SCALABLE BAYESIAN ACTIVE LEARNING WITH BATCH ACQUISITION UNDER DISTRIBUTION SHIFT

Anonymous authors
Paper under double-blind review

ABSTRACT

The performance of machine learning models may suffer from significant decline when evaluated on the data exhibiting distribution shift. Although extensive research on algorithm design have been proposed, the acquisition of new data points to enlarge training datasets has also been verified as a promising solution path. Starting from this idea, we built our research upon bayesian active learning and propose a method that can efficiently acquire samples from a candidate pool of diverse data sources for improving performance on the shifted target population. Specifically, our method designs a novel acquisition function characterizing a Lower Bound of Batch Information Gain (LB-BatchIG) for target distribution and formulates batch acquisition as a submodular optimization problem. By resolving it with a greedy algorithm, we can determine the data batch from the candidate pool for annotation and training. Empirical studies on synthetic datasets and real-world datasets, including tabular data and image data, demonstrate that our batch acquisition algorithm can contribute to greater performance improvement than other algorithms.

1 INTRODUCTION

With the remarkable advancement, machine learning has been widely applied in many scenarios and achieved promising performance. Owing to the prominent capability, the models can minimize the predictive loss and achieve impressive performance on the data population of the same distribution to training data. Unfortunately, when applied in the wild applications, machine learning models can encounter the test data from a shifted distribution, which violates independent and identically distributed (i.i.d) assumption and induces notably performance deterioration.

To improve the generalization ability towards out-of-distribution (OOD) data, a bunch of researches have been proposed and offer promising solutions by more provable and carefully designed learning models, such as invariant learning (Arjovsky et al., 2019) and distributionally robust optimization (DRO) (Mohajerin Esfahani & Kuhn, 2018; Duchi & Namkoong, 2021), while maintaining the original training data. In contrast, from a data-centered perspective, recent researches (Liu et al., 2024; Fu et al., 2021) uncovered the great significance of collecting more samples for generalization. Motivated by this, we explore methods to augment the training data with additional samples, thereby improving generalization to *the shifted target distribution*. Given the labor costs and ethical concerns associated with data annotation, it is important to investigate how to achieve optimal generalization performance under a limited acquisition budget.

To this end, active learning offers a promising approach by querying samples that are beneficial to model performance from a candidate pool. More specifically, the candidate pool is expected to comprise diverse samples whose contribution to model performance varies. Within the acquisition budget, conventional active learning algorithms attempt to select the most informative samples based on acquisition criteria from distinct categories, such as uncertainty (Ducoffe & Precioso, 2018; Joshi et al., 2009; Ržička et al., 2020), model influence (Fukumizu, 2000; Zhang & Oles, 2000; Ash et al., 2020; Gal et al., 2017), and representativeness (Liu et al., 2016; Sener & Savarese, 2017b; Qin et al., 2021; Chattopadhyay et al., 2013; Yu et al., 2006; Yang et al., 2017). However, they are primarily designed without accounting for the distributional shift from training to target data. To bridge this gap, recent efforts in Active Domain Adaptation (ADA) (Rangwani et al., 2021; Prabhu et al., 2021) acquire samples *directly from the target domain*. This approach, nevertheless, imposes

another restrictive requirements on the candidate pool, serving as a condition that may not hold in many practical scenarios.

To overcome the above issues, we aim to *propose a scalable and provable* criteria towards selecting a batch of samples for generalization to shifted target distribution. Fortunately, we observe that the development of Bayesian active learning (Sun et al., 2015; Haut et al., 2018; Kirsch et al., 2019; Gal et al., 2017; Houlsby et al., 2011) offers new opportunities to guide our expected sample acquisition. As a representative, a acquisition criteria, termed EPIG (Smith et al., 2023), is proposed to measure the predictive information gain on target data distribution brought by a single, unlabeled sample. However, such criteria suffers from the severe dilemma on the **scalability issue**, as it only *supports selecting sample with the highest criteria*. Specifically, directly extending it by selecting the top K ones with the highest EPIG (notated EPIG) or stochastically selecting samples with EPIG-based probability (notated PowerEPIG) will tend to select similar samples and suffer from redundant information problem. As shown in Figure 1, these straightforward extensions can induce prohibitively *high mutual predictive information gain* among the samples in selected batch, especially directly selecting top K ones.

Therefore, in this paper, we propose to facilitate a more scalable Bayesian active learning criteria at batch-level. Specifically, our approach theoretically characterizes a Lower Bound of Batch Information Gain(LB-BatchIG) for target population, thus avoiding the trivial sum of individual criteria on samples (Smith et al., 2023). As the number of potential batches is **combinatorial** relative to the pool size and batch size, identifying the optimal batch for our LB-BatchIG exhibits high time complexity if one enumerates the whole the candidate batches directly. To further improve the algorithm efficiency, we theoretically prove the submodular property of this combinational optimization problem. Consequently, an efficient greedy solution can be derived with the time complexity polynomial to the size of data pool and batch. Finally, the acquired batch by our LB-BatchIG can be guaranteed at least $(1 - \frac{1}{e})$-optimal (e is the natural constant) to the optimal (oracle) batch.

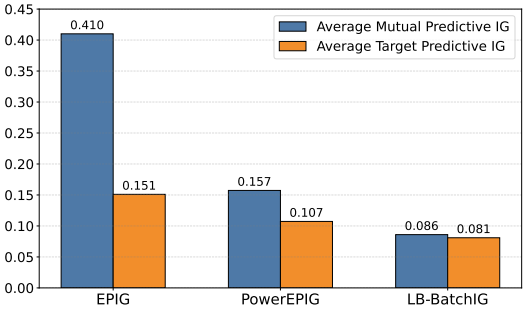


Figure 1: The results are obtained on the synthetic experiments (see details in section 4) with $\alpha = 1200$ and acquisition size $c = 50$. Mutual predictive IG refers to the predictive information gain provided to other samples within the batch, while target predictive IG denotes the individual information gain contributed to the target data distribution.

We conduct extensive experiments on synthetic datasets and real-world datasets, including tabular data and image data. In each experiment, the test dataset adheres to the target distribution distinct from both the training dataset and candidate pool. We compare our algorithm with baselines by iteratively acquiring new samples of given budget for annotation, adding them to the training data, and retraining the model. The experimental results show that with the same acquisition budget, the model trained on the samples selected by our algorithm achieves superior performance than other methods. The main contributions of this paper can be summarized as following:

- To the best of our knowledge, we are the first to investigate tractable acquisition function at batch-level for enhancing the model performance under distribution shift.
- We propose a novel acquisition function denoted as LB-BatchIG which supports acquiring data batch efficiently. We prove this function satisfies the submodular property and propose a greedy algorithm to maximize it.
- We conduct extensive experiments on synthetic datasets and real-world datasets to verify the effectiveness of our proposed algorithm.

2 PRELIMINARIES

In this section, we introduce our problem formulation and bayesian active learning which is the cornerstone of our method.

2.1 PROBLEM FORMULATION

We denote $\mathbf{X} \in \mathbb{R}^d$ as covariates and $\mathbf{y} \in \mathcal{Y}$ as label. This paper focuses on classification problem, thereby $\mathcal{Y} = [K]$. The initial training dataset is given as $\mathcal{D}_{tr} = \{(\mathbf{x}_i^{tr}, y_i^{tr})\}_{1 \le i \le n_{tr}}$. To enhance the model performance on the shifted target distribution $p^t(\mathbf{X}, \mathbf{y})$, we aim to select samples from a large unlabeled pool $\mathcal{D}_{po} = \{\mathbf{x}_i^{po}\}_{1 \le i \le n_{po}}$, annotate them and supplement them into training dataset. The acquisition process consists of t rounds, and in each round c samples are acquired. Therefore, the total acquisition budget is $g = t \cdot c$. The closed-form probability density of test distribution $p^t(\mathbf{X})$ is usually difficult to obtain. Hence, we denote the target distribution as the collection of test samples $\mathcal{D}_{te} = \{\mathbf{x}_i^{te}\}_{1 \le i \le n_{te}}$.

2.2 BAYESIAN ACTIVE LEARNING

Bayesian active learning borrows the idea of bayesian experimental design (Lindley, 1956; Chaloner & Verdinelli, 1995) which quantifies the information gain of interest variables from experiments.

To adapt bayesian experimental design to active learning, the designed experiment is defined as covariate $\mathbf{X}$ and the outcome of experiment is defined as the label $\mathbf{y}$. Assuming the predictive model $f_\theta(\mathbf{y}|\mathbf{X})$ with parameter θ characterizes the dependency between $\mathbf{X}$ and $\mathbf{y}$, BALD (Houlsby et al., 2011; Gal et al., 2017; Kirsch et al., 2019) takes the model parameter θ as the interest variables and search the sample $\mathbf{x}$ that can reduce the entropy of θ to the maximal extent. Formally, it can be formulated as:

$$\arg\max_{\mathbf{x}} H[\theta|\mathcal{D}] - \mathbb{E}_{p(\mathbf{y}|\mathbf{x};\mathcal{D})}[H[\theta|\mathcal{D}, \mathbf{x}, y]], \tag{1}$$

where $H(\cdot)$ is the shannon entropy, $\mathcal{D}$ is the dataset has been annotated and $p(\mathbf{y}|\mathbf{x};\mathcal{D}) = \mathbb{E}_{p(\theta|\mathcal{D})}[p(\mathbf{y}|\mathbf{x},\theta)]$ is the posterior predictive distribution marginalized over model parameter θ. However, (Smith et al., 2023) pointed out that predictive uncertainty is mismatched with parameter uncertainty, and therefore propose an acquisition function of prediction-oriented manner. The acquisition function expected predictive information gain (EPIG) is designed for single data point.

$$\text{EPIG}(\mathbf{x}) = \mathbb{E}_{p^t(\mathbf{x}^*), p(\mathbf{y},\mathbf{y}^*|\mathbf{x},\mathbf{x}^*;\mathcal{D})}\left[\log \frac{p(\mathbf{y}, \mathbf{y}^*|\mathbf{x}, \mathbf{x}^*; \mathcal{D})}{p(\mathbf{y}|\mathbf{x};\mathcal{D})p(\mathbf{y}^*|\mathbf{x}^*;\mathcal{D})}\right], \tag{2}$$

where $p^t(\mathbf{x}^*)$ represents the covariate distribution of target population.

Repeatedly acquiring only one sample with highest EPIG score would result in excessively many acquisition rounds. Though we can directly adapt this method to select samples of top-K EPIG value and consequently reduce the acquisition rounds, it would suffer from redundant information problem and lead to sub-optimal performance.

Summarily, it is essential to propose a new acquisition function for batch samples that considers simultaneously the reduction of predictive uncertainty on target population as well as the diversity of batch.

3 PROPOSED METHOD

In this section, we firstly propose our acquisition function LB-BatchIG. Then we conduct theoretical analysis to reveal that it satisfies the sub-modular property. Based on this theoretical finding, we can apply greedy algorithm to resolve the problem of searching batch with maximum LB-BatchIG.

3.1 ACQUISITION FUNCTION: LB-BATCHIG

By extending the idea of bayesian active learning, we characterize the information gain brought by batch samples $\{\mathbf{x}_i^b\}_{1 \le i \le c}$ to the prediction on the target population $p^t(\mathbf{x}^*)$ as follows:

$$\mathbb{E}_{p^t(\mathbf{x}^*)}\left[\text{BIG}(\{\mathbf{x}_i^b\}, \mathbf{x}^*)\right], \tag{3}$$

where the function BIG$(\cdot)$ is defined as:

$$\text{BIG}(\{\mathbf{x}_i^b\}, \mathbf{x}^*) = \mathbb{E}_{p(\mathbf{y}_1^b,\ldots,\mathbf{y}_K^b,\mathbf{y}^*|\mathbf{x}_1^b,\ldots,\mathbf{x}_K^b,\mathbf{x}^*;\mathcal{D})}\left[\log \frac{p(\mathbf{y}_1^b,\ldots,\mathbf{y}_K^b,\mathbf{y}^*|\mathbf{x}_1^b,\ldots,\mathbf{x}_K^b,\mathbf{x}^*;\mathcal{D})}{p(\mathbf{y}_1^b,\ldots,\mathbf{y}_K^b|\mathbf{x}_1^b,\ldots,\mathbf{x}_K^b;\mathcal{D})p(\mathbf{y}^*|\mathbf{x}^*;\mathcal{D})}\right]. \tag{4}$$

However, the numeric value of Equation 3 is difficult to exactly calculate since it involves the joint probability density estimation of multiple variables. Therefore, we propose an alternative measurement as the substitute, which bypasses the calculation of multi-dimensional probability and can be significantly easier to compute. Specifically, the formula of the substitute measurement is

$$\text{LB-BatchIG}(\{\mathbf{x}_i^b\}) = \mathbb{E}_{p^t(\mathbf{x}^*)}\left[\max_{1\le i\le c} \text{IG}(\mathbf{x}_i^b, \mathbf{x}^*)\right], \tag{5}$$

where the function $\text{IG}(\cdot)$ is

$$\text{IG}(\mathbf{x}, \mathbf{x}^*) = \mathbb{E}_{p(\mathbf{y},\mathbf{y}^*|\mathbf{x},\mathbf{x}^*;\mathcal{D})}\left[\log \frac{p(\mathbf{y}, \mathbf{y}^*|\mathbf{x}, \mathbf{x}^*; \mathcal{D})}{p(\mathbf{y}|\mathbf{x}; \mathcal{D})p(\mathbf{y}^*|\mathbf{x}^*; \mathcal{D})}\right]. \tag{6}$$

It can be proved that the measurement $\text{LB-BatchIG}(\{\mathbf{x}_i^b\})$ is a lower bound of Equation 3. Hence, we can achieve the high predictive information gain brought by batch samples $\{\mathbf{x}_i^b\}$ through maximizing the objective of LB-BatchIG. Formally, it can be theoretically revealed with the following theorem.

Theorem 3.1. *According to the information theory, we have*

$$\text{LB-BatchIG}(\{\mathbf{x}_i^b\}) \le \mathbb{E}_{p^t(\mathbf{x}^*)}\left[\text{BIG}(\{\mathbf{x}_i^b\}, \mathbf{x}^*)\right]. \tag{7}$$

The detailed proof can be found in the section of appendix B.

3.2 Sub-modular property of LB-BatchIG

The optimization of LB-BatchIG is a combinatorial search problem with the number of potential solutions $C_{n_{po}}^c$. The brute force method that directly enumerates the candidate batches and selects the one of the highest score is of exponential time complexity and computationally expensive. Therefore, it is in urgent need to design an efficient algorithm which produces a near-optimal solution with less computational cost.

The submodular property of the acquisition function brings the opportunity to solve the batch optimization problem with polynomial complexity. The definition of submodular function is as follows:

Definition 3.1. *Given a set $V = \{v_1, v_2, ..., v_m\}$, a function $f : 2^V \to \mathbb{R}$ taking subset of V as input is a submodular function if the inequality holds:*

$$f(A) + f(B) \ge f(A \cap B) + f(A \cup B), \forall A, B \subset V \tag{8}$$

The literature (Nemhauser et al., 1978) unveils that the optimization of normalized monotone non-decreasing submodular function with cardinality constraint, formally $\max_{S\subset V,|S|=c} f(S)$, can be resolved by greedy algorithm. The resulting solution approximates the optimal one with a factor at least $1 - 1/e \approx 0.632$. In this way, the time cost is significantly reduced.

Fortunately, through theoretical analysis, we can prove the proposed acquisition function LB-BatchIG satisfies the submodularity property.

Proposition 3.1. *Regarding the unlabeled pool as the element set V in Definition 3.1 and the sample batch as the subset, the acquisition function LB-BatchIG is a normalized monotone non-decreasing submodular function.*

The detailed proof can be found in the appendix B. Based on this promising property of our criterion, we design an efficient greedy algorithm to pursue a near-optimal sample batch.

3.3 Implementation

We successively introduce the details of our algorithms, including the acquisition function estimation and the batch construction process.

LB-BatchIG Estimation We firstly repeatedly draw a series of model parameters $\{\theta_l\}_{1\le l\le m}$ from posterior distribution and calculate two matrices $\mathbf{O} \in \mathbb{R}^{K\times K}$ and $\mathbf{Q} \in \mathbb{R}^{K\times K}$:

$$\begin{aligned} o_{i,j} &= \frac{1}{m}\sum_{l=1}^{m} p(\mathbf{y} = i|\mathbf{x}, \theta_l) \cdot p(\mathbf{y} = j|\mathbf{x}^*, \theta_l), \\ q_{i,j} &= \frac{1}{m}\sum_{l=1}^{m} p(\mathbf{y} = i|\mathbf{x}, \theta_l) \cdot \frac{1}{m}\sum_{l=1}^{m} p(\mathbf{y} = j|\mathbf{x}^*, \theta_l), \end{aligned}$$

where $\theta_l \sim p(\theta|\mathcal{D}), 1 \leq l \leq m$. The values of $o_{i,j}$ and $q_{i,j}$ are the empirical approximations of $p(\mathbf{y} = i, \mathbf{y}^* = j|\mathbf{x}, \mathbf{x}^*; \mathcal{D})$ and $p(\mathbf{y} = i|\mathbf{x}; \mathcal{D}) \cdot p(\mathbf{y}^* = j|\mathbf{x}^*; \mathcal{D})$ respectively.

Based on the calculated matrices $\mathbf{O}$ and $\mathbf{Q}$, we can empirically estimate the function $\text{IG}(\mathbf{x}, \mathbf{x}^*)$ as follows:

$$\hat{\text{IG}}(\mathbf{x}, \mathbf{x}^*) = \sum_{i=1}^{K} \sum_{j=1}^{K} o_{i,j} \cdot \log \frac{o_{i,j}}{q_{i,j}}. \tag{9}$$

By sampling $\mathbf{x}^*$ from the target distribution $p^t(\mathbf{x}^*)$, we can estimate the acquisition function for $\{\mathbf{x}_i^b\}_{1 \leq i \leq c}$ through the equation:

$$\text{LB-BatchIG}(\{\mathbf{x}_i^b\}) \approx \frac{1}{s} \sum_{j=1}^{s} \max_{1 \leq i \leq c} \hat{\text{IG}}(\mathbf{x}_i^b, \mathbf{x}_j^*) \tag{10}$$

Batch Construction Owing to the promising property of normalized monotone non-decreasing submodular function, we propose a greedy algorithm to construct the acquisition sample batch. The algorithm consists of three steps.

- Firstly, the acquisition batch is initialized as an empty set. Formally, $\mathcal{B}_0 = \varnothing$.
- We iteratively add c samples into the acquisition batch. In the i^{th} iteration, we search for the sample $\mathbf{x}_{b_i}^{po}$ with the largest improvement of LB-BatchIG. Formally, this means

  $$b_i = \arg \max_{1 \leq l \leq n_{po}} \text{LB-BatchIG}(\mathcal{B}_{i-1} \cup \mathbf{x}_l^{po}). \tag{11}$$

  Then the selected sample is incorporated in to the acquisition batch $\mathcal{B}_i = \mathcal{B}_{i-1} \cup \mathbf{x}_{b_i}^{po}$.
- Finally, the resulting batch $\mathcal{B}_c = \{\mathbf{x}_i^b\}_{1 \leq i \leq c}$ is the obtained batch for annotation.

The pseudo-code of the greedy algorithm can be found in the appendix D. After the batch $\mathcal{B}_c$ is obtained, the samples are removed from the unlabeled pool. The new pool is updated as $\mathcal{D}_{po} \leftarrow \mathcal{D}_{po} \setminus \mathcal{B}_c$. After annotating the oracle label $\{y_i^b\}_{1 \leq i \leq c}$ for $\mathcal{B}_c$, the samples are added into the training dataset, which means $\mathcal{D}_{tr} \leftarrow \mathcal{D}_{tr} \cup \{\mathbf{x}_i^b, y_i^b\}_{1 \leq i \leq c}$.

3.4 Time Complexity

The time complexity of calculating function $\text{IG}(\mathbf{x}, \mathbf{x}^*)$ is $\mathcal{O}(mK^2)$. Computing the acquisition function for the batch of size c requires $\mathcal{O}(cs)$ times of calculating IG function. Running our proposed algorithm consumes $\mathcal{O}(cn_{po})$ times of calculating LB-BatchIG. In contrast, the direct enumeration method needs to sweep all the $\mathcal{C}_{n_{po}}^c$ potential candidates. Since $c \ll n_{po}$, the times of LB-BatchIG calculation for enumeration method approximately equals $\mathcal{O}(n_{po}^c)$.

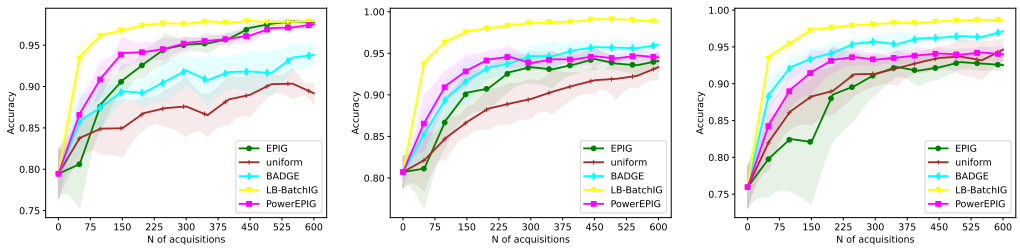
4 Experiment

We evaluate our proposed batch acquisition algorithm on diverse datasets, including synthetic data, tabular data and image data.

4.1 Experimental Setup

Baselines To demonstrate the effectiveness of our proposed method, we implement the following baselines for comparision:

- *Uniform*: This method randomly selects c samples from the candidate pool in each round without preference.
- *EPIG* (Smith et al., 2023): The original version of the method repeatedly select the sample with the highest expected predictive information gain (EPIG) score. To accommodate the batch acquisition setting, we simply adapt this method to rank the samples by EPIG score and choose the top-K samples in one round.



(a) $\alpha = 600$ (b) $\alpha = 1200$ (c) $\alpha = 1800$

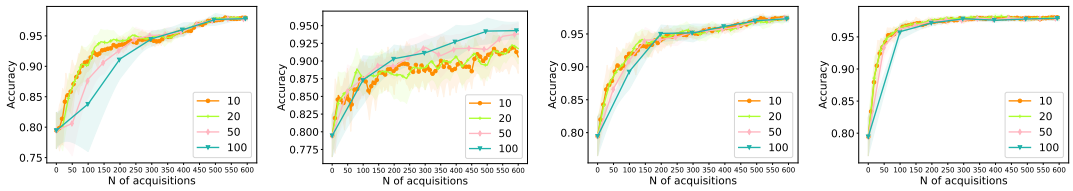
Figure 2: The results on the synthetic datasets under the settings with varied sample size α from the same subpopulation as the target population. The curves present the mean value of regret in 10 repeated experiments. The shaded region presents the interval $[\text{mean} - \text{std}, \text{mean} + \text{std}]$ of the regret.

- *PowerEPIG* (Kirsch, 2021): It stochastically selects the samples with the normalized probabilities proportional to the EPIG score to the γ^{th} power. Specifically, we set $\gamma = 5$ as the previous literature (Kirsch, 2021) suggested.
- *BADGE* (Ash et al., 2020): It computes the gradient-based embedding for each sample and runs k-means++ algorithm (Arthur & Vassilvitskii, 2007) to construct a batch of samples with diverse and representative embeddings.

Evaluation metric The data acquisition process consists of t rounds. After each round, we retrain the predictive model on the enlarged training dataset including the original training samples and the samples selected by algorithms from the candidate pool. The accuracy of the retrained model on the target population for each round is recorded.

We repeat the above process several times and calculate the mean value and standard deviation of the accuracy across the repeated experiments.

Model Setup To enable uncertainty estimation, we adopt the MC Dropout (Gal & Ghahramani, 2016) technology for the predictive models. Specifically, the dropout layers are kept activated during the inference. Therefore, the random activation of dropout unit can be viewed as sampling from the posterior parameter distribution, and the prediction result can be varied across multiple inferences.



(a) EPIG (b) BADGE (c) PowerEPIG (d) LB-BatchIG

Figure 3: The results on the synthetic datasets with varied batch size of acquisition. The experiments are conducted under the setting where $\alpha = 600$.

4.2 Synthetic Data

Experimental setup We generate the synthetic data of binary classification task. The samples come from three distinct data subpopulations. Specifically, the covariates $\mathbf{X} \in \mathbb{R}^d$ consists of two parts, the first $d_1 = d - 3$ elements of $\mathbf{X}$ are independently drawn from standard gaussian distribution:

$$x_{,1}, x_{,2}, ..., x_{,d_1} \overset{iid}{\sim} \mathcal{N}(0, 1).$$

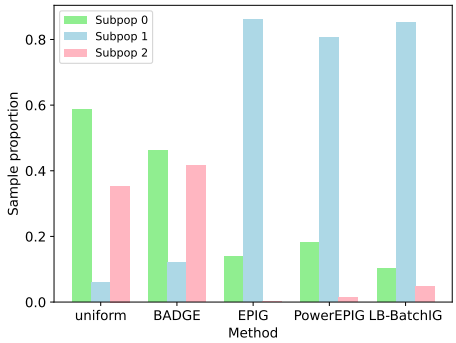


Figure 4: The average sample proportion from the subpopulations in an acquired batch. The results are calculated under the setting where $\alpha = 600, c = 50$.

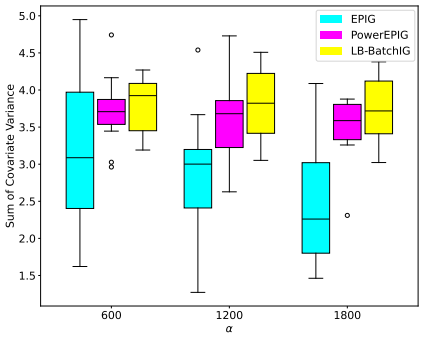


Figure 5: The value distribution of batch diversity measurements across the acquisition process of EPIG, PowerEPIG and LB-BatchIG.

Each sample belongs to a subpopulation $s_i \in \{0, 1, 2\}$. Based on the covariates and subpopulation index, the ground truth labels $y_i \in \{0, 1\}$ are generated by the functions: $y_i = \mathbb{I}\left(\sum_{j=1}^{d_1} x_{i,j}\beta_{s_i,j} > 0\right)$, where $\mathbb{I}(\cdot)$ is the indicator function, and $\beta_k \in \mathbb{R}^{d_1}, 0 \leq k \leq 2$ are coefficient vectors specific to each subpopulation.

The element of β_k is also drawn from standard gaussian distribution. The last three elements of covariates indicate the subpopulation index and represented as one-hot encoding. Specifically, we set $x_{i,d_1+s_i} = 1, x_{i,d_1+j} = 0, \forall j \neq s_i$.

We can induce distribution shift by adjusting the proportion of the subpopulations. Therefore, we construct the initial training datasets, candidate pools and target population with different compositions. The training datasets and target population are dominated by different subpopulations, while the candidate pool contains a large number of samples from different subpopulations. The detailed compositions are listed in the Appendix C.

For the model architecture, we adopt the neural networks consists of three fully connect layers with the hidden size equal to 32. We place dropout layer at the first hidden layer and set the dropout rate as 0.1.

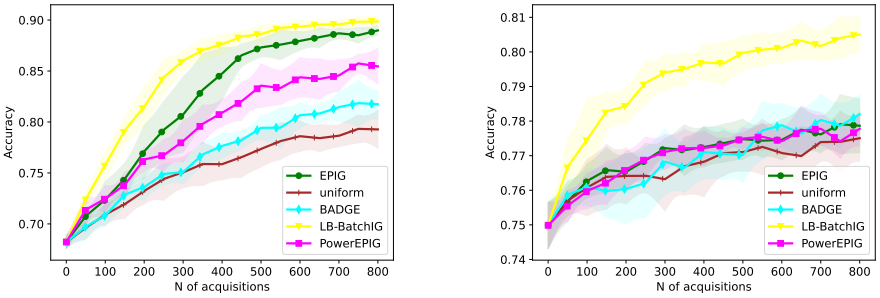
Results We conduct repeated experiments for 10 times with varied sample size α of subpopulation $s = 2$. The sample acquisition process consists of $t = 12$ rounds, and in each round $c = 50$ samples are selected. The results can be found in Figure 2.

From the results, we find that uniform acquisition method achieves worst performance across the methods. This is because it neglects the distinction of samples in improving models and fails to identify the beneficial samples for training.

The EPIG and BADGE method can achieve better performance than Uniform since they consider the contribution difference to model among samples and try to select more beneficial samples. However, the improvement brought by them is limited because of the redundant information and distribution shift problem respectively. When the sample size α increases, the proportion of samples belonging to the same subpopulation as target in the candidate pool also increases. The distribution shift problem is less severe. Therefore, the performances of BADGE and Uniform both are improved. The PowerEPIG method introduces diversity property into EPIG, and mitigate the redundant information problem. However, it does not explicitly optimize the predictive information gain of a batch, and thereby achieve sub-optimal performance.

We display the average sample proportion from the subpopulations in an acquired batch in Figure 4. From the results, we can observe that EPIG, PowerEPIG and LB-BatchIG focus on the predictive information gain on the target population and successfully identify the samples belonging to the same subpopulation (i.e. subpopulation $s = 1$).

However, when the sample size α increases, the performance of EPIG degrades. This is mainly because that the larger α leads to larger sample density in the same covariate region, which facilitate



(a) Income Prediction, Target State PR (b) Employment Prediction, Target State PR

Figure 6: The results on the tabular datasets under the different settings with different prediction tasks and state composition for training datasets, candidate pool and target population.

more severe redundant information problem. We aggregate the variance of the covariate elements $\{x_{,1}, x_{,2}, ..., x_{,d_1}\}$ of the acquired sample batch which measures the diversity of the selected samples. From the results in Figure 5, we observe that the diversity measurement of EPIG decreases with larger α which confirms our conjecture. In contrast, our proposed method achieves more diverse acquisition regardless of the composition of candidate pool, and thereby performs better than other methods.

Batch Size Analysis We change the batch size of single acquisition (as well as the acquisition rounds) to examine the effect of it on the performance curve. The results can be found in the Figure 3. EPIG and PowerEPIG underperforms with larger c due to more severe redundant information problem, while our proposed method is robust to the batch size variation and achieves superior performance.

4.3 Tabular Data

Ding et al. (2021) construct a series of datasets from available US Census sources spanning over multiple years, states of the United States and various prediction tasks for the research on algorithmic fairness and distribution shift.

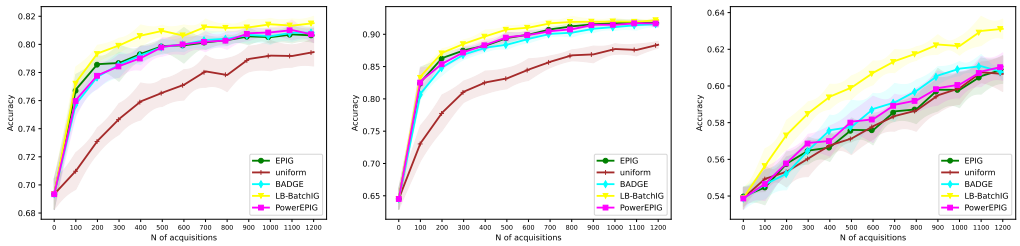
Experimental setup The data sources involve several prediction tasks. We choose the income and employment prediction as the benchmark to validate the effectiveness of methods. To create distribution shift, in this paper we leverage the available meta-information of states to constitute the different subpopulation composition. For each prediction task, we set up two experiment settings about different distribution shift respectively. The detailed information about the composition of the training dataset, candidate pool and target population can be found in the Appendix C.

We follow the same setup as the synthetic experiments and adopt the neural networks consists of three fully connect layers with the hidden size equal to 32. We place dropout layer at the first hidden layer and set the dropout rate as 0.1.

Results We repeat the experiments 10 times for each experimental setups about the prediction task and state compositions. The experimental results can be found in Figure 6. The overall trend is consistent with that of the synthetic dataset. On the whole, the Uniform performs worse than the other methods since it reflects no preference over the beneficial samples for performance improvement. Generally, the EPIG method achieves the second best performance especially in the income prediction task. However, in the employment prediction task, its advantage over other methods suffers from significant deterioration. It may be because in the employment prediction task, the samples with large predictive information gain tend to cluster and result in severe redundant information problem. In contrast, our method consistently accomplishes promising performance and outperforms the baselines.

4.4 Image Data

Domain generalization (Wang et al., 2022; Zhou et al., 2022) is an important branch of research developing the better performing models when encountering distribution shift. We leverage the



(a) VLCS (b) PACS (c) NICO++

Figure 7: The results on the image datasets, including VLCS, PACS and NICO++ benchmarks.

typical benchmarks of domain generalization in computer vision field to examine the effectiveness of our method.

Experimental setup We leverage three representative benchmarks respectively to construct datasets for experiments, that are VLCS (Fang et al., 2013), PACS (Li et al., 2017) and NICO++ (Zhang et al., 2023). The VLCS benchmark is composed by different data sources including PASCAL VOC (Everingham et al., 2010), Caltech101 (Griffin et al., 2007), LabelMe (Russell et al., 2008) and SUN09 (Choi et al., 2010) datasets. The PACS and NICO++ benchmarks consist of images from different domains. We control the proportion of different data sources or domains to manifest distribution shift among training dataset, candidate pool and target population. The detailed information about the composition of the training dataset, candidate pool and target population can be found in the Appendix C.

For the model architecutre, we adopt the Resnet18 (He et al., 2016) as the backbone and place dropout layer the last hidden layer with the dropout rate as 0.5.

Result We follow the same experiment settings as the previous experiments. Specifically, the batch size of single acquisition is set as $c = 50$, and we conduct the repeated experiments for 10 times.

The results are shown in Figure 7. The results further reinforce the conclusion we obtain in the previous experiments. Generally, the Uniform method underperforms the other methods since it ignores the distinction among the samples in improving model prediction and does not prefer the samples contributing more to the model performance. The BADGE and EPIG methods improve upon Uniform, but achieve suboptimal performance because of the distribution shift and redundant information problem respectively. PowerEPIG incorporate the diversity into acquisition process in a straightforward way. Our proposed acquisition function LB-BatchIG characterize the acquisition criteria at batch-level, which simultaneously considers the predictive information gain on the shifted target population and the diversity of acquired sample batch, and consistently achieves the best performance.

5 Conclusion and Limitation

In this paper, we investigate how to acquire new samples from the candidate data pool for improving the model performance on a shifted target population. We propose a novel acquisition function LB-BatchIG that is built upon predictive information gain on the target population to address the distribution shift, while considering the diversity of acquired batch to alleviate the redundant information problem. We also utilize the submodular property of the acquisition function to solve the optimization problem with greedy algorithm. Extensive experiments on the different datasets demonstrate the effectiveness of our method.

The greedy algorithm is an approximation solution to the batch acquisition. Therefore, proposing a more effective optimization algorithm is worthy to research for future work. Besides, the method is limited to classification task. The extension to more complex tasks, such as object detection, is also a valuable research problem in the future.

REFERENCES

Martin Arjovsky, Léon Bottou, Ishaan Gulrajani, and David Lopez-Paz. Invariant risk minimization. *arXiv preprint arXiv:1907.02893*, 2019.

David Arthur and Sergei Vassilvitskii. k-means++: the advantages of careful seeding. In *ACM-SIAM Symposium on Discrete Algorithms*, 2007. URL https://api.semanticscholar.org/CorpusID:1782131.

Jordan T Ash, Chicheng Zhang, Akshay Krishnamurthy, John Langford, and Alekh Agarwal. Deep batch active learning by diverse, uncertain gradient lower bounds. In *International Conference on Learning Representations*, 2020.

Shai Ben-David, John Blitzer, Koby Crammer, Alex Kulesza, Fernando Pereira, and Jennifer Wortman Vaughan. A theory of learning from different domains. *Machine learning*, 79:151–175, 2010.

José M Bernardo and Adrian FM Smith. *Bayesian theory*, volume 405. John Wiley & Sons, 2009.

Konstantinos Bousmalis, George Trigeorgis, Nathan Silberman, Dilip Krishnan, and Dumitru Erhan. Domain separation networks. *Advances in neural information processing systems*, 29, 2016.

Kathryn Chaloner and Isabella Verdinelli. Bayesian experimental design: A review. *Statistical science*, pp. 273–304, 1995.

Rita Chattopadhyay, Zheng Wang, Wei Fan, Ian Davidson, Sethuraman Panchanathan, and Jieping Ye. Batch mode active sampling based on marginal probability distribution matching. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 7(3):1–25, 2013.

Yuhua Chen, Wen Li, Christos Sakaridis, Dengxin Dai, and Luc Van Gool. Domain adaptive faster r-cnn for object detection in the wild. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3339–3348, 2018.

Myung Jin Choi, Joseph J Lim, Antonio Torralba, and Alan S Willsky. Exploiting hierarchical context on a large database of object categories. In *2010 IEEE computer society conference on computer vision and pattern recognition*, pp. 129–136. IEEE, 2010.

Nicolas Courty, Rémi Flamary, Devis Tuia, and Alain Rakotomamonjy. Optimal transport for domain adaptation. *IEEE transactions on pattern analysis and machine intelligence*, 39(9):1853–1865, 2016.

Frances Ding, Moritz Hardt, John Miller, and Ludwig Schmidt. Retiring adult: New datasets for fair machine learning. *Advances in neural information processing systems*, 34:6478–6490, 2021.

John C Duchi and Hongseok Namkoong. Learning models with uniform performance via distributionally robust optimization. *The Annals of Statistics*, 49(3):1378–1406, 2021.

Melanie Ducoffe and Frederic Precioso. Adversarial active learning for deep networks: a margin based approach. *arXiv preprint arXiv:1802.09841*, 2018.

Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. *International journal of computer vision*, 88: 303–338, 2010.

Chen Fang, Ye Xu, and Daniel N. Rockmore. Unbiased metric learning: On the utilization of multiple datasets and web images for softening bias. In *2013 IEEE International Conference on Computer Vision*, pp. 1657–1664, 2013. doi: 10.1109/ICCV.2013.208.

Ronald A Fisher. On the mathematical foundations of theoretical statistics. *Philosophical transactions of the Royal Society of London. Series A, containing papers of a mathematical or physical character*, 222(594-604):309–368, 1922.

Bo Fu, Zhangjie Cao, Jianmin Wang, and Mingsheng Long. Transferable query selection for active domain adaptation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 7272–7281, 2021.

Kenji Fukumizu. Statistical active learning in multilayer perceptrons. *IEEE Transactions on Neural Networks*, 11(1):17–26, 2000.

Yarin Gal and Zoubin Ghahramani. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *international conference on machine learning*, pp. 1050–1059. PMLR, 2016.

Yarin Gal, Riashat Islam, and Zoubin Ghahramani. Deep bayesian active learning with image data. In *International conference on machine learning*, pp. 1183–1192. PMLR, 2017.

Yaroslav Ganin and Victor Lempitsky. Unsupervised domain adaptation by backpropagation. In *International conference on machine learning*, pp. 1180–1189. PMLR, 2015.

Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario March, and Victor Lempitsky. Domain-adversarial training of neural networks. *Journal of machine learning research*, 17(59):1–35, 2016.

Gregory Griffin, Alex Holub, Pietro Perona, et al. Caltech-256 object category dataset. Technical report, Technical Report 7694, California Institute of Technology Pasadena, 2007.

Juan Mario Haut, Mercedes E Paoletti, Javier Plaza, Jun Li, and Antonio Plaza. Active learning with convolutional neural networks for hyperspectral image classification using a new bayesian approach. *IEEE Transactions on Geoscience and Remote Sensing*, 56(11):6440–6461, 2018.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

Neil Houlsby, Ferenc Huszár, Zoubin Ghahramani, and Máté Lengyel. Bayesian active learning for classification and preference learning. *arXiv preprint arXiv:1112.5745*, 2011.

Duojun Huang, Jichang Li, Weikai Chen, Junshi Huang, Zhenhua Chai, and Guanbin Li. Divide and adapt: Active domain adaptation via customized learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7651–7660, 2023.

Jiayuan Huang, Arthur Gretton, Karsten Borgwardt, Bernhard Schölkopf, and Alex Smola. Correcting sample selection bias by unlabeled data. *Advances in neural information processing systems*, 19, 2006.

Jing Jiang and Cheng Xiang Zhai. Instance weighting for domain adaptation in nlp. In *45th Annual Meeting of the Association for Computational Linguistics, ACL 2007*, pp. 264–271, 2007.

Ajay J Joshi, Fatih Porikli, and Nikolaos Papanikolopoulos. Multi-class active learning for image classification. In *2009 ieee conference on computer vision and pattern recognition*, pp. 2372–2379. IEEE, 2009.

Andreas Kirsch. Powerevaluationbald: Efficient evaluation-oriented deep (bayesian) active learning with stochastic acquisition functions. *ArXiv*, abs/2101.03552, 2021. URL `https://api.semanticscholar.org/CorpusID:231573455`.

Andreas Kirsch, Joost Van Amersfoort, and Yarin Gal. Batchbald: Efficient and diverse batch acquisition for deep bayesian active learning. *Advances in neural information processing systems*, 32, 2019.

Da Li, Yongxin Yang, Yi-Zhe Song, and Timothy M Hospedales. Deeper, broader and artier domain generalization. In *Proceedings of the IEEE international conference on computer vision*, pp. 5542–5550, 2017.

Mengxue Li, Yi-Ming Zhai, You-Wei Luo, Peng-Fei Ge, and Chuan-Xian Ren. Enhanced transport distance for unsupervised domain adaptation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 13936–13944, 2020.

Dennis V Lindley. On a measure of the information provided by an experiment. *The Annals of Mathematical Statistics*, 27(4):986–1005, 1956.

Jiashuo Liu, Tianyu Wang, Peng Cui, and Hongseok Namkoong. On the need for a language describing distribution shifts: Illustrations on tabular datasets. *Advances in Neural Information Processing Systems*, 36, 2024.

Peng Liu, Hui Zhang, and Kie B Eom. Active deep learning for classification of hyperspectral images. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 10 (2):712–724, 2016.

Peng Liu, Lizhe Wang, Rajiv Ranjan, Guojin He, and Lei Zhao. A survey on active deep learning: from model driven to data driven. *ACM Computing Surveys (CSUR)*, 54(10s):1–34, 2022.

Peyman Mohajerin Esfahani and Daniel Kuhn. Data-driven distributionally robust optimization using the wasserstein metric: Performance guarantees and tractable reformulations. *Mathematical Programming*, 171(1):115–166, 2018.

George L Nemhauser, Laurence A Wolsey, and Marshall L Fisher. An analysis of approximations for maximizing submodular set functions—i. *Mathematical programming*, 14:265–294, 1978.

Viraj Prabhu, Arjun Chandrasekaran, Kate Saenko, and Judy Hoffman. Active domain adaptation via clustering uncertainty-weighted embeddings. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 8505–8514, 2021.

Tian Qin, Tian-Zuo Wang, and Zhi-Hua Zhou. Budgeted heterogeneous treatment effect estimation. In *International Conference on Machine Learning*, pp. 8693–8702. PMLR, 2021.

Harsh Rangwani, Arihant Jain, Sumukh K Aithal, and R Venkatesh Babu. S3vaada: Submodular subset selection for virtual adversarial active domain adaptation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 7516–7525, 2021.

Bryan C Russell, Antonio Torralba, Kevin P Murphy, and William T Freeman. Labelme: a database and web-based tool for image annotation. *International journal of computer vision*, 77:157–173, 2008.

Vít Ržička, Stefano D'Aronco, Jan Dirk Wegner, and Konrad Schindler. Deep active learning in remote sensing for data efficient change detection. *arXiv preprint arXiv:2008.11201*, 2020.

Ozan Sener and Silvio Savarese. Active learning for convolutional neural networks: A core-set approach. *arXiv preprint arXiv:1708.00489*, 2017a.

Ozan Sener and Silvio Savarese. A geometric approach to active learning for convolutional neural networks. *arXiv preprint arXiv:1708.00489*, 7, 2017b.

Burr Settles and Mark Craven. An analysis of active learning strategies for sequence labeling tasks. In *proceedings of the 2008 conference on empirical methods in natural language processing*, pp. 1070–1079, 2008.

Freddie Bickford Smith, Andreas Kirsch, Sebastian Farquhar, Yarin Gal, Adam Foster, and Tom Rainforth. Prediction-oriented bayesian active learning. In *International Conference on Artificial Intelligence and Statistics*, pp. 7331–7348. PMLR, 2023.

Jamshid Sourati, Murat Akcakaya, Todd K Leen, Deniz Erdogmus, and Jennifer G Dy. Asymptotic analysis of objectives based on fisher information in active learning. *Journal of Machine Learning Research*, 18(34):1–41, 2017.

Jamshid Sourati, Ali Gholipour, Jennifer G Dy, Xavier Tomas-Fernandez, Sila Kurugol, and Simon K Warfield. Intelligent labeling based on fisher information for medical image segmentation using deep learning. *IEEE transactions on medical imaging*, 38(11):2642–2653, 2019.

Jong-Chyi Su, Yi-Hsuan Tsai, Kihyuk Sohn, Buyu Liu, Subhransu Maji, and Manmohan Chandraker. Active adversarial domain adaptation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 739–748, 2020.

Shujin Sun, Ping Zhong, Huaitie Xiao, and Runsheng Wang. An mrf model-based active learning framework for the spectral-spatial classification of hyperspectral imagery. *IEEE Journal of Selected Topics in Signal Processing*, 9(6):1074–1088, 2015.

Yi-Hsuan Tsai, Wei-Chih Hung, Samuel Schulter, Kihyuk Sohn, Ming-Hsuan Yang, and Manmohan Chandraker. Learning to adapt structured output space for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 7472–7481, 2018.

Eric Tzeng, Judy Hoffman, Ning Zhang, Kate Saenko, and Trevor Darrell. Deep domain confusion: Maximizing for domain invariance. *arXiv preprint arXiv:1412.3474*, 2014.

Jindong Wang, Cuiling Lan, Chang Liu, Yidong Ouyang, Tao Qin, Wang Lu, Yiqiang Chen, Wenjun Zeng, and S Yu Philip. Generalizing to unseen domains: A survey on domain generalization. *IEEE transactions on knowledge and data engineering*, 35(8):8052–8072, 2022.

Lin Yang, Yizhe Zhang, Jianxu Chen, Siyuan Zhang, and Danny Z Chen. Suggestive annotation: A deep active learning framework for biomedical image segmentation. In *Medical Image Computing and Computer Assisted Intervention- MICCAI 2017: 20th International Conference, Quebec City, QC, Canada, September 11-13, 2017, Proceedings, Part III 20*, pp. 399–407. Springer, 2017.

Kai Yu, Jinbo Bi, and Volker Tresp. Active learning via transductive experimental design. In *Proceedings of the 23rd international conference on Machine learning*, pp. 1081–1088, 2006.

Tong Zhang and F Oles. The value of unlabeled data for classification problems. In *Proceedings of the Seventeenth International Conference on Machine Learning,(Langley, P., ed.)*, volume 20, pp. 0. Citeseer, 2000.

Xingxuan Zhang, Yue He, Renzhe Xu, Han Yu, Zheyan Shen, and Peng Cui. Nico++: Towards better benchmarking for domain generalization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 16036–16047, 2023.

Kaiyang Zhou, Ziwei Liu, Yu Qiao, Tao Xiang, and Chen Change Loy. Domain generalization: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(4):4396–4415, 2022.

APPENDIX

A LARGE LANGUAGE MODEL USAGE

In this paper, we claim that large language models (LLMs) are used solely to support and refinethe writing process. Specifically, we use LLMs to provide word-level and sentence-level suggestions to enhancethe overall fluency of the text.

B PROOF

Theorem B.1. ***(Restated)*** *According to the information theory, we have*

$$\text{LB-BatchIG}(\{\mathbf{x}_i^b\}) \leq \mathbb{E}_{p^t(\mathbf{x}^*)}\left[\text{BIG}(\{\mathbf{x}_i^b\}, \mathbf{x}^*)\right].$$

Proof. For arbitrary $i \in [c]$, we can prove that $\text{IG}(\mathbf{x}_i^b, \mathbf{x}^*) \leq \text{BIG}(\{\mathbf{x}_i^b\}, \mathbf{x}^*)$. Without loss of the generality, we hypothesize $i = 1$. This is equivalent to

$$\begin{aligned}
&\int_{y_1^b, y_2^b, ..., y_c^b, y^*} p(y_1^b, y_2^b, ..., y_c^b, y^* | \mathbf{x}_1^b, \mathbf{x}_2^b, ..., \mathbf{x}_c^b, \mathbf{x}^*; \mathcal{D}) \cdot \\
&\log \frac{p(y_1^b, y_2^b, ..., y_c^b, y^* | \mathbf{x}_1^b, \mathbf{x}_2^b, ..., \mathbf{x}_c^b, \mathbf{x}^*; \mathcal{D})}{p(y_1^b, y_2^b, ..., y_c^b | \mathbf{x}_1^b, \mathbf{x}_2^b, ..., \mathbf{x}_c^b; \mathcal{D}) p(y^* | \mathbf{x}^*; \mathcal{D})} dy_1^b ... dy_c^b dy^* \\
&\geq \int_{y_1^b, y^*} p(y_1^b, y^* | \mathbf{x}_1^b, \mathbf{x}^*; \mathcal{D}) \log \frac{p(y_1^b, y^* | \mathbf{x}_1^b, \mathbf{x}^*; \mathcal{D})}{p(y_1^b | \mathbf{x}_1^b; \mathcal{D}) p(y^* | \mathbf{x}^*; \mathcal{D})} dy_1^b dy^* \\
&= \int_{y_1^b, y_2^b, ..., y_c^b, y^*} p(y_1^b, y_2^b, ..., y_c^b, y^* | \mathbf{x}_1^b, \mathbf{x}_2^b, ..., \mathbf{x}_c^b, \mathbf{x}^*; \mathcal{D}) \cdot \\
&\qquad \log \frac{p(y_1^b, y^* | \mathbf{x}_1^b, \mathbf{x}^*; \mathcal{D})}{p(y_1^b | \mathbf{x}_1^b; \mathcal{D}) p(y^* | \mathbf{x}^*; \mathcal{D})} dy_1^b dy^*.
\end{aligned} \tag{12}$$

By move the l.h.s to the right side of inequality, it becomes

$$\begin{aligned}
&\int_{y_1^b} p(y_1^b | \mathbf{x}_1^b; \mathcal{D}) dy_1^b \int_{y_2^b ..., y_c^b, y^*} p(y_2^b, ..., y_c^b, y^* | \mathbf{x}_1^b, y_1^b, \mathbf{x}_2^b, ..., \mathbf{x}_c^b, \mathbf{x}^*; \mathcal{D}) \cdot \\
&\qquad \log \frac{p(y_2^b, ..., y_c^b | \mathbf{x}_1^b, y_1^b, \mathbf{x}_2^b, ..., \mathbf{x}_c^b; \mathcal{D}) p(y^* | \mathbf{x}_1^b, y_1^b, \mathbf{x}^*; \mathcal{D})}{p(y_2^b, ..., y_c^b, y^* | \mathbf{x}_1^b, y_1^b, \mathbf{x}_2^b, ..., \mathbf{x}_c^b, \mathbf{x}^*; \mathcal{D})} dy_2^b ... dy_c^b dy^* \leq 0.
\end{aligned} \tag{13}$$

Because the function $\log(\cdot)$ is convex function, according to Jensen inequality, we can prove

$$\begin{aligned}
&\int_{y_2^b ..., y_c^b, y^*} p(y_2^b, ..., y_c^b, y^* | \mathbf{x}_1^b, y_1^b, \mathbf{x}_2^b, ..., \mathbf{x}_c^b, \mathbf{x}^*; \mathcal{D}) \\
&\log \frac{p(y_2^b, ..., y_c^b | \mathbf{x}_1^b, y_1^b, \mathbf{x}_2^b, ..., \mathbf{x}_c^b; \mathcal{D}) p(y^* | \mathbf{x}_1^b, y_1^b, \mathbf{x}^*; \mathcal{D})}{p(y_2^b, ..., y_c^b, y^* | \mathbf{x}_1^b, y_1^b, \mathbf{x}_2^b, ..., \mathbf{x}_c^b, \mathbf{x}^*; \mathcal{D})} dy_2^b ... dy_c^b dy^* \\
&\leq \log \int_{y_2^b ..., y_c^b, y^*} p(y_2^b, ..., y_c^b, y^* | \mathbf{x}_1^b, y_1^b, \mathbf{x}_2^b, ..., \mathbf{x}_c^b, \mathbf{x}^*; \mathcal{D}) \\
&\frac{p(y_2^b, ..., y_c^b | \mathbf{x}_1^b, y_1^b, \mathbf{x}_2^b, ..., \mathbf{x}_c^b; \mathcal{D}) p(y^* | \mathbf{x}_1^b, y_1^b, \mathbf{x}^*; \mathcal{D})}{p(y_2^b, ..., y_c^b, y^* | \mathbf{x}_1^b, y_1^b, \mathbf{x}_2^b, ..., \mathbf{x}_c^b, \mathbf{x}^*; \mathcal{D})} dy_2^b ... dy_c^b dy^* \\
&= \log \int_{y_2^b ..., y_c^b, y^*} p(y_2^b, ..., y_c^b | \mathbf{x}_1^b, y_1^b, \mathbf{x}_2^b, ..., \mathbf{x}_c^b; \mathcal{D}) \\
&\qquad p(y^* | \mathbf{x}_1^b, y_1^b, \mathbf{x}^*; \mathcal{D}) dy_2^b ... dy_c^b dy^* \\
&= \log 1 = 0
\end{aligned}$$

Therefore, the Equation 13 is satisfied. By analog, we can prove $\text{IG}(\mathbf{x}_i^b, \mathbf{x}^*) \leq \text{BIG}(\{\mathbf{x}_i^b\}, \mathbf{x}^*), \forall 1 \leq i \leq c$.

We can derive that

$$\begin{aligned}
&\max_{1\leq i\leq c} \text{IG}(\mathbf{x}_i^b, \mathbf{x}^*) \leq \text{BIG}(\{\mathbf{x}_i^b\}, \mathbf{x}^*) \\
&\Rightarrow \text{LB-BatchIG}(\{\mathbf{x}_i^b\}) = \mathbb{E}_{p^t(\mathbf{x}^*)}\left[\max_{1\leq i\leq c} \text{IG}(\mathbf{x}_i^b, \mathbf{x}^*)\right] \\
&\qquad\qquad \leq \mathbb{E}_{p^t(\mathbf{x}^*)}\left[\text{BIG}(\{\mathbf{x}_i^b\}, \mathbf{x}^*)\right]
\end{aligned}$$

□

Proposition B.1. ***(Restated)*** *Regarding the unlabeled pool as the element set V in Definition 3.1 and the sample batch as the subset, the acquisition function LB-BatchIG is a normalized monotone non-decreasing submodular function.*

Proof. We respectively proof the property of normalization, monotone non-decreasing, and submodular.

Normalization: According to Equation 5, we can easily obtain that $\text{LB-BatchIG}(\varnothing) = 0$. Hence the normalization property is satisfied.

Monotone non-decreasing: For a batch $\{\mathbf{x}_i^b\}_{1\leq i\leq d}$ and a sample $\mathbf{x}' \notin \{\mathbf{x}_i^b\}_{1\leq i\leq d}$, we have

$$\begin{aligned}
&\text{LB-BatchIG}(\{\mathbf{x}_i^b\}_{1\leq i\leq d} \cup \{\mathbf{x}'\}) \\
&= \mathbb{E}_{p^t(\mathbf{x}^*)}\left[\max\{\max_{1\leq i\leq d} \text{IG}(\mathbf{x}_i^b, \mathbf{x}^*), \text{IG}(\mathbf{x}', \mathbf{x}^*)\}\right] \\
&\geq \mathbb{E}_{p^t(\mathbf{x}^*)}\left[\max_{1\leq i\leq d} \text{IG}(\mathbf{x}_i^b, \mathbf{x}^*)\right] = \text{LB-BatchIG}(\{\mathbf{x}_i^b\}_{1\leq i\leq d})
\end{aligned} \tag{14}$$

Submodular: For two arbitrary sample batches A and B and sample $\mathbf{x}^*$, we denote

$$a = \max_{\mathbf{x}\in A} \text{IG}(\mathbf{x}, \mathbf{x}^*), \tag{15}$$

$$b = \max_{\mathbf{x}\in B} \text{IG}(\mathbf{x}, \mathbf{x}^*), \tag{16}$$

$$c = \max_{\mathbf{x}\in A\cap B} \text{IG}(\mathbf{x}, \mathbf{x}^*). \tag{17}$$

Then we have $\max\{a, b\} = \max_{\mathbf{x}\in A\cup B} \text{IG}(\mathbf{x}, \mathbf{x}^*)$ and $a \geq c, b \geq c$.

If $a \geq b$, then $a = \max\{a, b\} \cap b \geq c \Rightarrow a + b \geq \max\{a, b\} + c$.

If $a < b$, then $b = \max\{a, b\} \cap a \geq c \Rightarrow a + b \geq \max\{a, b\} + c$.

Therefore, we have

$$\begin{aligned}
&\mathbb{E}_{p^t(\mathbf{x}^*)}[\max_{\mathbf{x}\in A} \text{IG}(\mathbf{x}, \mathbf{x}^*) + \max_{\mathbf{x}\in B} \text{IG}(\mathbf{x}, \mathbf{x}^*)] \\
&\geq \mathbb{E}_{p^t(\mathbf{x}^*)}[\max_{\mathbf{x}\in A\cap B} \text{IG}(\mathbf{x}, \mathbf{x}^*) + \max_{\mathbf{x}\in A\cup B} \text{IG}(\mathbf{x}, \mathbf{x}^*)].
\end{aligned} \tag{18}$$

Based on this, we conclude

$$\begin{aligned}
&\text{LB-BatchIG}(A) + \text{LB-BatchIG}(B) \\
&\geq \text{LB-BatchIG}(A \cap B) + \text{LB-BatchIG}(A \cup B)
\end{aligned}$$

□

C Data Composition of Experiments

The data composition in the experiments of synthetic data, tabular data and image data are as follows:

Synthetic Data

- Training dataset: Subpop-0 1200 + Subpop-1 20 + Subpop-2 180
- Candidate pool: Subpop-0 6000 + Subpop-1 α + Subpop-2 3400
- Target population: Subpop-1 3000

Tabular Data

Income Prediction

- Training dataset: CA 1000 + AL 100
- Candidate pool: CA 10000 + AL 10000 + PR 5000
- Target population: PR 2000

Employment Prediction

- Training dataset: CA 400 + AL 200
- Candidate pool: CA 10000 + AL 10000 + PR 5000
- Target population: PR 3000

Image Data

VLCS benchmark

- Training dataset: LABELME 400+CALTECH 400+ SUN 400
- Candidate pool: LABELME 1000 + CALTECH 400+ PASCAL 1000+SUN 1000
- Target population: PASCAL 1000

PACS benchmark

- Training dataset: photo 400+ art 400+sketch 400
- Candidate pool: photo 1000+ art 1000 cartoon 1000+ sketch 1000
- Target population: cartoon 1000

Nico++ benchmark

- Training dataset: autumn 400 +rock 400+dim 400 +grass 400
- Candidate pool: autumn 1000 + rock 1000 + dim 1000 + grass 1000 + outdoor 1000 + water 1000
- Target population: outdoor 1000 + water 1000

D Pseudo-code of Our Algorithm

The pseudo-code of our algorithm is presented as Algorithm 1.

E Related Work

In this section, we briefly review the related research of the unsupervised domain adaptation, active learning and active DA.

Algorithm 1 Greedy algorithm for optimizing LB-BatchIG

Input: Training dataset $\mathcal{D}_{tr} = \{(\mathbf{x}_i^{tr}, y_i^{tr})\}_{1 \le i \le n_{tr}}$, unlabeled pool $\mathcal{D}_{po} = \{\mathbf{x}_i^{po}\}_{1 \le i \le n_{po}}$, the test samples representing the test distribution $\mathcal{D}_{te} = \{\mathbf{x}_i^{te}\}_{1 \le i \le n_{te}}$ and the model f_θ with parameter θ.
Output: the sample batch $\mathcal{B} = \{\mathbf{x}_i^b\}_{1 \le i \le c}$
Learning the model f_θ on the training dataset $\mathcal{D}_{tr}$ and sample a collection of parameters $\{\theta_l\}_{1 \le l \le m}$ from posterior distribution $p(\theta|\mathcal{D}_{tr})$.
Initialize the sample batch $\mathcal{B}_0 = \varnothing$.
for $i = 1$ **to** c **do**
  Sample a batch of test samples $\{x_j^*\}_{1 \le j \le s}$
  $b_i \leftarrow 1$
  $r \leftarrow 0$
  for $w = 2$ **to** n_{po} **do**
    if LB-BatchIG$(\mathcal{B}_{i-1} \cup \mathbf{x}_w^{po}) > r$ **then**
      $b_i \leftarrow w$
      $r \leftarrow$ LB-BatchIG$(\mathcal{B}_{i-1} \cup \mathbf{x}_w^{po})$
    end if
    $\mathcal{B}_i = \mathcal{B}_{i-1} \cup \{\mathbf{x}_{b_i}^{po}\}$
  end for
end for
return the acquisition batch $\mathcal{B}_c$

E.1 Unsupervised Domain Adaptation

When the target domain for generalization is known, a bunch of domain adaptation (DA) (Ben-David et al., 2010) works can be proposed. To match the feature distribution of source domains and target domains, some approaches re-weight or select the training samples (Jiang & Zhai, 2007; Huang et al., 2006). Besides, learning a feature transformation (Ganin & Lempitsky, 2015; Bousmalis et al., 2016; Tzeng et al., 2014) is an alternative method to align the feature distributions. Specifically, (Tzeng et al., 2014) leverage Maximum Mean Discrepancy (MMD) which characterizes the difference of distribution mean in reproducing kernel Hilbert space. (Ganin & Lempitsky, 2015) and (Ganin et al., 2016) train a domain classifier and applied the separability between domains as the discrepancy measurement. And some literature (Li et al., 2020; Courty et al., 2016) use transport distance to learn the domain-aligned transformation. Although noteworthy advancements have been made from the perspective of algorithms to enhance model performance on target domains, they still fall behind the supervised learning counterpart (Chen et al., 2018; Tsai et al., 2018). Therefore, it can play a significant role to enlarge the training dataset with beneficial samples for training the models.

E.2 Active Learning

Active learning investigates how to acquire data for annotation to optimize the model. The proposed data acquisition criterions cover many aspects (Liu et al., 2022), including uncertainty, model influence, representativeness. For criterions based on uncertainty, (Joshi et al., 2009) calculate the prediction uncertainty by the entropy of classification probability and the difference of the highest two probability, (Ržička et al., 2020) take the gap between the highest probability and 1.0 as the uncertainty. For criterion based on representativeness, the samples central to the data distribution are acquired. (Settles & Craven, 2008) calculate the similarity to other samples as the representativeness metric. Core-set methods (Qin et al., 2021; Sener & Savarese, 2017a) try to choose center samples so that the largest distance between samples and the nearest center is minimized. As for criterions based on model influence, the methods select the samples having great impact on the model parameters if incorporated into training dataset. Some active learning algorithms (Sourati et al., 2017; 2019) apply Fisher information (Fisher, 1922) as the measurement of the impact on model parameters. Although Fisher information is theoretically grounded, it is computationally intensive in practice. BADGE (Ash et al., 2020) use the magnitude of gradient constituting the metric of impact on model parameters. Inspired by the bayesian inference (Bernardo & Smith, 2009), bayesian active learning (Gal et al., 2017; Kirsch et al., 2019; Smith et al., 2023) make assumptions on the

prior distribution of model parameters, and calculate the entropy reduction after adding samples as the model influence metric.

E.3 Active Domain Adaptation

Active DA focuses on acquiring samples from the target domain to accomplish domain adaptation. To be concrete, AADA (Su et al., 2020) selects samples based on uncertainty and domainness measured by a domain discriminator. (Fu et al., 2021) propose a unified criterion incorporating transferable committee, transferable uncertainty, and transferable domainness. CLUE (Prabhu et al., 2021) design a clustering algorithm weighted by uncertainty to select samples from target domain. DiaNA (Huang et al., 2023) propose a Divide-And-Adapt protocol which partitions the target samples into four types and selects the uncertain and inconsistent ones. These methods hypothesize a candidate pool with the same distribution to target domains can be acquired for annotation. However, the commercial restriction, privacy concerns and other issues can make this hypothesis unrealistic.

F Experimental compute resource

All experiments are conducted with the following settings:

- Operating Systems: Ubuntu 14.04.1 LTS
- CPU: Intel(R) Xeon(R) CPU E5-2630 v2 @ 2.60GHz
- GPU: Nvidia RTX 3090 $\times$ 1
- Memory: 256GB