Deep Active Learning: Unified and Principled Method for Query and Training [Supplementary Material]

Changjian Shui*[1], Fan Zhou[1], Christian Gagné[1,2], and Boyu Wang[3,4]

[1]Université Laval
[2]Canada CIFAR AI Chair
[3]University of Western Ontario
[4]Vector Institute

1 Theorem 1: Proof

Theorem 1. *Supposing $\mathcal{D}$ is the data generation distribution and $\mathcal{Q}$ is the querying distribution, if the loss ℓ is symmetric, L-Lipschitz; $\forall h \in \mathcal{H}$ is at most H-Lipschitz function and underlying labeling function $h^\star$ is $\phi(\lambda)$-$(\mathcal{D}, \mathcal{Q})$ Joint Probabilistic Lipschitz, then the expected risk w.r.t. $\mathcal{D}$ can be upper bounded by:*

$$R_{\mathcal{D}}(h) \leq R_{\mathcal{Q}}(h) + L(H + \lambda)W_1(\mathcal{D}, \mathcal{Q}) + L\phi(\lambda)$$

1.1 Notations

We define the hypothesis $h : \mathcal{X} \to \mathcal{Y} = [0, 1]$ and loss function $\ell : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}^+$, then the expected risk w.r.t. $\mathcal{D}$ is $R_{\mathcal{D}}(h) = \mathbb{E}_{x\sim\mathcal{D}}\ell(h(x), h^\star(x))$ and empirical risk $\hat{R}_{\mathcal{D}}(f) = \frac{1}{N}\sum_{i=1}^{N}\ell(h(x_i), y_i)$. We assume the loss ℓ is symmetric, L-Lipschitz and bounded by M.

1.2 Transfer risk

The first step is to bound the the gap $R_{\mathcal{D}}(h) - R_Q(h)$:

$$\begin{aligned} R_{\mathcal{D}}(h) - R_Q(h) &\leq |R_{\mathcal{D}}(h) - R_{\mathcal{Q}}(h)| = |E_{x\sim\mathcal{D}}\ell(h(x), h^\star(x)) - E_{x\sim\mathcal{Q}}\ell(h(x), h^\star(x))| \\ &= |\int_{x\in\Omega}\ell(h(x), h^\star(x))d(\mathcal{D} - \mathcal{Q})| \end{aligned} \tag{1}$$

From the Kantorovich - Rubinstein duality theorem and combing Eq. (1), for **any** distribution coupling $\gamma \in \Pi(\mathcal{D}, \mathcal{Q})$, we have:

$$\begin{aligned} &= |\int_{\Omega\times\Omega}\big(\ell(h(x_{\mathcal{D}}), h^\star(x_{\mathcal{D}})) - \ell(h(x_{\mathcal{Q}}), h^\star(x_{\mathcal{Q}}))\big)d\gamma(x_{\mathcal{D}}, x_{\mathcal{Q}})| \\ &\leq \int_{\Omega\times\Omega}|\ell(h(x_{\mathcal{D}}), h^\star(x_{\mathcal{D}})) - \ell(h(x_{\mathcal{Q}}), h^\star(x_{\mathcal{Q}}))|d\gamma(x_{\mathcal{D}}, x_{\mathcal{Q}}) \\ &\leq \int_{\Omega\times\Omega}|\ell(h(x_{\mathcal{D}}), h^\star(x_{\mathcal{D}})) - \ell(h(x_{\mathcal{D}}), h^\star(x_{\mathcal{Q}}))| + |\ell(h(x_{\mathcal{D}}), h^\star(x_{\mathcal{Q}})) - \ell(h(x_{\mathcal{Q}}), h^\star(x_{\mathcal{Q}}))|d\gamma(x_{\mathcal{D}}, x_{\mathcal{Q}}) \end{aligned}$$

Since we assume ℓ is symmetric and L-Lipschitz, then we have:

$$\leq L\int_{\Omega\times\Omega}|h^\star(x_{\mathcal{D}}) - h^\star(x_Q)|d\gamma(x_{\mathcal{D}}, x_Q) + L\int_{\Omega\times\Omega}|h(x_{\mathcal{D}}) - h(x_Q)|d\gamma(x_{\mathcal{D}}, x_Q) \tag{2}$$

*changjian.shui.1@ulaval.ca

From Eq.(2) the risk gap is controlled by two terms, the property of labeling function and property of predictor. Moreover, we assume the learner is H-Lipschitz function, then we have:

$$\leq L \int_{\Omega\times\Omega} |h^\star(x_\mathcal{D}) - h^\star(x_\mathcal{Q})| d\gamma(x_\mathcal{D}, x_\mathcal{Q}) + LH \int_{\Omega\times\Omega} ||x_\mathcal{D} - x_\mathcal{Q}||_2 d\gamma(x_\mathcal{D}, x_\mathcal{Q})$$

Labeling function assumption As mentioned before, the goodness of the underlying labeling function decides the level of risk. [1] formalize the such a property as *Probabilistic Lipschitz* condition in AL, in which relaxes the condition of Lipschitzness condition and formalizes the intuition that *under suitable feature representation the probability of two close points having different labels is small* [2]. We adopt the joint probabilistic Lipschitz property, which is coherent with [3].

Definition 1. *The labeling function $h^\star$ satisfies $\phi(\lambda)$-$(\mathcal{D}, \mathcal{Q})$ Joint Probabilistic Lipschitz if* $\mathrm{supp}(\mathcal{Q}) \subseteq \mathrm{supp}(\mathcal{D})$ *and for all $\lambda > 0$:*

$$\mathbb{P}_{(x_\mathcal{D}, x_\mathcal{Q})\sim\gamma}[|h^\star(x_\mathcal{D}) - h^\star(x_\mathcal{Q})| > \lambda \|x_\mathcal{D} - x_\mathcal{Q}\|_2] \leq \phi(\lambda) \tag{3}$$

Where $\phi(\lambda)$ reflects the decay rate. [1] showed that the faster the decay of $\phi(\lambda)$ with $\lambda \to 0$, the nicer the distribution and the easier it is to learn the task.

Combining with Eq.(3), the labeling function term can be decomposed and upper bounded by:

$$\begin{aligned}
&\leq L \int_{\Omega\times\Omega} \mathbf{1}\{|h^\star(x_\mathcal{D}) - h^\star(x_\mathcal{Q})| \leq \lambda \|x_\mathcal{D} - x_\mathcal{Q}\|_2\} |h^\star(x_\mathcal{D}) - h^\star(x_\mathcal{Q})| d\gamma(x_\mathcal{D}, x_\mathcal{Q}) \\
&+ L \int_{\Omega\times\Omega} \mathbf{1}\{|h^\star(x_\mathcal{D}) - h^\star(x_\mathcal{Q}) > \lambda \|x_\mathcal{D} - x_\mathcal{Q}\|_2\} |h^\star(x_\mathcal{D}) - h^\star(x_\mathcal{Q})| d\gamma(x_\mathcal{D}, x_\mathcal{Q}) \\
&\leq L\lambda \int_{\Omega\times\Omega} \|x_\mathcal{D} - x_\mathcal{Q}\|_2 d\gamma(x_\mathcal{D}, x_\mathcal{Q}) + L\phi(\lambda)
\end{aligned}$$

The first term is upper bounded through the probability of this event at most 1 and second term adopts the definition of Joint Probabilistic Lipschitz with restricting the output space $h^\star(\cdot) \in [0, 1]$. Plugging in the aforementioned results, we have:

$$\leq L(H + \lambda) \int_{\Omega\times\Omega} \|x_\mathcal{D} - x_\mathcal{Q}\|_2 d\gamma(x_\mathcal{D}, x_\mathcal{Q}) + L\phi(\lambda)$$

Since this inequality satisfies with any distribution coupling γ, then it is also satisfies with the optimal coupling, w.r.t. the Wasserstein-1 distance with the cost function ℓ_2 distance: $\|\cdot\|_2$. Then we have:

$$R_\mathcal{D}(h) - R_\mathcal{Q}(h) \leq L(H + \lambda) \inf_\gamma \int_{\Omega\times\Omega} \|x_\mathcal{D} - x_\mathcal{Q}\|_2 d\gamma(x_\mathcal{D}, x_\mathcal{Q}) + L\phi(\lambda)$$

Finally we can derive:

$$R_\mathcal{D}(h) \leq R_\mathcal{Q}(h) + L(H + \lambda) W_1(\mathcal{D}, \mathcal{Q}) + L\phi(\lambda) \tag{4}$$

2 Corollary 1: Proof

2.1 Basic statistical learning theory

According to the standard statistical learning theory such as [4], w.h.p. $1 - \delta/2$, $\forall h \in \mathcal{H}$ we have:

$$R_\mathcal{D}(h) \leq \hat{R}_\mathcal{D}(h) + 2L\mathrm{Rad}_N(h) + \kappa_1(\delta, N) \tag{5}$$

Where $\mathrm{Rad}_N(h) = \mathbb{E}_{S\sim\mathcal{D}^N} \mathbb{E}_{\sigma_1^N}[\sup_h \frac{1}{N} \sum_{i=1}^N \sigma_i h(x_i)]$ is the expected Rademacher complexity with $\mathrm{Rad}_N(h) = \mathcal{O}(\sqrt{\frac{1}{N}})$, and $\kappa_1(\delta, N) = \mathcal{O}(\sqrt{\frac{M \log(2/\delta)}{N}})$ is the confidence term.
In the Active learning, the goal is to control the generalization error w.r.t. $(\mathcal{D}, h^\star)$, thus from Eq.5 we have:

$$R_\mathcal{D}(h) \leq (R_\mathcal{D}(h) - R_\mathcal{Q}(h)) + \hat{R}_\mathcal{Q}(h) + 2L\mathrm{Rad}_{N_q}(h) + \kappa_1(\delta, N_q)$$

Combining with Eq.(4), we have

$$R_{\mathcal{D}}(h) \leq \hat{R}_{\mathcal{Q}}(h) + L(H+\lambda)W_1(\mathcal{D},\mathcal{Q}) + L\phi(\lambda) + 2L\mathrm{Rad}_{N_q}(h) + \kappa_1(\delta, N_q)$$

In general we have finite observations (Supposing we have the sample i.i.d. sampled from query distribution $\mathcal{Q}$) with Dirac distribution: $\hat{D} = \frac{1}{N}\sum_{i=1}^{N}\delta\{x_{\mathcal{D}}^i\}$ and $\hat{Q} = \frac{1}{N_q}\sum_{i=1}^{N_q}\delta\{x_{\mathcal{Q}}^i\}$ with $N_q \leq N$. Several recent works show the concentration bound between empirical and expected Wasserstein distance such as [5, 6]. We just adopt the conclusion from [6] and apply to bound the empirical measures in Wasserstein-1 distance.

Lemma 1. *[6] [Definition 3,4] Given a measure μ on X, the (ϵ,τ)-covering number on a given set $S \subseteq X$ is:*

$$\mathcal{N}_\epsilon(\mu,\tau) := \inf\{\mathcal{N}_\epsilon(S) : \mu(S) \geq 1-\tau\}$$

and the (ϵ,μ)-dimension is:

$$d_\epsilon(\mu,\tau) := \frac{\log \mathcal{N}_\epsilon(\mu,\epsilon)}{-\log\epsilon}$$

Then the upper Wasserstein-1 dimensions can be defined as:

$$d_1^\star(\mu) = \inf\{s \in (2,+\infty) : \limsup_{\epsilon\to 0} d_\epsilon(\mu, \epsilon^{-\frac{s}{s-2}}) \leq s\}$$

Lemma 2. *[6][Theorem 1, Proposition 20] For $p=1$ and $s \geq d_1^\star(\mu)$, there exists a positive constant C with probability at least $1-\delta$, we have:*

$$W_1(\mu,\hat{\mu}_N) \leq CN^{-1/s} + \sqrt{\frac{1}{2N}\log(\frac{1}{\delta})}$$

Since $s > 2$ thus the convergence rate of Wasserstein distance is slower than $\mathcal{O}(N^{-1/2})$, also named as *weak convergence*. Then according to the triangle inequality of Wasserstein-1 distance, we have:

$$W_1(\mathcal{D},\mathcal{Q}) \leq W_1(\mathcal{D},\hat{\mathcal{D}}) + W_1(\hat{\mathcal{D}},\mathcal{Q}) \leq W_1(\mathcal{D},\hat{\mathcal{D}}) + W_1(\hat{\mathcal{D}},\hat{\mathcal{Q}}) + W_1(\hat{\mathcal{Q}},\mathcal{Q}) \tag{6}$$

Combing the conclusion with Lemma 2, there exist some constants (C_d, s_d) and (C_q, s_q) we have with probability $1-\delta/2$:

$$W_1(\mathcal{D},\mathcal{Q}) \leq W_1(\hat{\mathcal{D}},\hat{\mathcal{Q}}) + C_d N^{-1/s_d} + C_q N_q^{-1/s_q} + \sqrt{\frac{1}{2}\log(\frac{2}{\delta})}(\sqrt{\frac{1}{N}} + \sqrt{\frac{1}{N_q}}) \tag{7}$$

Combining Eq.7 and Eq.5, we have

$$R_{\mathcal{D}}(h) \leq \hat{R}_{\mathcal{Q}}(h) + L(H+\lambda)W_1(\hat{\mathcal{D}},\hat{\mathcal{Q}}) + L\phi(\lambda) + 2L\mathrm{Rad}_{N_q}(h) + \kappa(\delta, N, N_q)$$

Where $\kappa(\delta, N, N_q) = \mathcal{O}(N^{-1/s_d} + N_q^{-1/s_q} + \sqrt{\frac{\log(1/\delta)}{N}} + \sqrt{\frac{\log(1/\delta)}{N_q}})$

3 Computing $\mathcal{H}$-divergence and Wasserstein distance

3.1 $\mathcal{H}$ divergence

$d_{\mathcal{H}}(\mathcal{D}_1,\mathcal{D}_3)$ We can discuss the discrepancy with different values of p, since $\mathcal{D}_3 \subseteq \mathcal{D}_1$ then we have $x_0 \in [a+b, 2a-b]$:

1. If $p \leq x_0 - b$, then the area of mis-classification will be $(2a-p)+2b$. If we select $p = x_0 - b$, then the optimal mis-classification area will be $2a+2b-x_0 \geq 2a+2b-2a+b = 3b$
2. If $p \in [x_0-b, x_0+b]$, then the area of mis-classification will be $(2a-p)+(p-(x_0-b)) = 2a+b-x_0 \geq 2a+b-2a+b = 2b$
3. If $p \geq x_0 + b$, then the area of mis-classification will be $2b + \max(0, 2a-p)$, if we select $p \geq 2a$, then the optimal mis-classification area will be $2b$.

Then the minimal mis-classification area is $2b$, corresponding the optimal risk $\frac{b}{a+b}$, then $d_{\mathcal{H}}(\mathcal{D}_1,\mathcal{D}_3) = \frac{b}{a+b}$

$d_{\mathcal{H}}(\mathcal{D}_1, \mathcal{D}_2)$ We can discuss the discrepancy with different values of p, since $\mathcal{D}_2 \subseteq \mathcal{D}_1$, then we have $x_0 \in [a + b/2, 2a - b/2]$:

1. If $p \leq -x_0 - b/2$, then the mis-classification area will be $a + \max(0, p + a) + 2b$ with optimal value $2a + b/2 - x_0 + 2b \geq a + 2b$;
2. If $p \in [-x_0 - b/2, -x_0 + b/2]$, then the mis-classification area will be $p - (-x_0 - b/2) + (-a - p) + a + b = x_0 + 3b/2 \geq a + 2b$;
3. If $p \in [-x_0 + b/2, x_0 - b/2]$, then the mis-classification area will be $b + a$;
4. If $p \in [x_0 - b/2, x_0 + b/2]$, then the mis-classification area will be $b + p - (x_0 - b/2) + (2a - p) = 2a + 3b/2 - x_0 \geq 2a + b/2 - 2a + b/2 + b = 2b$
5. If $p \geq x_0 + b/2$, then the mis-classification area will be $2b + \max(0, 2a - p) \geq 2b$

Then the minimal mis-classification area is $2b$, corresponding the optimal risk $\frac{b}{a+b}$, then $d_{\mathcal{H}}(\mathcal{D}_1, \mathcal{D}_2) = \frac{b}{a+b}$. From the previous example $d_{\mathcal{H}}(\mathcal{D}_1, \mathcal{D}_2) = d_{\mathcal{H}}(\mathcal{D}_1, \mathcal{D}_3)$, we show the $\mathcal{H}$ divergence is not good metric for measuring the representative in the data space. Since we want the query distribution more diverse spread in the space, then $\mathcal{H}$ may not be a good indicator.

3.2 Wasserstein-1 distance

We can also estimate the distribution distance through Wasserstein-1 metric. From [7] we have:

$$W_1(P, Q) = \int_0^1 |F^{-1}(z) - G^{-1}(z)| dz$$

where $F(z)$ and $G(z)$ is the CDF (cumulative density function) of distribution P and Q, respectively.

CDF of $\mathcal{D}_1$, $\mathcal{D}_2$ and $\mathcal{D}_3$

1.
$$F_1(z) = \begin{cases} \frac{1}{2a}(z + 2a) & -2a \leq z \leq -a \\ \frac{1}{2} & -a \leq z \leq a \\ \frac{1}{2a} z & a \leq z \leq 2a \end{cases}$$

$$F_1^{-1}(z) = \begin{cases} 2a(z - 1) & 0 \leq z < 1/2 \\ [-a, a] & z = 1/2 \\ 2az & 1/2 < z \leq 1 \end{cases}$$

2.
$$F_2(z) = \begin{cases} \frac{1}{2b}(z + x_0 + b/2) & -x_0 - b/2 \leq z \leq -x_0 + b/2 \\ \frac{1}{2} & -x_0 + b/2 \leq z \leq x_0 - b/2 \\ \frac{1}{2b}(z - x_0 + 3b/2) & x_0 - b/2 \leq z \leq x_0 + b/2 \end{cases}$$

$$F_2^{-1}(z) = \begin{cases} 2bz - x_0 - b/2 & 0 \leq z < 1/2 \\ [-x_0 + b/2, x_0 - b/2] & z = 1/2 \\ 2bz + x_0 - 3b/2 & 1/2 < z \leq 1 \end{cases}$$

3.
$$F_3(z) = \frac{1}{2b}(z - x_0 + b) \;\; z \in [x_0 - b, x_0 + b]$$

$$F_3^{-1}(z) = 2bz + x_0 - b \;\; z \in [0, 1]$$

Computing $W_1(\mathcal{D}_1, \mathcal{D}_2)$ According to the definition, we can compute

$$W_1(\mathcal{D}_1, \mathcal{D}_2) = \int_0^{1/2} |2a(z-1) - 2bz - x_0 - \frac{b}{2}|dz + \int_{1/2}^{1} |2az - 2bz - x_0 + \frac{3}{2}b|dz$$

We firstly compute $\int_0^{1/2} |2a(z-1) - 2bz - x_0 - \frac{b}{2}|dz$, since $2a(z-1) - 2bz - x_0 - \frac{b}{2} < 0$ for $z \in [0, 1/2]$ (since $-a - b - x_0 - b/2 < 0$. Then we have:

$$\begin{aligned}\int_0^{1/2} |2a(z-1) - 2bz - x_0 - \frac{b}{2}|dz &= \int_0^{1/2} \{-2a(z-1) + 2bz + x_0 + \frac{b}{2}\}dz \\ &= \frac{3}{4}a + \frac{1}{2}x_0 + \frac{1}{2}b\end{aligned}$$

Then we compute the second part:

$$\begin{aligned}&\int_{1/2}^{1} |2az - 2bz - x_0 + \frac{3}{2}b|dz \\ &= \int_{1/2}^{z_0} \{(x_0 - \frac{3}{2}b) - 2(a-b)z\}dz + \int_{z_0}^{1} \{2(a-b)z - x_0 + \frac{3}{2}b\}dz \\ &= \frac{1}{2(a-b)}(x_0 - \frac{3}{2}b)^2 - \frac{3}{2}(x_0 - \frac{3}{2}b) + \frac{3}{4}(a-b)\end{aligned}$$

with $z_0 = \frac{x_0 - 3b/2}{2(a-b)}$. Therefore we can compute the wasserstein-1 distance between distribution $\mathcal{D}_1$ and $\mathcal{D}_2$:

$$= \frac{1}{2(a-b)}(x_0 - \frac{3}{2}b)^2 - x_0 + 2b + \frac{3}{2}a$$

With $x_0 \in [a + b/2, 2a - b/2]$. If we take $x_0 = 2a - b/2$, we can get the maximum:

$$\max_{x_0} W_1(\mathcal{D}_1, \mathcal{D}_2) = \frac{3}{2}a - \frac{b}{2}$$

Computing $W_1(\mathcal{D}_1, \mathcal{D}_3)$ According to definition, we can compute

$$W_1(\mathcal{D}_1, \mathcal{D}_3) = \int_0^{1/2} |2a(z-1) - 2bz - x_0 + b|dz + \int_{1/2}^{1} |2az - 2bz - x_0 + b|dz$$

We firstly compute $\int_0^{1/2} |2a(z-1) - 2bz - x_0 + b|dz$, since $2(a-b)z - 2a - x_0 + b \leq 0$ for $z \in [0, 1/2]$. (easy to verify: $2(a-b)z - 2a - x_0 + b \leq (a-b) - 2a - x_0 + b = -a - x_0 < 0$), then

$$\begin{aligned}&\int_0^{1/2} |2a(z-1) - 2bz - x_0 + b|dz = \int_0^{1/2} (x_0 + 2a - b) - 2(a-b)zdz \\ &= \frac{1}{2}(x_0 + 2a - b) - \frac{1}{4}(a-b) = \frac{3}{4}a + x_0 - \frac{1}{4}b\end{aligned}$$

Then we compute the second term $\int_{1/2}^{1} |2az - 2bz - x_0 + b|dz$, we define $z_0 = \frac{x_0 - b}{2(a-b)}$ and we can verify that $z_0 \in [1/2, 1]$, then this term can be decomposed as we can rewrite it as:

$$\int_{1/2}^{z_0} -2(a-b)z + (x_0 - b)dz + \int_{z_0}^{1} 2(a-b)z - (x_0 - b)dz$$

$$= \frac{(x_0 - b)^2}{2(a-b)} - \frac{3}{2}(x_0 - b) + \frac{5}{4}(a-b)$$

Then $W_1(\mathcal{D}_1, \mathcal{D}_3) = \frac{(x_0-b)^2}{2(a-b)} - \frac{3}{2}(x_0 - b) + \frac{5}{4}(a-b) + \frac{1}{2}x_0 + \frac{3}{4}a - \frac{b}{4} = \frac{1}{2(a-b)}(x_0 - b)^2 - x_0 + 2a = \frac{1}{2(a-b)}(x_0 - b)^2 - x_0 + 2a$ since $x_0 \in [a + b, 2a - b]$, then we have:

$$\min_{x_0} W_1(\mathcal{D}_1, \mathcal{D}_3) = \frac{(x_0 - b)^2}{2(a-b)} - (a+b) + 2a = \frac{a^2}{2(a-b)} + a - b$$

We can verify: $\frac{a^2}{2(a-b)} + a - b > \frac{3}{2}a - \frac{b}{2}$ when $a > b$, then we have:

$$\min_{x_0} W_1(\mathcal{D}_1, \mathcal{D}_3) > \max_{x_0} W_1(\mathcal{D}_1, \mathcal{D}_2)$$

which means in Wasserstein-1 distance metric, the diversity of two distribution can be much better measured.

4 Developing loss in deep batch active learning

We have the original loss:

$$\min_{\boldsymbol{\theta}^f, \boldsymbol{\theta}^h, \hat{B}} \max_{\boldsymbol{\theta}^d} \mathbb{E}_{(x,y)\sim\hat{L}\cup\hat{B}} \ell(h(x,y)) + \mu\big(\mathbb{E}_{x\sim\hat{\mathcal{D}}}[g(x)] - \mathbb{E}_{x\sim\hat{L}\cup\hat{B}}[g(x)]\big). \tag{8}$$

Since $\hat{L}$, $\hat{B}$ and $\hat{\mathcal{D}}$ are Dirac distributions, then we have:

$$\begin{aligned}
&\frac{1}{L+B}\sum_{(x,y)\in\hat{L}\cup\hat{B}} \ell(h(x,y)) + \mu\big(\frac{1}{L+U}\sum_{x\in\hat{D}} g(x) - \frac{1}{L+B}\sum_{x\in\hat{L}\cup\hat{B}} g(x)\big) \\
&= \frac{1}{L+B}\sum_{(x,y)\in\hat{L}} \ell(h(x,y)) + \frac{1}{L+B}\sum_{(x,y^?)\in\hat{B}} \ell(h(x,y^?)) \\
&+ \mu\big(\frac{1}{L+U}\sum_{x\in\hat{L}} g(x) + \frac{1}{L+U}\sum_{x\in\hat{U}} g(x) - \frac{1}{L+B}\sum_{x\in\hat{L}} g(x) - \frac{1}{L+B}\sum_{x\in\hat{B}} g(x)\big) \\
&= \frac{1}{L+B}\sum_{(x,y)\in\hat{L}} \ell(h(x,y)) + \frac{1}{L+B}\sum_{(x,y^?)\in\hat{B}} \ell(h(x,y^?)) \\
&+ \mu\big(\frac{1}{L+U}\sum_{x\in\hat{U}} g(x) - (\frac{1}{L+B} - \frac{1}{L+U})\sum_{x\in\hat{L}} g(x)\big) - \frac{\mu}{L+B}\sum_{x\in\hat{B}} g(x) \\
&= \underbrace{\Big(\frac{1}{L+B}\sum_{(x,y)\in\hat{L}} \ell(h(x,y)) + \mu\big(\frac{1}{L+U}\sum_{x\in\hat{U}} g(x) - (\frac{1}{L+B} - \frac{1}{L+U})\sum_{x\in\hat{L}} g(x)\big)\Big)}_{\text{Training Stage}} \\
&+ \underbrace{\Big(\frac{1}{L+B}\sum_{(x,y^?)\in\hat{B}} \ell(h(x,y^?)) - \frac{\mu}{L+B}\sum_{x\in\hat{B}} g(x)\Big)}_{\text{Querying Stage}}
\end{aligned}$$

We note that $x \in \hat{D}$ means enumerating all samples from the observations (empirical distribution).

5 Redundancy trick: Computation

$$\begin{aligned}
&\mu\big(\frac{1}{L+U}\sum_{x\in\hat{U}} g(x) - (\frac{1}{L+B} - \frac{1}{L+U})\sum_{x\in\hat{L}} g(x)\big) \\
&= \mu\big(\frac{\gamma}{1+\gamma}\frac{1}{U}\sum_{x\in\hat{U}} g(x) - (\frac{1}{1+\alpha} - \frac{1}{1+\gamma})\frac{1}{L}\sum_{x\in\hat{L}} g(x)\big) \\
&= \mu'\big(\frac{1}{U}\sum_{x\in\hat{U}} g(x) - \frac{1}{\gamma}(\frac{1+\gamma}{1+\alpha} - 1)\frac{1}{L}\sum_{x\in\hat{L}} g(x)\big) \\
&= \mu'\big(\frac{1}{U}\sum_{x\in\hat{U}} g(x) - \frac{1}{\gamma}\frac{\gamma-\alpha}{1+\alpha}\frac{1}{L}\sum_{x\in\hat{L}} g(x)\big)
\end{aligned}$$

6 Uniform Output Arrives the Minimal loss

For the abuse of notation, we suppose the output of classifier $h(x, \cdot) = [p_1, \ldots, p_K] \equiv \mathbf{p}$ with $p_i > 0$ and $\sum_{i=1}^{K} p_i = 1$. Then we tried to minimize

$$\min_{\mathbf{p}} \sum_{i=1}^{K} -\log p_i$$

By applying the Lagrange Multiplier approach, we have

$$\min_{\mathbf{p}, \lambda > 0} \sum_{i=1}^{K} -\log p_i + \lambda (\sum_{i=1}^{K} p_i - 1)$$

Then we do the partial derivative w.r.t. p_i, then we have $\forall i$:

$$\frac{-1}{p_i} + \lambda = 0 \ \rightarrow \ p_i = \frac{1}{\lambda}$$

Given $\sum_{i=1}^{K} p_i = 1$, then we can compute $p_i = \frac{1}{K}$ arriving the minimial, i.e the uniform distribution.

7 Experiments

7.1 Dataset Descriptions

Dataset	#Classes	Train + Validation	Test	Initially labelled	Query size	Image size
Fashion-MNIST [8]	10	40K + 20K	10K	1K	500	28×28
SVHN [9]	10	40K + 33K	26K	1K	1K	32×32
CIFAR10 [10]	10	45K + 5K	10K	2K	2K	32×32
STL10* [11]	10	8K + 1K	4K	0.5K	0.5K	96×96

Table 1: Dataset descriptions

*We used a variant instead of the original STL10 dataset with arranging the training size to 8K (each class 800) and validation 1K and test 4K. We do not use the unlabeled dataset in our training or test procedure.

7.2 Implementation details

FashionMNIST For the FashionMNIST dataset, we adopted the LeNet5 as feature extractor, then we used two-layer MLPs for the classification (320-50-relu-dropout-10) and critic function (320-50-relu-dropout-1-sigmoid).

SVHN, CIFAR10 We adopt the VGG16 with batch normalization as feature extractor. then we used two-layer MLPs for the classification (512-50-relu-dropout-10) and critic function (512-50-relu-dropout-1-sigmoid).

STL10 We adopt the VGG16 with batch normalization as feature extractor. then we used two-layer MLPs for the classification (4096-100-relu-dropout-10) and critic function (4096-100-relu-dropout-1-sigmoid).

7.3 Hyper-parameter setting

Dataset	lr	Momentum	Mini-Batch size	μ	Selection coefficient	Mixture coefficient**
Fashion-MNIST	0.01*	0.5	64	1e-2	5	0.5
SVHN	0.01*	0.5	64	1e-2	5	0.5
CIFAR10	0.01*	0.3	64	1e-2	10	0.5
STL10	0.01*	0.3	64	1e-3	10	0.5

Table 2: Hyper-parameter setting

* We set the initial learning rate as 0.01, then at 50% epoch we decay to 1e-3, after 75% epoch we decay to 1e-4.

** The mixture coefficient means the convex combination coefficient in the two uncertainty based approach.

	Random	LeastCon	Margin	Entropy	KMedian	DBAL	Core-set	DeepFool	WAAL
1K	58.03±2.81	57.93±1.62	57.81±2.19	57.40±1.75	57.62±2.5	58.01±2.75	58.14±2.19	58.19±2.4	72.29±1.16
1.5K	66.81±1.02	64.24±2.49	65.61± 2.5	66.37±0.62	67.13±2.87	66.53±2.5	68.79±1.99	66.21±1.78	76.99±1.05
2K	71.21±2.35	68.36±1.09	70.05±2.77	69.70±0.88	71.57±0.79	69.77±0.93	71.22±1.38	70.14±1.32	79.85±0.49
2.5K	73.12±2.1	71.68±1.67	72.74±1.55	71.60±1.42	73.84±0.98	72.60±0.60	72.61±1.16	71.77±1.49	81.08±0.68
3K	75.80±0.64	75.03±1.56	76.55±1.01	74.84±1.29	75.79±0.44	74.75±1.04	73.77±1.74	73.69±1.21	82.04±0.58
3.5K	77.34±0.67	77.73±1.04	78.99±1.11	76.66±1.26	77.44±0.97	75.86±1.02	75.10±1.11	74.00±0.71	82.74±0.79
4K	78.68±0.41	79.26±0.47	81.77±0.51	79.00±0.24	77.97±0.65	77.02±0.42	76.28 ±0.98	74.93±2.05	83.25±0.62
4.5K	79.58±0.47	80.08±0.82	82.32±0.47	79.89±0.78	79.49±0.7	77.90±0.58	77.30±0.61	76.64±0.97	83.96±0.54
5K	80.02±0.45	81.32±0.64	83.89±0.84	80.85±0.87	79.97±0.59	78.87±0.58	78.34±0.37	77.24±0.69	84.45±0.45
5.5K	80.93±0.33	83.21±0.42	84.87±0.18	82.26±0.77	81.11±0.41	79.47±2.9	78.42±0.66	77.72±0.57	85.20±0.44
6K	81.30±0.25	84.50±0.73	85.52±0.27	83.66±0.98	81.86±0.6	80.43±0.76	79.66±0.34	78.99±0.33	85.99±0.43

Table 3: Result of FashionMNIST (Average ± std)

	Random	LeastCon	Margin	Entropy	KMedian	DBAL	Core-set	DeepFool	WAAL
1K	63.97±2.04	63.40±2.16	63.10±2.3	63.49±2.79	63.50±2.53	63.76±0.73	63.90±1.07	63.62±2.34	75.18±1.41
2K	75.85±1.16	74.86±2.44	75.27± 1.7	72.78±3.15	76.17±3.2	77.07±1.57	77.9±1.25	76.29±1.62	80.69±2.00
3K	80.83± 1.04	81.87± 0.64	80.9± 2.22	80.88± 1.26	81.36± 1.29	81.17± 1.72	81.7 ± 0.84	80.92± 0.79	83.89± 2.08
4K	82.70±1.18	84.00±0.88	83.10±1.38	83.19±0.95	83.41±1.58	83.95±1.87	84.81±1.3	83.79±0.64	86.82±1.11
5K	85.10±0.73	85.68±0.94	85.02±1.1	84.75±0.83	84.93±0.94	86.34±1.1	86.52±0.95	85.32±0.58	88.71±1.08
6K	86.20±0.48	87.23±0.97	87.53±0.63	87.51±0.50	87.04±0.45	87.61±0.72	88.00 ±0.44	87.02±0.64	89.71±0.83

Table 4: Result of SVHN (Average ± std)

	Random	LeastCon	Margin	Entropy	KMedian	DBAL	Core-set	DeepFool	WAAL
2K	46.33±3.18	46.43±3.17	46.69±3.87	46.79±3.62	46.53±3.39	46.48±3.11	46.38±4.03	46.54±3.77	55.00±0.40
4K	56.33±3.40	53.26±3.84	55.52± 2.69	53.13±2.99	53.58±2.57	56.18±2.37	56.09±3.89	54.48±1.62	62.32±0.36
6K	59.63± 4.17	59.00± 2.19	63.05± 1.78	62.63± 1.29	61.25±1.76	62.48± 1.38	59.56 ± 1.17	60.80± 0.70	66.67± 0.60
8K	62.85±3.37	66.46±1.33	66.44±1.85	65.23±1.89	63.73±1.34	65.84±0.78	65.84±1.27	64.87±1.98	69.33±1.47
10K	68.13±2.53	68.91±1.10	69.86±0.24	69.72±1.53	68.92±2.33	68.94±1.96	69.11±0.80	69.39±0.47	72.39±1.21
12K	70.41±1.02	71.90±1.35	72.25±0.68	71.58±0.77	72.65±0.64	72.25±1.24	72.60 ±0.79	71.17±1.03	75.11±0.49

Table 5: Result of CIFAR10 (Average ± std)

	Random	LeastCon	Margin	Entropy	KMedian	DBAL	Core-set	DeepFool	WAAL
0.5K	41.78±2.42	41.69±3.22	41.81±2.27	41.12±1.67	41.24±1.41	41.30±1.45	41.41±2.30	41.82±2.67	47.01±1.09
1K	48.24±1.37	47.05±1.42	46.7±0.85	46.38±2.31	46.45±1.11	47.45±3.71	47.58±2.06	45.15±0.74	52.47±1.62
1.5K	51.78± 2.5	50.87± 1.24	50.44± 2.57	50.24± 1.21	49.91±1.74	52.53± 1.29	51.2 ± 1.63	48.64± 2.43	57.25± 1.78
2K	56.52±1.78	56.25±1.58	55.54±1.09	55.15±2.13	54.92±2.19	57.54±1.70	58.13±1.57	54.26±2.40	60.08±1.63
2.5K	58.42±1.42	58.49±2.05	57.62±1.42	57.81±2.87	57.87±1.51	59.25±2.89	57.66±1.79	57.05±2.53	62.58±1.44
3K	61.13±1.67	60.80±2.64	59.42±1.49	60.88±0.72	60.00±0.65	62.11±1.65	61.02 ±0.48	59.74±1.74	65.42±1.33

Table 6: Result of STL10 (Average ± std)

7.4 Detailed results with numerical values

We report the accuracy in the form of percentage (%), showing in Tab. 3, 4, 5, 6.

8 Ablation study

In this part, we will conduct $\mathcal{H}$-divergence based adversarial training for the parameters of DNN.

$$\min_{\boldsymbol{\theta}^f,\boldsymbol{\theta}^h}\max_{\boldsymbol{\theta}^d}\sum_{(x,y)\in\hat{L}}\ell(h(x,y))-\mu\big(\sum_{x\in\hat{U}}\log(g(x))+\sum_{x\in\hat{L}}\log(1-g(x))\big) \tag{9}$$

Where the g is defined as the discriminator function [1]. In the adversarial training, the discriminator parameter aims at discriminating the empirical unlabeled and labeled data via the binary classification, while the feature

[1] This notation is slightly different from the critic function [12]

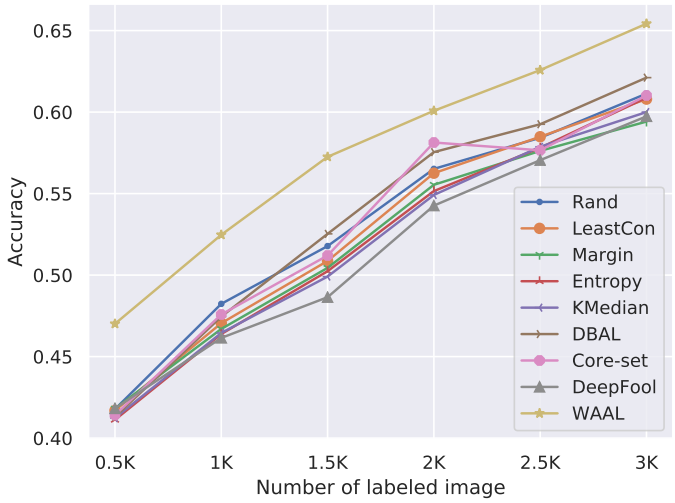

Figure 1: Empirical performance on STL10, over five repetitions.

	Random	LeastCon	Margin	Entropy	KMedian	DBAL	Core-set	DeepFool	WAAL
2K	49.85±0.32	50.00±1.81	49.8±2.28	49.92±1.08	49.31±2.76	49.58±1.05	49.87±4.03	49.85±1.36	55.00±0.40
4K	56.63±3.27	59.11±0.85	61.93± 2.12	59.15±0.41	60.6±0.72	58.55±1.99	60.97±1.62	58.80±2.59	62.32±0.36
6K	62.30± 2.54	63.15± 2.21	63.04± 1.98	63.74± 0.94	64.73±1.37	63.82± 2.33	64.95 ± 1.66	64.80± 1.4	66.67± 0.60
8K	66.97±0.76	64.32±2.58	68.30±1.02	67.67±1.04	65.98±1.45	66.65±1.00	67.54±2.16	67.65±1.27	69.33±1.47
10K	69.23±1.97	69.74±2.52	69.98±0.25	69.92±1.17	70.95±1.93	69.96±1.74	70.62±0.74	70.55±0.80	72.39±1.21
12K	71.78±1.34	71.60±1.25	71.56±1.53	72.90±1.37	72.56±1.39	73.53±1.71	71.83 ±1.20	71.86±0.33	75.11±0.49

Table 7: Ablation study of CIFAR10 (Average ± std)

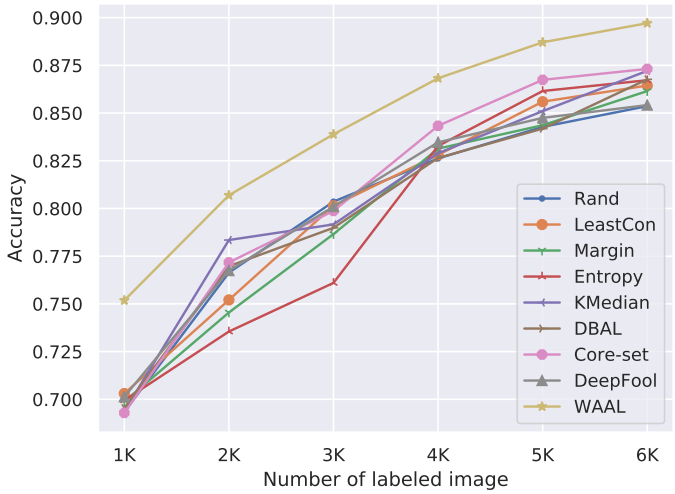

Figure 2: Ablation study in SVHN: the baselines are all trained by leveraging the unlabeled information through $\mathcal{H}$-divergence.

extractor parameter aims at not being correctly classified. By this manner, the unlabeled dataset will be used for constructing a better feature representation in the adversarial training. As for the query part, we directly used baseline strategies. The numerical values will show in Tab. 7. Moreover, we also evaluated the ablation study for the SVHN dataset, showing in Tab. 8 and Fig. 2.

	Random	LeastCon	Margin	Entropy	KMedian	DBAL	Core-set	DeepFool	WAAL
1K	68.34±0.96	70.3±1.75	68.88±1.19	68.94±1.17	68.38±0.92	68.54±3.65	69.29±0.71	70.14±1.84	75.18±1.41
2K	76.63±3.14	75.21±2.45	74.55± 3.16	73.55±2.49	78.35±1.63	76.97±1.19	77.17±1.8	76.74±2.15	80.69±2.00
3K	80.36± 0.46	80.14±1.66	78.66± 1.54	76.10± 1.46	79.16± 1.47	78.99± 1.57	79.87 ± 0.33	80.10± 1.27	83.89± 2.08
4K	82.62±1.15	82.81±0.66	83.13±1.01	83.27±0.18	82.89±0.73	82.65±1.61	84.33±0.72	83.47±0.74	86.82±1.11
5K	84.27±0.77	85.59±0.74	84.36±0.75	86.15±0.23	85.10±0.57	84.18±0.25	86.74±0.34	84.75±0.57	88.71±1.08
6K	85.36±0.36	86.44±0.93	86.15±0.89	86.72 ± 0.66	87.21±0.52	86.77±1.26	87.31±0.71	85.42 ±0.55	89.71±0.83

Table 8: Ablation study of SVHN (Average ± std)

References

[1] Ruth Urner, Sharon Wulff, and Shai Ben-David. Plal: Cluster-based active learning. In *Conference on Learning Theory*, pages 376–397, 2013.

[2] Ruth Urner and Shai Ben-David. Probabilistic lipschitzness a niceness assumption for deterministic labels. In *NIPS 2013*, 2013.

[3] Nicolas Courty, Rémi Flamary, Amaury Habrard, and Alain Rakotomamonjy. Joint distribution optimal transportation for domain adaptation. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 3730–3739. Curran Associates, Inc., 2017.

[4] Mehryar Mohri, Afshin Rostamizadeh, and Ameet Talwalkar. *Foundations of machine learning*. MIT Press, 2018.

[5] François Bolley, Arnaud Guillin, and Cédric Villani. Quantitative concentration inequalities for empirical measures on non-compact spaces. *Probability Theory and Related Fields*, 137(3-4):541–593, 2007.

[6] Jonathan Weed and Francis Bach. Sharp asymptotic and finite-sample rates of convergence of empirical measures in wasserstein distance. *arXiv preprint arXiv:1707.00087*, 2017.

[7] Larry Wasserman. Lecture note: Statistical methods for machine learning, 2019.

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

[9] Yuval Netzer, Tao Wang, Adam Coates, Alessandro Bissacco, Bo Wu, and Andrew Y Ng. Reading digits in natural images with unsupervised feature learning. 2011.

[10] Alex Krizhevsky et al. Learning multiple layers of features from tiny images. Technical report, Citeseer, 2009.

[11] Adam Coates, Andrew Ng, and Honglak Lee. An analysis of single-layer networks in unsupervised feature learning. In *Proceedings of the fourteenth international conference on artificial intelligence and statistics*, pages 215–223, 2011.

[12] Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein gan. *arXiv preprint arXiv:1701.07875*, 2017.