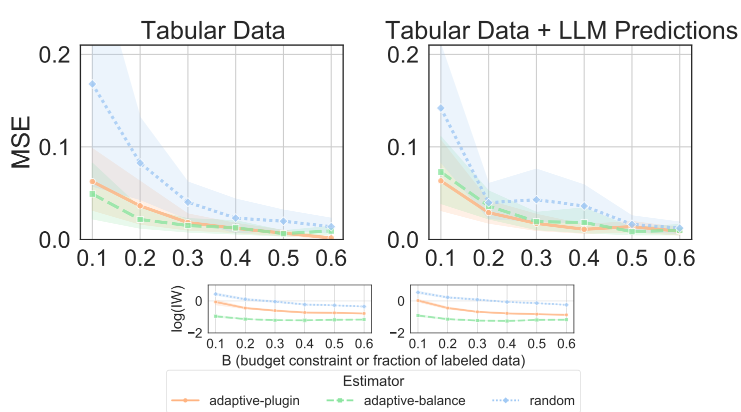

Figure 1: **Street Outreach Data.** Mean squared error and 95% confidence interval width averaged over 20 trials across budget percentages of the data. This plot makes use of tabular data and the best-performing random forest outcome model (left) and text-encoded outcomes using LLMs (right).

References

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

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

[3] Wenbin Cai, Ya Zhang, and Jun Zhou. Maximizing expected model change for active learning in regression. In *2013 IEEE 13th International Conference on Data Mining*, pages 51–60, 2013. doi: 10.1109/ICDM.2013.104.

[4] Kamalika Chaudhuri, Sham M Kakade, Praneeth Netrapalli, and Sujay Sanghavi. Convergence rates of active learning for maximum likelihood estimation. In C. Cortes, N. Lawrence, D. Lee, M. Sugiyama, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 28. Curran Associates, Inc., 2015. URL `https://proceedings.neurips.cc/paper_files/paper/2015/file/ca9c267dad0305d1a6308d2a0cf1c39c-Paper.pdf`.

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

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

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

[8] David A. Cohn, Zoubin Ghahramani, and Michael I. Jordan. Active learning with statistical models. *J. Artif. Int. Res.*, 4(1):129–145, March 1996. ISSN 1076-9757.

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

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

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

[12] Naoki Egami, Musashi Hinck, Brandon Stewart, and Hanying Wei. Using imperfect surrogates for downstream inference: Design-based supervised learning for social science applications of large language models. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, editors, *Advances in Neural Information Processing Systems*, volume 36, pages 68589–68601. Curran Associates, Inc., 2023. URL `https://proceedings.neurips.cc/paper_files/paper/2023/file/d862f7f5445255090de13b825b880d59-Paper-Conference.pdf`.

[13] Claudio Gentile, Zhilei Wang, and Tong Zhang. Fast rates in pool-based batch active learning. *J. Mach. Learn. Res.*, 25(1), January 2024. ISSN 1532-4435.

[14] Jinyong Hahn. On the role of the propensity score in efficient semiparametric estimation of average treatment effects. *Econometrica*, 66:315–331, 1998.

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

[16] M.A. Hernan and J.M. Robins. *Causal Inference: What If.* Chapman & Hall/CRC Monographs on Statistics & Applied Probab. CRC Press, 2025. ISBN 9781420076165. URL `https://books.google.com/books?id=_KnHIAAACAAJ`.

[17] Guido W. Imbens. Nonparametric estimation of average treatment effects under exogeneity: A review. *The Review of Economics and Statistics*, 86(1):4–29, 02 2004. ISSN 0034-6535. doi: 10.1162/003465304323023651. URL https://doi.org/10.1162/003465304323023651.

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

[19] Edward H. Kennedy. Efficient nonparametric causal inference with missing exposure information. *The International Journal of Biostatistics*, 2020. URL https://doi.org/10.1515/ijb-2019-0087.

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

[21] Donald B. Rubin. Inference and missing data. *Biometrika*, 63(3):581–592, 1976. ISSN 00063444, 14643510. URL http://www.jstor.org/stable/2335739.

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

[23] Burr Settles. Active learning literature survey, 2009. URL https://api.semanticscholar.org/CorpusID:324600.

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

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

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

[27] Dongrui Wu, Chin-Teng Lin, and Jian Huang. Active learning for regression using greedy sampling. *Information Sciences*, 474:90–105, 2019. ISSN 0020-0255. doi: https://doi.org/10.1016/j.ins.2018.09.060. URL https://www.sciencedirect.com/science/article/pii/S0020025518307680.

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

[29] Tijana Zrnic and Emmanuel Candes. Active statistical inference. In *Forty-first International Conference on Machine Learning*, 2024.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: Yes, we provide our main results in Sections 4 and 5, along with any necessary assumptions. We include the proofs for all of our theorem statements in Appendix E. We also substantiate our claims by providing empirical evidence using synthetic and real world data.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: Yes, we discuss some limitations of our work in the last subsection of Section 6. We also state the assumptions made by our framework in Sections 3, 4 and 5.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

   Justification: Yes, we state all main theorems and the full set of assumptions that accompany them in Sections 3, 4, and 5. We include any additional results, that are not essential to the main argument, but still interesting in Appendix X. We provide the full proof derivations and any additional lemmas used in proofs in Appendix F and G.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: We include the experimental details needed to reproduce the synthetic datasets. We provide the empirical results in Section 6, and include further details such as a description of the data generating process for the simulated data, description of the datasets, prompts used to query language models, and computer specifications to run experiments on in Appendix H. The data set provided by nonprofit collaborators cannot be released for privacy reasons, so instead we describe the features of the data and detailed instructions on how each variable was constructed, but we cannot release this dataset publicly. To compensate for this, we run experiments on simulated data and a second real-world dataset that is publicly available.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: We plan to release the code for reproducing all the experimental results on synthetic and semi-synthetic data, along with scripts for reproducing the simulate data and running our algorithm. The street outreach data is private data that was released to us under a data use agreement, but it cannot be released publicly for privacy reasons. However, we will release our other experimental results and code with the final version of the paper.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

Justification: Yes, we describe the data details and construction in the experimental results Sections 6 of the main paper. We go into much more detail about the data generation process and model tuning details in Appendix H.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: We include error bands in all of our main results in Figures 2 and 3.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: In Section 6, we provide sufficient information on what is needed to reproduce the experiments, such as running LLM predictions offline and in batch and reference the models used to run each experiment, such as random forest. We specify the type of compute resources in more detail in Appendix H.

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: We conform to all aspects of the NeurIPS Code of Ethics. We ensure author anonymity by also removing identifying information about our nonprofit collaboration.

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: Throughout the paper, we highlight the potential positive impacts especially in the introduction and motivation of this work. We discuss limitations of our method that could potentially have negative societal impacts in the limitations section of the paper. We go into more detail about the potential negative effects in an Impact Statement and the steps that we take to mitigate these impacts in Appendix A.

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [Yes]

    Justification: Yes, we mention how the data was released to our research team through a collaboration with a nonprofit reviewed by Institutional Review Boards at author universities. We discuss more about the steps taken to preserve the data privacy when training models in the Impact Statement in Appendix A. We do not release any large models, but we do plan to release scripts that reproduce our results on the sythethic data and publicly available data.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: Yes, we cite all the original owners of the code, data, and models in the main text and in Appendix H.

13. **New assets**

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [Yes]

Justification: Yes, we include well documented code to run our algorithm and reproduce our experimental results in an anonymized zip file included in supplementary material.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: While one of our experiments includes real data about a vulnerable community, this data was historical data collected by a nonprofit organization. None of this research included crowdsourcing or directly involved human subjects.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [Yes]

    Justification: Yes, we mention that IRB approval was obtained for the use of the street outreach data in Section 6.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

    Answer: [Yes]

    Justification: Yes, we describe our use of LLMs to get predictions from complex embedded outcome data, i.e. text data in Section 6. We explain in detail how we run the experiments on a secure HIPAA compliant cloud platform in Appendix H.

A Impact Statement

Our work deals with sensitive information about a vulnerable community so care must be taken when deploying our methods. The case notes are redacted by the organization, and any sensitive information is removed from the notes. Furthermore, we use local LLMs accessed through a HIPAA-compliant fire-walled cloud instance to mitigate ensure the privacy of clients. We work in collaboration with a nonprofit to ensure that the necessary guardrails are in place and that their data is used responsibly and in line with their mission.

B Notation

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
$f(\tilde{Y}_i)$	Estimated function of complex embedded outcomes, e.g. zero-shot LLM prediction from raw text
$\mu_z(X_i, f(\tilde{Y}_i))$	Estimated model predicting Y as function of $f(\tilde{Y})$ alone or $(X, f(\tilde{Y}))$

C 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 [1] 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.

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 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 [28], Chaudhuri et al. [5]. 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.

Relationship to causal inference and NLP There is a large and rapidly growing literature on causal inference with text data [11, 24, 25]. Throughout, we have deliberately used the terminology of measurement error to characterize our approach: that text measures outcomes of interest. [10] 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" [22], 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) [18]. Analogously, we view the underlying ground-truth outcomes Y as causing the measurement thereof, $\tilde{Y}$.

D Diagram of Cross-fitting Procedure

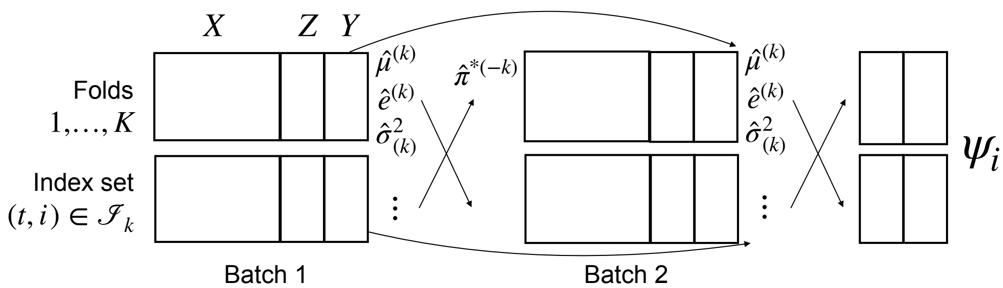

Figure 2: Illustration of cross-fitting (K folds within batches)

E Algorithm

Algorithm 1 (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 $X, \tilde{Y}$, 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. (OPT (global budget)), 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 **??** and out of fold nuisances $\pi^{*(-k)}$, $\hat{\mu}_z^{(-k)}$, $\hat{\sigma}_z^{2(-k)}$, and $\hat{e}_z^{(-k)}$

F Additional Results

F.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:

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

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

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

.

G Proofs

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 is

$$\begin{aligned}\mathrm{Var}[\psi_z - \psi_{z'}] &= \mathrm{Var}\left[\frac{Z \cdot R \cdot [Y - \mu_z(X)]}{e_z(X) \cdot \pi(z, X)} + \mu_z(X) - \frac{Z' \cdot R \cdot [Y - \mu_{z'}(X)]}{e_{z'}(X) \cdot \pi(z', X)} + \mu_{z'}(X)\right] \\ &= \underbrace{\mathrm{Var}\left[\frac{Z \cdot R \cdot [Y - \mu_z(X)]}{e_z(X) \cdot \pi(z, X)} + \mu_z(X)\right]}_{V_1} + \underbrace{\mathrm{Var}\left[\frac{Z' \cdot R \cdot [Y - \mu_{z'}(X)]}{e_{z'}(X) \cdot \pi(z', X)} + \mu_{z'}(X)\right]}_{V_2} \\ &\quad - \underbrace{2\mathrm{Cov}\left[\frac{Z \cdot R \cdot [Y - \mu_z(X)]}{e_z(X) \cdot \pi(z, X)} + \mu_z(X), \frac{Z' \cdot R \cdot [Y - \mu_{z'}(X)]}{e_{z'}(X) \cdot \pi(z', X)} + \mu_{z'}(X)\right]}_{V_3}\end{aligned}$$

For V_3:

$$\begin{aligned}&2\mathrm{Cov}\left[\frac{Z \cdot R \cdot [Y - \mu_z(X)]}{e_z(X) \cdot \pi(z, X)} + \mu_z(X), \frac{Z' \cdot R \cdot [Y - \mu_{z'}(X)]}{e_{z'}(X) \cdot \pi(z', X)} + \mu_{z'}(X)\right] \\ =&2\Bigg[\mathbb{E}\left[\frac{Z \cdot R}{e_z(X) \cdot \pi(z, X)} \underbrace{[\mathbb{E}[Y|Z, R=1, X] - \mu_z(X)]}_{=0}\right]\Bigg] \\ &+\Bigg[\mathbb{E}\left[\mu_z(X) \cdot \frac{Z' \cdot R}{e_{z'}(X) \cdot \pi(z', X)} \underbrace{[\mathbb{E}[Y|Z', R=1, X] - \mu_{z'}(X)]}_{=0} + \mu_{z'}(X)\right]\Bigg] \\ &- \mathbb{E}\left[\frac{Z \cdot R}{e_z(X) \cdot \pi(z, X)} \underbrace{[\mathbb{E}[Y|Z, R=1, X] - \mu_z(X)]}_{=0} + \mu_z(X)\right] \\ &\times \mathbb{E}\left[\frac{Z' \cdot R}{e_{z'}(X) \cdot \pi(z', X)} \underbrace{[\mathbb{E}[Y|Z', R=1, X] - \mu_{z'}(X)]}_{=0} + \mu_{z'}(X)\right]\Bigg] \\ =&2\Big[\mathbb{E}[\mu_z(X) \cdot \mu_{z'}(X)] - \mathbb{E}[\mu_z(X)\mu_{z'}(X)]\Big]\end{aligned}$$

For V_1**:**

$$\begin{aligned}&\mathrm{Var}\left[\frac{Z \cdot R \cdot [Y - \mu_z(X)]}{e_z(X) \cdot \pi(z, X)} + \mu_z(X)\right] \\ &= \mathrm{Var}\left[\frac{Z \cdot R \cdot [Y - \mu_z(X)]}{e_z(X) \cdot \pi(z, X)}\right] + \mathrm{Var}[\mu_z(X)] + 2\underbrace{\mathrm{Cov}\left[\frac{Z \cdot R \cdot [Y - \mu_z(X)]}{e_z(X) \cdot \pi(z, X)}, \mu_z(X)\right]}_{=0} \\ &= \mathbb{E}\left[\left[\frac{Z \cdot R \cdot [Y - \mu_z(X)]}{e_z(X) \cdot \pi(z, X)}\right]^2\right] - \left[\frac{Z \cdot R \cdot}{e_z(X) \cdot \pi(z, X)} \underbrace{[\mathbb{E}[Y|Z, R=1, X] - \mu_z(X)]}_{=0}\right]^2 \\ &+ \mathbb{E}\left[\mu_z(X)^2\right] - \mathbb{E}[\mu_z(X)]^2 \\ &= \mathbb{E}\left[\left[\frac{Z^2 \cdot R^2}{e_z^2(X) \cdot \pi^2(z, X)} \cdot [Y - \mu_z(X)]^2\right]\right] + \mathbb{E}\left[\mu_z(X)^2\right] - \mathbb{E}[\mu_z(X)]^2 \\ &= \mathbb{E}\left[\frac{Z \cdot R}{e_z^2(X) \cdot \pi^2(z, X)} \cdot [Y - \mu_z(X)]^2\right] + \mathbb{E}\left[\mu(z, 1, X)^2\right] - \mathbb{E}[\mu_z(X)]^2 \\ &= \mathbb{E}\left[\frac{1}{e_z(X) \cdot \pi(z, X)} \cdot [Y - \mu_z(X)]^2\right] + \mathbb{E}\left[\mu_z(X)^2\right] - \mathbb{E}[\mu_z(X)]^2\end{aligned}$$

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

$$\begin{aligned}
\mathrm{Var}[\psi_z - \psi_{z'}] &= \mathbb{E}\left[\frac{1}{e_z(X)\cdot\pi(z,X)}\cdot[Y-\mu_z(X)]^2\right] + \mathbb{E}\left[\frac{1}{e_{z'}(X)\cdot\pi(z',X)}\cdot[Y-\mu_{z'}(X)]^2\right] \\
&+ \mathbb{E}\left[(\mu_z(X)-\mu_{z'}(X))^2\right] - \mathbb{E}\left[\mu_z(X)-\mu_{z'}(X)\right]^2 \\
&= \mathbb{E}\left[\frac{1}{e_z(X)\cdot\pi(z,X)}\cdot[Y-\mu_z(X)]^2\right] + \mathbb{E}\left[\frac{1}{e_{z'}(X)\cdot\pi(z',X)}\cdot[Y-\mu_{z'}(X)]^2\right] \\
&+ \mathrm{Var}\Bigg[\mu_z(X)-\mu_{z'}(X)\Bigg]
\end{aligned}$$

Rewriting the bound from Hahn (1998), we get

$$\begin{aligned}
V \geq\ & \mathbb{E}\left[\frac{1}{e_z(X)\cdot\pi(z,X)}\cdot[Y-\mu_z(X)]^2\right] + \mathbb{E}\left[\frac{1}{e_{z'}(X)\cdot\pi(z',X)}\cdot[Y-\mu_{z'}(X)]^2\right] \\
&+ \mathrm{Var}\Bigg[\mu_z(X)-\mu_{z'}(X)\Bigg]
\end{aligned}$$

□

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:

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

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 $p(X=x \mid Z=z) = \frac{e_z(x)p(x)}{p(Z=z)}$

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

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 $\lambda_z^* = {p(Z=z)\mathbb{E}\left[\sqrt{\mathbb{I}[Z=z]\sigma^2(X)/e_z^2(X)}\right]^2}/{B_z^2}$ in the prior parametrized expression.

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

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:

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

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 **??** . The objective function arises from the asymptotic variance expression in [9, 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.

Then the Lagrangian is

$$\int \frac{\sigma^2(z \mid x)}{e(z,x)\pi(z,x)} f(x)dx + \lambda \left(\int \int \pi(z,x) K_h(z'-z) e(z',x) dz' f(x) dx \right)$$

Define the kernel localization of $e(z,x)$ around z under the kernel function $K_h(z'-z)$:

$$\tilde{e}_h(z,x) = \int K_h(z'-z) e(z',x) dz'$$

Taking derivatives with respect to $\pi(z,x)$, we obtain the FOC

$$\nabla_{\pi(t|x)} \mathcal{L} = \frac{-\sigma^2(z,x)}{e(z,x)\pi(z,x)^2} f(x) + \lambda \tilde{e}_h(z,x) f(x) = 0$$

Solving the FOC, we obtain

$$\frac{-\sigma^2(z,x)}{e(z,x)\pi(z,x)^2} + \lambda \tilde{e}_h(z,x) = 0 \implies \pi^*(z,x) = \frac{1}{\lambda} \frac{\sqrt{\sigma^2(z,x)}}{e(z,x)} \sqrt{\frac{e(z,x)}{\tilde{e}_h(z,x)}}$$

We conclude that

$$\pi^*(z,x) \propto \frac{\sqrt{\sigma^2(z,x)}}{e(z,x)} \sqrt{\frac{e(z,x)}{\tilde{e}_h(z,x)}}$$

□

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, $\hat{\tau}_{AIPW}^{(t,k)}$, as follows:

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

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

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

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.

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

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

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

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:

$$
\begin{aligned}
&\frac{Z_i\tilde{R}_i(Y_i - \hat{\mu}_1(X_i))}{\hat{e}_1(X_i)\hat{\pi}(1, X_i)} - \frac{Z_i\tilde{R}_i(Y_i - \mu_1(X_i))}{e_1(X_i)\pi(1, X_i)} + (\hat{\mu}_1(X_i) - \mu_1(X_i)) \\
&= (\mu_1(X_i) - \hat{\mu}_1(X_i))\left(\frac{Z_i\tilde{R}_i}{e_1(X_i)\pi(1, X_i)} - 1\right) + Z_i\tilde{R}_i(Y_i - \hat{\mu}_1(X_i))\left(\frac{1}{\hat{e}_1(X_i)\hat{\pi}(1, X_i)} - \frac{1}{e_1(X_i)\pi(1, X_i)}\right) \\
&\qquad\qquad \text{(by } \pm\tfrac{Z_i\tilde{R}_i(Y_i - \hat{\mu}_1(X_i))}{e_1(X_i)\pi(1,X_i)}\text{)} \\
&= (\mu_1(X_i) - \hat{\mu}_1(X_i))\left(\frac{Z_i\tilde{R}_i}{e_1(X_i)\pi(1, X_i)} - 1\right) + Z_i\tilde{R}_i(Y_i - \mu_1(X_i))\left(\frac{1}{\hat{e}_1(X_i)\hat{\pi}(1, X_i)} - \frac{1}{e_1(X_i)\pi(1, X_i)}\right) \\
&\quad + Z_i\tilde{R}_i(\mu_1(X_i) - \hat{\mu}_1(X_i))\left(\frac{1}{\hat{e}_1(X_i)\hat{\pi}(1, X_i)} - \frac{1}{e_1(X_i)\pi(1, X_i)}\right) \\
&\qquad\qquad \text{(by } \pm Z_i\tilde{R}_i\mu_1(X_i)(\tfrac{1}{\hat{e}_1(X_i)\hat{\pi}(1,X_i)} - \tfrac{1}{e_1(X_i)\pi(1,X_i)})\text{)} \\
&= (\mu_1(X_i) - \hat{\mu}_1(X_i))\left(\frac{Z_i\tilde{R}_i}{e_1(X_i)\pi(1, X_i)} - 1\right) \\
&\quad + Z_i\tilde{R}_i(Y_i - \mu_1(X_i))\left(\hat{\pi}(1, X_i)^{-1}(\hat{e}_1(X_i)^{-1} - e_1(X_i)^{-1}) + e_1(X_i)^{-1}(\hat{\pi}(1, X_i)^{-1} - \pi(1, X_i)^{-1})\right) \\
&\quad + Z_i\tilde{R}_i(\mu_1(X_i) - \hat{\mu}_1(X_i))\left(\hat{\pi}(1, X_i)^{-1}(\hat{e}_1(X_i)^{-1} - e_1(X_i)^{-1}) + e_1(X_i)^{-1}(\hat{\pi}(1, X_i)^{-1} - \pi(1, X_i)^{-1})\right) \\
&\qquad\qquad \text{(by } \pm\tfrac{1}{e\hat{\pi}}\text{)}
\end{aligned}
$$

We want to show that

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

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:

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

Bound for third term:

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

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.

$$
\begin{aligned}
&\mathbb{E}\Big[\frac{1}{\sqrt{n_{t,k}}}\sum_{i:(t,i)\in\mathcal{I}_k}\Big((\mu_1(X_i)-\hat{\mu}_1^{(-k)}(X_i))\Big(\frac{Z_i\tilde{R}_i}{e_1(X_i)\pi(1,X_i)}-1\Big)\Big)^2 \mid \mathcal{I}_{(-k)},\{X_i\}\Big]\\
&\quad=\mathrm{Var}\Big[\frac{1}{\sqrt{n_{t,k}}}\sum_{i:(t,i)\in\mathcal{I}_k}(\mu_1(X_i)-\hat{\mu}_1^{(-k)}(X_i))\Big(\frac{Z_i\tilde{R}_i}{e_1(X_i)\pi(1,X_i)}-1\Big) \mid \mathcal{I}_{(-k)},\{X_i\}\Big]\\
&\quad=\frac{1}{n_{t,k}}\sum_{i:(t,i)\in\mathcal{I}_k}\mathbb{E}\Big[(\mu_1(X_i)-\hat{\mu}_1^{(-k)}(X_i))^2\Big(\frac{Z_i\tilde{R}_i}{e_1(X_i)\pi(1,X_i)}-1\Big)^2 \mid \mathcal{I}_{(-k)},\{X_i\}\Big]\\
&\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\text{(expectation of } (\tfrac{Z_i\tilde{R}_i}{e_1(X_i)\pi(1,X_i)}-1)^2)\\
&\quad=\frac{1}{n_{t,k}}\sum_{i:(t,i)\in\mathcal{I}_k}\frac{1-e_1(X_i)\pi(z,X_i)}{e_1(X_i)\pi(1,X_i)}(\mu_1(X_i)-\hat{\mu}_1^{(-k)}(X_i))^2\\
&\quad\le\frac{1-\nu_e\lambda_\pi}{\nu_e\lambda_\pi}\frac{1}{n_{t,k}}\sum_{i:(t,i)\in\mathcal{I}_k}((\mu_1(X_i)-\hat{\mu}_1^{(-k)}(X_i))^2=o_p(\frac{1}{n^{1+2r_\mu}})
\end{aligned}
$$

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 $r_\mu\ge 0$, 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.

$$
\begin{aligned}
&\mathbb{E}\Big[\frac{1}{\sqrt{n_{t,k}}}\sum_{i:(t,i)\in\mathcal{I}_k}\Big(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)\Big)^2 \mid \mathcal{I}_{(-k)},\{X_i\}\Big]\\
&+\mathbb{E}\Big[\frac{1}{\sqrt{n_{t,k}}}\sum_{i:(t,i)\in\mathcal{I}_k}\Big(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)\Big)^2 \mid \mathcal{I}_{(-k)},\{X_i\}\Big]\\
&\quad=\mathrm{Var}\Big[\frac{1}{\sqrt{n_{t,k}}}\sum_{i:(t,i)\in\mathcal{I}_k}\Big(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\}\Big]\\
&+\mathrm{Var}\Big[\frac{1}{\sqrt{n_{t,k}}}\sum_{i:(t,i)\in\mathcal{I}_k}\Big(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\}\Big]\\
&\quad=\frac{1}{n_{t,k}}\sum_{i:(t,i)\in\mathcal{I}_k}\mathbb{E}\Big[\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\}\Big]\\
&+\frac{1}{n_{t,k}}\sum_{i:(t,i)\in\mathcal{I}_k}\mathbb{E}\Big[\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\}\Big]\\
&\quad=\frac{1}{n_{t,k}}\sum_{i:(t,i)\in\mathcal{I}_k}\frac{e_1^2(X_i)\pi^2(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_1^2(X_i)(\pi^{(-k)}(z,X_i))^2}{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\\
&\quad\le\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\\
&\quad=o_p(\frac{1}{n^{1+2r_e+2r_\pi}})
\end{aligned}
$$

where the last inequality is because $\sigma^2(X)$ is bounded above, $\sigma^2(X)\le B_{\sigma^2}$, 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. □

H Additional Lemmas

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

Lemma 1 (Conditional convergence implies unconditional convergence, from [6]). *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), [26]). *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*

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

H.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 **??**. 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,

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

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

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

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:

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

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 (2), which bounds the first term of Equation (1). 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 $\mathbb{E}[\cdot]$ instead of $\mathbb{E}_n[\cdot]$. Then the second term of Equation (1) 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),

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

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

I Additional Experiment, Details and Discussion

I.1 Additional details

All experiments using our full algorithm 1 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 SVM model 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 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.

I.2 Synthetic Data

Before running our batch adaptive algorithm, we split the data into a validation set (35% of data) in which we estimate the ATE on. Then we use the remainder 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 $\epsilon_0 \sim \mathcal{N}(0, \sigma_0(X))$ and $\epsilon_1 \sim \mathcal{N}(0, \sigma_1(X))$. The observed outcomes are $Y = Z \cdot Y(1) + (1 - Z) \cdot Y(0)$.

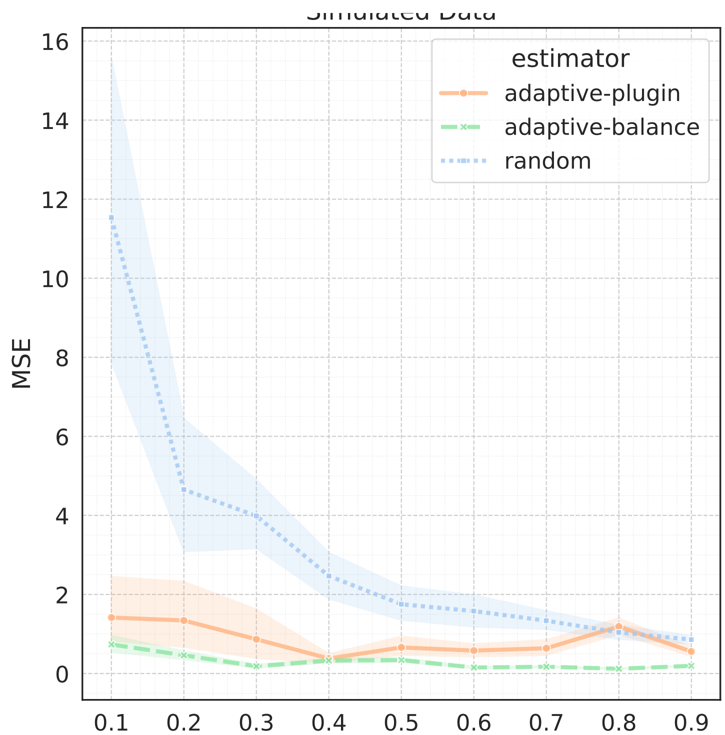

Figure 3: Mean squared error between estimated ATE and true ATE averaged over 100 trials across varying budgets.

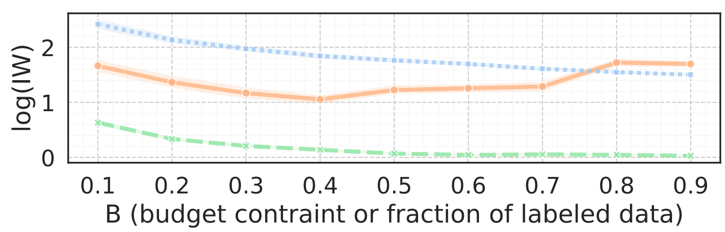

Figure 4: Average confidence interval width averaged over 100 trials across varying budgets.

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.

To stabilize the estimation of the inverse annotation probabilities, we use the plug-in estimator following **??** and the ForestReisz method to estimate the balancing weights [7].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.

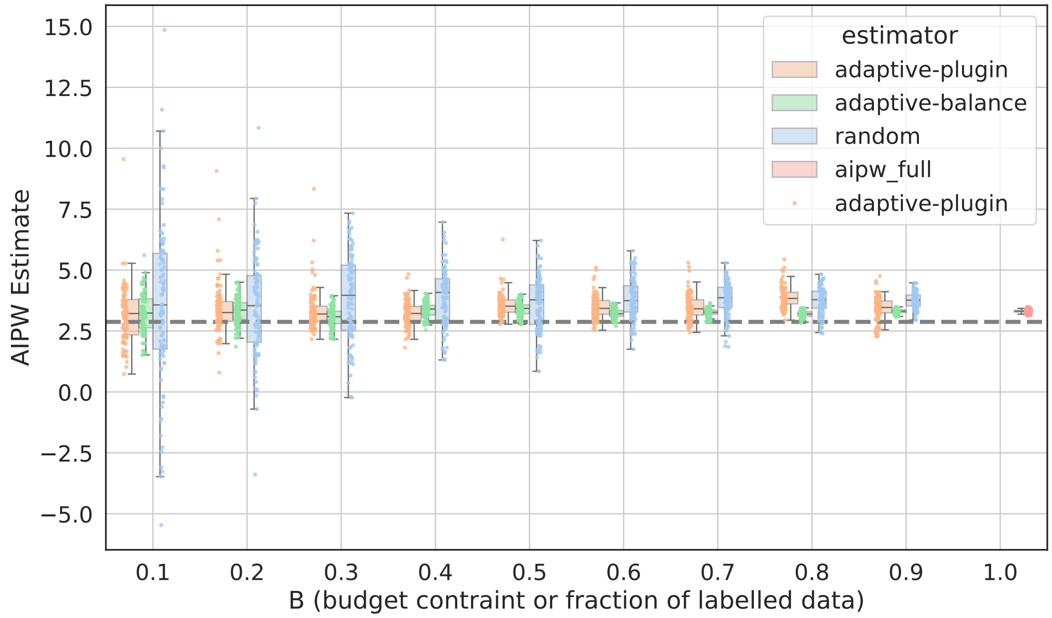

Figure 5: Boxplots of ATE estimates compared to skyline $\hat{\tau}_{AIPW}$ when the labeling budget is the entire dataset in red and the grey dotted line is τ.

I.3 Real-world Dataset Details

We provide further details about the treatment, covariates and outcomes for each dataset. Table 1 and table 2 describe the variables in the retail hero and outreach datasets, respectively. We refer the reader to [10] 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 **??**, our method does well even when the propensity scores do not have good overlap.

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 1: Covariate, treatment, and outcome descriptions and discrete category definitions for Retail-Hero dataset.

I.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

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

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

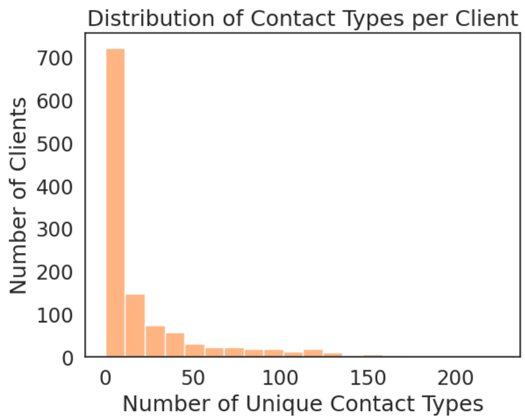

Figure 6: Distribution of street outreach engagements for client population.

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.

I.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 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.

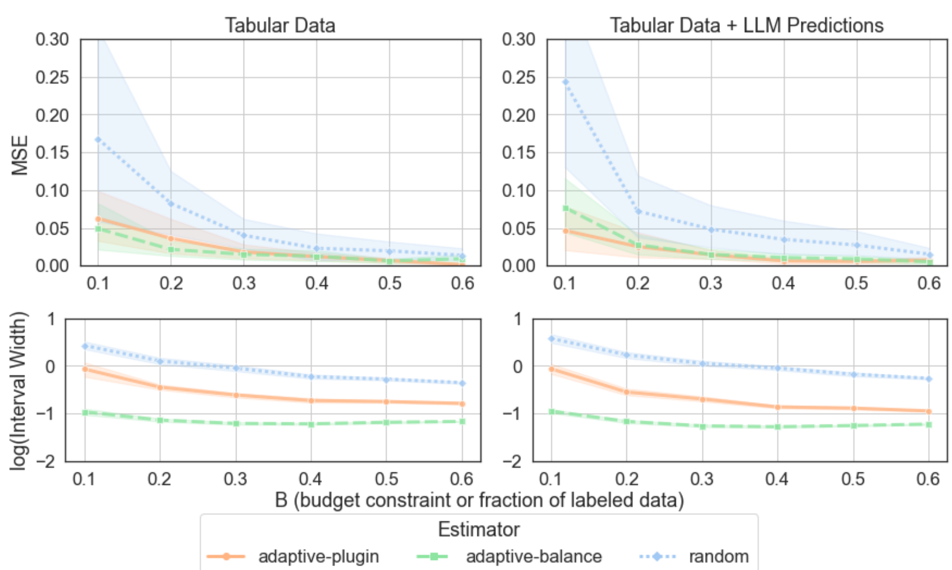

Figure 7: **Street outreach data with post-treatment summaries only.** Mean squared error and 95% confidence interval width averaged over 20 trials across budget percentages of the data. This plot makes use of tabular data and the best-performing random forest outcome model (left) and text-encoded outcomes using LLMs (right).

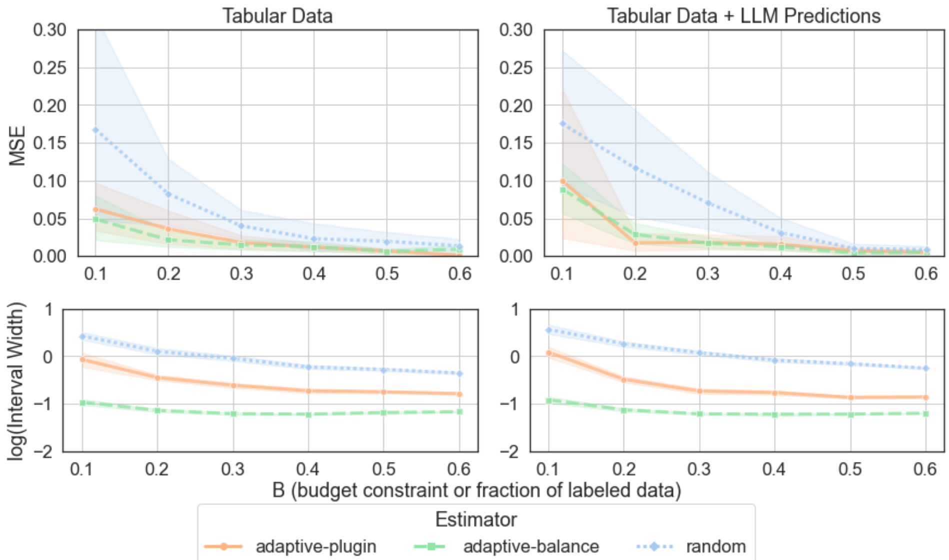

Figure 8: **Street outreach data with pre- and post-treatment summaries.** Mean squared error and 95% confidence interval width averaged over 20 trials across budget percentages of the data. This plot makes use of tabular data and the best-performing random forest outcome model (left) and text-encoded outcomes using LLMs (right).

In Figure 7 and Figure 8, 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 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.

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

I.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.

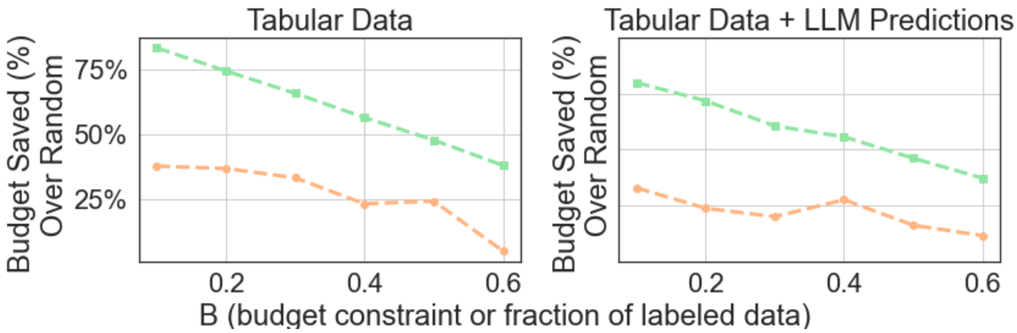

Figure 9: **RetailHero Data.** Budget saved due to batch adaptive annotation. The reduction in annotation sample size needed to achieve the same confidence interval width with batch adaptive annotation on tabular data (left) and on tabular data + complex embedded outcomes (right) compared to random sampling.

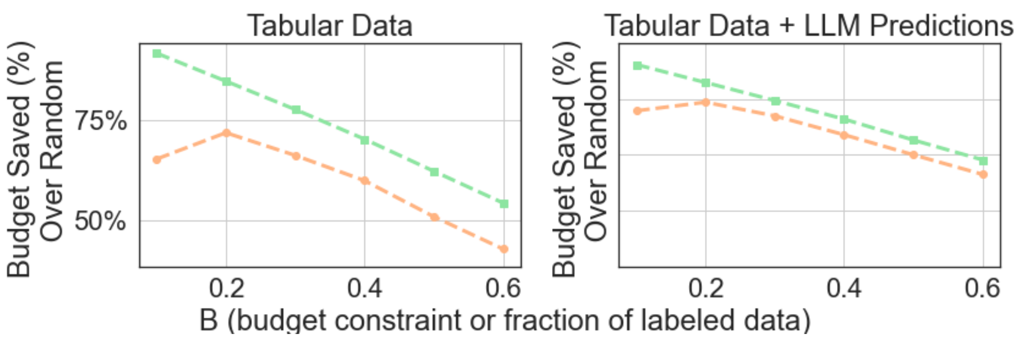

Figure 10: **Street Outreach Data.** Budget saved due to batch adaptive annotation. The reduction in annotation sample size needed to achieve the same confidence interval width with batch adaptive annotation on tabular data (left) and on tabular data + complex embedded outcomes (right) compared to random sampling.

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. [4] and Gentile et al. [13] (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: $\frac{\sqrt{100}}{0.5} = 10$

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 $P(\hat{Y} = 1 \mid x)$ [23]. For the regression tasks, we use Expected Model Variance Reduction [8], Expected Model Change Maximization [3], and Improved Greedy Sampling [27]; 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 3: 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 4: 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 5: Averaged MSEs for Street Outreach Data.

Gentile et al. [13] 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.

I.8 LLM Prompts

Prompt 1 (Retail Hero):

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.
—
$\{post\}$
–_

Prompt 2 (Street Outreach Casenote Summaries) :

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.

Prompt 2 (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: $\{examples\}$