Unbiased Active Semi-supervised Binary Classification Models

JooChul Lee[1*], **Weidong Ma**[2] and **Ziyang Wang**[3]
[1]Department of Mathematics and Statistics, Auburn University, USA
[2]Department of Biostatistics, Epidemiology and Informatics, University of Pennsylvania Perelman School of Medicine, USA
[3]Department of Statistics, University of Connecticut, USA

Abstract

Active learning is known to be a well-motivated algorithm that aims to maximize model performance with relatively small data, but it introduces sampling bias due to active selection. To adjust the bias, current literature utilizes corrective weights in a supervised learning approach. However, those methods consider only a small amount of actively sampled data and thus estimation efficiency can be improved using unsampled data together. In this paper, we develop an actively improved augmented estimation equation (AI-AEE) based on corrective weights as well as imputation models that allow us to leverage unlabeled data. The asymptotic distribution of the proposed estimator as the solution to the AI-AEE is derived, and an optimal sampling scheme to minimize the asymptotic mean squared error of the estimator is proposed. We then propose a general practical algorithm for training prediction models in the active and semi-supervised learning framework. The superiority of our method is demonstrated on synthetic and real data examples.

1 Introduction

With the advancement of technology, big data has improved the performance of modern machine learning and statistical models. However, dealing with a huge amount of unlabeled data is a key challenge in many fields, such as electronic health records [Gronsbell *et al.*, 2022], speech recognition [Zhu, 2005], and text extraction [Settles *et al.*, 2008]. Since labeling massive data is time-consuming, expensive, and labor-intensive, it is important to acquire a subset of reliable data points from domain experts. Active learning (AL) is an algorithm aiming for maximizing model performance with sampled data. By selecting potentially more informative data points, models can be trained more efficiently in AL setting. AL has a close connection with sampling designs in that subsamples are drawn from a pool dataset (it is known as $subsampling$). In many cases, sampling designs necessitate prior information that we aim to estimate. Regarding subsampling, we can implement the sampling designs using knowledge obtained from actively labeled data within the AL setting.

Although AL could be a promising sample-efficient learning algorithm, a major limitation is the sampling bias caused by active selection where data points are selected from a pool dataset in sequence. Since the actively sampled data points are not drawn from a common population distribution, they may lead to biased training models for the target population unless the sampling bias is appropriately adjusted. Inverse probability weighting [Horvitz and Thompson, 1952; Ganti and Gray, 2012] is a standard method for removing sampling bias under importance sampling, but this cannot be applied directly to actively sampled data. To adjust the bias introduced by active selection, Farquhar *et al.*[2021] corrected sampling weights in a manner where modified weights are assigned to the data points selected at earlier steps. By applying corrective weighting, they proposed an unbiased estimator for a general loss function. However, they utilized only a small amount of actively sampled data for model training in the manner of supervised learning and therefore, there is room for improvement in estimation efficiency by leveraging unsampled data.

Relevant work on semi-supervised learning (SSL) has demonstrated that SSL algorithms yields better performance than supervised learning in many cases by constructing imputation models using labeled data, which are then employed to impute outcomes for unlabeled data.[Krijthe and Loog, 2017; Chakrabortty and Cai, 2018]. Subsequently, prediction models are built using the labeled data and the imputed data. Motivated by these results, we focus on leveraging unlabeled data under the AL setting in this work. We propose an actively improved augmented estimating equation (AI-AEE) based on corrective weights and imputation models. The main idea behind AI-AEE is to automatically annotate unsampled data by using an imputation model constructed from actively sampled data. Moreover, we propose sampling schemes to actively select informative data points. Several recent works have investigated optimal sampling probability under binary classifiers aiming to minimize the asymptotic variance of the resultant estimator [Wang *et al.*, 2018; Zhang *et al.*, 2021]. Adopting this idea, we derive the asymptotic distribution of our proposed estimator and an optimal sampling scheme by minimizing the asymptotic mean square

*Corresponding Author

error of the estimator. The major contributions are as follows.

1. We propose the AI-AEE constructed from actively labeled data and unlabeled data. To leverage the unlabeled data, an actively improved imputation model is considered. The AI-AEE is an unbiased estimator for true target population risk, and is robust even when the imputation model deviates from for the true model.
2. We derive asymptotic distributions of the proposed estimator obtained from AI-AEE and an existing estimator and compare the efficiency of the estimators.
3. We propose an optimal sampling scheme to minimize the asymptotic mean squared error of the proposed estimator.
4. Based on the proposed estimator and sampling scheme, we propose a practical batch-mode algorithm for training prediction models in the active and semi-supervised learning setting. By applying the algorithm to synthetic and real data examples, we demonstrate the superiority of our methods compared to others.

The paper is organized as follows. Section 2 describes the problem setup in this work and examines theoretical results of the estimator based on the existing method. Section 3 proposes the estimator as the solution to AI-AEE and provides its theoretical properties and insights. Also, we propose an optimal sampling scheme for the proposed estimator and develop the practical algorithm in the active semi-supervised learning setting. In Section 4, we investigate previous works relevant to AL, SSL, and optimal subsampling schemes. Section 5 presents results of numerical studies. Section 6 concludes the paper and discusses future works.

2 Related Works

2.1 Unbiased Active Learning and Testing

In machine learning fields, AL has been a powerful tool for developing sample-efficient algorithms in a manner that informative data points are labeled throughout multiple steps using the information from earlier steps. However, many related works with active learning algorithms did not address bias due to active selection [Gal *et al.*, 2017; Yoo and Kweon, 2019]. To overcome this problem, unbiased AL algorithms were proposed for training models under sampling with replacement [Ganti and Gray, 2012] and active selection [Farquhar *et al.*, 2021]. A few recent works developed unbiased model evaluation methods in the AL setting (it is also known as $active\ testing$). Yilmaz *et al.* proposed an unbiased estimator of test metrics under Poisson sampling. Kossen *et al.*[2021] developed an estimator for a model test risk and Kossen *et al.*[2022] improved the efficiency of model evaluation using a surrogate model under active selection.

2.2 Semi-Supervised Learning (SSL)

SSL can lead to efficiency gains in training models by using labeled and unlabeled data together. Recent relevant works have investigated the classification model training with high dimensional covariates [Chakrabortty *et al.*, 2019], data shift [Cai *et al.*, 2022], and surrogate variables [Hou *et al.*, 2021], as well as model validation with classification accuracy metrics [Gronsbell and Cai, 2018] and data shift [Wang *et al.*, 2022b; Zhou *et al.*, 2022] under the SSL setting. Those works considered imputation models to replace unlabeled data with imputed values. However, they have studied under the simple sampling setting where the labeled data were derived from a random sampling. Gronsbell *et al.*[2022] selected a small subset of data under the stratified sampling and improved estimation efficiency of Brier score and overall misclassification rate leveraging unlabelled data together.

2.3 Optimal Subsampling

Subsampling strategy is important to improve estimation efficiency by labeling informative subsets of the pool data. In recent works faced with massive data, optimal subsampling strategies have been developed for machine learning and statistical models, such as classification models [Wang *et al.*, 2018; Yao and Wang, 2019; Wang *et al.*, 2021], generalized linear models [Ai *et al.*, 2018; Lee *et al.*, 2021], and mixture models [Lee *et al.*, 2022]. Those papers and our work have different goals. Assuming fully labeled data are available, the above-mentioned work selects subsample with the goal of mitigating computational burden. Also, the proposed subsampling probabilities in those papers depend on outcomes which cannot be used in our paper. Imberg *et al.*[2020] and Zhang *et al.*[2021] constructed optimal subsampling designs for generalized linear models under sampling with replacement when outcomes are not available. However, since we considered sampling design under the active learning setting and an extension of generalized linear models, we cannot directly apply their sampling design to our setting. Farquhar *et al.*[2021] introduced an optimal subsampling distribution which is proportional to the expectation of a one dimensional loss function. Imberg *et al.*[2022] developed optimal active sampling schemes for finite population characteristics based on machine learning tools. However, since both designs were derived based on models different from our target model, we cannot directly apply them to our specific setting.

3 Problem Setup

Let y be the binary outcome variable and $\mathbf{x}$ be the p dimensional vector of covariates including the intercept term. We consider a possibly misspecified working model $P(y = 1|\mathbf{x}) = g(\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta})$ where $g(\cdot)$ is a known smooth function of $\boldsymbol{\beta}$ and $\boldsymbol{\beta}$ is the unknown parameter. That is, the working model might deviate from the conditional density of y given $\mathbf{x}$ due to the invalid model assumptions.

Let $\boldsymbol{\beta}_t$ be the unknown parameter satisfying the following estimating equation,

$$\mathbf{E}[\mathbf{x}\{y - g(\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta})\}] = \mathbf{0}. \tag{1}$$

The equation is commonly used to obtain quasi-likelihood estimators for generalized linear models. Although this model may not be correctly specified for the true model, the working model is commonly used for the purpose of interpretability to examine the association between the outcome and covariates in the statistical field. Under the AL setting, we start with a dataset including only fully labeled covariates. Let

$\mathcal{D}_N = \{\mathbf{x}_i\}_{i=1}^N$ be the dataset of size N with labeled covariates where $\mathbf{x}_i$'s are independent and identically distributed. If data for y is fully observed, we can obtain $\hat{\boldsymbol{\beta}}_f$ for $\boldsymbol{\beta}_t$ as the solution to the full data based estimating equation,

$$\sum_{i=1}^{N} \mathbf{x}_i\{y_i - g(\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta})\} = \mathbf{0}. \tag{2}$$

However, since we cannot acquire all labeled outcomes in practice, we sample a subset of data for labeling. Let $\mathcal{D}_n = \{(y_1, \mathbf{x}_1), ..., (y_n, \mathbf{x}_n)\}$ denote as labeled outcomes, covariates selected from $\mathcal{D}_N$. For $1 \le s \le n$, let $\mathcal{D}_s$ be the selected data from the first to the s^{th} sampling step from $\mathcal{D}_N$, and $\mathcal{D}_{-s}$ be the remaining data excepting $\mathcal{D}_s$. Let $R_s = \{i : \mathbf{x}_i \notin \mathcal{D}_{-s}, 1 \le i \le N\}$ and $R_{-s} = \{i : \mathbf{x}_i \in \mathcal{D}_{-s}, 1 \le i \le N\}$ be the index sets indicating the sampled data and the unsampled data from the first to the s^{th} sampling step, respectively. Let $\pi(\mathbf{x}_k, R_{(s-1)})$ be the sampling probability used for the s^{th} sampling step for $k \in R_{-(s-1)}$. If we select data points randomly, we can consider the sampled data based estimating equation to obtain an estimator,

$$\sum_{i=1}^{n} \mathbf{x}_i\{y_i - g(\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta})\} = \mathbf{0}. \tag{3}$$

Since the non-uniform sampling probability is considered in the AL setting, the application of the estimating equation in (3) with the actively sampled data can lead to a biased estimator. If the model $g(\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta})$ is correctly specified, then we can obtain (asymptotically) an unbiased estimator from (3) for $\boldsymbol{\beta}_t$ since the sampling probability depends on only covariates [Wang and Kim, 2022; Wang *et al.*, 2022a]. However, when the prediction models are misspecified, the unbiased estimator from (3) is not guaranteed. In this work, we wish to unbiasedly estimate $\hat{\boldsymbol{\beta}}_f$ with the actively selected subdata accommodating the model misspecification.

4 Estimation

4.1 Corrective Weighting Estimator

Under the AL framework, Farquhar *et al.*[2021] adjusted the sampling bias based on corrective weights. Adopting the approach, we can obtain the corrective weighting (CW) estimator $\tilde{\boldsymbol{\beta}}^{cw}$ from the following estimating equation,

$$Q^{cw}(\boldsymbol{\beta}) \equiv \sum_{i=1}^{n} w_i \mathbf{x}_i\{y_i - g(\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta})\} = \mathbf{0}, \tag{4}$$

where $w_i = 1 + [\{(N-i+1)\pi(\mathbf{x}_i, R_{(i-1)})\}^{-1} - 1](N-n)/(N-i)$. The corrective weight w_i is readjusted at each step in an iterative manner to remove the bias. When the sample size n increases, $n \to N$, the corrective weights goes to 1 and $Q^{cw}(\boldsymbol{\beta})$ is closer to the full data based estimating equation in (2). In addition, if the sampling probability is uniform, w_i's are equal to one and $Q^{cw}(\boldsymbol{\beta})$ is the same as the estimating equation in (3). The following result shows that $Q^{cw}(\boldsymbol{\beta})$ is unbiased.

Proposition 1. *The estimating equation $Q^{CW}(\boldsymbol{\beta})$ is an unbiased estimator of $\mathbf{E}[\mathbf{x}\{y - g(\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta})\}]$.*

To further investigate the asymptotic distribution of $\tilde{\boldsymbol{\beta}}^{cw}$, we need the following assumptions.

Assumption 1. *The matrix $\sum_{i=1}^{N} \dot{g}(\mathbf{x}_i^{\mathrm{T}}\hat{\boldsymbol{\beta}}_f)\mathbf{x}_i\mathbf{x}_i^{\mathrm{T}}/N$ goes to a positive-definite matrix in probability where $\dot{g}(\eta) = \partial g(\eta)/\partial\eta$.*

Assumption 2. $\max_{k \in R_{(s-1)}} \|\mathbf{x}_k\|^4/\{N\pi(\mathbf{x}_k, R_{(s-1)})\} = O_p(1)$ *for $1 \le s \le n$.*

Assumption 3. *Assume that $\dot{g}(\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta})$ is Lipschitz continuous in $\boldsymbol{\beta}$. There exists $\varphi(\mathbf{x}_i)$ with $\mathbf{E}(\varphi(\mathbf{x}_i)^2) < \infty$ such that $|\dot{g}(\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}_1) - \dot{g}(\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}_2)| \le \varphi(\mathbf{x}_i)\|\boldsymbol{\beta}_1 - \boldsymbol{\beta}_2\|$ for every $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$.*

Assumption 1 is a mild condition to ensure that the target function has an unique maximum solution. However, this assumption may not hold in the high-dimensional setting where the number of covariates is much larger than the total data size and subdata size. Assumption 2 is a condition on sampling probabilities and the distribution of covariates. It imposes moment constraints. For example, Assumption 2 holds if $\mathbf{E}(\mathbf{x}^4) < \infty$ for equal sampling probabilities. Assumption 3 restrict the gradient of the function $g(\cdot)$ to ensure that we can use a martingale central limit theorem to establish the asymptotic normality of the estimator.

Theorem 1. *Under Assumptions* (1)- (3)*, if $N, n \to \infty$*

$$\sqrt{n}\mathbf{V}_{cw}^{-1/2}(\tilde{\boldsymbol{\beta}}^{cw} - \hat{\boldsymbol{\beta}}_f) \longrightarrow N(0, \mathbf{I}), \tag{5}$$

in distribution, where $\mathbf{V}_{cw} = \Sigma_N^{-1}\Lambda^{cw}\Sigma_N^{-1}$, $\Lambda^{cw} = \Lambda_1^{cw} - \Lambda_2^{cw}$,

$$\Sigma_N = \frac{1}{N}\sum_{i=1}^{N} \dot{g}(\mathbf{x}_i^{\mathrm{T}}\hat{\boldsymbol{\beta}}_f)\mathbf{x}_i\mathbf{x}_i^{\mathrm{T}},$$

$$\Lambda_1^{cw} = \frac{1}{nN^2}\sum_{i=1}^{n} c_i \sum_{k \in R_{-(i-1)}} \frac{\mathbf{x}_k\mathbf{x}_k^{\mathrm{T}}\{y_k - g(\mathbf{x}_k^{\mathrm{T}}\hat{\boldsymbol{\beta}}_f)\}^2}{\pi(\mathbf{x}_k, R_{(i-1)})},$$

$$\Lambda_2^{cw} = \frac{1}{nN^2}\sum_{i=1}^{n} c_i \left(\sum_{k \in R_{-(i-1)}} \mathbf{x}_k\{y_k - g(\mathbf{x}_k^{\mathrm{T}}\hat{\boldsymbol{\beta}}_f)\}\right)^{\otimes}.$$

and $\mathbf{A}^{\otimes} = \mathbf{A}\mathbf{A}^{\mathrm{T}}$ for any vector $\mathbf{A}$.

In Theorem 1, the matrix Λ^{cw} can be viewed as the variation due to subsampling and Λ_1^{cw} depends on the sampling probabilities.

Remark 1. We note that under some regularity conditions, $\hat{\boldsymbol{\beta}}_f - \boldsymbol{\beta}_t = O_p(1/\sqrt{N})$ [McCullagh, 1983]. Then, if $n/N \to 0$, we have $\sqrt{n}(\tilde{\boldsymbol{\beta}}^{cw} - \hat{\boldsymbol{\beta}}_f)$ converges in distribution to a normal distribution with mean $\mathbf{0}$ and variance-covariance matrix $\mathbf{V}_{cw}$.

4.2 Proposed Estimator

Although the estimating equation in (4) yields the unbiased estimator, it ignores the unsampled data. In the SSL literature, improved estimation efficiency was gained by augmenting imputed outcomes [Robins *et al.*, 1994; Carpenter *et al.*,

2006; Cao *et al.*, 2009]. Inspired by these results, we leverage the unlabeled data through an imputation approach. Let $m(\mathbf{x})$ be the imputation model used for labeling unsampled outcomes. Let $\{\delta_i^{1:n}\}_{i=1}^N$ be the indicator function where $\delta_i^{1:n} = 1$ if i^{th} data point is selected in the first through the n^{th} sampling and $\delta_i^{1:n} = 0$ otherwise. We consider the augmented estimating equation (AEE),

$$Q^{aee}(\boldsymbol{\beta}) \equiv \sum_{i=1}^{N} \delta_i^{1:n} w(\mathbf{x}_i)\mathbf{x}_i\{y_i - g(\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta})\} + \{1/N - w(\mathbf{x}_i)\delta_i^{1:n}\}\mathbf{x}\{m(\mathbf{x}_i) - g(\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta})\} = \mathbf{0}.$$

We show that the $Q^{aee}(\boldsymbol{\beta})$ is unbiased.

Proposition 2. *The AEE, $Q^{aee}(\boldsymbol{\beta})$ is an unbiased estimator of $\mathbf{E}[\mathbf{x}\{y - g(\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta})\}]$.*

From Proposition 2, we observe that the proposed AEE is unbiased although the imputation model $m(\mathbf{x})$ is misspecified for the true model. We can consider the imputation models using all covariates without loosing information of covariates such as addictive models, nonparametric models, random forest, and gradient boosting to deal with more complex structures between outcome and covariates. With $\hat{m}_n(\cdot)$ the imputation model developed by the actively sampled data $\mathcal{D}_n$, we propose the actively improved (AI) estimator $\tilde{\boldsymbol{\beta}}^{ai}$ as the solution to actively improved AEE (AI-AEE),

$$Q^{ai}(\boldsymbol{\beta}) \equiv \sum_{i=1}^{N} \delta_i^{1:n} w(\mathbf{x}_i)\mathbf{x}_i\{y_i - g(\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta})\} + \{1/N - w(\mathbf{x}_i)\delta_i^{1:n}\}\mathbf{x}_i\{\hat{m}_n(\mathbf{x}_i) - g(\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta})\} = \mathbf{0}.$$

We present an additional assumption to investigate the asymptotic distribution of $\tilde{\boldsymbol{\beta}}^{ai}$.

Assumption 4. *Assume that* $\sup_{\mathbf{x}\in\mathcal{X}} |\hat{m}_n(\mathbf{x}) - m(\mathbf{x})| = o_p(1)$.

Assumption 4 imposes the condition on the imputation model $m(\mathbf{x})$ to ensure that the difference between $\hat{m}_n(\mathbf{x})$ and the limiting of $m(\mathbf{x})$ is small when the subdata size is large enough.

Theorem 2. *Under Assumptions 1, 2 and 4, if $N, n \to \infty$*

$$\sqrt{n}\mathbf{V}_{ai}^{-1/2}(\tilde{\boldsymbol{\beta}}^{ai} - \hat{\boldsymbol{\beta}}_f) \longrightarrow N(0, \mathbf{I}), \tag{6}$$

in distribution, where $\mathbf{V}_{ai} = \Sigma_N^{-1}\Lambda^{ai}\Sigma_N^{-1}$, $\Lambda^{ai} = \Lambda_1^{ai} - \Lambda_2^{ai}$,

$$\Lambda_1^{ai} = \sum_{i=1}^{n} \frac{c_i^2}{nN^2} \sum_{k\in R_{-(i-1)}} \frac{\mathbf{x}_k\mathbf{x}_k^{\mathrm{T}}\{y_k - m(\mathbf{x}_k)\}^2}{\pi(\mathbf{x}_k, R_{(i-1)})},$$

$$\Lambda_2^{ai} = \frac{1}{nN^2}\sum_{i=1}^{n} c_i^2 \left(\sum_{k\in R_{-(i-1)}} \mathbf{x}_k\{y_k - m(\mathbf{x}_k)\}\right)^{\otimes}.$$

As discussed in the previous section, the matrix Λ^{ai} can be viewed as the variation due to subsampling and Λ_1^{ai} depends on the sampling probabilities. Also, if $n/N \to 0$, we have $\sqrt{n}(\tilde{\boldsymbol{\beta}}^{ai} - \boldsymbol{\beta}_t)$ converges in distribution to a normal distribution with mean $\mathbf{0}$ and variance-covariance matrix $\mathbf{V}_{ai}$ under some regularity conditions. From the asymptotic results, we compare the proposed estimator $\tilde{\boldsymbol{\beta}}^{ai}$ with $\tilde{\boldsymbol{\beta}}^{cw}$.

Theorem 3. *Under Assumption 2, if $m(\mathbf{x}) = \mathbf{E}(y|\mathbf{x})$, $\mathbf{V}_{cw} + o_p(1) \geq \mathbf{V}_{ai}$, where $\mathbf{B}_1 \geq \mathbf{B}_2$ if and only if $\mathbf{B}_1 - \mathbf{B}_2$ is positive semi-definite for two positive semi-definite matrices $\mathbf{B}_1$ and $\mathbf{B}_2$.*

From Theorem 3, when the imputation model is correctly specified, the estimation efficiency of $\tilde{\boldsymbol{\beta}}^{ai}$ is asymptotically higher, compared to $\tilde{\boldsymbol{\beta}}^{cw}$.

5 Subsampling Probability and Algorithm

5.1 Self-Learning Based Subsampling Probability

A key challenge in the AL setting is to select informative data points. Theorem 2 shows that the asymptotic variance of the proposed estimator depends on the sampling probability. Thus, we aim to minimize $\mathbf{V}_{ai}$ for achieving higher estimation efficiency with less data. To the this end, we consider the A-optimality criterion minimizing the trace of asymptotic variance matrix [Kiefer, 1959; Wang *et al.*, 2018].

Theorem 4. *The optimal subsampling probabilities at s^{th} sampling step given $R_{(s-1)}$ that minimize* $\mathrm{tr}(\mathbf{V}_{ai})$ *are*

$$\pi_{k,s}^{os} = \frac{|y_k - m(\mathbf{x}_k)|\|\Sigma_N^{-1}\mathbf{x}_k\|}{\sum_{j\in R_{-(s-1)}} |y_j - m(\mathbf{x}_j)|\|\Sigma_N^{-1}\mathbf{x}_j\|}, \tag{7}$$

for $k \in R_{-(s-1)}$. In (4), we give preferences to data points with larger quantities of $|y_i - m(\mathbf{x}_i)|$. The closer data points are to classification boundary, the more they are likely to be sampled. However, we cannot directly calculate the sampling probability since it depends on unobserved outcome y_i. Thus, we propose surrogate sampling probabilities. By replacing y_i by $g(\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta})$, we can consider the following self-learning based sampling (SBS) probabilities at the s^{th} sampling step given $R_{(s-1)}$,

$$\pi_{k,s}^{sbs} = \frac{|g(\mathbf{x}_k^{\mathrm{T}}\boldsymbol{\beta}) - m(\mathbf{x}_k)|\|\Sigma_N^{-1}\mathbf{x}_k\|}{\sum_{j\in R_{-(s-1)}} |g(\mathbf{x}_j^{\mathrm{T}}\boldsymbol{\beta}) - m(\mathbf{x}_j)|\|\Sigma_N^{-1}\mathbf{x}_j\|}, \tag{8}$$

for $k \in R_{-(s-1)}$. The SBS probability is proportional to the quantity $|g(\mathbf{x}_k^{\mathrm{T}}\boldsymbol{\beta}) - m(\mathbf{x}_k)| = |\{y_k - g(\mathbf{x}_k^{\mathrm{T}}\boldsymbol{\beta})\} - \{y_k - m(\mathbf{x}_k)\}|$. Either the data points are close to the boundary from the model $g(\cdot)$ but are not close to the boundary from the model $m(\cdot)$ or visa versa, they are selected with high probability.

Remark 2. We note that $\mathrm{tr}(\mathbf{V}_{ai}) \leq \mathrm{tr}(\mathbf{V}_{cw}) + \mathrm{tr}(\mathbf{V}_u)$ where $\mathbf{V}_u = \Sigma_N^{-1}\Lambda_u\Sigma_N^{-1}$ and

$$\Lambda_u = \frac{1}{nN^2}\sum_{i=1}^{n}\sum_{k\in R_{-(i-1)}} \frac{\mathbf{x}_k\mathbf{x}_k^{\mathrm{T}}\{g(\mathbf{x}_k^{\mathrm{T}}\boldsymbol{\beta}) - m(\mathbf{x}_i)\}^2}{\pi(\mathbf{x}_k|R_{(i-1)}, \mathcal{D}_N)}.$$

The SBS probability can be obtained by minimizing $\mathrm{tr}(\mathbf{V}_u)$. Thus, we can view that SBS probability aims for minimizing the upper bound of the asymptotic variance $\mathbf{V}_{ai}$.

Algorithm 1 Unbiased Active Semi-supervised Learning Algorithm

• Select a pilot subsample of size n_0 randomly from $\mathcal{D}_N$ for the initial step. Using the subsample, build the imputation model $\hat{m}_0(\cdot)$, and calculate $\tilde{\boldsymbol{\beta}}_0^{ai}$ from the equation (3) and $\Sigma_{\tilde{\boldsymbol{\beta}}_0^{ai},N}$.

• For $b = 1, 2, ...,$ repeat until cost of labelling is regulated

1. Calculate the sampling probability based on $\tilde{\boldsymbol{\beta}}_{b-1}^{ai}, \Sigma_{\tilde{\boldsymbol{\beta}}_{b-1}^{AI},N}$, and $\hat{m}_{b-1}(\cdot)$,

$$\tilde{\pi}_{k,b}^{sbs} \propto |g(\mathbf{x}_k^{\mathrm{T}}\tilde{\boldsymbol{\beta}}_{b-1}^{ai}) - \hat{m}_{b-1}(\mathbf{x}_k)| \|\Sigma^{-1}_{\hat{\boldsymbol{\beta}}_{b-1}^{AI},N}\mathbf{x}_k\|,$$

   for $k \in R_{-(b-1)}^{bat}$. According to $\tilde{\pi}_{k,b}^{sbs}$, select data points without replacement and label the outcomes, $\mathcal{B}_b = \{(y_{bi}, \mathbf{x}_{bi}) : i = 1, ..., n_b\}$

2. With the combined sub-data $\mathcal{B}_{1:b}$ of size N_b, update the imputation model $\hat{m}_b(\cdot)$

3. Obtain the estimate $\tilde{\boldsymbol{\beta}}_b^{ai}$ from the AI-AEE $Q^{ai}(\boldsymbol{\beta})$ with $\hat{m}_b(\cdot)$

5.2 Practical Algorithm

To specify the proposed sampling probability in (8) under the AL framework, quantities to replace $\boldsymbol{\beta}, m(\cdot)$ and Σ_N are required in practice. To deal with this, we propose a general practical algorithm based on batch-mode active selection. For $b = 1, 2, ...,$ we denote $\mathcal{B}_b = \{(y_{bi}, \mathbf{x}_{bi}) : i = 1, ..., n_b\}$ as a sub-data selected from $\mathcal{B}_{N/N_{b-1}}$ at b^{th} batch where $\mathcal{B}_{N/N_{b-1}}$ is the remaining data with the covariates excepting the data selected at 1,..., $b-1$ batch. Denote $\mathcal{B}_{1:b} = \{\mathcal{B}_1, ..., \mathcal{B}_b\}$ as a cumulative sub-data collected from 1^{st} to b^{th} batch. $R_{-b}^{bat} = \{i : \mathbf{x}_i \in \mathcal{B}_{N/N_b}, 1 \le i \le N\}$ be the index sets. $N_b = \sum_{i=1}^{b} n_i$ is the cumulative sub-data size.

The basic idea of the algorithm is that we construct the quantities using the sampled data acquired at previous steps and update the sampling probability to select additional data points. Let $\hat{m}_{b-1}(\cdot)$ and $\tilde{\boldsymbol{\beta}}_{b-1}^{ai}$ be the imputation model and the coefficient estimator constructed by $\mathcal{B}_{1:(b-1)}$. To select b^{th} sub-data, we replace $\boldsymbol{\beta}, m(\cdot)$ and Σ_N by $\tilde{\boldsymbol{\beta}}_{b-1}^{ai}, \hat{m}_{b-1}(\cdot)$ and $\Sigma_{\tilde{\boldsymbol{\beta}}_{b-1}^{ai},N}$ in (8) where $\Sigma_{\tilde{\boldsymbol{\beta}}_{b-1}^{ai},N} = \sum_{i=1}^{N} \dot{g}(\mathbf{x}_i^{\mathrm{T}}\tilde{\boldsymbol{\beta}}_{b-1}^{ai})\mathbf{x}_i\mathbf{x}_i^{\mathrm{T}}/N$. Then, we obtain the updated estimator $\tilde{\boldsymbol{\beta}}_b^{ai}$ based on the cumulative sampled data $\mathcal{B}_{1:b}$. The summary of the algorithm is presented in Algorithm 1.

6 Numerical Studies

In this section, we conduct numerical studies to assess the performance of the proposed estimator with synthetic data and four real data examples. The codes used for the numerical studies are available on a Github repository https://github.com/IJCAI-24/ActiveSemiPrediction.

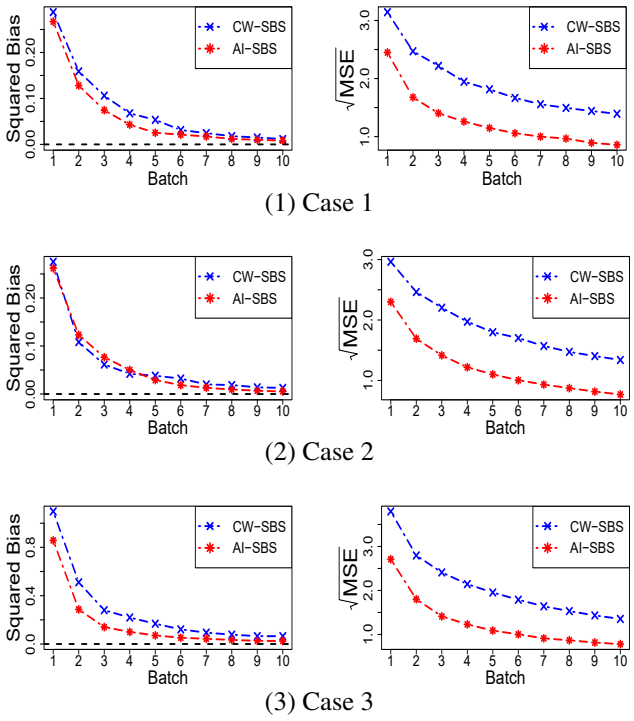


Figure 1: The sum of squared bias and the square root of MSEs over 10 batches for three different cases under the proposed self-learning based sampling probability. Each batch size is 100 and a pilot subsample of size is 150 for the initial step. CW and AI mean the corrective weighting esti- mator and the actively improved estimator, respectively.

6.1 Synthetic Data

We generate synthetic data to evaluate the performance of the proposed algorithm. We consider 7 dimensional covariates $\mathbf{x}_i = (x_{1,i}, ..., x_{7,i})$. The covariates $(x_{1,i}, ..., x_{5,i})$ are generated from a multivariate normal distribution $N(\mathbf{0}, \boldsymbol{\Sigma})$ and $x_{6,i}$ and $x_{7,i}$ are generated from a uniform distribution $Unif(0, 0.5)$ where $\boldsymbol{\Sigma}_{jk} = 2 * 0.4^{I(j \neq k)}$ for $j, k = 1, \ldots, 5$ and $I()$ is the indicator function. With $\beta_1 = ... = \beta_7 = 0.7$, we consider three different models to generate the outcome y_i;

Case 1. $y_i \sim Bern(\theta_i)$ with $\text{logit}(\theta_i) = -3 + \mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}$

Case 2. $y_i \sim Bern(\theta_i)$ with $\text{logit}(\theta_i) = -3 + \mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta} + h_1(\mathbf{x})$, where $h_1(\mathbf{x}) = 0.5 * \sin(0.5 * x_{1,i}) - 0.5 * \sin(0.5 * x_{2,i}) + 0.2 * \sin(0.2 * x_{7,i})$.

Case 3. $y_i \sim Bern(\theta_i)$ with $\text{logit}(\theta_i) = -5.4 + \mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta} + h_2(\mathbf{x})$, where $h_2(\mathbf{x}) = 0.5 * x_{5,i}^2 - 0.5 * x_{7,i}^2 + \exp(0.5 * x_{1,i} + 0.5 * x_{2,i})$.

For all cases, about 25% of outcomes is $y = 1$. We generate full training data of size $N = 10^5$ and consider 10 batches. In each batch, we select the subdata of size 100. For the initial values in the proposed algorithm, uniform samples of size 150 are used. Natural spline models with 2 degree of freedom is considered for the imputation model in each repe-

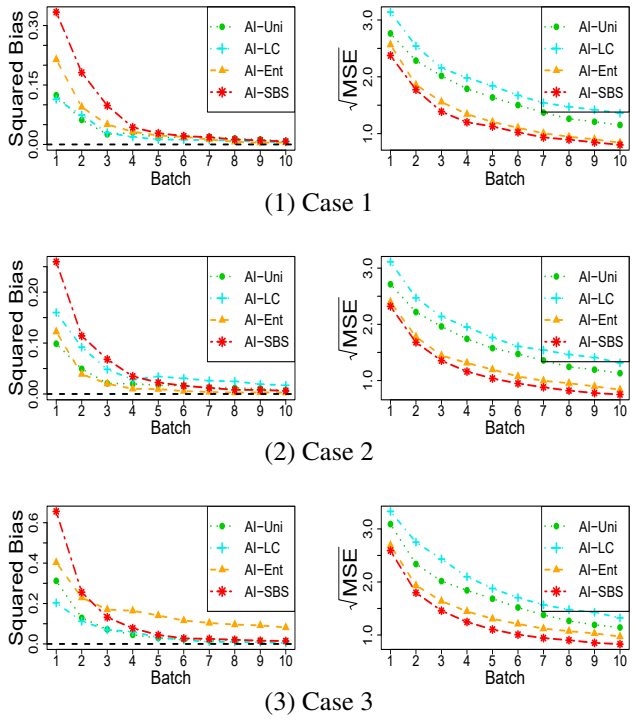


Figure 2: The sum of squared bias and the square root of MSEs over 10 batches for three different cases. Each batch size is 100 and a pilot subsample of size is 150 for the initial step. Uni, LC, Ent and SBS mean the uniform sampling probability, the least confidence sampling probability, the entropy sampling probability and the proposed self-learning based sampling probability, respectively.

tition, and the models estimate non-linear effects for all continuous covariates. The repetition is 300 times and calculate empirical MSEs and the sum of squared bias for coefficients based on $\sum_{s=1}^{S} \|\tilde{\boldsymbol{\beta}}_b^{(s)} - \hat{\boldsymbol{\beta}}_f\|^2/S$ and $\|\sum_{s=1}^{S} \tilde{\boldsymbol{\beta}}_b^{(s)}/S - \hat{\boldsymbol{\beta}}_f\|^2$, respectively where S is the number of replications, $\tilde{\boldsymbol{\beta}}_b^{(s)}$ is the estimate provided from the b^{th} batch at the s^{th} repetition, and $\hat{\boldsymbol{\beta}}_f$ is the full data estimate.

We compare the proposed AI estimator with the CW estimator. In addition, we investigate the efficiency of sampling schemes considering four different sampling probabilities; uniform sampling probability (Uni), least confidence (LC), entropy sampling probability (Ent), proposed self-learning based sampling probability, $\tilde{\pi}^{sbs}$ (SBS). We used sampling probabilities proportional to $1-g(\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta})$ for LC and $g(\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta})\log\{g(\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta})\}$ for Ent.

Comparison of estimators We first compare the proposed AI estimator with the CW estimator under the SBS scheme. The results are shown in Figure 1. As the cumulative batch size increases, the MSEs of all methods become smaller and the biases are reduced. The AI method leveraging unlabeled data outperforms the CW estimator using only labeled data for both of the sampling schemes in all cases.

Comparison of sampling schemes As shown in Figure 2, the SBS scheme is always preferred for the AI estimator for all cases in terms of the MSE. When combining different estimators and sampling schemes, it is clear that the AI estimator with the SBS performs best in terms of the MSE. As we expected, the statistical bias decreases for all sampling schemes as the cumulative labeled size increases.

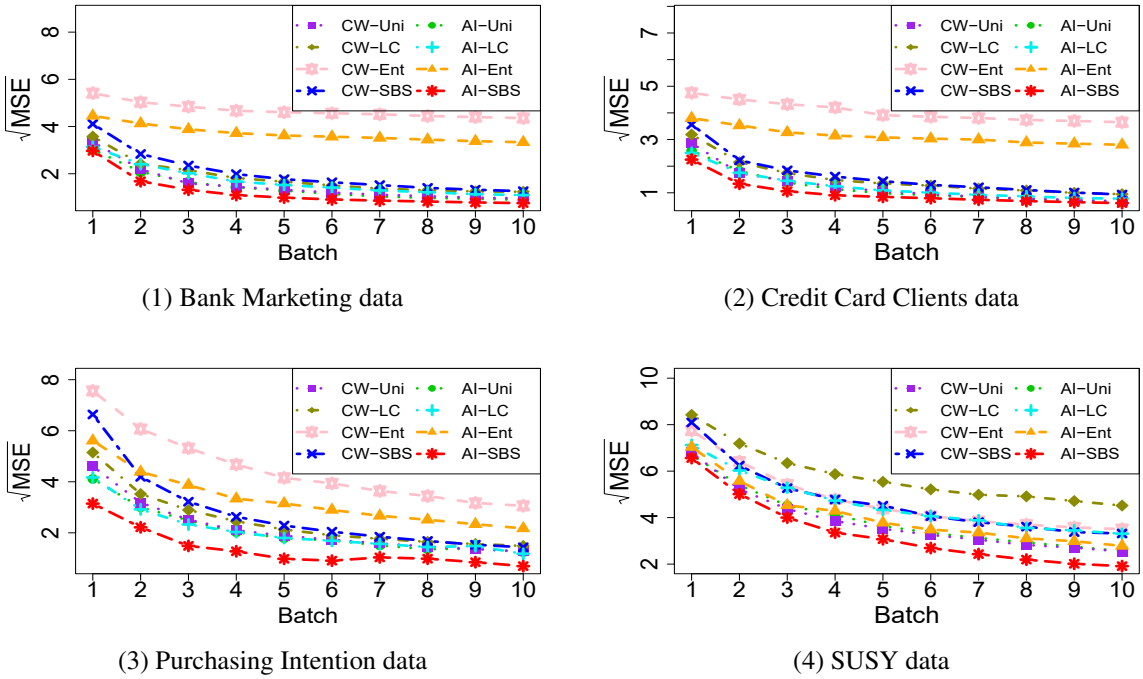
Smaller subsample size We conduct additional numerical studies using synthetic data from Case 1 , 2 and 3 with 120 pilot sample size and 80 subdata size. Figure S.1 and S.2 in Section 7 of Supplementary Material show the results. Overall, the results are similar to those in Figure 1 and 2. Across all cases, the proposed AI method produces better results for MSE than the CW method. Moreover, the AI method under the proposed SBS sampling scheme tends to have better performance than the others in terms of MSE. In general, the results also indicate that the larger the pilot sample size and subdata size are, the smaller the MSE tends to be over the batches.

Effect of imputation models To investigate the impact of imputation models, we use a simple natural spline model considering a non-linear effect of only x_4 and linear effects of the remaining covariates (OnlyX4). We compared the simple imputation model with the natural spline models considering non-linear effects of all continuous covariates (AllXs). Figure S.3 in Section 7 of Supplementary Material presents the results of the bias and MSE over batches for the CW and AI method under the SBS sampling. Regardless imputation models, the proposed AI methods shows better MSE performance than the CW method. Also we observe that OnlyX4 and AllXs imputation models with the AI method show similar performance for Case 1 and 2, while allXs yields lower MSEs than onlyX4 for Case 3.

6.2 Real Data Examples

We apply the proposed algorithm to four real datasets; 1) Bank Marketing data, 2) SUSY data, 3) Credit Card Clients data and 4) Purchasing Intention data. The datasets are available on the UCI Machine Learning repository:
1) https://archive.ics.uci.edu/ml/datasets/bank+marketing,
2) https://archive.ics.uci.edu/ml/datasets/SUSY,
3) https://archive.ics.uci.edu/ml/datasets/default+of+credit+card+clients, and
4) https://archive.ics.uci.edu/ml/datasets/Online+Shoppers+Purchasing+Intention+Dataset.

For Bank Marketing dataset, 14 covariates with client information is considered to predict whether the client will agree to a term deposit. The data size is 41,188 and about 11.27% of the responses is $y = 1$. SUSY dataset includes 18 features to classify a signal process which produces supersymmetric particles. We consider the last 200,000 examples in the dataset. The percent of the response $y = 1$ is about 45.93%. We use Credit Card Clients dataset with 30,000 customers to classify default payment (yes = 1, no = 0) using 23 predictors such as demographics, history of past payment, amount of bill statement and previous payment. The percent of the response $y = 1$ is 22.12%. Purchasing Intention dataset has 12,330 observations with about 15.47% of the response $y = 1$ ($y = 1$; ending with shopping, $y = 0$; not end with shopping). The dataset contains 17 covariates related to users information in e-commerce market. We build a natural spline model with 2 or 3 degree of freedom for the imputation mod-



(1) Bank Marketing data

(2) Credit Card Clients data

(3) Purchasing Intention data

(4) SUSY data

Figure 3: The square root of MSEs over 10 batches for four real data examples. Uni, LC, Ent and SBS mean the uniform sampling probability, the least confidence sampling probability, the entropy sampling probability and the proposed self-learning based sampling probability, respectively. CW and AI mean the corrective weighting estimator and the actively improved estimator, respectively.

els. For initial values and the subdata size, we consider 150 and100 for the first two examples, and 200 and 200 for the other examples, respectively. The total number of batches is 10 and the repetition is 300 times.

Figure 3 shows the performance of the CW and AI estimators with four different sampling schemes. In general, the results are similar to those in the experiments with the synthetic data. The AI method outperforms the CW method over the batches under the identical non-uniform sampling scheme. It is worth noting that the performance of the AI estimator with the SBS scheme achieves the lowest MSE for all real data examples. As shown in Figure S.4 in Section 7 of Supplementary Material, the biases tend to decrease as the labeled data is larger.

7 Conclusion and Limitations

In the AL setting, we proposed the AI-AEE to estimate the unknown parameters in the target prediction models using labelled and unlabelled data based on an imputation model. We found that even when the imputation models are misspecified, AI-AEE is unbiased and robust. We derived the asymptotic results for the CW estimator and the proposed AI estimator and showed that the AI estimator has a higher efficiency gain than the CW estimator. Furthermore, by minimizing the asymptotic mean squared errors of the AI estimator, we derived the optimal sampling probability for each sampling step. However, since the sampling probability depends on the unlabelled outcomes and the full data-based estimator, we proposed the surrogate SBS probability that is actively updated with sampled data. We demonstrated that our methods provide better performance than other methods in the numerical studies.

There are some interesting topics that need to be further investigated. In this paper, we found that the AI estimator is more efficient than the CW estimator when the imputation model is correctly specified. Under misspecification of the imputation model, however, this is not guaranteed. In a recent paper, Deng *et al.* developed a 'safe' estimator in the linear regression problem under the SSL setting. They showed that it is no worse than the supervised estimators even when the imputation model is not correctly specified. Using this idea, we can build more robust prediction models in the AL setting even when the imputation model is misspecified. Also, it is challenging to train prediction models on rare event data in practice. The scarcity of rare event data can lead to poor performance. Also, the low prevalence of the rare event cases may require tedious annotation work to collect data points in the minor class under the AL setting. Therefore, it would be important to collect the rare events for labelling. One possible solution is to use surrogate variables that are highly correlated with the rare cases [Liu *et al.*, 2022]. The other solution is to label data points with high risk prediction based on the trained models to enrich the rare cases [Tan and Heagerty, 2020].

8 Supplementary Material

Supplementary Material includes all proofs of propositions and theorems in the main manuscript, additional numerical experiments and codes used for numerical studies.

References

[Ai *et al.*, 2018] Mingyao Ai, Jun Yu, Huiming Zhang, and HaiYing Wang. Optimal subsampling algorithms for big data regressions. *arXiv preprint arXiv:1806.06761*, 2018.

[Cai *et al.*, 2022] Tianxi Cai, Mengyan Li, and Molei Liu. Semi-supervised triply robust inductive transfer learning. *arXiv preprint arXiv:2209.04977*, 2022.

[Cao *et al.*, 2009] Weihua Cao, Anastasios A Tsiatis, and Marie Davidian. Improving efficiency and robustness of the doubly robust estimator for a population mean with incomplete data. *Biometrika*, 96(3):723–734, 2009.

[Carpenter *et al.*, 2006] James R Carpenter, Michael G Kenward, and Stijn Vansteelandt. A comparison of multiple imputation and doubly robust estimation for analyses with missing data. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 169(3):571–584, 2006.

[Chakrabortty and Cai, 2018] Abhishek Chakrabortty and Tianxi Cai. Efficient and adaptive linear regression in semi-supervised settings. *The Annals of Statistics*, 46(4):1541–1572, 2018.

[Chakrabortty *et al.*, 2019] Abhishek Chakrabortty, Jiarui Lu, T Tony Cai, and Hongzhe Li. High dimensional m-estimation with missing outcomes: A semi-parametric framework. *arXiv preprint arXiv:1911.11345*, 2019.

[Deng *et al.*, 2020] Siyi Deng, Yang Ning, Jiwei Zhao, and Heping Zhang. Optimal semi-supervised estimation and inference for high-dimensional linear regression. *arXiv preprint arXiv:2011.14185*, 2020.

[Farquhar *et al.*, 2021] Sebastian Farquhar, Yarin Gal, and Tom Rainforth. On statistical bias in active learning: How and when to fix it. *arXiv preprint arXiv:2101.11665*, 2021.

[Gal *et al.*, 2017] Yarin Gal, Riashat Islam, and Zoubin Ghahramani. Deep bayesian active learning with image data. In *International Conference on Machine Learning*, pages 1183–1192. PMLR, 2017.

[Ganti and Gray, 2012] Ravi Ganti and Alexander Gray. Upal: Unbiased pool based active learning. In *Artificial Intelligence and Statistics*, pages 422–431. PMLR, 2012.

[Gronsbell and Cai, 2018] Jessica L Gronsbell and Tianxi Cai. Semi-supervised approaches to efficient evaluation of model prediction performance. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 80(3):579–594, 2018.

[Gronsbell *et al.*, 2022] Jessica Gronsbell, Molei Liu, Lu Tian, and Tianxi Cai. Efficient evaluation of prediction rules in semi-supervised settings under stratified sampling. *Journal of the Royal Statistical Society. Series B, Statistical Methodology*, 84(4):1353–1391, 2022.

[Horvitz and Thompson, 1952] Daniel G Horvitz and Donovan J Thompson. A generalization of sampling without replacement from a finite universe. *Journal of the American statistical Association*, 47(260):663–685, 1952.

[Hou *et al.*, 2021] Jue Hou, Zijian Guo, and Tianxi Cai. Surrogate assisted semi-supervised inference for high dimensional risk prediction. *arXiv preprint arXiv:2105.01264*, 2021.

[Imberg *et al.*, 2020] Henrik Imberg, Johan Jonasson, and Marina Axelson-Fisk. Optimal sampling in unbiased active learning. In *International Conference on Artificial Intelligence and Statistics*, pages 559–569. PMLR, 2020.

[Imberg *et al.*, 2022] Henrik Imberg, Xiaomi Yang, Carol Flannagan, and Jonas Bärgman. Active sampling: A machine-learning-assisted framework for finite population inference with optimal subsamples. *arXiv preprint arXiv:2212.10024*, 2022.

[Kiefer, 1959] Jack Kiefer. Optimum experimental designs. *Journal of the Royal Statistical Society: Series B (Methodological)*, 21(2):272–304, 1959.

[Kossen *et al.*, 2021] Jannik Kossen, Sebastian Farquhar, Yarin Gal, and Tom Rainforth. Active testing: Sample-efficient model evaluation. In *International Conference on Machine Learning*, pages 5753–5763. PMLR, 2021.

[Kossen *et al.*, 2022] Jannik Kossen, Sebastian Farquhar, Yarin Gal, and Tom Rainforth. Active surrogate estimators: An active learning approach to label-efficient model evaluation. *arXiv preprint arXiv:2202.06881*, 2022.

[Krijthe and Loog, 2017] Jesse H Krijthe and Marco Loog. Projected estimators for robust semi-supervised classification. *Machine Learning*, 106(7):993–1008, 2017.

[Lee *et al.*, 2021] JooChul Lee, Elizabeth D Schifano, and HaiYing Wang. Fast optimal subsampling probability approximation for generalized linear models. *Econometrics and Statistics*, 2021.

[Lee *et al.*, 2022] JooChul Lee, Elizabeth D Schifano, and HaiYing Wang. Sampling-based gaussian mixture regression for big data. *Journal of Data Science*, pages 1–15, 2022.

[Liu *et al.*, 2022] Xiaokang Liu, Jessica Chubak, Rebecca A Hubbard, and Yong Chen. Sat: a surrogate-assisted two-wave case boosting sampling method, with application to ehr-based association studies. *Journal of the American Medical Informatics Association*, 29(5):918–927, 2022.

[McCullagh, 1983] Peter McCullagh. Quasi-likelihood functions. *The Annals of Statistics*, 11(1):59–67, 1983.

[Robins *et al.*, 1994] James M Robins, Andrea Rotnitzky, and Lue Ping Zhao. Estimation of regression coefficients when some regressors are not always observed. *Journal of the American statistical Association*, 89(427):846–866, 1994.

[Settles *et al.*, 2008] Burr Settles, Mark Craven, and Lewis Friedland. Active learning with real annotation costs. In *Proceedings of the NIPS workshop on cost-sensitive learning*, volume 1. Vancouver, CA:, 2008.

[Tan and Heagerty, 2020] W Katherine Tan and Patrick J Heagerty. Predictive case control designs for modification learning. *arXiv preprint arXiv:2011.14529*, 2020.

[Wang and Kim, 2022] HaiYing Wang and Jae Kwang Kim. Maximum sampled conditional likelihood for informative subsampling. *Journal of Machine Learning Research*, 23:1–50, 2022.

[Wang *et al.*, 2018] HaiYing Wang, Rong Zhu, and Ping Ma. Optimal subsampling for large sample logistic regression. *Journal of the American Statistical Association*, 113(522):829–844, 2018.

[Wang *et al.*, 2021] HaiYing Wang, Aonan Zhang, and Chong Wang. Nonuniform negative sampling and log odds correction with rare events data. *Advances in Neural Information Processing Systems*, 34:19847–19859, 2021.

[Wang *et al.*, 2022a] Jing Wang, HaiYing Wang, and Shifeng Xiong. Unweighted estimation based on optimal sample under measurement constraints. *Canadian Journal of Statistics*, 2022.

[Wang *et al.*, 2022b] Linshanshan Wang, Xuan Wang, Katherine P Liao, and Tianxi Cai. Semi-supervised transfer learning for evaluation of model classification performance. *arXiv preprint arXiv:2208.07927*, 2022.

[Yao and Wang, 2019] Yaqiong Yao and HaiYing Wang. Optimal subsampling for softmax regression. *Statistical Papers*, 60(2):585–599, 2019.

[Yilmaz *et al.*, 2021] Emine Yilmaz, Peter Hayes, Raza Habib, Jordan Burgess, and David Barber. Sample efficient model evaluation. *arXiv preprint arXiv:2109.12043*, 2021.

[Yoo and Kweon, 2019] Donggeun Yoo and In So Kweon. Learning loss for active learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 93–102, 2019.

[Zhang *et al.*, 2021] Tao Zhang, Yang Ning, and David Ruppert. Optimal sampling for generalized linear models under measurement constraints. *Journal of Computational and Graphical Statistics*, 30(1):106–114, 2021.

[Zhou *et al.*, 2022] Doudou Zhou, Molei Liu, Mengyan Li, and Tianxi Cai. Doubly robust augmented model accuracy transfer inference with high dimensional features. *arXiv preprint arXiv:2208.05134*, 2022.

[Zhu, 2005] Xiaojin Zhu. *Semi-supervised learning with graphs*. Carnegie Mellon University, 2005.