# Active Learning Using Discrepancy

**Zhenghang Cui** cui@ms.k.u-tokyo.ac.jp
*University of Tokyo, Tokyo, Japan*
*RIKEN Center for Advanced Intelligence Project, Tokyo, Japan*

**Issei Sato** sato@k.u-tokyo.ac.jp
*University of Tokyo, Tokyo, Japan*
*RIKEN Center for Advanced Intelligence Project, Tokyo, Japan*

## Abstract

Developing active learning methods using deep neural networks to select a batch of unlabeled data points at each step are piratically important. However, existing such methods either rely on heuristic objectives such as uncertainty, diversity and representativity; or are computationally troublesome. Inspired by the discrepancy measures studied in the field of domain adaption that consider the hypotheses set and the loss function, we define a variant of discrepancy that suits the setting of deep batch active learning. We show that the newly defined discrepancy can establish a valid generalization error bound, which is comparable to the bound established by the Wasserstein distance that are recently adapted to active learning. Learning using the newly defined discrepancy results in a principled deep batch active learning method with an objective which can be stably optimized. We empirically confirm the performance of the proposed algorithm against the state-of-the-art method.

**Keywords:** active learning, deep neural networks, discrepancy measure

## 1. Introduction

Active learning aims to relieve the burden of annotation by allowing the algorithm to choose unlabeled data points to query. Traditional analysis on the generalization error of pool-based active learning heavily relies on the properties of the hypotheses set. Thus, the existing theory can hardly provide insights for the emerging deep batch active learning setting where an active learning algorithm uses a deep neural network as the classifier and selects a batch of unlabeled data points at each step.

Recently developed methods focus on finding the criterion for selecting the batch that can benefit deep neural network models. Sener and Savarese (2018) choose to construct a coreset, which minimizes an upper bound of the generalization error but are very computationally expensive. Ash et al. (2020) choose to use gradients as the selection criterion, which lacks evidence of connection to the generalization error. Shui et al. (2020) recognize two important distributions in active learning: (1) the underlying data distribution with respect to which the generalization error is defined, and (2) the empirical distribution of already labeled data points and the batch of unlabeled data points we are going to select. Shui et al. (2020) then propose a principled algorithm that minimizes an upper bound of the generalization error, including a term of the Wasserstein distance (Villani, 2009) between the two aforementioned distributions. However, because the Wasserstein distance does not

take into account the hypotheses set and the loss function, it is possible that a tighter bound can be established. Moreover, because of the intrinsic definition of the Wasserstein distance, the implied minimax objective has to be optimized in an adversarial way, which is unstable and needs further practical treatments.

On the other hand, minimizing the generalization error on one distribution with access to data sampled from another distribution resembles the problem setting of domain adaptation (Ben-David et al., 2007). In this field, discrepancy measures (Mansour et al., 2009) have been extensively studied to capture the difference between two distributions. However, learning using theoretically guaranteed discrepancy measures either implies a minimax optimization problem, or requires the underlying labeling function of at least one distribution. Therefore, a naive adaption cannot significantly improve existing deep batch active learning methods.

In this paper, we first define an appropriate discrepancy measure for deep batch active learning in Section 2. Next we establish a generalization error bound and compare it with the one established using the Wasserstein distance in Section 3. Then, being guided by the theory, we derive a practical active learning algorithm in Section 4. Finally, we empirically confirm the performance of learning using the proposed discrepancy in Section 5.

## 2. Active learning using discrepancy

In this section, we introduce the proposed discrepancy, show its theoretical properties, and derive an algorithm from the theoretical result.

### 2.1 Preliminaries

Let $\mathcal{X} \subset \mathbb{R}^d$ be the sample space and $\mathcal{Y} \triangleq \{0, 1\}$ be the binary label space. Let $\mathbb{P}_{\mathcal{D}}$ denote the underlying distribution over $\mathcal{X}$. For the pool-based active learning problem setting, we are given an i.i.d. sample $D \triangleq \{x_i\}_{i=1}^n$ drawn from $\mathbb{P}_{\mathcal{D}}$. For a score function $f \in \mathcal{F} : \mathcal{X} \to \mathbb{R}$ in the hypotheses set $\mathcal{F}$, we define its induced prediction function as $h_f : x \mapsto \mathbb{1}_{f(x)>0}$. Let $f^*$ denote the corresponding scoring function for the the underlying labeling function $h^* : \mathcal{X} \to \mathcal{Y}$. We denote the generalized difference between two functions $f, f'$ on the distribution $\mathcal{D}$ using a loss function $\ell : \mathbb{R} \times \mathbb{R} \to \mathbb{R}^+$ as $R^\ell_{\mathcal{D}}(f, f') \triangleq \mathbb{E}_{x\sim\mathbb{P}}[\ell(f(x), f'(x))]$. Our goal is to find a scoring function $f \in \mathcal{F}$ such that its induced prediction function $h_f$ has low generalization error on $\mathcal{D}$, which can be expressed as $\arg\min_{f\in\mathcal{F}} R^\ell_{\mathcal{D}}(f, f^*)$.

At the beginning of each step, for the labeled set $L \subset D$, we have their corresponding labels $Y \triangleq \{h^*(x) : x \in L\}$. As we consider active learning as an instance of distribution matching, the source distribution $\mathbb{P}_{\mathcal{S}}$ is defined as the empirical distribution of $L$ and the batch of unlabeled samples to be selected $B \subset D \setminus L$.

### 2.2 Discrepancy measures for domain adaptation

We briefly review several important discrepancy measures that are extensively used in the filed of domain adaption, paving the path to the newly defined discrepancy for active learning in Section 2.3.

The general discrepancy measure (Mansour et al., 2009) is defined as

$$\sup_{f,f' \in \mathcal{F}} \left| R^{\ell}_{\mathrm{T}}(f, f') - R^{\ell}_{\mathrm{S}}(f, f') \right|,$$

where T and S denote the target distribution and the source distribution, respectively. In this common form of discrepancy, the supremum is taken over a function space.

Mohri and Medina (2012) proposes a discrepancy measure to provide a tighter generalization error bound. It is defined as $\sup_{f \in \mathcal{F}} \left| R^{\ell}_{\mathrm{T}}(f, f_{\mathrm{T}}) - R^{\ell}_{\mathrm{S}}(f, f_{\mathrm{S}}) \right|$, where $f_{\mathrm{T}}$ and $f_{\mathrm{S}}$ denote the underlying labeling functions for each distribution, respectively. The supremum is taken over one function now, but it cannot be approximated without knowing the underlying labeling functions. However, this gives us a hint that we can substitute and fix one function, and still achieve a proper discrepancy measure. Following the same intuition, Kuroki et al. (2019) proposes the source discrepancy that does not require the underlying labeling function for the target domain. It substitutes $f_{\mathrm{T}}$ and $f_{\mathrm{S}}$ with $f^*_{\mathrm{S}} \triangleq \arg\min_{f \in \mathcal{F}} R^{\ell}_{\mathrm{S}}(f, f_{\mathrm{S}})$, the true risk minimizer in the source domain. This gives us another hint that we can use the classifier that suits one distribution to achieve a discrepancy measure that enjoys better theoretical guarantees.

### 2.3 Active learning using discrepancy

Active learning is inherently different from domain adaptation as it is adaptive to user feedback and has multiple steps during execution. After each step, we end up having an imperfect but somehow satisfactory classifier for the moment. However, existing methods discard this imperfect classifier and train from scratch for each step. To this end, we propose to substitute one function in the generalization error terms with this imperfect classifier. Moreover, we usually want to minimize the discrepancy term in generalization error bounds, which will imply a minimax optimization problem that is unstable to solve. To this end, we define the following discrepancy for active learning.

**Definition 1 (Discrepancy for active learning)** *Let $\mathcal{F}$ be a scoring function set and $\ell : \mathbb{R} \times \mathbb{R} \to \mathbb{R}_+$ be a loss function. Then, the discrepancy for active learning between the underlying distributions $\mathbb{P}$ and the empirical distribution of $\mathbb{Q}$, being parameterized by $f, f' \in \mathcal{F}$ is defined as*

$$\mathrm{disc}_{f,f'}(\mathbb{P}, \mathbb{Q}) = \left| R^{\ell}_{\mathbb{P}}(f, f') - R^{\ell}_{\mathbb{Q}}(f, f') \right|. \tag{1}$$

This discrepancy is symmetric and satisfies the triangular inequality for distributions. However, it may that the discrepancy becomes 0 when $\mathbb{P}$ and $\mathbb{Q}$ are not identical.

## 3. Theoretical analysis

In this section, we investigate some theoretical properties of the newly defined discrepancy measure. First, we establish the following generalization bound for the discrepancy.

**Theorem 2 (Generalization error bound)** *For any $f, f' \in \mathcal{F}$, we have*

$$R_{\mathbb{P}}(f, f^*) \leq R_{\mathbb{Q}}(f, f^*) + \mathrm{disc}_{f,f'}(\mathbb{P}, \mathbb{Q}) + R_{\mathbb{P}}(f', f^*) + R_{\mathbb{Q}}(f', f^*). \tag{2}$$

**Proof**

$$R_{\mathbb{P}}(f, f^*) = R_{\mathbb{Q}}(f, f^*) - R_{\mathbb{Q}}(f, f^*) + R_{\mathbb{P}}(f, f') + R_{\mathbb{P}}(f', f^*) \tag{3}$$
$$\leq R_{\mathbb{Q}}(f, f^*) - R_{\mathbb{Q}}(f, f') + R_{\mathbb{P}}(f, f') + R_{\mathbb{Q}}(f', f^*) + R_{\mathbb{P}}(f', f^*) \tag{4}$$
$$\leq R_{\mathbb{Q}}(f, f^*) + \mathrm{disc}_{f,f'}(\mathbb{P}, \mathbb{Q}) + R_{\mathbb{P}}(f', f^*) + R_{\mathbb{Q}}(f', f^*). \tag{5}$$

■

From the above theory, we know that we can minimize the generalization error corresponding to the underlying distribution by minimizing the one corresponding to another distribution adding the discrepancy between the two distributions. This will imply a novel active learning algorithm, as we will show in Section 4. Although we cannot directly minimize the last two terms of the generalization errors including the plugin function $f'$, we believe that as the training step of active learning goes on, we would have a classifier that implies smaller generalization errors indicated by these two terms.

Then, we show the above discrepancy provides a potentially tighter generalization error bound than the Wasserstein distance. First we show the following lemma.

**Lemma 3 (Comparison to the Wasserstein distance)** *Assume the loss function $\ell$ is $\rho$-Lipschitz with respect to both arguments where $0 < \rho < \infty$, and all functions in the hypotheses set $\mathcal{F}$ are at most $\gamma$-Lipschitz and bounded, i.e., there exists a constant $C$ such that $||f||_\infty \leq C$ for any $f \in \mathcal{F}$. Then, for any $f, f' \in \mathcal{F}$, it holds that* $\mathrm{disc}_{f,f'}(\mathbb{P}, \mathbb{Q}) \leq \rho\gamma\, W(\mathbb{P}, \mathbb{Q})$.

The proof can be found in Appendix.

In the previous work (Shui et al., 2020), the generalization error bound is established as $R^\ell_{\mathbb{Q}}(f, f^*) + \rho(\gamma + \lambda)\mathrm{W}(\mathbb{P}, \mathbb{Q}) + \phi(\lambda)$, where $\lambda > 0$ and $\phi : \mathbb{R} \to (0, 1)$ is a function indicating the decay property of the underlying labeling function. Thus, considering the edge condition that $\mathrm{disc}_{f,f'}(\mathbb{P}, \mathbb{Q}) = \rho\gamma\mathrm{W}(\mathbb{P}, \mathbb{Q})$, the difference between two bounds are $(R_{\mathbb{P}}(f, f^*) + R_{\mathbb{Q}}(f, f^*)) - (\rho\lambda\mathrm{W}(\mathbb{P}, \mathbb{Q}) + \phi(\lambda))$. As the learning step proceeds, the classifier becomes more accurate and has the potential to drive this difference to be negative. As shown in Section 5, the algorithm guided by our bound shows significant improvement even from the first step.

## 4. Algorithm

In this section, we show a practical active learning algorithm that are directly guided by the theory.

We use the same definition as Shui et al. (2020) for the two distributions in active learning. We let $\mathbb{P}$ be the underlying data distribution $\mathbb{P}_{\mathcal{D}}$ and $\mathbb{Q}$ be the empirical distribution of $L \cup B$. Focusing on the first two tractable terms in Equation (2) and substituting the two distributions, the problem becomes

$$\min_{f,f',B} R^\ell_{L\cup B}(f, f^*) + \left| R^\ell_{L\cup U}(f, f') - R^\ell_{L\cup B}(f, f') \right|. \tag{6}$$

First, we optimize the model $f$ from initialization and the model $f'$ from its parameters at the time. Then, we use the trained models to select the batch of unlabeled data points for querying labels.

**Optimization stage** Note that there is a term involving the query batch in the absolute value. As in this stage, our task is to optimize the models without finding the query batch, we assume the dataset is flexible enough that there exists a batch of data points with $R^{\ell}_{L\cup B}(f, f')$ approaching 0. Excluding the query batch this term, we design the objective for $f'$ as

$$\frac{1}{|L| + |B|} \sum_{(x,y)\in L} \ell(f'(x), y). \tag{7}$$

Then, we optimize $f$ by minimizing

$$\frac{1}{|L| + |B|} \sum_{(x,y)\in L} \ell(f(x), y) + \frac{1}{|L| + |U|} \sum_{x\in L\cup U} \ell(f(x), f'(x)), \tag{8}$$

where taking the absolute value is omitted as the loss is always positive.

**Query stage** This stage aims to find the batch of unlabeled data points that minimizing Equation (6) with fixed models $f$ and $f'$. The first term including the query batch requires the underlying label $y$ for the unlabeled data point, which we do not know. Therefore, we adapt the same inequality as (Shui et al., 2020): $\ell(f(x), y) \leq \max_{y'\in[K]} - \log(f(x, y'))$, where $K$ denotes the number of classes and $f(x, y')$ denotes the $y'$-th softmax score of $x$ predicted by $f$. We then calculate the following score for every unlabeled data points $x \in U$ and select those with the most small values:

$$\left(\max_{y'\in[K]} - \log(f(x, y'))\right) + \ell(f(x), f'(x)). \tag{9}$$

We then formally describe the procedure in Algorithm 1.

**Algorithm 1** Active learning algorithm using discrepancy for one step.

**Input:** Labeled data $L$, unlabeled data $U$, query budget —B—, an initialized classifier $f$ and the classifier $f'$ inherited from last step. If it is the first step, prepare another initialized classifier as $f'$.

1: **while** a stopping criterion is not met **do**
2:  **for** mini-batches of $U$ **do**
3:  Select a mini-batch of $L$.
4:  Update parameters of $f'$ by minimizing Equation (7).
5:  Update parameters of $f$ by minimizing Equation (8).
6:  **end for**
7: **end while**
8: Calculate scores for $U$ using Equation (9).

**Output:** The batch $B$ of data points with lowest scores.

Optimization tricks such as adjusting the weights of different terms in an objective can be further applied. However, they are not necessarily required and are not included in the implementation for Section 5.

## 5. Experiments

In this section, we confirmed the performance of Algorithm 1 using the Fashion-MNIST (Xiao et al., 2017) dataset. We used the exactly same setting and budget scheduling as the 'WALL' algorithm (Shui et al., 2020). We repeated the same experiment for 10 times and report the mean and standard deviation values of test accuracy and execution time.

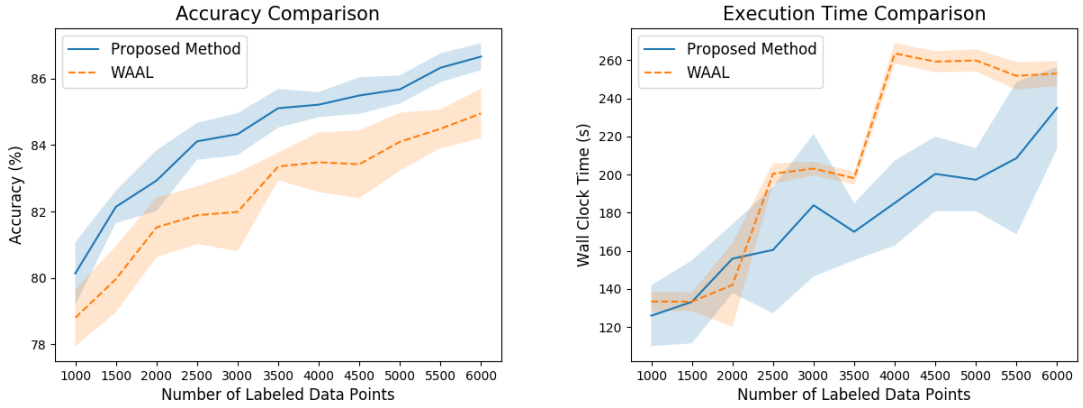


Figure 1: Accuracy and execution time.

Figure 1 reports the test accuracy and execute time for each step. We observe that learning according to our proposed discrepancy offers a consistently better generalization performance with shorter wall clock time and thus fewer computation resource. Note that our implementation of the 'WAAL' algorithm showed a higher test accuracy than reported (Shui et al., 2020) for the exactly same setting.

## 6. Conclusion

In this paper, we define a discrepancy measure for active learning and propose a principled learning algorithm that minimizes the generalization error bound established by the proposed discrepancy measure. The proposed discrepancy measure instantiates a minimization problem instead of a potentially unstable minimax problem. Thus, this algorithm can be easily implemented and shows promising performance in experiments on real-world datasets against the state-of-the-art deep batch active learning method. Conducting more comparison experiments and more analysis on the theoretical properties of active learning related discrepancy measures are left to be future work.

## Acknowledgments

ZC was supported by the IST-RA program, the University of Tokyo and the JST AIP Network Laboratory AIP Challenge program, Japan. IS was supported by the JST CREST Grant Number JPMJCR17A1, Japan.

## Proof of Lemma 3

**Proof** According to the Kantorovich-Rubinstein duality (Villani, 2009), the Wasserstein distance can be expressed as

$$\mathrm{W}(\mathbb{P}, \mathbb{Q}) = \sup_{||h||_L \leq 1} \left( \mathop{\mathbb{E}}_{x \sim \mathbb{P}}[h(x)] - \mathop{\mathbb{E}}_{x \sim \mathbb{Q}}[h(x)] \right), \tag{10}$$

where the supremum is take over all 1-Lipschitz functions. Define $h(x) = \ell(f(x), f'(x))$, which is a $\rho\gamma$-Lipschitz function. Then we have

$$\mathrm{disc}_{f,f'}(\mathbb{P}, \mathbb{Q}) = \left| \mathop{\mathbb{E}}_{x \sim \mathbb{P}}[h(x)] - \mathop{\mathbb{E}}_{x \sim \mathbb{Q}}[h(x)] \right| \tag{11}$$

$$\leq \sup_{||h||_L \leq \rho\gamma} \left( \mathop{\mathbb{E}}_{x \sim \mathbb{P}}[h(x)] - \mathop{\mathbb{E}}_{x \sim \mathbb{Q}}[h(x)] \right) = \rho\gamma \mathrm{W}(\mathbb{P}, \mathbb{Q}). \tag{12}$$

■

## References

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. URL `https://openreview.net/forum?id=ryghZJBKPS`.

Shai Ben-David, John Blitzer, Koby Crammer, and Fernando Pereira. Analysis of representations for domain adaptation. In *Advances in neural information processing systems*, pages 137–144, 2007.

Seiichi Kuroki, Nontawat Charoenphakdee, Han Bao, Junya Honda, Issei Sato, and Masashi Sugiyama. Unsupervised domain adaptation based on source-guided discrepancy. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 4122–4129, 2019.

Yishay Mansour, Mehryar Mohri, and Afshin Rostamizadeh. Domain adaptation: Learning bounds and algorithms. *arXiv preprint arXiv:0902.3430*, 2009.

Mehryar Mohri and Andres Munoz Medina. New analysis and algorithm for learning with drifting distributions. In *International Conference on Algorithmic Learning Theory*, pages 124–138. Springer, 2012.

Ozan Sener and Silvio Savarese. Active learning for convolutional neural networks: A core-set approach. In *International Conference on Learning Representations*, 2018. URL `https://openreview.net/forum?id=H1aIuk-RW`.

Changjian Shui, Fan Zhou, Christian Gagné, and Boyu Wang. Deep active learning: Unified and principled method for query and training. In *International Conference on Artificial Intelligence and Statistics*, 2020.

Cédric Villani. The wasserstein distances. In *Optimal Transport*, pages 93–111. Springer, 2009.

Han Xiao, Kashif Rasul, and Roland Vollgraf. Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms. *preprint*, 2017.