Progressive Generalization Risk Reduction for Data-Efficient Causal Effect Estimation

Hechuan Wen The University of Queensland Brisbane, Australia h.wen@uq.edu.au

Li Kheng Chai Health and Wellbeing Queensland Brisbane, Australia likheng.chai@hw.qld.gov.au Tong Chen
The University of Queensland
Brisbane, Australia
tong.chen@uq.edu.au

Shazia Sadiq The University of Queensland Brisbane, Australia shazia@eecs.uq.edu.au Guanhua Ye Beijing University of Posts and Telecommunications, China g.ye@bupt.edu.cn

Hongzhi Yin*
The University of Queensland
Brisbane, Australia
h.yin1@uq.edu.au

Abstract

Causal effect estimation (CEE) provides a crucial tool for predicting the unobserved counterfactual outcome for an entity. As CEE relaxes the requirement for "perfect" counterfactual samples (e.g., patients with identical attributes and only differ in treatments received) that are impractical to obtain and can instead operate on observational data, it is usually used in high-stake domains like medical treatment effect prediction. Nevertheless, in those high-stake domains, gathering a decently sized, fully labelled observational dataset remains challenging due to hurdles associated with costs, ethics, expertise and time needed, etc., of which medical treatment surveys are a typical example. Consequently, if the training dataset is small in scale, low generalization risks can hardly be achieved on any CEE algorithms.

Unlike existing CEE methods that assume the constant availability of a dataset with abundant samples, in this paper, we study a more realistic CEE setting where the labelled data samples are scarce at the beginning, while more can be gradually acquired over the course of training - assuredly under a limited budget considering their expensive nature. Then, the problem naturally comes down to actively selecting the best possible samples to be labelled, e.g., identifying the next subset of patients to conduct the treatment survey. However, acquiring quality data for reducing the CEE risk under limited labelling budgets remains under-explored until now. To fill the gap, we theoretically analyse the generalization risk from an intriguing perspective of progressively shrinking its upper bound, and develop a principled label acquisition pipeline exclusively for CEE tasks. With our analysis, we propose the Model Agnostic Causal Active Learning (MACAL) algorithm for batchwise label acquisition, which aims to reduce both the CEE model's uncertainty and the post-acquisition distributional imbalance simultaneously at each acquisition step. Extensive experiments are

*Corresponding Author

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

KDD '25, August 3-7, 2025, Toronto, Canada

© 2025 Copyright held by the owner/author(s). Publication rights licensed to ACM. ACM ISBN XXX-X-XXXXX-XXXXX-X/25/08 https://doi.org/XXXXXXXXXXXXXXX

gain from MACAL is observed over state-of-the-art active learning baselines. The implementation repository is open-sourced at: https://github.com/uqhwen2/MACAL.

conducted on three datasets, where a clear empirical performance

CCS Concepts

Computing methodologies → Machine learning.

Keywords

Causal Effect Estimation, Active Learning, Generalization Risk Reduction

ACM Reference Format:

1 Introduction

Understanding causal effects to support decision-making in high-stake domains is crucial, where typical examples include randomized control trials in medication [26], A/B testing for business decision-making [21], and the potential in advancing big data management [13, 24, 25, 50]. As performing large-scale and statistically reliable human tests is prohibitively costly, algorithms for causal effect estimation (CEE) using passively observed data samples have become a promising solution [18, 40, 48]. In short, a CEE algorithm is trained with observational data to predict the counterfactual outcome for an entity, e.g., what the outcome will be if a patient received the other treatment, instead of the one already had.

To perform CEE, a common practice is to build a regression model that estimates a continuous effect value [4, 18, 33, 48, 49], which is trained on the observational data containing two groups of samples. The groups are formed based on the treatment imposed on each sample, where each sample consists of raw attributes drawn from a well-defined feature space (e.g., a patient's health indicators), and a label that corresponds to the observed outcome after receiving the binary treatment (e.g., blood sugar concentration after taking one diabetes medicine).

 $^{^{1}\}mathrm{Depending}$ on the context of applications, treatments can also be interpreted as interventions, services, or information provided to an individual.

For training a capable CEE model, a quality observational dataset with diverse and abundant samples is highly desirable. On the one hand, as in many other tasks, richer training data enables the model to better capture predictive patterns. On the other hand, this also helps maintain some pivotal CEE assumptions [14] on the training data, where the positivity (a.k.a. overlapping) assumption is arguably a very fundamental one. Practically, as each distinct sample only receives one treatment, positivity requires statistically identical attribute distributions between two treatment groups, such that counterfactual predictions can be confidently made. Given that, the majority of CEE models [23, 33, 42, 49] are trained on a fix-sized dataset with sufficient samples, where the positivity assumption can easily hold. However, such a setting oversimplifies the data availability in high-stake domains - the major adopters of CEE. The challenge often lies in obtaining the ground truth label on the treatment outcome of each sample. For instance, though clinics record patients' health-related attributes when performing a treatment, the real post-treatment outcome can only be obtained through longitudinal surveys [30] over a long time period, and is subject to ethical concerns. Furthermore, in a business context, the effect of a treatment (e.g., a sales campaign) cannot be reliably quantified without ample expertise and evidence. As a result, the sufficiency of labelled training data in CEE is not always guaranteed, hindering the real-life practicality of existing CEE pipelines.

As a response, in this paper, we subsume CEE under a more realistic setting: the availability of labelled samples in both treatment groups monotonically grows. Essentially, this translates into an active learning (AL) paradigm [32] for CEE tasks. Considering the expensive nature of labelling the treatment outcome of all samples, we allow a CEE model's training to start with a very small portion of labelled data within both treatment groups, then gradually and selectively extend to the remaining unlabelled samples by assigning post-treatment outcome labels. As a side effect of this more practical setting, a CEE model with low generalization risk is harder to obtain, especially at early training stages where the labelled dataset is small in scale, limiting the informativeness and compliance to positivity. With a standard AL algorithm [9, 29, 31, 41, 51], during the progress of label acquisition, informative samples can be selected for labelling and enriching the training data. In scenarios where counterfactual predictions are not needed, such active label acquisition is proven useful [29]. However, in CEE, with the existence of two treatment groups, the direct adoption of AL will incur sub-optimal results within the given labelling budget. This is because the acquisition criterion is not designed to account for the crucial positivity assumption, thus failing to align the sample distributions between the two groups. Consequently, the generalizability of the actively trained CEE algorithm will be harmed by the ill-posed data distribution.

Bearing this motivation, we aim to answer the important question: how to label the most informative samples in CEE tasks? Assuming the label availability of the samples in both treatment groups, active learning for CEE should meet two desiderata: 1) maximize the positivity among the chosen samples to be labelled during dataset expansion; 2) improve the generalizability of the CEE model. Recently, this niche area of study has started drawing more attention, however, as we will discuss later, only a few models [1, 17, 28] are suited for CEE task with AL. In this paper, we propose an

intriguing perspective to conduct efficient selective labelling exclusively for the CEE task under the AL paradigm. Unlike the other closely-related approaches [28], we look directly into the theoretical analysis of the risk upper bound without loosening it, where we then propose a theory-inspired, simplified yet effective label acquisition criterion for batch-mode AL with paired samples. We summarize our contributions as follows:

- We study the well-under-explored yet important and practical topic active learning for causal effect estimation, where the conventional active learning paradigm failed to obtain the optimal label acquisition scheme, and the existing studies are yet able to well solve the violation of positivity during the process of the label acquisition.
- We propose a theoretical framework for causal effect estimation under the active learning paradigm, where a more informative risk upper bound is decomposed and leveraged for algorithm design. Inspired by the proposed theory, we come up with a simplified yet effective label acquisition criterion, namely Model Agnostic Causal Active Learning (MACAL) for label acquisition by promoting the individual sample diversity in different treatment groups and penalizing the treatment pair dissimilarity. Also, the mathematically guaranteed risk convergence is given under certain conditions to justify the acquisition algorithm.
- We compete against numerous SOTA baselines by fixing the label acquisition criterion as the only variable during benchmarking. Extensive experiments are conducted on various combinations of different datasets and downstream CEE models, and demonstrable performance gain from MACAL is observed across all comparisons.

2 Preliminaries

2.1 Causal Effect Estimation

Under the potential outcome framework [14], the individual treatment effect (ITE) is expected to be estimated with the tabular dataset $\mathcal{D} = \{\mathbf{x}_i, t_i, y_i\}_{i=1}^N$, where \mathbf{x}_i, t_i, y_i are respectively the raw feature variables, observed treatment, treatment outcome that correspond to the i-th individual. For simplicity, we consider the binary treatment t of 1 and 0 to denote the different treatment statuses, respectively. The ground truth ITE for an individual with feature vector \mathbf{x} is defined as:

$$\tau(\mathbf{x}) = \mathbb{E}[Y^{t=1} - Y^{t=0}|\mathbf{x}],\tag{1}$$

where $Y^{t=1}$ and $Y^{t=0}$ are the unobserved potential outcomes with treatment t=1 and t=0 respectively. Generally, under the deep neural network learning framework [23, 33], the common practice is to transform the raw variable \mathbf{x}_i into the latent representation \mathbf{z}_i for individual i, then such representation is utilised for ITE prediction. To evaluate the performance of the CEE model, the generalization risk of the model denoted by ϵ_{PEHE} , is defined in (2) according to the well-recognized literature [11]. The lower the value of ϵ_{PEHE} , the better the performance of the predictor.

DEFINITION 1. The expected Precision in Estimation of Heterogeneous Effect (PEHE) of the CEE model $f = \{\phi, \Phi\}$ with squared loss

metric $L(\cdot, \cdot)$ is defined as:

$$\epsilon_{PEHE}(f) = \int_{\mathcal{X}} L_f(\mathbf{x}) p(\mathbf{x}) d\mathbf{x},$$
 (2)

where we denote $L(\hat{\tau}(\mathbf{x}), \tau(\mathbf{x}))$ as $L_f(\mathbf{x})$ for notation simplicity. The $\tau(\mathbf{x})$ is the ground truth treatment effect defined in (1) and $\hat{\tau}(\mathbf{x})$ is its estimation.

For clarity, we formally define the CEE problem as follows:

DEFINITION 2 (CAUSAL EFFECT ESTIMATION). Given the dataset \mathcal{D} , the common pipeline is to train an estimator f that can give the causal effect estimation $\hat{\tau}(\mathbf{x})$ to be as accurate as possible to the ground truth $\tau(\mathbf{x})$ for entity \mathbf{x} , i.e., we aim to empirically minimize the evaluation metric $\epsilon_{PEHE}(f)$.

To validate the CEE $\hat{\tau}(\mathbf{x})$, three common assumptions from the causal inference literature are needed to lay the theoretical foundation. These assumptions are stated as follows:

ASSUMPTION 1 (STABLE UNIT TREATMENT VALUE ASSUMPTION (SUTVA) [14]). For any individual: (a) the potential outcomes for the individual do not vary with the treatment assigned to other individuals; and (b) there are no different forms or versions of each treatment that may lead to different potential outcomes.

ASSUMPTION 2 (UNCONFOUNDEDNESS). The independence relation $\{Y^{t=0}, Y^{t=1}\} \perp t \mid x$ holds, where treatment assignment t is independent to the potential outcomes $\{Y^{t=0}, Y^{t=1}\}$ given the covariate x.

Assumption 3 (Positivity). For every individual with feature covariate \mathbf{x} , the treatment assignment mechanism obeys: $0 < p(t = 1|\mathbf{x}) < 1$.

The causal effect identifiability stated in Proposition 1 is needed to finalize the validation of the estimation results. The proof of the proposition is provided in Appendix A.4.

Proposition 1 (Identifiability). The causal effect is identifiable if and only if the SUTVA, the unconfoundedness, and the positivity assumptions hold.

2.2 Active Learning as the Challenge and Opportunity

When CEE meets active learning, the general logistics of the query steps become: 1) Let the CEE model f get trained on the labelled training set $\mathcal{D}_{\text{train}} = \{\mathbf{x}_i, t_i, y_i\}_{i=1}^{N_{\text{train}}}$. 2) Given the unlabelled pool set $\mathcal{D}_{\text{pool}} = \{\mathbf{x}_i, t_i\}_{i=1}^{N_{\text{pool}}}$, the pre-defined label acquisition criterion (normally has trained model f embedded, e.g., uncertainty-aware CEE model [17]) examines through the pool set and returns a subset of it, i.e., $\tilde{\mathcal{D}}$, for the oracle to label. 3) The labelled subset $\tilde{\mathcal{D}}$ is added to the training set for which the CEE model f can get updated before the upcoming querying round starts, then return to Step 1). Note, that the samples with attributes are already available in the pool set but without the labels, the process only attaches ground truth labels to them. Such a recursive procedure terminates mostly when the desired performance is reached or the labelling budget gets exhausted. Subsequently, let's take one step further from the conventional CEE problem defined in Definition 2, we form the research problem of CEE under the AL paradigm as follows:

Definition 3 (Active Causal Effect Estimation). Causal effect estimations with active learning aims to expand the current dataset with more informative samples such that the trained model's estimation risk ϵ_{PEHE} can be significantly reduced before the exhaustion of the labelling budget.

It is noted that the Assumption 3 regarding positivity is a vulnerable one in the real-world scenario for the CEE research field. The risk of deviating from such an assumption and thus leading to the unidentifiable causal effect has been widely discussed in [16, 43]. As a result of the active causal effect estimation defined in Definition 3, if the selective labelling process keeps introducing more imbalance (e.g., a non-negative crucial part of the CEE risk upper bound [33]) to the current treatment groups' distributions, the use of AL to expand the labelled dataset will hardly help obtain optimal CEE model with significant estimation risk reduction. However, as we will discuss in the following, AL also sheds light on reducing the estimation risk by expanding the dataset if it is properly configured.

Assume that we initially have a large enough pool set \mathcal{D}_{pool} , but labelling all the samples, i.e., obtaining all ITEs, is infeasible due to the considerable cost of time and capital. Ideally, there exists a smallest optimal subset \mathcal{D}_{opt} where the positivity assumption holds across the sample space X. Additionally, treatment groups' distributions, $p_{\mathrm{opt}}^{t=1}$ and $p_{\mathrm{opt}}^{t=0}$, are identical such that the distributional discrepancy measured by integral probability metric (IPM) is statistically zero, i.e., $IPM(p_{opt}^{t=1}, p_{opt}^{t=0}) = 0$. Given a sparse warm-up set (at i = 0 query step), the current distributions of the treatment groups, i.e., $p_{i=0}^{t=1}$ and $p_{i=0}^{t=0}$, are realistically not the same, i.e., $\text{IPM}(p_{i=0}^{t=1}, p_{i=0}^{t=0}) = \mathcal{I}_{i=0} \neq 0$. During the recursive selective labelling process, a growing number of samples are added into the training set \mathcal{D}_{train} – the challenge is, the current disparity between different treatment groups' distributions can be amplified, e.g., $I_{i=10} \gg I_{i=0}$ after 10 uncontrolled query steps, thus further countering the positivity even with more data. Meanwhile, the opportunity is, with proper acquisition setup, we can not only reduce the imbalance in the training set after every query step, i.e., $I_{i+1} \ll I_i$, but also quickly converge to the optimal set by using the smallest budget, e.g., for each of the treatment group t, $\text{IPM}(p_{\text{opt}}^t, p_{i=1}^t) \to 0$ after I iterations, reaching the lowest risk.

Therefore, keep reconciling the positivity assumption during the active learning process plays a crucial role in obtaining a lower risk for CEE. In what follows, we detail our label acquisition design in every query step to fulfil this principle.

3 Methodology

3.1 Theory and Practice

In this paper, we focus on batch-mode active learning (BMAL). The reason for conducting batch-mode active learning is to acquire more samples at one query step for the oracle to label them, thus reducing the frequency of retraining the model in case the model training is costly. In the following, we propose a maximum risk upper bound reduction theorem for CEE with active learning, and the main proof of the general theorem – Theorem 1 is provided in Appendix A.1, followed by the sub-proofs for each of the convergence analysis in Appendix A.2 and A.3.

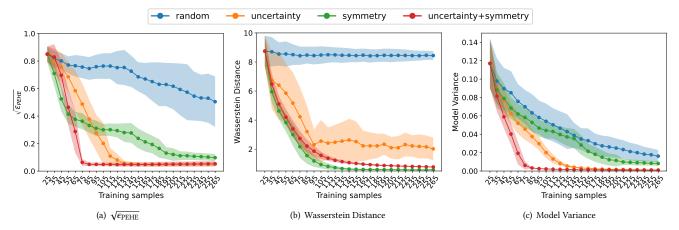


Figure 1: Toy example indicating the importance of considering the reduction of both distributional discrepancy and model variance to help quickly achieve the lowest risk.

Theorem 1. With budget \mathcal{M} , the maximum risk upper bound reduction $\Delta_{\mathcal{B}_{overall}}$ is achieved at the termination of the entire I data query steps given that the generalization risk upper bound shrinkage $\Delta_{\mathcal{B}_i}$ is maximized at each query step $\forall i$, i.e.:

$$\underset{\tilde{\mathcal{D}}_{overall}}{\arg \max} \Delta_{\mathcal{B}_{overall}} = \bigcup_{i=1}^{I} \underset{\tilde{\mathcal{D}}_{i}}{\arg \max} \Delta_{\mathcal{B}_{i}}$$

$$s.t. \ |\tilde{\mathcal{D}}_{overall}| \leq \mathcal{M},$$
(3)

where $\tilde{\mathcal{D}}_{overall}$ is the overall acquired data, $\tilde{\mathcal{D}}_i$ is the acquired batch at i-th query step, and $\Delta_{\mathcal{B}_i} = \sum_{t \in \{0,1\}} \Delta^t_{Var_i}(\tilde{\mathcal{D}}_i) + C_\phi \Delta_{IPM_i}(\tilde{\mathcal{D}}_i)$. The convergence rate of the risk upper bound has the following guaranteed behaviours under certain circumstances:

- i) When variance reduction $\sum_{t \in \{0,1\}} \Delta^t_{Var_i}(\tilde{\mathcal{D}}_i)$ becomes the dominant part of the risk upper bound, the risk convergence is lower-bounded by $\Omega(\beta^i)$ with constant $\beta \in [0,1)$.
- ii) While, with dominant constant C_{ϕ} , the risk convergence is upper-bounded by $O(\frac{1}{i+\gamma_0})$ with constant $\gamma_0 \in \mathbb{R}^+$.

Remark 1. The bound shrinkage $\Delta_{\mathcal{B}_i}$ at i-th query step constitutes the variance difference and the distributional discrepancy difference, solely focusing on the reduction of one term would not contribute to optimal risk reduction for active causal effect estimation, while a proper combination of such two terms can lead to the optimal result.

Toy Dataset: We design a fully synthetic 1-dimensional toy dataset and conduct experiments on four kinds of acquisition functions, i.e., Random, Uncertainty, Symmetry and Uncertainty + Symmetry (our proposed method MACAL), to illustrate the importance of considering both model variance and data distributional discrepancy reduction in each query step. The simulation of the toy dataset are described in Appendix B.2.

In Figure 1, we present the empirical evaluation of the four methods in terms of the PEHE – $\sqrt{\epsilon_{\text{PEHE}}}$, Wasserstein distance, and model variance respectively. As shown in Figure 1(b) and Figure 1(c), the Symmetry and Uncertainty acquisition strategies achieve the fastest reduction solely in distributional discrepancy and model variance respectively among the three naive methods. When mapping their performance into the empirical risk reduction shown in

Figure 1(a), Symmetry has faster empirical risk reduction than Uncertainty in the early stage, but it saturates due to the incapability of capturing the informative uncertain samples in the late stage. While MACAL combines both aspects, it brings consistent and significant risk reduction before convergence, and it is the quickest one to achieve the lowest risk.

3.2 Algorithm

According to Theorem 1, we aim to optimize the upper bound shrinkage at each query step, additionally, it is well recognized that in BMAL, sample diversity in the acquired batch is crucial [17, 20], thus we modify the shrinkage without loss of generality to cater the BMAL and aim to maximize the following at *i*-th query step:

$$\tilde{\mathcal{D}}_{i}^{*} = \underset{\tilde{\mathcal{D}}_{i}=\tilde{\mathcal{D}}_{i}^{t=1} \cup \tilde{\mathcal{D}}_{i}^{t=0} \subseteq \mathcal{D}_{\text{pool}}}{\arg \max} \sum_{t \in \{0,1\}} \mathbb{H}(\tilde{\mathcal{D}}_{i}^{t}) \cdot \Delta_{\text{Var}_{i}}^{t}(\tilde{\mathcal{D}}_{i}^{t}) + C_{\phi} \Delta_{\text{IPM}_{i}}(\tilde{\mathcal{D}}_{i}) \cdot \prod_{t \in \{0,1\}} \mathbb{H}(\tilde{\mathcal{D}}_{i}^{t}),$$

$$(4)$$

where $\mathbb{H}(\cdot)$ measures the entropy of the set, and the union of the batches for each of the treatment groups, $\tilde{\mathcal{D}}_i^{t=1}$ and $\tilde{\mathcal{D}}_i^{t=0}$, renders the acquired batch $\tilde{\mathcal{D}}_i$ at i-th query step.

Note, that the optimization in (4) is a combinatorial problem. For example, to label $|\tilde{\mathcal{D}}_i| = N_i$ samples out of the pool samples $|\mathcal{D}_{\text{pool}}| = N_{\text{pool}}$ at a time, we face a combinatorial search space which takes $O(\frac{N_{\text{pool}}!}{N_i!(N_{\text{pool}}-N_i)!})$ time to get the optimum. The bruteforce suffers from such time complexity is prohibitive as N_{pool} goes up given fixed $1 \ll N_i \ll N_{\text{pool}}$. Thus, instead of leveraging the prototype criterion in (4), we propose a model agnostic method to approximate the terms in (4) to reduce the NP-hard problem to one that can be solved in polynomial time. In the following, we analyse the optimization objective in (4) separately and combine them to conquer afterwards.

Diversity. To deal with the diversity term $\mathbb{H}(\tilde{\mathcal{D}}_i^t)$ with combinatorial nature, the key step here is to select the sample which is most distinguished from the acquired data in the batch iteratively one at a time. We use the Euclidean distance $d(\cdot, \cdot)$ to measure the similarity between two points for label acquisition. For a batch

selection on the treatment group t, we do $|\tilde{\mathcal{D}}_i^t|$ times iteratively. Since $|\tilde{\mathcal{D}}_i^t| \ll N_{\text{pool}}^t$, the time complexity of the batch acquisition is capped by $O(N_{\text{pool}}^t)$.

Lemma 2. Given the variance counted by Gaussian process regression model f^t on treatment group t, by acquiring the most uncertain samples that have the maximum predictive variance $\sigma^2_{f^t}$, the slowest convergence rate of the model variance is lower-bounded by $\Omega(\beta^i)$, where $0 \le \beta < 1$.

Uncertainty. To gain the maximum variance reduction over the sample space \mathcal{X} , labelling the most uncertain sample gives the highest variance reduction and the model variance can converge as depicted in Lemma 2, where the proof is provided in Appendix A.2.

However, the variance term is model-dependent, after one sample point is added into the acquired batch $\mathcal{\tilde{D}}_{i}^{t}$, the model should ideally get retrained and update its confidence on the pool set for the next selection to maintain the batch diversity $\mathbb{H}(\tilde{\mathcal{D}}_i^t)$ and variance reduction $\Delta_{\mathrm{Var}_i}^t(\tilde{\mathcal{D}}_i^t)$ at a high level, but retraining $|\tilde{\mathcal{D}}_i^t|$ times is not cost-efficient. To overcome this issue, we perform an approximation for the variance term to leave it model-independent since we care less about its exact value but more about its relative magnitude for comparison, i.e., acquiring the most distinct point from the pool set $\mathcal{D}_{\mathrm{pool}}^t$ with the highest Euclidean distance from its nearest neighbour in the training set (with previously acquired unlabelled samples integrated). Thus, we firstly calculate the minimum distance between every candidate sample from pool set $\mathcal{D}_{\mathrm{pool}}^t$ and the acquired sample from training set $\mathcal{D}_{\text{train}}^t,$ which results in a set of distance $\{d_i^{\min}\}$ of size N_{pool}^t . This set of values embeds both diversity and uncertainty in terms of Euclidean distance since the higher the minimum value, the more distinct the candidate point from the training set. Subsequently, the maximum one is selected from the set as follows:

$$(\tilde{x}_i^t)^* = \arg\max_{\tilde{x}_i^t \in \mathcal{D}_{\text{pool}}^t} \min_{x_i' \in \mathcal{D}_{\text{train}}^t} d(\tilde{x}_i^t, x_i'). \tag{5}$$

Note, that the unlabelled sample $(\tilde{x}_i^t)^*$ should be added into the training set before the next selection starts because a similar or repetitive sample is redundant. It should also be noted that (5) does not necessarily return the most uncertain sample, but the larger distance from its in-sample nearest neighbour is positively correlated with higher uncertainty. Thus, we take this trade-off to approximately approach the NP-hard term $\mathbb{H}(\tilde{\mathcal{D}}_i^t)\Delta^t_{\mathrm{Var}_i}(\tilde{\mathcal{D}}_i^t)$ by iterating $|\tilde{\mathcal{D}}_i^t|$ times. By considering the enumeration through both of the treatment groups, the time complexity of one query step for (5) is $O(\max\{(N_{\mathrm{pool}}^{t-1})^2,(N_{\mathrm{pool}}^{t=0})^2\})$, which is obviously capped by $O(N_{\mathrm{pool}}^2)$.

Distributional Discrepancy. In light of the above-mentioned iterative acquisition within a single query step, we can apply a similar mentality to avoid the combinatorial nature of the second term in (4) for the batch acquisition. To obtain a high-level reduction $\Delta_{\mathrm{IPM}_i}(\tilde{\mathcal{D}}_i)$, an effective labelling in terms of reducing the imbalance would lead to symmetrical acquisition, namely labelling the identical sample from different treatment groups to make a pair. Thus the local distributional discrepancy (within the acquired dataset $\tilde{\mathcal{D}}_i$) becomes zero if identical (or very similar) samples can be collected

to counteract the violation of positivity locally. Subsequently, the accumulated global distributional discrepancy gains an asymptotic behaviour approaching zero as more symmetrical (or similar) samples are added into the training set (as also empirically observed in Figure 1(b)). We propose Lemma 3 concerning the convergence rate with the proof provided in Appendix A.3:

Lemma 3. Given two empirical distributions $p^{t=1}$ and $p^{t=0}$ for different treatment groups, the distributional discrepancy given by 1-Wasserstein distance $W_1(p^{t=1},p^{t=0})$ has a convergence rate of $O(\frac{1}{i+\gamma_0})$ if the identical samples from two groups can always be found throughout the query steps.

Note, that under the deep learning framework [23, 33], the distributional discrepancy is calculated over the latent space with $\phi: \mathcal{X} \to \mathcal{Z}$, where ϕ is one-on-one mapping. Thus, samples that are identical or similar in the original space \mathcal{X} should still preserve their semantic manifold in the latent space, such that labelling similar points over the raw space is the same as the one in latent space. To label a pair, we calculate the Euclidean distance between each sample from different treatment groups, and the optimal pair is selected with the smallest distance. This selection is flexible since it does not constrain the identical acquisition but the most similar pair. Thus, the time complexity to get the optimal pair takes $O(N_{\text{pool}}^{t=1}, N_{\text{pool}}^{t=0})$, which is capped by $O(N_{\text{pool}}^2)$. At a single iteration, we acquire a pair as follows:

$$\{(\tilde{x}_i^{t=1}, \tilde{x}_j^{t=0})^*\} = \underset{\tilde{x}_i^{t=1} \in \mathcal{D}_{\text{pool}}^{t=1}, \tilde{x}_i^{t=0} \in \mathcal{D}_{\text{pool}}^{t=0}}{\arg \max} -d(\tilde{x}_i^{t=1}, \tilde{x}_j^{t=0}).$$
 (6)

MACAL. In this paper, we combine the optimization in (5) and (6) altogether, a pair $(\tilde{x}^{t=1}, \tilde{x}^{t=0})^*$ that maximizes the following term is selected from both treatment groups:

$$\begin{split} &(\tilde{x}^{t=1}, \tilde{x}^{t=0})^* = \\ &\underset{\tilde{x}^{t=1} \in \mathcal{D}_{\text{pool}}^{t=1}, \tilde{x}^{t=0} \in \mathcal{D}_{\text{pool}}^{t=0}}{\arg\max} \sum_{t \in \{0,1\}} \min_{x' \in \mathcal{D}_{\text{train}}^t} d(\tilde{x}^t, x') - \alpha \cdot d(\tilde{x}^{t=1}, \tilde{x}^{t=0}). \end{split}$$

$$(7)$$

Note, that we require the batch size $|\tilde{\mathcal{D}}_i|$ to be an even number, which is quite easy to satisfy. Thus, we do $|\tilde{\mathcal{D}}_i|/2$ efficient iterations to obtain the batch in one query step. Also, we set the coefficient α to penalize the acquisition that violates the positivity assumption, this regularization constant is taken as a hyperparameter and its impact is further discussed in Appendix 5.3 since it is hardly possible to compute the exact value for the bounded constant C_{ϕ} [33]. We visualize the dynamic selection process in Figure 2 according to the selection criterion defined in (7). For the case where one of the treatment pool sets is exhausted, the acquisition is down to only one side by simply updating the $\alpha=0$ since no counterpart can be acquired anymore. The full algorithm's pseudo code is provided in Appendix A.6.

In summary, MACAL promotes variance and discrepancy reduction by labelling diverse uncertain samples and it penalizes the dissimilarity of the paired samples via the Euclidean distance. To label a batch of samples of size $|\mathcal{\tilde{D}}_i|$, we take $O(N_{\text{pool}}^2)$ time complexity to obtain the optimal batch at each query step, which is significantly lower than the cost to solve the NP-hard problem by brute-force.

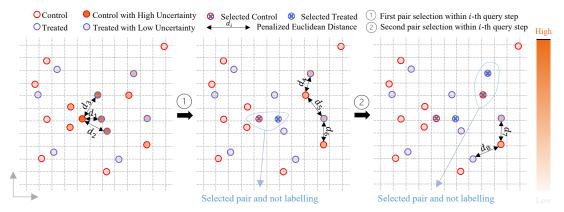


Figure 2: At *i*-th query step, the entire batch selection is divided into several pairs selection. Left: The most uncertain points are the candidates for selection, and the pair score is penalized by the distance between two points; Mid: The two most uncertain and closest points are selected (not labelled yet) and objectively bringing down other unlabelled points' uncertainty in proximity; Right: Select the next pair recursively until the batch is filled up.

4 Related Work

Active Learning. The history of active learning can be traced back to over a century ago [35], with such a long time progress till nowadays, it has become a frontier research branch of machine learning [29, 32, 51]. The core of active learning is to make model performance cost-efficient, i.e., obtaining relatively low model risk with as few labelled samples as possible. Generally, the active learning approach can be portioned into three scenarios: query synthesis [41], stream-based [8] and pool-based [47]. In this paper, we focus on pool-based active learning, especially on regression problems, where the uncertainty-based sampling [9], density-based querying [31], and hybrid strategies [3] are three key acquisition methods under such setting. For instance, the information-theoretic based Bayesian Active Learning by Disagreement (BALD) [9] leverages the epistemic uncertainty to acquire unlabelled samples. Core-Set [31] selects the greatest distance to its nearest neighbour in the hidden space. ACS-FW [27] is a hybrid between Core-Set and Bayesian approaches which balances the sample diversity and uncertainty in batch-mode acquisition. Note, that although the general active learning methods are not designated for CEE, benchmarking on these methods provides insightful results.

Causal Effect Estimation with Active Learning. Some noticeable advances have been made in this area of research. [36] approximates the decision-making reliability via the estimated S-type error rate (the probability of the model inferring the sign of the treatment effect wrong) of the prediction model, which is then used as the querying criterion. However, [36] focuses on estimating the correct sign of the treatment effect, which is different from the risk metric in our setting. For works focusing on the same risk metric, OHTE [28] integrates the Core-Set concept [31, 37] to form a theoretical framework, for which a theory-based optimization is proposed. However, the QHTE relaxes the tightness of the bound given by Shalit et. al. [33], where a covering radius r = 0 from the relaxed bound cannot even obtain the original tightness, while, we propose a more informative theory which does not undermine the tightness of the original bound. More importantly, OHTE does not consider the distribution imbalance during sample acquisition, which is what

Table 1: Summary of the Acquisition Setup and Testing

Dataset	Start S.	Step S.	Queries	Pool S.	Val S.	Test S.
IHDP	10	10	46	470	75	202
IBM	50	50	50	9540	3180	6250
CMNIST	50	50	50	31500	10500	18000

our method can prominently distinguish from. To fix the acquisition imbalance issue, Causal-BALD [17] cut into the problem from the information-theoretic perspective, its most representative criterion $\mu\rho$ BALD accounts for the overlapping by especially scaling the criterion with the inverse of counterfactual variance, leaving the acquisition toward pairing up similar samples if its counterfactual were missing, which is a non-trivial improvement from its base - μBALD (an uncertainty-based softmax-BALD method [19]). Our proposed method is distinguished from [17] in several points. First, Causal-BALD relies on model-dependent variance estimation, e.g., the deep kernel learning model [38, 46], while our criterion is model agnostic. Second, we cut into the problem from an intriguing perspective to maximize the upper bound shrinkage at each query step instead of the mutual information perspective. Finally, taking the inverse of the counterfactual variance is undesirable and causes numerical instability, while our method leverages the simple but effective addition and subtraction operations to form the label acquisition criterion. It is also noted that some existing literature [1, 7] uses active learning to take the initiative for efficient experimental trials design, i.e., the pool set does not include the treatment information but enforcing treatment after sample acquisition, which is different from our setting.

5 Experiments

Due to the unique nature of the CEE tasks, the counterfactual effect is hardly observed in the real world. Thus, in this paper, we take the common practice to utilise the fully-synthetic and semi-synthetic datasets for algorithm evaluations.

Dataset: IHDP [11] - an imbalanced dataset based on 747 samples (among them 139 with treatment status t = 1 and 608 with

status t=0) and 25 covariates, with 100 times simulated treatment outcomes by [11]. **IBM** [34] - uses a cohort of 100k individuals from the publicly available Linked Births and Infant Deaths Database with 177 real-world covariates. Each original simulation randomly takes 25k out of such 100k samples and the potential outcomes are simulated 10 times according to [34], we create more imbalance by omitting additional samples from treatment group t=1. **CMNIST** [15] - is of size 60k adapted from MINIST [22] dataset. The input from CMNIST is the handwritten digit of size 28×28, which is distinct from the previous tabular datasets. The potential outcomes are simulated 10 times and generated by projecting the digits into a 1-dimensional latent manifold as described in [15].

Metric: We use precision in estimation of heterogeneous effect (PEHE) [33], a well-established metric with the empirical formulation: $\sqrt{\epsilon_{\text{PEHE}}} = \sqrt{\Sigma_{i=1}^{N}((y_i^{t=1}-y_i^{t=0})-\tau_i)^2/N}$ for measuring the accuracy of the treatment effect estimation at the individual level. The lower the value of $\sqrt{\epsilon_{\text{PEHE}}}$, the better the performance.

Baselines: We set the Random method as the benchmark acquisition function, as this is the most naive method that selects the data purely at random. We also compare our proposed method against many SOTA baselines from the general AL research field, that is, BADGE [3], BAIT [2], and LCMD [12]. We argue that a good comparison to these methods from the broad AL research is indispensable and this paper also fills the blank for such comparisons. Moreover, the most related work - QHTE [28], and especially Causal-Bald [17] is the designated algorithm proposed to deal with the active causal effect estimation. Thus, we compare three representative variants of the Causal-Bald, namely μBALD, ρBALD, and μρBALD.

Prediction Backbone: DUE-DNN[38]. It is one of the SOTA deep kernel learning frameworks with the multi-layer perceptron as the common feature extractor and two sparse Gaussian process regressions defined over the extracted latent features as the downstream estimators for different treatment groups' effect estimations. **DUE-CNN**[38]. It is a variant of the DUE model especially catering for the image-as-input experiment. It has a similar structure as DUE-DNN besides the latent feature extractor being replaced by the convolutional neural network (CNN), e.g., the ResNet [10] is embedded. The computation resources and hyperparameter selection are described in B.3.

Acquisition setup: We begin with a small Start Size (Start S.) to simulate the real-world scenario where only sparse labelled data can be accessed at the beginning. Then, a fixed Step Size (Step S.) is enforced at each query step, and the entire AL sessions (Queries), which consist of many query steps comes to an end when the model converges or the sample pool has been exhausted. The detail of the label acquisition setup is summarized in Table 1.

5.1 Baseline Risk Evaluation

Generally, across all figures, MACAL's performance set an empirical risk lower bound for all the other baselines. On the IHDP dataset, our proposed method obtains a lower risk till 160 training size (turning point). Then, it performs similarly to the other methods till the exhaustion of the pool set. This phenomenon is explainable due to the fact that the samples from treatment group t=1 get exhausted at the turning point, and MACAL can only acquire the

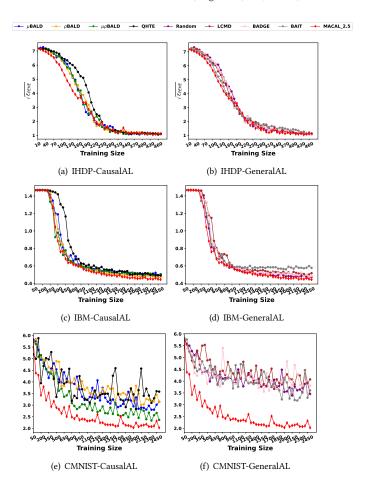


Figure 3: Benchmarking with the available SOTAs on three datasets, i.e., IHDP (average with 100 simulations), IBM (average with 10 simulations) and CMNIST (average with 10 simulations). The first column concludes the comparisons against the baselines designated for active causal effect estimation. The second column shows the comparisons with the baselines from general active learning approaches. All of the results are given by the same downstream CEE model by DUE-DNN for IHDP & IBM, and DUE-CNN for CMNIST.

samples from the treatment group (t=0) without benefiting from the reduction of distributional discrepancy by labelling similar pairs. It is also arguable that when deploying the general AL methods, e.g., BADGE, directly into active causal effect estimation, it is interesting to see that even the Random method can perform similarly to these SOTAs. We overall see a better performance of causal AL methods, e.g., Causal-BALD, and MACAL, over the general AL methods by additionally considering reducing the violation of positivity during label acquisition.

Moreover, none of the current SOTAs focusing on active causal effect estimation can consistently outperform the other methods from the general active learning research field across all the datasets. QHTE mostly underperforms because it only focuses on selecting the point that has the maximum distance from its closest neighbour in the current training set without meaningful constraints

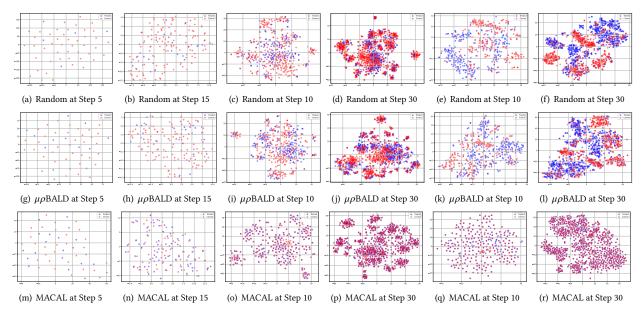


Figure 4: Visualization of the post-acquisition dataset (IHDP: First two columns, IBM: Middle two columns, and CMNIST: Last two columns) via t-SNE for treatment group t = 1, treatment group t = 0, and overlapping for Random, $\mu\rho$ BALD, and MACAL.

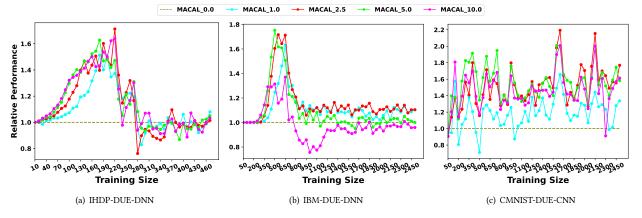


Figure 5: Hyperparameter α representing various levels of symmetrical regularization for label acquisition. The relative performance of MACAL_ α_i is calculated as $\epsilon_{\text{PEHE},\alpha=0}/\epsilon_{\text{PEHE},\alpha=\alpha_i}$, the higher the better.

on post-acquisition imbalance on raw features. Also, even though $\mu\rho \rm BALD$ is the most representative method from [17] incorporating the imbalance penalization in its query criterion, such indirect regularization via counterfactual uncertainty is not as optimal as ours. Because MACAL directly reduces the post-acquisition distributional imbalance by acquiring more similar pairs based on raw features. It is also noted that the division-form criterion of Causal-BALD can embed immense variation (the performance of all its variants fluctuates drastically) in estimations as shown in Figure 3(e), while our proposed simple addition-form criterion is significantly more stable.

5.2 Acquisition Visualization

To give a direct comparison of acquisition quality, we visualize the results in Figure 4 by projecting the post-acquisition training set

from three acquisition criteria, i.e., Random, $\mu\rho$ BALD, and MACAL, on each dataset at two different query stages into the 2-dimensional latent space via t-SNE [39]. From Figure 4(a) to 4(f) across three datasets, we observe that the Random draw from the original distribution inherently expresses a strong violation of positivity across three different datasets, such that we barely see large overlapping regions. While, $\mu\rho$ BALD shown in Figure 4(g) to 4(l) looks slightly better than the Random method by being more spread out, but it still cannot well resolve the violation of the positivity issue at large scale, rendering a significantly imbalanced label acquisition for different treatment groups. As for MACAL, we observe an exceptional acquisition result from Figure 4(m) to 4(r), each of the samples from both of the treatment groups can mostly find its (close) counterfactual such that the violation of positivity is significantly reduced. The

acquisition by MACAL also shows high diversity instead of clustering. Hence, the remarkable performance gap shown in Figure 3(e) can also be explained in essence by the acquisition results shown in 4(q) and 4(r). Additional visualizations for the other baselines, and on different datasets are accessible in Appendix B.1.

5.3 Symmetrical Regularization Study

We conduct extensive ablation experiments for $\alpha \in \{0, 1, 2.5, 5, 10\}$, a clear observation is that, even though the benefit of setting the symmetrical regularization is non-trivial, there is no single hyperparameter α that can consistently outperform all the others throughout the entire label acquisition process. Also, the stronger regularization, e.g., $\alpha = 10$, delivers better performance at the early stage of the acquisition, but such an advantage cannot be maintained across the whole acquisition process. Interestingly, during the course of the acquisition, a decreasing coefficient empirically grants an increasing relative performance, e.g., $\alpha = 2.5$ underperforms $\alpha = 10$ at the early stage, but it outperforms $\alpha = 10$ in the later stage. This is explainable because when the key set of the overlapping samples is mostly collected, there is less information can be obtained from acquiring the repetitive samples even though these are from the overlapping region. The criterion should bias its acquisition toward the uncertain non-overlapping area to gain more information to reduce the risk of the model.

6 Conclusion

In this paper, we study the well-under-explored yet important and practical active causal effect estimation problem and construct a theoretical framework from a novel and intriguing perspective, i.e., decompose a more informative risk upper bound without loosening it and give mathematically guaranteed risk convergence analysis under certain conditions. Therefore, in theory we maximize the decomposed terms at each query step in order to minimize the generalization risk. Subsequently, we derive a theory-inspired simplified yet effective label acquisition algorithm, i.e., MACAL, which considers the joint reduction of the model's variance and post-acquisition distributional imbalance via a simplified yet effective label acquisition criterion. Moreover, reaching data-efficient labelling is never an NP-hard problem via MACAL, and thus the optimum can be obtained in polynomial time with $O(N^2)$. It is generally demonstrated that our proposed method consistently outperforms the other baselines across all the datasets with a nontrivial performance gain.

Acknowledgments

This work is supported by the Australian Research Council under the streams of Future Fellowship (No. FT210100624), Discovery Early Career Researcher Award (No. DE230101033), Industrial Transformation Training Centre (No. IC200100022), Discovery Project (No. DP240101108 and No. DP240101814), and Linkage Project (No. LP230200892). Partial support from Health and Wellbeing Queensland is gratefully acknowledged. HW would like to additionally thank the invaluable support from Carrie Chen.

References

- Raghavendra Addanki, David Arbour, Tung Mai, Cameron Musco, and Anup Rao. 2022. Sample constrained treatment effect estimation. Advances in Neural Information Processing Systems 35 (2022), 5417–5430.
- [2] Jordan Ash, Surbhi Goel, Akshay Krishnamurthy, and Sham Kakade. 2021. Gone fishing: Neural active learning with fisher embeddings. Advances in Neural Information Processing Systems 34 (2021), 8927–8939.
- [3] Jordan T Ash, Chicheng Zhang, Akshay Krishnamurthy, John Langford, and Alekh Agarwal. 2019. Deep Batch Active Learning by Diverse, Uncertain Gradient Lower Bounds. In International Conference on Learning Representations.
- [4] Vinod Kumar Chauhan, Jiandong Zhou, Ghadeer Ghosheh, Soheila Molaei, and David A Clifton. 2024. Dynamic inter-treatment information sharing for individualized treatment effects estimation. In *International Conference on Artificial Intelligence and Statistics*. PMLR, 3529–3537.
- [5] Philippe Clement and Wolfgang Desch. 2008. An elementary proof of the triangle inequality for the Wasserstein metric. Proc. Amer. Math. Soc. 136, 1 (2008), 333– 339.
- [6] David A Cohn, Zoubin Ghahramani, and Michael I Jordan. 1996. Active learning with statistical models. Journal of artificial intelligence research 4 (1996), 129–145.
- [7] Kun Deng, Joelle Pineau, and Susan Murphy. 2011. Active learning for personalizing treatment. In 2011 IEEE Symposium on Adaptive Dynamic Programming and Reinforcement Learning (ADPRL). IEEE, 32–39.
- [8] Kaito Fujii and Hisashi Kashima. 2016. Budgeted stream-based active learning via adaptive submodular maximization. Advances in Neural Information Processing Systems 29 (2016).
- [9] Yarin Gal, Riashat Islam, and Zoubin Ghahramani. 2017. Deep bayesian active learning with image data. In *International conference on machine learning*. PMLR, 1183–1192.
- [10] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In Proceedings of the IEEE conference on computer vision and pattern recognition. 770–778.
- [11] Jennifer L Hill. 2011. Bayesian nonparametric modeling for causal inference. Journal of Computational and Graphical Statistics 20, 1 (2011), 217–240.
- [12] David Holzmüller, Viktor Zaverkin, Johannes Kästner, and Ingo Steinwart. 2023. A framework and benchmark for deep batch active learning for regression. *Journal* of Machine Learning Research 24, 164 (2023), 1–81.
- [13] Nguyen Quoc Viet Hung, Huynh Huu Viet, Nguyen Thanh Tam, Matthias Weidlich, Hongzhi Yin, and Xiaofang Zhou. 2017. Computing crowd consensus with partial agreement. *IEEE Transactions on Knowledge and Data Engineering* 30, 1 (2017), 1–14.
- [14] Guido W Imbens and Donald B Rubin. 2015. Causal inference in statistics, social, and biomedical sciences. Cambridge University Press.
- [15] Andrew Jesson, Sören Mindermann, Yarin Gal, and Uri Shalit. 2021. Quantifying ignorance in individual-level causal-effect estimates under hidden confounding. In *International Conference on Machine Learning*. PMLR, 4829–4838.
- [16] Andrew Jesson, Sören Mindermann, Uri Shalit, and Yarin Gal. 2020. Identifying causal-effect inference failure with uncertainty-aware models. Advances in Neural Information Processing Systems 33 (2020), 11637–11649.
- [17] Andrew Jesson, Panagiotis Tigas, Joost van Amersfoort, Andreas Kirsch, Uri Shalit, and Yarin Gal. 2021. Causal-bald: Deep bayesian active learning of outcomes to infer treatment-effects from observational data. Advances in Neural Information Processing Systems 34 (2021), 30465–30478.
- [18] Fredrik Johansson, Uri Shalit, and David Sontag. 2016. Learning representations for counterfactual inference. In *International conference on machine learning*. PMLR, 3020–3029.
- [19] Andreas Kirsch, Sebastian Farquhar, Parmida Atighehchian, Andrew Jesson, Frederic Branchaud-Charron, and Yarin Gal. 2021. Stochastic Batch Acquisition: A Simple Baseline for Deep Active Learning. arXiv preprint arXiv:2106.12059 (2021).
- [20] Andreas Kirsch, Joost Van Amersfoort, and Yarin Gal. 2019. Batchbald: Efficient and diverse batch acquisition for deep bayesian active learning. Advances in neural information processing systems 32 (2019).
- [21] Ron Kohavi and Roger Longbotham. 2015. Online controlled experiments and A/B tests. Encyclopedia of machine learning and data mining (2015), 1–11.
- [22] Yann LeCun. 1998. The MNIST database of handwritten digits. http://yann. lecun. com/exdb/mnist/ (1998).
- [23] Christos Louizos, Uri Shalit, Joris M Mooij, David Sontag, Richard Zemel, and Max Welling. 2017. Causal effect inference with deep latent-variable models. Advances in neural information processing systems 30 (2017).
- [24] Quoc Viet Hung Nguyen, Chi Thang Duong, Thanh Tam Nguyen, Matthias Weidlich, Karl Aberer, Hongzhi Yin, and Xiaofang Zhou. 2017. Argument discovery via crowdsourcing. The VLDB Journal 26 (2017), 511–535.
- [25] Thanh Tam Nguyen, Chi Thang Duong, Matthias Weidlich, Hongzhi Yin, and Quoc Viet Hung Nguyen. 2017. Retaining data from streams of social platforms with minimal regret. In Twenty-sixth International Joint Conference on Artificial Intelligence.

- [26] C Pilat, T Frech, A Wagner, K Krüger, A Hillebrecht, J Pons-Kühnemann, C Scheibelhut, R-H Bödeker, and F-C Mooren. 2015. Exploring effects of a natural combination medicine on exercise-induced inflammatory immune response: A double-blind RCT. Scandinavian Journal of Medicine & Science in Sports 25, 4 (2015), 534–542.
- [27] Robert Pinsler, Jonathan Gordon, Eric Nalisnick, and José Miguel Hernández-Lobato. 2019. Bayesian batch active learning as sparse subset approximation. Advances in neural information processing systems 32 (2019).
- [28] Tian Qin, Tian-Zuo Wang, and Zhi-Hua Zhou. 2021. Budgeted heterogeneous treatment effect estimation. In *International Conference on Machine Learning*. PMI R. 8693–8702
- [29] Pengzhen Ren, Yun Xiao, Xiaojun Chang, Po-Yao Huang, Zhihui Li, Brij B Gupta, Xiaojiang Chen, and Xin Wang. 2021. A survey of deep active learning. ACM computing surveys (CSUR) 54, 9 (2021), 1–40.
- [30] Bent Rosenbaum, Kristian Valbak, Susanne Harder, Per Knudsen, Anne Køster, Matilde Lajer, Anne Lindhardt, Gerda Winther, Lone Petersen, Per Jørgensen, et al. 2005. The Danish National Schizophrenia Project: prospective, comparative longitudinal treatment study of first-episode psychosis. The British Journal of Psychiatry 186, 5 (2005), 394–399.
- [31] Ozan Sener and Silvio Savarese. 2018. Active Learning for Convolutional Neural Networks: A Core-Set Approach. In International Conference on Learning Representations.
- [32] Burr Settles. 2009. Active learning literature survey. Computer Sciences Technical Report (2009).
- [33] Uri Shalit, Fredrik D Johansson, and David Sontag. 2017. Estimating individual treatment effect: generalization bounds and algorithms. In *International Confer*ence on Machine Learning. PMLR, 3076–3085.
- [34] Yishai Shimoni, Chen Yanover, Ehud Karavani, and Yaara Goldschmnidt. 2018. Benchmarking framework for performance-evaluation of causal inference analysis. arXiv preprint arXiv:1802.05046 (2018).
- [35] Kirstine Smith. 1918. On the standard deviations of adjusted and interpolated values of an observed polynomial function and its constants and the guidance they give towards a proper choice of the distribution of observations. *Biometrika* 12, 1/2 (1918), 1–85.
- [36] Iiris Sundin, Peter Schulam, Eero Siivola, Aki Vehtari, Suchi Saria, and Samuel Kaski. 2019. Active learning for decision-making from imbalanced observational data. In *International conference on machine learning*. PMLR, 6046–6055.
- [37] Ivor W Tsang, James T Kwok, Pak-Ming Cheung, and Nello Cristianini. 2005. Core vector machines: Fast SVM training on very large data sets. *Journal of Machine Learning Research* 6, 4 (2005).
- [38] Joost Van Amersfoort, Lewis Smith, Andrew Jesson, Oscar Key, and Yarin Gal. 2021. On feature collapse and deep kernel learning for single forward pass uncertainty. arXiv preprint arXiv:2102.11409 (2021).
- [39] Laurens Van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-SNE. JMLR (2008).
- [40] Hao Wang, Jiajun Fan, Zhichao Chen, Haoxuan Li, Weiming Liu, Tianqiao Liu, Quanyu Dai, Yichao Wang, Zhenhua Dong, and Ruiming Tang. 2024. Optimal transport for treatment effect estimation. Advances in Neural Information Processing Systems 36 (2024).
- [41] Liantao Wang, Xuelei Hu, Bo Yuan, and Jianfeng Lu. 2015. Active learning via query synthesis and nearest neighbour search. *Neurocomputing* 147 (2015), 426–434.
- [42] Hechuan Wen, Tong Chen, Li Kheng Chai, Shazia Sadiq, Junbin Gao, and Hongzhi Yin. 2023. Variational Counterfactual Prediction under Runtime Domain Corruption. IEEE Transactions on Knowledge and Data Engineering (2023).
- [43] Hechuan Wen, Tong Chen, Li Kheng Chai, Shazia Sadiq, Kai Zheng, and Hongzhi Yin. 2023. To Predict or to Reject: Causal Effect Estimation with Uncertainty on Networked Data. In 2023 IEEE International Conference on Data Mining (ICDM). IEEE, 1415–1420.
- [44] Christopher KI Williams and Carl Edward Rasmussen. 2006. Gaussian processes for machine learning. MIT press Cambridge, MA.
- [45] Christopher KI Williams and Francesco Vivarelli. 2000. Upper and lower bounds on the learning curve for Gaussian processes. Machine Learning 40 (2000), 77–102.
- [46] Andrew Gordon Wilson, Zhiting Hu, Ruslan Salakhutdinov, and Eric P Xing. 2016. Deep kernel learning. In Artificial intelligence and statistics. PMLR, 370–378.
- [47] Dongrui Wu. 2018. Pool-based sequential active learning for regression. IEEE transactions on neural networks and learning systems 30, 5 (2018), 1348–1359.
- [48] Liuyi Yao, Zhixuan Chu, Sheng Li, Yaliang Li, Jing Gao, and Aidong Zhang. 2021. A survey on causal inference. ACM TKDD 15, 5 (2021), 1–46.
- [49] Liuyi Yao, Sheng Li, Yaliang Li, Mengdi Huai, Jing Gao, and Aidong Zhang. 2018. Representation learning for treatment effect estimation from observational data. Advances in Neural Information Processing Systems 31 (2018).
- [50] Hongzhi Yin, Qinyong Wang, Kai Zheng, Zhixu Li, and Xiaofang Zhou. 2020. Overcoming data sparsity in group recommendation. IEEE Transactions on Knowledge and Data Engineering 34, 7 (2020), 3447–3460.
- [51] Xueying Zhan, Qingzhong Wang, Kuan-hao Huang, Haoyi Xiong, Dejing Dou, and Antoni B Chan. 2022. A comparative survey of deep active learning. arXiv preprint arXiv:2203.13450 (2022).

A Theory

A.1 Convergence Behaviour of Risk Upper Bound

Theorem 1. With budget M, the maximum risk upper bound reduction $\Delta_{\mathcal{B}_{overall}}$ is achieved at the termination of the entire I data query steps given that the generalization risk upper bound shrinkage $\Delta_{\mathcal{B}_{i}}$ is maximized at each query step $\forall i$, i.e.:

$$\underset{\tilde{\mathcal{D}}_{overall}}{\arg \max } \Delta_{\mathcal{B}_{overall}} = \bigcup_{i=1}^{I} \underset{\tilde{\mathcal{D}}_{i}}{\arg \max } \Delta_{\mathcal{B}_{i}}$$

$$s.t. |\tilde{\mathcal{D}}_{overall}| \leq \mathcal{M},$$
(8)

where $\tilde{\mathcal{D}}_{overall}$ is the overall acquired data, $\tilde{\mathcal{D}}_i$ is the acquired batch at i-th query step, and $\Delta_{\mathcal{B}_i} = \sum_{t \in \{0,1\}} \Delta^t_{Var_i}(\tilde{\mathcal{D}}_i) + C_\phi \Delta_{IPM_i}(\tilde{\mathcal{D}}_i)$. The convergence rate of the risk upper bound has the following guaranteed behaviours under certain circumstances:

- i) When variance reduction $\sum_{t \in \{0,1\}} \Delta^t_{Var_i}(\tilde{\mathcal{D}}_i)$ becomes the dominant part of the risk upper bound, the risk convergence is lower-bounded by $\Omega(\beta^i)$ with constant $\beta \in [0,1)$.
- ii) While, with dominant constant C_{ϕ} , the risk convergence is upper-bounded by $O(\frac{1}{1+\gamma_0})$ with constant $\gamma_0 \in \mathbb{R}^+$.

PROOF. Shalit et al. [33] provide the well-known upper-bound for the expected PEHE as:

$$\epsilon_{\text{PEHE}} \leq 2 \left[\epsilon_{\text{F}}^{t=1} + \epsilon_{\text{F}}^{t=0} + C_{\phi} (\text{IPM}_{\mathcal{F}}(p_{\phi}^{t=1}, p_{\phi}^{t=0})) \right]. \tag{9}$$

We denote such upper-bound at i-th query step as \mathcal{B}_i with post-acquisition dataset \mathcal{D}_i (which includes the labelled optimal batch $\tilde{\mathcal{D}}_i^*$). By factual error decomposition in Proposition 2, we have:

$$\epsilon_{\mathbf{F}}^{t} = \mathbb{E}_{\mathcal{X}}[(f^{t}(x) - \mathbb{E}[\hat{f}^{t}(x)])^{2}] + \mathbb{E}_{\mathcal{X}}[\mathbb{E}[(\hat{f}^{t}(x) - \mathbb{E}[\hat{f}^{t}(x)])^{2}]] + \sigma_{\xi t}^{2}. \tag{10}$$

We leave out the constant multiplier 2 in the original bound for notation simplicity during deduction (as the calculation for the shrinkage cancels off the constant), and derive the upper bound \mathcal{B}_i at *i*-th query step in a brand-new form but with the same tightness as it is:

$$\mathcal{B}_i = \epsilon_{i,\mathrm{F}}^{t=1} + \epsilon_{i,\mathrm{F}}^{t=0} + C_\phi(\mathrm{IPM}_{\mathcal{F}}(p_{i,\phi}^{t=1},p_{i,\phi}^{t=0})) \tag{11a}$$

$$= \sum_{t \in \{0,1\}} \mathbb{E}_{\mathcal{X}} \left[\operatorname{Bias}_{i}^{t} [\hat{f}_{i}^{t}(x; \mathcal{D}_{i})] \right] + \tag{11b}$$

$$\sum_{t \in \{0,1\}} \mathbb{E}_{\mathcal{X}} \left[\operatorname{Var}_{i}^{t} [\hat{f}_{i}^{t}(x; \mathcal{D}_{i})] \right] + \tag{11c}$$

$$\sum_{t \in \{0,1\}} \sigma_{\xi_i^t}^2 + C_{\phi} \text{IPM}_{\mathcal{F}}(p_{i,\phi}^{t=1}, p_{i,\phi}^{t=0}), \tag{11d}$$

where $\operatorname{Bias}_i^t[\hat{f}_i^t(x;\mathcal{D}_i)] = (f_i^t(x) - \mathbb{E}[\hat{f}_i^t(x)])^2$, and $\operatorname{Var}_i^t[\hat{f}_i^t(x;\mathcal{D}_i)] = \mathbb{E}[(\hat{f}_i^t - \mathbb{E}[\hat{f}_i^t])^2]$.

Analogously, at (i-1)-th query step, namely \mathcal{B}_{i-1} with dataset \mathcal{D}_{i-1} , we have:

$$\mathcal{B}_{i-1} = \sum_{t \in \{0,1\}} \mathbb{E}_{\mathcal{X}} \left[\text{Bias}_{i-1}^{t} [\hat{f}_{i-1}^{t}(x; \mathcal{D}_{i-1})] \right] +$$
 (12a)

$$\sum_{t \in \{0,1\}} \mathbb{E}_{\mathcal{X}} \left[\operatorname{Var}_{i-1}^{t} [\hat{f}_{i-1}^{t}(x; \mathcal{D}_{i-1})] \right] + \tag{12b}$$

$$\sum_{t \in \{0,1\}} \sigma_{\xi_{i-1}^t}^2 + C_{\phi} \text{IPM}_{\mathcal{F}}(p_{i-1,\phi}^{t=1}, p_{i-1,\phi}^{t=0}). \tag{12c}$$

Subsequently, the shrinkage $\Delta_{\mathcal{B}_i}$ at i-th query step is defined as:

$$\Delta_{\mathcal{B}_{i}} = \mathcal{B}_{i-1} - \mathcal{B}_{i}$$

$$= \sum_{t \in \{0,1\}} \mathbb{E}_{\mathcal{X}} \left[\operatorname{Bias}_{i-1}^{t} [\hat{f}_{i-1}^{t}(x; \mathcal{D}_{i-1})] - \operatorname{Bias}_{i}^{t} [\hat{f}_{i}^{t}(x; \mathcal{D}_{i})] \right] +$$
(13a)

$$\sum_{t \in \{0,1\}} \mathbb{E}_{\mathcal{X}} \left[\operatorname{Var}_{i-1}^{t} [\hat{f}_{i-1}^{t}(x; \mathcal{D}_{i-1})] - \operatorname{Var}_{i}^{t} [\hat{f}_{i}^{t}(x; \mathcal{D}_{i})] \right] + (13c)$$

$$\sum_{t \in \{0,1\}} (\sigma_{\xi_{i-1}^t}^2 - \sigma_{\xi_i^t}^2) + C_\phi \left(\text{IPM}_{\mathcal{F}}(p_{i-1,\phi}^{t=1}, p_{i-1,\phi}^{t=0}) - \text{IPM}_{\mathcal{F}}(p_{i,\phi}^{t=1}, p_{i,\phi}^{t=0}) \right) \text{ where } \zeta = \sum_{t \in \{0,1\}} \sigma_{\xi_{i-1}^t}^2 \text{, and the approximation } \approx \text{ is given by the a small enough } C_\phi, \text{ and } G_i \text{ is the shrinkage in terms of the variance}$$

$$=0+$$
 (13e)

$$\sum_{t \in \{0,1\}} \underbrace{\mathbb{E}_{\mathcal{X}} \left[\operatorname{Var}_{i-1}^{t} \left[\hat{f}_{i-1}^{t}(x; \mathcal{D}_{i-1}) \right] - \operatorname{Var}_{i}^{t} \left[\hat{f}_{i}^{t}(x; \mathcal{D}_{i}) \right] \right]}_{\Delta_{\operatorname{Var}:}^{t}} + (13f)$$

$$0 + C_{\phi} \underbrace{\left(\text{IPM}_{\mathcal{F}}(p_{i-1,\phi}^{t=1}, p_{i-1,\phi}^{t=0}) - \text{IPM}_{\mathcal{F}}(p_{i,\phi}^{t=1}, p_{i,\phi}^{t=0}) \right)}_{\Delta_{\text{IPM}_{i}}}$$
(13g)

$$= \sum_{t \in \{0,1\}} \Delta_{\operatorname{Var}_i}^t + C_\phi \Delta_{\operatorname{IPM}_i}. \tag{13h}$$

The second equality holds because the models' bias only depends on the selection of the model class [6], or more empirically models' bias is negligible for models with enough complexity [32]. Either way two bias terms cancelled off. Also, the data generation process has the same noise assumption, thus two noise variance terms

The overall bound shrinkage $\Delta_{\mathcal{B}}$ after the termination of the entire *I* query steps is thus:

$$\Delta_{\mathcal{B}_{\text{overall}}} = \mathcal{B}_0 - \mathcal{B}_I$$

$$= \mathcal{B}_0 - \mathcal{B}_1 + \mathcal{B}_1 - \mathcal{B}_2 + \dots + \mathcal{B}_{I-2} - \mathcal{B}_{I-1} + \mathcal{B}_{I-1} - \mathcal{B}_I$$
(14a)
(14b)

$$= \Delta_{\mathcal{B}_0} + \Delta_{\mathcal{B}_1} + \dots + \Delta_{\mathcal{B}_{I-1}} + \Delta_{\mathcal{B}_I}$$
 (14c)

$$=\sum_{i=1}^{I}\Delta_{\mathcal{B}_{i}}.$$
(14d)

Therefore, to maximize the bound reduction and return the optimal set:

$$\underset{\tilde{\mathcal{D}}_{\text{overall}}}{\arg\max} \Delta_{\mathcal{B}_{\text{overall}}} = \underset{\tilde{\mathcal{D}}_{\text{overall}}=\bigcup_{i=1}^{I} \tilde{\mathcal{D}}_{i}}{\arg\max} \sum_{i=1}^{I} \Delta_{\mathcal{B}_{i}} = \bigcup_{i=1}^{I} \arg\max_{\tilde{\mathcal{D}}_{i}} \Delta_{\mathcal{B}_{i}},$$
(15)

where the entire optimal set $\mathcal{\tilde{D}}_{\mathrm{overall}}^{*}$ is a union of the optimal set $ilde{\mathcal{D}}_i^*$ which is acquired at every query step to maximize the shrinkage $\Delta_{\mathcal{B}_i}$ in (13h), thus we can conclude that the maximum risk upper bound reduction is obtained after the termination of entire label acquisition process.

Now that we define the shrinkage $A_i = \Delta_{\mathcal{B}_i}/\mathcal{B}_{i-1}$ at *i*-th step, and discuss two extreme contexts for the total risk upper bound since it is hardly to compute the exact value for the bounded constant C_{ϕ} as discussed in [33].

Scenario 1: Lower-bounded convergence rate with negligible C_{ϕ}

$$A_{i} = \frac{\Delta_{\mathcal{B}_{i}}}{\mathcal{B}_{i-1}}$$

$$= \frac{\sum_{t \in \{0,1\}} \Delta_{\operatorname{Var}_{i}}^{t} + C_{\phi} \Delta_{\operatorname{IPM}_{i}}}{\sum_{t \in \{0,1\}} \mathbb{E}_{X} \left[\operatorname{Var}_{i-1}^{t} \left[\hat{f}^{t}(x; \mathcal{D}_{i-1}) \right] \right] + C_{\phi} \operatorname{IPM}_{\mathcal{F}}(p_{i-1,\phi}^{t=1}, p_{i-1,\phi}^{t=0}) + \zeta}$$

$$\approx G_{i}.$$

$$(16a)$$

$$\approx G_{i}.$$

$$(16b)$$

a small enough C_{ϕ} , and G_i is the shrinkage in terms of the variance reduction at *i*-query step defined in Eq. 22.

When the variance term becomes the dominant part of the risk upper bound and leaving out the distributional discrepancy, we have the shrinkage difference $|A_i - G_i| < \epsilon$, for a small $\epsilon > 0$. Under such circumstances, the convergence rate of the risk upper bound cannot go slower than $\Omega(\beta^i)$ as stated in Lemma 2 with proof given

Scenario 2: Upper-bounded convergence rate with dominant C_{ϕ}

$$A_{i} = \frac{\Delta_{\mathcal{B}_{i}}}{\mathcal{B}_{i-1}}$$

$$= \frac{\sum_{t \in \{0,1\}} \Delta_{\text{Var}_{i}}^{t} + C_{\phi} \Delta_{\text{IPM}_{i}}}{\sum_{t \in \{0,1\}} \Delta_{\text{Var}_{i}}^{t} + C_{\phi} \Delta_{\text{IPM}_{i}}}$$

$$= \frac{\sum_{t \in \{0,1\}} \Delta_{\text{Var}_{i-1}}^{t} [\hat{f}^{t}(x; \mathcal{D}_{i-1})]] + C_{\phi} \text{IPM}_{\mathcal{F}}(p_{i-1,\phi}^{t=1}, p_{i-1,\phi}^{t=0}) + \zeta}$$

$$\leq \frac{\sum_{t \in \{0,1\}} \Delta_{\text{Var}_{i}}^{t} + C_{\phi} \Delta_{\text{IPM}_{i}}}{C_{\phi} \text{IPM}_{\mathcal{F}}(p_{i-1,\phi}^{t=1}, p_{i-1,\phi}^{t=0})}$$

$$\leq \frac{\sigma_{f^{t}}^{2} + C_{\phi} \Delta_{\text{IPM}_{i}}}{C_{\phi} \text{IPM}_{\mathcal{F}}(p_{i-1,\phi}^{t=1}, p_{i-1,\phi}^{t=0})}$$

$$= \frac{\sigma_{f^{t}}^{2}}{C_{\phi}} + \Delta_{\text{IPM}_{i}}}{\text{IPM}_{\mathcal{F}}(p_{i-1,\phi}^{t=1}, p_{i-1,\phi}^{t=0})} \approx S_{i},$$

$$(17d)$$

where $\zeta = \sum_{t \in \{0,1\}} \sigma_{\xi_t^t}^2$, and the approximation \approx is given by the the dominant C_{ϕ} , and \dot{S}_{i} is the shrinkage in terms of the discrepancy reduction at i-query step defined in Eq. 29.

When the discrepancy term becomes the dominant part of the risk upper bound and leaving out the variance, we have the shrinkage difference $|A_i - S_i| < \epsilon$, for a small $\epsilon > 0$. Under such circumstances, the convergence rate of the risk upper bound cannot exceed $O(\frac{1}{i+v_0})$ as stated in Lemma 3 with proof given in A.3

Thus we conclude the proof for the two convergence behaviours under two extreme circumstances from the influence of the distributional discrepancy.

A.2 Convergence Rate of Model Variance

Lemma 2. Given the variance counted by Gaussian process regression model f^t on treatment group t, by acquiring the most uncertain samples that have the maximum predictive variance $\sigma^2_{f^t}$, the slowest convergence rate of the model variance is lower-bounded by $\Omega(\beta^i)$, where $0 \le \beta < 1$.

PROOF. We denote the model variance $\mathbb{E}_{\mathcal{X}}\left[\operatorname{Var}^t[\hat{f}^t(x;\mathcal{D}_{\mathrm{train}}^t)]\right] = \mathbb{E}_{\mathcal{X}}\left[(\hat{f}^t(x) - \mathbb{E}[\hat{f}^t(x)])^2\right]$ on treatment group t. With the empirical distribution on sample space \mathcal{X}^t , the empirical realization of the model variance with N_{pool} samples is:

$$\mathbb{E}_{\mathcal{X}}\left[\operatorname{Var}^{t}\left[\hat{f}^{t}(x; \mathcal{D}_{\text{train}}^{t})\right] = \frac{1}{N_{\text{pool}}} \sum_{i=1}^{N_{\text{pool}}} \sigma^{2}(f^{t}(x_{i})), \quad (18)$$

where the predictive variance for observation x_i is denoted as $\sigma^2(f^t(x_i))$.

For the model that counts the gold standard variance, i.e., Gaussian process [44], will cap the variance by the constant signal variance $\sigma_{f^t}^2$, e.g., $\sigma_{f^t}^2 = 1$. Intuitively for samples far away from the training set, the model's belief reverts back to the prior. Mathematically, for noiseless observations, we have the following bounded predictive variance $\sigma^2(f^t(x_*))$ for any x_* :

$$0 \le \sigma^2(f^t(x_*)) = k(x_*, x_*) - \mathbf{k}_*^T K^{-1} \mathbf{k}_* \le \sigma_{ft}^2$$
 (19)

where the RBF kernel $k(x_i, x_j) = \sigma_{f^t}^2 \exp\left(-\frac{1}{2\theta}\|x_i - x_j\|^2\right)$. We have zero variance estimation if $x_* \in \mathcal{D}_{\text{train}}^t$, and maximally $\sigma_{f^t}^2$ if x_* is far away since $\forall x_i \in \mathcal{D}_{\text{train}}^t$, $\exp\left(-\frac{1}{2\theta}\|x_i - x_*\|^2\right) \to 0$.

LEMMA 4. Denote the predictive variance of the Gaussian process regression model f, trained on the dataset of size m, for any test point x_* as $\sigma_m^2(f(x_*))$, the predictive variance will not grow when the training set is expanding with size $m_0 \ge 0$, i.e.: [45]

$$\sigma_{m+m_0}^2(f(x_*)) \le \sigma_m^2(f(x_*)). \tag{20}$$

Thus, assuming the training size $N_{i-1} = |\mathcal{D}_{i-1}^t|$, the fixed batch size $b_0 = |\tilde{\mathcal{D}}_i^t|$, and the post-acquisition training size $N_i = |\mathcal{D}_{i-1}^t| + |\tilde{\mathcal{D}}_i^t|$. Let's denote N_{pool} to be the pool set at (i-1)-th query step (containing the acquired batch $\tilde{\mathcal{D}}_i^t$), the variance reduction at i-th

query step is:

$$\Delta_{\text{Var.}}^t(\tilde{\mathcal{D}}_i^t) \tag{21a}$$

$$= \mathbb{E}_{\mathcal{X}} \left[\operatorname{Var}_{i-1}^{t} [\hat{f}^{t}(x; \mathcal{D}_{i-1}^{t})] - \operatorname{Var}_{i}^{t} [\hat{f}^{t}(x; \mathcal{D}_{i-1}^{t} \cup \tilde{\mathcal{D}}_{i}^{t})] \right]$$
(21b)

$$= \frac{1}{N_{\text{pool}}} \sum_{k=1}^{N_{\text{pool}}} \sigma_{N_{i-1}}^{2}(\hat{f}^{t}(x_{k})) - \frac{1}{N_{\text{pool}}} \sum_{k=1}^{N_{\text{pool}}} \sigma_{N_{i}}^{2}(\hat{f}^{t}(x_{k}))$$
(21c)

$$= \frac{1}{N_{\text{pool}}} \left(\sum_{k=1}^{N_{\text{pool}}-b_0} \sigma_{N_{i-1}}^2(\hat{f}^t(x_k)) + \sum_{j=N_{\text{pool}}-b_0+1}^{N_{\text{pool}}} \sigma_{N_{i-1}}^2(\hat{f}^t(x_j)) \right) -$$
(21d)

$$\frac{1}{N_{\text{pool}}} \left(\sum_{k=1}^{N_{\text{pool}} - b_0} \sigma_{N_i}^2(\hat{f}^t(x_k)) + \sum_{j=N_{\text{pool}} - b_0 + 1}^{N_{\text{pool}}} \sigma_{N_i}^2(\hat{f}^t(x_j)) \right)$$
(21e)

$$= \frac{1}{N_{\text{pool}}} \sum_{k=1}^{N_{\text{pool}}-b_0} \underbrace{\left(\sigma_{N_{i-1}}^2(\hat{f}^t(x_k)) - \sigma_{N_i}^2(\hat{f}^t(x_k))\right)}_{\text{None Negative by Largest}} + \tag{21f}$$

$$\frac{1}{N_{\text{pool}}} \sum_{j=N_{\text{pool}}-b_0+1}^{N_{\text{pool}}} \left(\sigma_{N_{i-1}}^2(\hat{f}^t(x_j)) - \sigma_{N_i}^2(\hat{f}^t(x_j)) \right)$$
(21g)

$$\geq \frac{1}{N_{\text{pool}}} \left(\sum_{j=N_{\text{pool}}-b_0+1}^{N_{\text{pool}}} \left(\sigma_{N_{i-1}}^2(\hat{f}^t(x_j)) - \sigma_{N_i}^2(\hat{f}^t(x_j)) \right) \right) \tag{21h}$$

$$= \frac{1}{N_{\text{pool}}} \cdot b_0 \cdot \sigma_{ft}^2, \tag{21i}$$

where the first inequality is given by **Lemma** 4, and the last equality is given by acquiring the most uncertain samples at *i*-th query step with maximum predictive variance $\sigma_{f^t}^2$, and together with the fact that the observed sample has zero variance by the Gaussian process model

Now that at i-th query step, we calculate the shrinkage G_i as follows:

$$G_i = \frac{\Delta_{\text{Var}_i}^t(\tilde{\mathcal{D}}_i^t)}{\mathbb{E}_{\mathcal{X}}\left[\text{Var}_{i-1}^t[\hat{f}^t(x;\mathcal{D}_{i-1}^t)]\right]}$$
(22a)

$$\geq \frac{\frac{1}{N_{\text{pool}}} \cdot b_0 \cdot \sigma_{f^t}^2}{\frac{1}{N_{\text{pool}}} \sum_{k=1}^{N_{\text{pool}}} \sigma_{N_{i-1}}^2(\hat{f}^t(x_k))}$$
(22b)

$$= \frac{b_0 \cdot \sigma_{ft}^2}{\sum_{k=1}^{N_{\text{pool}}} \sigma_{N_{i-1}}^2(\hat{f}^t(x_k))}$$
(22c)

$$\geq \frac{b_0 \cdot \sigma_{ft}^2}{N_{\text{pool}} \cdot \sigma_{ft}^2} = \frac{b_0}{N_{\text{pool}}} = \frac{b_0}{\omega b_0} = \frac{1}{\omega}, \tag{22d}$$

where $\omega \in \mathbb{R}^+$ as N_{pool} can be arbitrary larger than b_0 , i.e., $N_{\text{pool}} = \omega b_0$.

Thus, we derive the total shrinking coefficient G^i after i iterations as follows:

$$G^{i} = \prod_{n=1}^{i} (1 - G_{n})$$
 (23a)

$$\leq \prod_{n=1}^{l} \left(1 - \frac{1}{\omega} \right) \tag{23b}$$

$$= (1 - \frac{1}{\omega})^i = \beta^i, \tag{23c}$$

where the first inequality is straightforwardly by the inequality in (22), and furthermore we have the coefficient $0 \le \beta = (1 - \frac{1}{\omega}) < 1$.

Subsequently, with the initial model variance I_0' , and after i iterations (at i-th query step literally means accumulated i iterations for label acquisition), we have the upper bound for the variance as $I_0'G^i \leq g(i) = I_0'\beta^i$, where g(i) obeys the asymptotic behaviour in the following:

$$\lim_{i \to \infty} g(i) = \lim_{i \to \infty} \beta^i = 0.$$
 (24)

Thus, we can conclude that, by acquiring the most uncertain samples that have the maximum predictive variance $\sigma_{f^t}^2$, the slowest convergence rate is lower-bounded by $\Omega(\beta^i)$ where $0 \le \beta < 1$ since $I_0'G^i \le g(i)$.

A.3 Convergence Rate of Distributional Discrepancy

DEFINITION 4. Let $\mathcal{J}(P,Q)$ be all the joint distribution \mathcal{J} for (X,Y) that respectively have the marginal distribution P and Q. Then, the p-Wasserstein distance is defined as:

$$W_{p}(P,Q) = \left(\inf_{J \in \mathcal{J}(P,Q)} \int \|x - y\|^{p} dJ(x,y)\right)^{\frac{1}{p}}$$
(25)

Lemma 3. Given two empirical distributions $p^{t=1}$ and $p^{t=0}$ for different treatment groups, the distributional discrepancy given by 1-Wasserstein distance $W_1(p^{t=1},p^{t=0})$ has a convergence rate of $O(\frac{1}{i+\gamma_0})$ if the identical samples from two groups can always be found throughout the query steps.

PROOF. Given the empirical distribution for P and Q with N_i multi-dimensional observations at query step i, i.e., $p_i^{t=1}$ and $p_i^{t=0}$, the 1-Wasserstein distance is analogously by the Definition 4 reduced to the following:

$$W_1^i(p_i^{t=1}, p_i^{t=0}) = \inf_{\pi} \left(\sum_{k=1}^{N_i} \|x_k^{t=1} - x_{\pi(k)}^{t=0}\| \right), \tag{26}$$

where the infimum runs over all the possible permutations $\pi.$

Since there exists an optimal permutation π_i^* at i-th query step, by plugging in π_i^* to (26), we obtain the 1-Wasserstein distance between two empirical distributions. For now we just denote this value by $W_1^i(p_i^{t=1},p_i^{t=0})$ without knowing what exactly the number is. Therefore, for the distributional difference at i-th query step, we have two optimal permutation π_{i-1}^* and π_i^* to help us calculate the difference $\Delta_{\mathrm{IPM}_i}(\tilde{\mathcal{D}}_i)$:

$$\Delta_{\mathrm{IPM}_i}(\tilde{\mathcal{D}}_i) \tag{27a}$$

$$=W_1^{i-1}(p_{i-1}^{t=1}, p_{i-1}^{t=0}) - W_1^{i}(p_i^{t=1}, p_i^{t=0})$$
(27b)

$$= \frac{1}{N_{i-1}} \sum_{k=1}^{N_{i-1}} \|x_k^{t=1} - x_{\pi_{i-1}^*(k)}^{t=0}\| - \frac{1}{N_{i-1} + b_0} \sum_{k=1}^{N_{i-1} + b_0} \|x_k^{t=1} - x_{\pi_i^*(k)}^{t=0}\|$$
(27c)

$$= \frac{1}{N_{i-1}} \sum_{t=1}^{N_{i-1}} \|x_k^{t=1} - x_{\pi_{i-1}^*(k)}^{t=0}\| -$$
 (27d)

$$\frac{1}{N_{i-1} + b_0} \left(\sum_{k=1}^{N_{i-1}} \| x_k^{t=1} - x_{\pi_i^*(k)}^{t=0} \| + \sum_{l=N_{i-1}+1}^{N_{i-1}+b_0} \| x_l^{t=1} - x_{\pi_i^*(l)}^{t=0} \| \right)$$
(27e)

$$= \frac{N_{i-1} + b_0}{N_{i-1}(N_{i-1} + b_0)} \sum_{k=1}^{N_{i-1}} \|x_k^{t=1} - x_{\pi_{i-1}^*(k)}^{t=0}\| -$$
(27f)

$$\frac{N_{i-1}}{N_{i-1}(N_{i-1}+b_0)} \sum_{k=1}^{N_{i-1}} \|x_k^{t=1} - x_{\pi_i^*(k)}^{t=0}\|$$
 (27g)

$$= \left(\frac{N_{i-1} + b_0}{N_{i-1}(N_{i-1} + b_0)} - \frac{N_{i-1}}{N_{i-1}(N_{i-1} + b_0)}\right) \sum_{k=1}^{N_{i-1}} \|x_k^{t-1} - x_{\pi_{i-1}^*(k)}^{t=0}\|$$
(27h)

$$= \frac{b_0}{N_{i-1} + b_0} \cdot \frac{1}{N_{i-1}} \sum_{k=1}^{N_{i-1}} ||x_k^{t=1} - x_{\pi_{i-1}^*(k)}^{t=0}||,$$
 (27i)

where $b_0 = |\tilde{\mathcal{D}}_i|/2$, i.e., half of the batch size at each query step. Given that the added identical samples, the optimal permutation π_i^* at i-th query step will match these identical pairs due to the cost $\sum_{l=N_{i-1}+b_0}^{N_{i-1}+b_0} \|x_l^{t=1} - x_{\pi^*(l)}^{t=0}\| = 0$ introduce zero distributional discrepancy, thus the forth equality holds. Thus, for the rest of the N_{i-1} samples, the optimal permutation π_i^* must have the same transportation strategy as π_{i-1}^* does to obtain the lowest cost on the rest N_{i-1} samples, such that the fifth equality holds. We believe the equality claim resonates with the rigorously proved triangular inequality nature of Wasserstein metric [5].

Therefore, at i-th query step, the discrepancy shrinkage S_i is defined as:

$$S_i = \frac{\Delta_{\text{IPM}_i}(\tilde{\mathcal{D}}_i)}{W_1^{i-1}(p_{i-1}^{t=1}, p_{i-1}^{t=0})} = \frac{b_0}{N_{i-1} + b_0}$$
 (28)

Since the number of samples in one treatment group (training) at (i-1)-th query step, N_{i-1} , can be reformulated as $N_{i-1} = \gamma_{i-1} \cdot b_0$ with arbitrary $\gamma_{i-1} \in \mathbb{R}^+$. Subsequently S_i is reduced to:

$$S_i = \frac{b_0}{\gamma_{i-1} \cdot b_0 + b_0} = \frac{b_0}{b_0(\gamma_{i-1} + 1)} = \frac{1}{\gamma_{i-1} + 1}$$
(29)

Thus, we derive the total shrinking coefficient S^{i} after iiterations as follows:

$$S^{i} = \prod_{n=1}^{i} (1 - S_{n}) \tag{30a}$$

$$= \prod_{n=1}^{i} \left(1 - \frac{1}{\gamma_{n-1} + 1} \right) \tag{30b}$$

$$= \left(1 - \frac{1}{\gamma_0 + 1}\right) \cdot \left(1 - \frac{1}{\gamma_1 + 1}\right) \cdots \left(1 - \frac{1}{\gamma_{i-1} + 1}\right)$$
(30c)
$$= \left(1 - \frac{1}{\gamma_0 + 1}\right) \cdot \left(1 - \frac{1}{(\gamma_0 + 1) + 1}\right) \cdots \left(1 - \frac{1}{(\gamma_0 + i - 1) + 1}\right)$$

$$= \frac{\gamma_0}{\gamma_0 + 1} \cdot \frac{\gamma_0 + 1}{(\gamma_0 + 1) + 1} \cdot \cdot \cdot \frac{\gamma_0 + i - 2}{\gamma_0 + i - 1} \cdot \frac{\gamma_0 + i - 1}{\gamma_0 + i}$$
(30e)

$$=\frac{\gamma_0}{\gamma_0+i},\tag{30f}$$

where the arbitrary constant $\gamma_0 \in \mathbb{R}^+$ means the first ratio constant for N_0/b_0 , such that the distributional discrepancy can be initialized as $I_0 = W_1^0(p_0^{t=1}, p_0^{t=0}).$

Define $s(i) = I_0 S^i = I_0 \cdot \gamma_0 / (\gamma_0 + i)$, and for s(i) we have the following asymptotic behaviour:

$$\lim_{i \to \infty} s(i) = \lim_{i \to \infty} \frac{I_0 \gamma_0}{\gamma_0 + i} = 0.$$
 (31)

With identical acquisition from both of the treatment groups, we can conclude that the convergence rate of discrepancy is as fast as $O(\frac{1}{i+v_0}).$

Causal Effect Identifiability

Proposition 1 (Identifiability). The causal effect is identifiable if and only if the SUTVA, the unconfoundedness, and the positivity assumptions hold.

PROOF. Under SUTVA (Assumption 1) and unconfoundedness (**Assumption** 2), the ITE for instance i with cocvariate \mathbf{x}_i is:

$$\mathbb{E}[Y^{t=1} - Y^{t=0} | \mathbf{x}_i] = \mathbb{E}[Y^{t=1} | \mathbf{x}_i] - \mathbb{E}[Y^{t=0} | \mathbf{x}_i]$$

$$= \mathbb{E}[Y^{t=1} | \mathbf{x}_i, t_i = 1] - \mathbb{E}[Y^{t=0} | \mathbf{x}_i, t_i = 0]$$

$$= \mathbb{E}[y_i | \mathbf{x}_i, t_i = 1] - \mathbb{E}[y_i | \mathbf{x}_i, t_i = 0],$$
(32)

where y_i denotes the observed outcomes after the intervention t = 1/0 has been taken. The first equality is the rewritten expectation, the second equality is based on the unconfoundedness, and the third equality states that the expected values of the observed outcomes $\{y_1, y_0\}$ equal the unobserved potential outcomes. The last two terms are identifiable as we assume $0 < p(t = 1|\mathbf{x}) < 1$ (Assumption 3).

Factual Error Decomposition

Definition 5. *The expected treatment risk with status t in terms* of the expected squared loss function $\ell(x,t)$ with density $p^t(x)$ are defined respectively as follows:

$$\epsilon_F^{t=1} = \int_X \ell(x, 1) p^{t=1}(x) dx, \ \epsilon_F^{t=0} = \int_X \ell(x, 0) p^{t=0}(x) dx$$
 (33)

Proposition 2. Assume the potential effect y^t has the form $y^t =$ $f^{t}(x) + \xi^{t}$, and the estimated effect is set to $\hat{y}^{t} = \hat{f}^{t}(x)$. With the expected square loss function $\ell(x,t) = \mathbb{E}[(y-\hat{y})^2]$ The expected risk for either of the treatment groups can be decomposed into the following:

$$\epsilon_F^t = \mathbb{E}_{\mathcal{X}}[(f^t(x) - \mathbb{E}[\hat{f}^t(x)])^2] + \mathbb{E}_{\mathcal{X}}[\mathbb{E}[(\hat{f}^t(x) - \mathbb{E}[\hat{f}^t(x)])^2]] + \sigma_{\xi^t}^2$$
(34)

where $f^{t}(x)$ is the true function, ξ^{t} is the noise with 0 mean and constant variance $\sigma_{\xi t}^2$, and $\hat{f}^t(x)$ is the approximation.

PROOF. We start the proof toward a single data point x, then simply extending to the domain X can conclude the proof.

$$\ell(x,t) \tag{35a}$$

$$=\mathbb{E}[(y^t - \hat{y}^t)^2] \tag{35b}$$

$$=\mathbb{E}[(f^{t}(x) + \xi^{t} - \hat{f}^{t}(x))^{2}]$$
(35c)

$$=\mathbb{E}[(f^{t}(x) - \hat{f}^{t}(x) + \xi^{t})^{2}]$$
(35d)

$$=\mathbb{E}[(f^{t}(x) - \hat{f}^{t}(x))^{2} + 2(f^{t}(x) - \hat{f}^{t}(x))\xi^{t} + (\xi^{t})^{2})]$$
(35e)

$$= \mathbb{E}[(f^t(x) - \hat{f}^t(x))^2] + 2\mathbb{E}[(f^t(x) - \hat{f}^t(x))] \underbrace{\mathbb{E}[\xi^t]}_{0} + \mathbb{E}[(\xi^t)^2]$$

(35f)

$$= \mathbb{E}[(f^{t}(x) - \mathbb{E}[\hat{f}^{t}(x)] + \mathbb{E}[\hat{f}^{t}(x)] - \hat{f}^{t}(x))^{2}] + \sigma_{\varepsilon t}^{2}$$
 (35g)

$$= \mathbb{E}[(f^{t}(x) - \mathbb{E}[\hat{f}^{t}(x)])^{2}] + \mathbb{E}[(\mathbb{E}[\hat{f}^{t}(x)] - \hat{f}^{t}(x))^{2}] +$$
(35h)

$$2\mathbb{E}\left[\underbrace{(f^{t}(x) - \mathbb{E}[\hat{f}^{t}(x)])}_{\text{Constant w.r.t x}}\underbrace{(\mathbb{E}[\hat{f}^{t}(x)] - \hat{f}^{t}(x))}_{0 \text{ expectation}}\right] + \sigma_{\xi^{t}}^{2}$$
(35i)

$$= \underbrace{(f^{t}(x) - \mathbb{E}[\hat{f}^{t}(x)])^{2}}_{\text{Bias}} + \underbrace{\mathbb{E}[(\hat{f}^{t}(x) - \mathbb{E}[\hat{f}^{t}(x)])^{2}]}_{\text{Variance}} + \sigma_{\xi^{t}}^{2}$$
(35j)

Then, the expected factual loss across the domain X is:

$$\begin{split} &\int_{X} \ell(x,t) p(x) dx \\ &= \int_{X} \left((f^{t}(x) - \mathbb{E}[\hat{f}^{t}(x)])^{2} + \mathbb{E}[(\hat{f}^{t}(x) - \mathbb{E}[\hat{f}^{t}(x)])^{2}] + \sigma_{\xi^{t}}^{2} \right) p(x) dx \\ &= \underbrace{\int_{X} (f^{t}(x) - \mathbb{E}[\hat{f}^{t}(x)])^{2} p^{t}(x) dx}_{\text{Expected Bias}} + \end{split}$$

$$\underbrace{\int_{\mathcal{X}} \mathbb{E}[(\hat{f}^t(x) - \mathbb{E}[\hat{f}^t(x)])^2] p^t(x) dx}_{\text{Expected Variance}} + \int_{\mathcal{X}} \sigma_{\xi^t}^2 p^t(x) dx$$

$$= \mathbb{E}_{\mathcal{X}}[(f^{t}(x) - \mathbb{E}[\hat{f}^{t}(x)])^{2}] + \mathbb{E}_{\mathcal{X}}[\mathbb{E}[(\hat{f}^{t}(x) - \mathbb{E}[\hat{f}^{t}(x)])^{2}]] + \sigma_{\xi^{t}}^{2}$$
(36)

A.6 Algorithm

The full algorithm of MACAL is concluded in Algorithm 1 with additional consideration to the scenario where one of the treatment group's pool set is exhausted, such that the active learning process down to one-sided label acquisition without the accessibility to the other side.

Algorithm 1 MACAL

```
1: Input: Initializing the training set \mathcal{D}_{\text{train}} = \mathcal{D}_{\text{train}}^{t=1} \cup \mathcal{D}_{\text{train}}^{t=0}, pool set \mathcal{D}_{\text{pool}} = \mathcal{D}_{\text{pool}}^{t=1} \cup \mathcal{D}_{\text{pool}}^{t=0}, batch size S, initial query step Q = 1,
         symmetrical penalization \alpha, and maximum query step Q_{\text{max}}.
   2: for Q < Q_{\text{max}} do
                  \tilde{\mathcal{D}} = \emptyset
   3:
                  while |\mathcal{\tilde{D}}| < S do
                                                                                                                                                                     ▶ Keep acquiring the samples if the batch is not filled up in each query step
   4:
                           if |\mathcal{D}| < 3 do

if \mathcal{D}_{\text{pool}}^{t=1} \neq \emptyset and \mathcal{D}_{\text{pool}}^{t=0} \neq \emptyset then

\tilde{x}^{t=1}, \tilde{x}^{t=0} \leftarrow \arg\max_{(\tilde{x}^{t=1}, \tilde{x}^{t=0})} \sum_{t \in \{0,1\}} \min_{x' \in \mathcal{D}_{\text{train}}^t} d(\tilde{x}^t, x') - \alpha d(\tilde{x}^{t=1}, \tilde{x}^{t=0})
                                                                                                                                                                                            ▶ If both treatment groups still have available unlabelled samples
   5:
                                                                                                                                                                                                                                                                                           ▶ Acquire the sample pairs
   6:
                                      \mathcal{D}_{\text{train}}^{t=1}, \mathcal{D}_{\text{train}}^{t=0} \leftarrow \mathcal{D}_{\text{train}}^{t=1} \cup \{\tilde{x}^{t=1}\}, \mathcal{D}_{\text{train}}^{t=0} \cup \{\tilde{x}^{t=0}\} 
 \mathcal{D}_{\text{pool}}^{t=1}, \mathcal{D}_{\text{pool}}^{t=0} \leftarrow \mathcal{D}_{\text{pool}}^{t=1} \setminus \{\tilde{x}^{t=1}\}, \mathcal{D}_{\text{pool}}^{t=0} \setminus \{\tilde{x}^{t=0}\} 
 \tilde{\mathcal{D}} \leftarrow \tilde{\mathcal{D}} \cup \{\tilde{x}^{t=1}, \tilde{x}^{t=0}\} 
                                                                                                                                                                                                                                ▶ Add the unlabelled samples into the training sets
   7:
                                                                                                                                                                                                                                   ▶ Exclude the acquired samples from the pool sets
                                                                                                                                                                                                                                                                                        ▶ Update the acquired batch
                                                                                                                                                                                                                           ▶ If one of the treatment groups' pool set is exhausted
   8:
                                    \begin{aligned} \mathbf{f} & & \mathcal{D}_{\mathrm{pool}}^{t=1} \neq \varnothing \mathbf{then} \\ & & \tilde{x}^{t=1} \leftarrow \arg\max_{\tilde{x}^{t=1} \in \mathcal{D}^{t=1}} \min_{x' \in \mathcal{D}_{\mathrm{train}}^{t=1}} d(\tilde{x}^t, x') \\ & & \mathcal{D}_{\mathrm{train}}^{t=1}, \mathcal{D}_{\mathrm{pool}}^{t=1}, \tilde{\mathcal{D}} \leftarrow \mathcal{D}_{\mathrm{train}}^{t=1} \cup \{\tilde{x}^{t=1}\}, \mathcal{D}_{\mathrm{pool}}^{t=1} \setminus \{\tilde{x}^{t=1}\}, \tilde{\mathcal{D}} \cup \{\tilde{x}^{t=1}\} \end{aligned}
                                                                                                                                                                                                                                         ▶ If treatment group with t = 1 is not exhausted
   9:
 10:
 11:
                                                                                                                                                                                                                                        ▶ If treatment group with t = 0 is not exhausted
 12:
                                              \begin{split} \tilde{x}^{t=0} \leftarrow \arg\max_{\tilde{x}^{t=0} \in \mathcal{D}^{t=0}} \min_{x' \in \mathcal{D}^{t=0}_{\text{train}}} d(\tilde{x}^t, x') \\ \mathcal{D}^{t=0}_{\text{train}}, \mathcal{D}^{t=0}_{\text{pool}}, \tilde{\mathcal{D}} \leftarrow \mathcal{D}^{t=1}_{\text{train}} \cup \{\tilde{x}^{t=0}\}, \mathcal{D}^{t=0}_{\text{pool}} \setminus \{\tilde{x}^{t=0}\}, \tilde{\mathcal{D}} \cup \{\tilde{x}^{t=0}\} \end{split}
 13:
 14:
 15:
 16:
                            end if
                  end while
 17:
                  Reveal the acquired unlabelled samples in the training set via the Oracle.
                                                                                                                                                                                                                                                   ▶ Label all the acquired samples all at once
 18:
                                                                                                                                                                                                                                                                                               ▶ Move to next query step
 19:
20: end for
21: Output: Dtrain.
```

B Additional Experiments and Setup

B.1 Additional Visualizations of Post-Acquisition Dataset

We visualize the post-acquisition dataset distribution via t-SNE for the five most representative models: MACAL, Random, LCMD, QHTE, and $\mu\rho$ BALD, on all three datasets, i.e., CMNIST, IBM, and IHDP. Across all Figure 6, 7, and 8, we consistently observe that MACAL can significantly outperform the other methods in terms of acquiring the pairs to avoid the violation of positivity, and also expanding the data boundary for not being clustering at a small area to avoid repetitive samples. Interestingly, we notice that in the IHDP dataset, the pair acquisition by MACAL terminates at Step 15 as shown in Figure 8(b) since samples from the treatment group with t=1 are exhausted, such that, MACAL can only label the other treatment samples and leave a quite imbalanced dataset at Step 35 as shown in Figure 8(c).

B.2 Toy Dataset

We simulate the one-dimensional toy dataset for a simple demonstration of the importance of considering minimizing the model variance and distributional discrepancy altogether during label acquisition.

For samples with treatment status t=1: the first 100 samples are from the interval of [-12, 10] with equal spacing, and the second 400 samples are from the normal distribution with mean -2.5 and variance 1.

For samples with treatment status t = 0: the first 500 samples are from the interval of [-10, 11] with equal spacing, and the second 2000 samples are from the normal distribution with mean 2.5 and variance 1.

We have in total 500 samples with treatment status t=1 and 2500 samples with treatment status t=0 to form the imbalanced treatment groups as the entire dataset, then we do train/test split with 3:1 ratio for the model evaluation. The data-generating process is described mathematically as follows:

$$\begin{cases} x_i^{t=1} = -12 + i \cdot \frac{10 - (-12)}{100 - 1}, \text{ for } i \in [1, 100] \\ x_i^{t=1} \sim \mathcal{N}(-2.5, 1), \text{ for } i \in [101, 400] \end{cases}$$
(37a)

$$\begin{cases} x_i^{t=0} = -10 + j \cdot \frac{11 - (-10)}{500 - 1}, & \text{for } j \in [1, 500] \\ x_j^{t=0} \sim \mathcal{N}(2.5, 1), & \text{for } j \in [101, 2000] \end{cases}$$
(37b)

$$\begin{cases} y_i^{t=1} = \sin(2 \cdot x_i^{t=1}), \forall i \\ y_j^{t=0} = \cos(2 \cdot x_j^{t=1}), \forall j \end{cases}$$
(37c)

B.3 Hyperparameters

We conduct all the experiments with 48GB NVIDIA A40 on Ubuntu 22.04 LTS platform where GPU training is enabled, otherwise the 12th Gen Intel i7-12700K 12-Core 20-Thread CPU is used. The standard hyperparameter tuning on the validation set which is further split from the train set with 3:1 ratio, the best hyperparameters are selected with the smallest validation loss. Since the DUE models are borrowed from [17], we acknowledge the model set up from

Hyperparameters	Search Space	DUE-DNN	DUE-CNN
Kernel	[RBF, Matern, RQ]	RBF	Matern
Inducing Points	[50, 100, 200]	100	100
Hidden Neurons	[100, 200, 500]	200	200
Depth	[2,3,5]	3	2
Dropout Rate	[0.05, 0.1, 0.25]	0.1	0.05
Spectral Norm	[0.95, 1.5, 3.0]	0.95	3.0
Batch Size	[64, 100, 200]	100	64
Learning Rate	[1e-3, 1e-4]	1e-3	1e-3

Table 2: Search Space and Tuned Hyperparameters for DUE-DNN and DUE-CNN

the previous literature and adopt a similar search space as shown in Table 2.

C Limitation and Future Work

In our proposed risk upper reduction theory, we make further claims for the risk convergence behaviour under two extreme circumstances due to the negligibility of the bounded constant C_{ϕ} . We believe, the convergence analysis for each of the extreme situations can help justify the algorithm design, i.e., with negligible C_{ϕ} the risk upper bound shrinks to the variance term, where keep acquiring the most uncertain samples can enable the rate of convergence is lower-bounded by $\Omega(\beta^i)$, while, with dominant C_{ϕ} , the rate of convergence is upper-bounded by $O(\frac{1}{i+\gamma_0})$. We also empirically observe these situations by setting different C_{ϕ} via the ablation study in Appendix 5.3, where it is clearly observed that MACAL with dominant C_{ϕ} performs the best at the start, but in the mid of the acquisition, a smaller C_ϕ (not negligible yet) obtains the best performance. However, the limitation of the convergence analysis is, due to technical difficulties, we do not obtain the risk convergence for the entire risk upper bound, i.e., when the C_{ϕ} sitting in the middle and making both the variance and the distributional discrepancy comparably important (which can be more realistic). We believe this point of research remains a import direction to be figured out in future work.

Additionally, our designed algorithm MACAL, even though bring down the NP-hard combinatorial optimization to be approximately solved in polynomial time, i.e., $O(N_{pool}^2)$. When facing a significant large pool set with hundreds of millions of samples, the squared time complexity still suffers from considerable computational problems and become undesirable. Thus, future research on how to further reduce the algorithm time complexity is also an important direction to go when facing large real-world datasets.

D Broader Impacts

Causal effect estimation with active Learning could potentially have broader impacts on society if the algorithm is leveraged to deal with the treatment effect estimation in reality. One of the representative examples can be the hospital scenario, where patients' information is used for the training of the treatment effect estimator.

When doing the active learning to selectively screen the samples and label them, once the AL algorithm identifies the informative sample to be labelled, the patient's individual information (features), and the corresponding treatment effect would be revealed. Subsequently, by labelling more informative samples, the positive impact is that a more precise treatment effect estimator can be trained on the ongoing growing training set, and help make more precise decision on the patient's treatment plan. However, the negative impact is, that the identified patients need to reveal their treatment information which can introduce privacy concerns and go against their will. Thus, when the causal effect active learning algorithm is used in the real world, the conductors should strictly consider the negative impact on the patient's privacy and its willing during the label acquisition process.

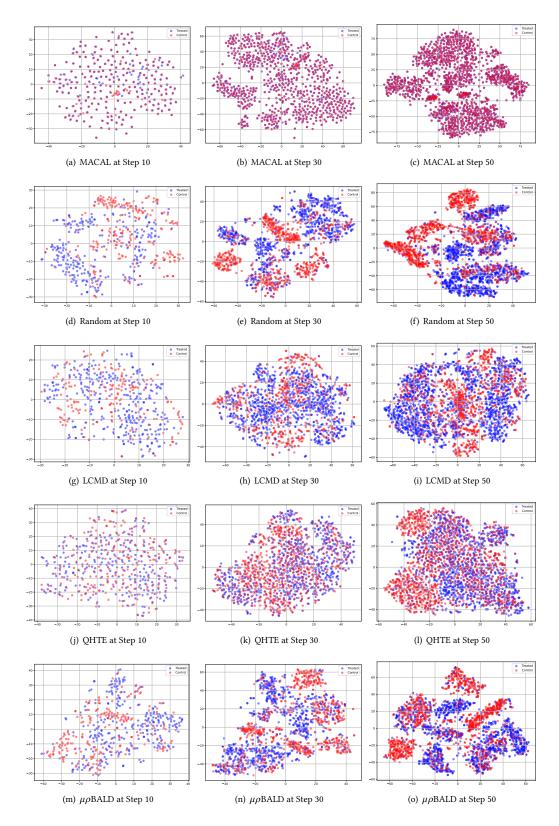


Figure 6: Visualization of the post-acquisition training set at query step 10, 30, and 50 via t-SNE on CMNIST dataset.

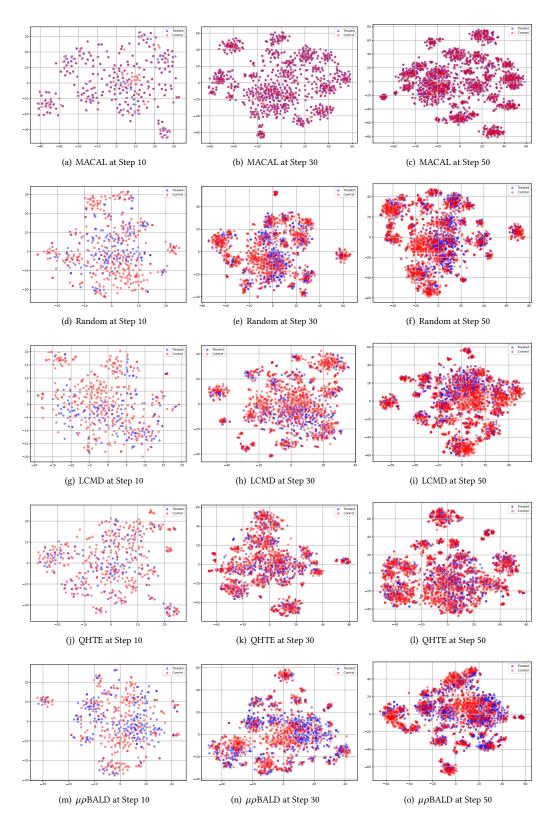


Figure 7: Visualization of the post-acquisition training set at query step 10, 30, and 50 via t-SNE on IBM dataset.

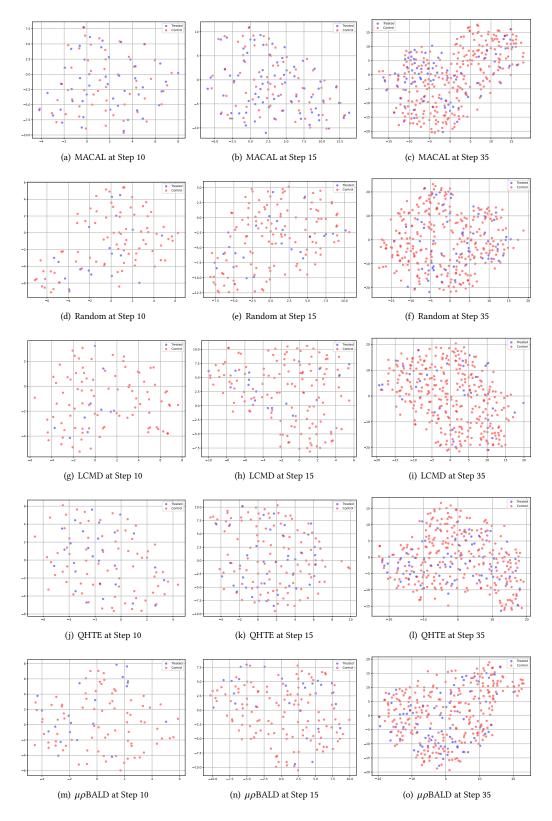


Figure 8: Visualization of the post-acquisition training set at query step 10, 15, and 35 via t-SNE on IHDP dataset.