Appendix

Table of Contents

A Appendix Overview **13**

B Summary of Notations **13**

C Proof and Additional Analysis of Main Theoretical Results **13**
- C.1 Proof of Theorem 1 . . . 13
  - Corollary 1 . . . 14
- C.2 Proposition 1 . . . 15
- C.3 Proof of Theorem 2 and Active Feedback Analysis . . . 16

D Additional Experiment Results **17**
- D.1 Synthetic Experiment . . . 17
- D.2 Real-world Experiments . . . 18
  - D.2.1 Experimental Settings . . . 18
  - D.2.2 Random Hold-out Test and Random Feedback . . . 19
  - D.2.3 Additional Active Feedback Comparisons . . . 19
  - D.2.4 Feedback Size Study . . . 20
  - D.2.5 Single Feedback Round Comparison . . . 20
  - D.2.6 Early Stopping in AL . . . 21

E Details of Hardware for Experiments **22**

F Limitation, Future work, and Social Impact **22**
- F.1 Limitation and Future Directions . . . 22
  - Specifying AL Strategies . . . 22
  - Feedback Size Control and Balancing the Learning-Testing Objective . . . 22
  - Connecting Learning-Testing-Feedback Proposals . . . 22
- F.2 Social Impact . . . 22

G Source Code **22**

A Appendix Overview

Organization. The Appendix is organized as follows. We first provide a summary of notations in Appendix B. Then, we present the details of the theoretical analysis in the main paper in Appendix C. Next, we show additional experimental results in Appendix D, including a synthetic visualization for different feedback approaches in Appendix D.1, the risk and estimation error comparison on real-world datasets in Appendix D.2.3 and the early stopping experiments in Appendix D.2.6. We discuss the limitation, future direction, and social impact of the proposed work in Appendix F. We provide the link to the source code in Appendix G.

B Summary of Notations

Table 4: Summary of key notations with definitions

Notation	Definition
$(\mathbf{x}, y) \in \mathcal{X} \times \mathcal{Y}$	Data points
$D; p(\mathbf{x}, y)$	Data distribution
$\mathcal{L}_\theta$	Loss function of model $f_\theta()$
R	True risk
$R(f_{\theta\|D})$	True risk evaluated on model f whose parameter θ is learned from dataset D.
$\mathcal{S}_L, \mathcal{S}_U$	Labeled set and unlabeled pool
$\mathcal{Q}_t = \{\mathbf{x}_t\}^{n_t}$	The t-th quiz set
$q(\mathbf{x}), q^*(\mathbf{x})$	Test sample selection proposal and the optimal proposal
$\hat{R}_q, \hat{R}_t$	Risk estimator indexed by the test proposal q or time step t
$\tilde{R}$	Integrated risk estimator
C_t, v_t	Model confidence of f_t and the weight coefficient for time step t in final $\tilde{R}$
$\mathcal{S}_{\text{FB}}$	Active feedback set
N_L, N_T, N_{FB}	Number of samples in learning, testing and feedback sets
$d(\cdot,\cdot), A_L, \epsilon$	Diversity metric, diversity norm matrix, small positive value ϵ to avoid singular issues
$q_{\text{FB}}(\mathbf{x}), \eta$	Feedback proposal, balancing parameter between the proposal-loss term and the diversity term in the feedback proposal
λ	Balancing parameter for the risk estimation in unlabeled-information-combined early stopping criterion

C Proof and Additional Analysis of Main Theoretical Results

C.1 Proof of Theorem 1

Proof. We start by presenting the asymptotic convergence of the active risk estimator and the solution for the optimal testing selection proposal $q^*(\mathbf{x})$. From [19], we know that using the risk estimator $\hat{R}_{n,q}$ we would get an unbiased estimate of the true risk R because it is essentially an importance sampling based estimator. Then from the central limit theorem, $\hat{R}^0_{n,q} = \sum_{i=1}^n w^{(i)} l^{(i)}$ and $W_n = \sum_{i=1}^n w^{(i)}$ are asymptotically normally distributed with

$$\sqrt{n}\left(\frac{1}{n}\hat{R}^0_{n,q} - R\right) \xrightarrow{n\to\infty} \mathcal{N}(0, \text{var}[w^{(i)} l^{(i)}]) \tag{11}$$

$$\sqrt{n}\left(\frac{1}{n}W_n - 1\right) \xrightarrow{n\to\infty} \mathcal{N}(0, \text{var}[w^{(i)}]) \tag{12}$$

Then, with the multivariate delta method, we know that if $Y_n = (Y_{n1}, ..., Y_{nk})$ is a sequence and $\sqrt{n}(Y_n - \mu) \xrightarrow{n\to\infty} \mathcal{N}(0, \Sigma)$, then

$$\sqrt{n}(g(Y_n) - g(\mu)) \xrightarrow{n\to\infty} \mathcal{N}(0, \triangledown g(y)^\top \Sigma \triangledown g(y)) \tag{13}$$

Here the function is $g(x,y)=\frac{x}{y}$ with $x=\frac{1}{n}\hat{R}^0_{n,q}$ and $y=\frac{1}{n}W_n$. The result is

$$\sqrt{n}\left(\frac{\frac{1}{n}\hat{R}^0_{n,q}}{\frac{1}{n}W_n}-R\right)\xrightarrow{n\to\infty}\mathcal{N}(0,\sigma_q^2) \tag{14}$$

where $\sigma_q^2=\int\frac{p(\mathbf{x})}{q(\mathbf{x})}\left(\int[\mathcal{L}(f(\mathbf{x}),y)-R(f)]^2p(y|\mathbf{x})dy\right)p(\mathbf{x})\mathrm{d}\mathbf{x}$.

Then, the optimal test proposal is obtained by minimizing σ_q^2. By introducing a Lagrange multiplier β for the constraint $\int q(\mathbf{x})d\mathbf{x}=1$, we have

$$L(q,\beta)=\sigma_q^2+\beta\left(\int q(\mathbf{x})d\mathbf{x}-1\right) \tag{15}$$

$$\frac{\partial L}{\partial q}=-\frac{p(\mathbf{x})^2\int[\mathcal{L}(f(\mathbf{x}),y)-R(f)]^2p(y|\mathbf{x})dy}{q(\mathbf{x})^2}+\beta=0 \tag{16}$$

Thus, we have $q^*(\mathbf{x})\propto p(\mathbf{x})\sqrt{\int[\mathcal{L}(f(\mathbf{x}),y)-R(f)]^2p(y|\mathbf{x})\mathrm{d}y}$.

Now, we provide the detailed proof for Theorem 1. As shown in Section 3.4, $\widehat{\mathbf{R}}$ satisfies

$$\sqrt{n_t}(\widehat{\mathbf{R}}-R\mathbb{1})\sim\mathcal{N}\left(\mathbf{0},\mathrm{diag}\left[\sigma_1^2,..\sigma_T^2)^\top\right]\right) \tag{17}$$

Next, we apply the multi-variant delta method. Define $g:\mathbb{R}^T\to\mathbb{R}$, $g(\widehat{\mathbf{R}})=\sum_{t=1}^T v_t\widehat{R}_{\mathcal{Q}_t}$. Then, we have $\bigtriangledown g=(v_1,...,v_t)^\top$. Given the diagonal covariance matrix, the final variance is:

$$\begin{aligned}\sigma_T^2&=(v_1,...,v_t)\begin{pmatrix}\sigma_1^2&&\\&...&\\&&\sigma_T^2\end{pmatrix}(v_1,...,v_t)^\top\\&=\sum_{t=1}^T\int\frac{p(\mathbf{x})}{q_t(\mathbf{x})}v_t^2\left(\int[\mathcal{L}(f_T(\mathbf{x}),y)-R(f_T)]^2p(y|\mathbf{x})dy\right)p(\mathbf{x})d\mathbf{x}\end{aligned} \tag{18}$$

When we perform the "final exam" estimation after gathering all quizzes $\{\mathcal{Q}_1,...,\mathcal{Q}_T\}$, the other factors including testing proposals are fixed. We analyze the optimal solution for v_t by constructing the Lagrangian objective $\sigma_T^2+\gamma(\sum_t v_t-1)$ (where γ is a Lagrangian multiplier). By taking the derivative w.r.t each v_t along with the Lagrangian, we have

$$\frac{\partial[\sum_{t=1}^T v_t^2(\sigma_t^2)+\gamma(v_t-1/T)]}{\partial v_t}=0 \tag{19}$$

which leads to $v_t=\frac{C_t}{\sum_{t=1}^T C_t}$. □

The Corollary below provides an alternative view of Theorem 1.

Corollary 1. *If we do not change individual q_t but still combine all available test samples, then adjusting their importance weight by ${w'}_t^{(i)}=v_t\times w_t^{(i)}$ gives the optimal estimator.*

Proof. In the alternative view, we have:

$$\widetilde{R}=\frac{\widetilde{R}^0}{W'}=\frac{\sum_{t=1}^T\sum_{i=1}^{n_t}v_tw_t^{(i)}l_t^{(i)}}{\sum_{t=1}^T\sum_{i=1}^{n_t}v_tw_t^{(i)}} \tag{20}$$

where $w_i^{(t)}=\frac{p(\mathbf{x}^{(i)})}{q_t(\mathbf{x}^{(i)})}$. We can view the final estimate $\widetilde{R}$ as a function of $\widetilde{R}^0$ and W' that has the form $f(X,Y)=\frac{X}{Y}$. Then we directly analyze the expectation and variance of $\widetilde{R}$ using the delta method: First we have

$$\begin{aligned}\mathbb{E}(f(X,Y))&=\mathbb{E}[f(\mu_X,\mu_Y)+f'_Y(\mu_X,\mu_Y)(X-\mu_X)+f'_Y(\mu_X,\mu_Y)(Y-\mu_Y)+R]\\&\approx\mathbb{E}[f(\mu_X,\mu_Y)]+\mathbb{E}[f'_X(\mu_X,\mu_Y)(X-\mu_X)]+\mathbb{E}[f'_Y(\mu_X,\mu_Y)(Y-\mu_Y)]\\&=\mathbb{E}[f(\mu_X,\mu_Y)]+f'_X(\mu_X,\mu_Y)\mathbb{E}[(X-\mu_X)]+f'_Y(\mu_X,\mu_Y)\mathbb{E}[(Y-\mu_Y)]\\&=f(\mu_X,\mu_Y)\end{aligned} \tag{21}$$

where $\mu_X = \mathbb{E}[X]$ and $\mu_Y = \mathbb{E}[Y]$. Applying (21) on our estimate, and we get:

$$\mathbb{E}[\tilde{R}(f_T)] = \mathbb{E}\left[\frac{\sum_{t=1}^{T}\sum_{i=1}^{n} v_t w_t^{(i)} l_{it}}{\sum_{t=1}^{T}\sum_{i=1}^{n} v_t w_t^{(i)}}\right] = \frac{\sum_{t=1}^{T} v_t \mathbb{E}[\tilde{R}_t^0]}{\sum_{t=1}^{T} v_t W_{n,t}} = R$$

where we utilize $\sum_{t=1}^{T} v_t = 1$ and $\mathbb{E}[\frac{\widetilde{R}_t^0}{W_t'}] = R$. For the variance, we have:

$$\begin{aligned}
\text{Var}[f(X,Y)] &= \mathbb{E}[(f(X,Y) - \mathbb{E}[f(X,Y)])^2] \\
&\approx \mathbb{E}[(f(X,Y) - f(\mu_X,\mu_Y))^2] \\
&\approx \mathbb{E}[(f(\mu_X,\mu_Y) + f_X'(\mu_X,\mu_Y)(X-\mu_X) + f_Y'(\mu_X,\mu_Y)(Y-\mu_Y) - f(\mu_X,\mu_Y))^2] \\
&= \mathbb{E}[f_X'^2(\mu_X,\mu_Y)(X-\mu_X)^2 + 2f_X'(\mu_X,\mu_Y)(X-\mu_X)f_Y'(\mu_X,\mu_Y)(Y-\mu_Y) \\
&\quad + f_Y'^2(\mu_X,\mu_Y)(Y-\mu_Y)^2] \\
&= f_X'^2(\mu_X,\mu_Y)\text{Var}[X] + 2f_X'(\mu_X,\mu_Y)f_Y'(\mu_X,\mu_Y)\text{Cov}[X,Y] + f_Y'^2(\mu_X,\mu_Y)\text{Var}[Y]
\end{aligned} \tag{22}$$

Applying to our estimate leads to

$$\begin{aligned}
\text{Var}[\widetilde{R}] &\approx R^2\text{Var}[W'] + \text{Var}[\widetilde{R}^0] - 2R\text{Cov}[W', \widetilde{R}^0] \\
&= R^2(\mathbb{E}[W'^2] - \mathbb{E}^2[W']) + (\mathbb{E}[(\widetilde{R}^0)^2] - \mathbb{E}^2[\widetilde{R}^0]) - 2R(\mathbb{E}[W'\widetilde{R}^0] - \mathbb{E}[\widetilde{R}^0]\mathbb{E}[W']) \\
&= R^2\mathbb{E}[W'^2] - 2R\mathbb{E}[W'\widetilde{R}^0] + \mathbb{E}[(\widetilde{R}^0)^2] \\
&= \sum_{t=1}^{T}\int \frac{p(\mathbf{x})}{q_t(\mathbf{x})} v_t^2 \left(\int [\mathcal{L}(f_T(\mathbf{x}), y) - R(f_T)]^2 p(y|\mathbf{x}) dy\right) p(\mathbf{x}) d\mathbf{x}
\end{aligned} \tag{23}$$

where we utilize $f(X,Y) = \frac{X}{Y} \to f_X' = \frac{1}{Y}$, $f_Y' = -\frac{X}{Y^2}$, $\mu_X = R, \mu_Y = 1$. Note that since we assume $q_t(\mathbf{x})$ are fixed, we have $\mathbb{E}[W'\widetilde{R}^0] = \mathbb{E}_{p(y|\mathbf{x})}\mathbb{E}_{q_1}...\mathbb{E}_{q_T}[\sum_{t=1}^{T} v_t \sum_{i=1}^{n}(w_t^{(i)})^2 l(\mathbf{x}_t^{(t)})] = \mathbb{E}_{p(y|\mathbf{x})}[\sum_{t=1}^{T} v_t \mathbb{E}_{q_t} \sum_{i=1}^{n}(w_t^{(i)})^2 l(\mathbf{x}_t^{(t)})]$. □

C.2 Proposition 1

We show two concrete examples for Proposition 1. In each case, the estimated introspective loss is analogous to an uncertainty measure.

- The estimation of 0-1 loss is:

$$R_\theta = \frac{1}{|\mathcal{S}_U|}\sum_{\mathbf{x}\in\mathcal{S}_U}\sum_{y} \mathbb{1}(f_\theta(\mathbf{x}) \neq y) p(y|\mathbf{x};\theta) \tag{24}$$

  which is the sum of the predicted probability of all classes other than the most probable class.

- The estimation of cross-entropy loss is:

$$R_\theta = \frac{1}{|\mathcal{S}_U|}\sum_{\mathbf{x}\in\mathcal{S}_U}\sum_{y} p(y|\mathbf{x};\theta)\log(p(y|\mathbf{x};\theta)) \tag{25}$$

  which is the entropy of the predicted probability.

When we use deep learning models, R_θ usually largely underestimates the risk over the entire pool. In other works such as [13, 14], the surrogate risk acts in a similar way. For the final risk estimator to be accurate, the introspective risk estimation or the surrogate risk first needs to be accurate, which somewhat beats the purpose of active risk estimation. However, we still try to improve this intermediate step without assuming that we have access to an unrealistically accurate estimation, leading to our proposed R_θ in Section 3.3.

C.3 Proof of Theorem 2 and Active Feedback Analysis

In Theorem 2, we formalize the combined learning-testing objective as a joint optimization problem with the variable being a subset $\mathcal{S}_{\text{FB}}$ that can be transferred from the testing set $\mathcal{S}_T$ to the learning set $\mathcal{S}_L$. We define the process of selecting the subset as the "active feedback" process, which connects the learning and testing objectives through a balancing parameter C given in (8). Performing exact optimization of the subset along with a parameter C would require more detailed knowledge on the learning model and the AL strategy. We instead provide a general analysis to show that active feedback could indeed provide an optimal solution for the joint optimization problem, where C scales as $\mathcal{O}(1)$. Following our theoretical result, we empirically demonstrate the effectiveness of an intuitive feedback approach in the experimental sections (Section 4.3, Appendix D).

Proof overview. We apply some generic generalization bound (*e.g.,* [16] for CNN or similar models) to the learning objective (I) in the joint optimization problem given by (8), which gives $\mathcal{O}(1/\sqrt{n})$. We then leverage the confidence interval to get a high probability bound for the testing objective (II), which also gives $\mathcal{O}(1/\sqrt{n})$ [4, 9, 26]. We use the formalized results on the convergence of the estimate as introduced in [19]. With that, we continue to show that both the learning and testing objectives share the same dependency on n. These common dependencies on n give us the foundation to further analyze the feedback process. We offer an intuitive justification of active feedback as follows. The risk estimators are importance weighted estimates of the true risk. The estimate converges to the true risk asymptotically, so fewer samples might hurt the quality of the estimate (due to a large variance), but does not change the fact that the expected average of the estimate is still the true risk. With the confidence interval conversions, we can see that except for the change of constants, the objective's dependency on the number of samples does not change. (This also provides guidance for the feedback proposal later: if we can keep the change of the estimate to the minimum, meanwhile using the samples discarded from the test set to improve the AL model as much as possible, it would be the ideal use of available labels.) Following these high-level ideas as described above, we present the detailed proof below.

Proof. We first break the joint (I) learning-(II) testing objective into two parts and approach each part separately:

$$\begin{aligned} R(f_{\theta|(\mathcal{S}_L\cup\mathcal{S}_{\text{FB}})}) &\leq R_{CNN}(f^*_{\theta|(\mathcal{S}_L\cup\mathcal{S}_{\text{FB}})}) + \mathcal{O}\left(1/\sqrt{N_L+N_{\text{FB}}}\right) \\ &\lesssim R_{CNN}(f^*_{\theta|(\mathcal{S}_L)}) + \mathcal{O}\left(1/\sqrt{N_L+N_{\text{FB}}}\right) \end{aligned} \tag{26}$$

$$\begin{aligned} \|R - \tilde{R}_{(\{Q_1,\ldots,Q_T\}\setminus\mathcal{S}_{\text{FB}})}\| \leq \;& ||\tilde{R}_T(\{Q_1,\ldots,Q_T\}) - \tilde{R}_T(\{Q_1,\ldots,Q_T\}\setminus\mathcal{S}_{\text{FB}})|| \\ &+ ||\tilde{R}_T(\{Q_1,\ldots,Q_T\}) - R|| \end{aligned} \tag{27}$$

The learning objective. As mentioned earlier, (26) is a common generalization error bound for CNN or similar models. For example, given a training set $\mathcal{S}_L$ with N_L samples, we can draw from the basic bound (*e.g.,* according to Theorem 2.1 in [16]):

$$R(f_{\theta|\mathcal{S}_L}) = \mathbb{E}_{\mathcal{D}}[l_{f_{\theta|\mathcal{S}_L}}(\cdot)] \leq \mathbb{E}_{\mathcal{S}_L}[l_{f_{\theta|\mathcal{S}_L}}(\cdot)] + C'\left(\beta'\lambda'\sqrt{\frac{|\theta|}{N_L}} + \sqrt{\frac{\log(1/\delta)}{N_L}}\right) \tag{28}$$

with probability of at least $1-\delta$, where C', β', and λ' are constants and $|\theta|$ is the total number of trainable parameters in the network. In our case, we do not make further assumptions about the constants and $|\theta|$ is fixed for evaluating a certain model. Similarly, we can substitute N_L with $N_L + N_{\text{FB}}$ and arrive at:

$$\begin{aligned} R(f_{\theta|(\mathcal{S}_L\cup\mathcal{S}_{\text{FB}})}) = \mathbb{E}_{\mathcal{D}}[l_{f_{\theta|(\mathcal{S}_L\cup\mathcal{S}_{\text{FB}})}}(\cdot)] \leq& \mathbb{E}_{(\mathcal{S}_L\cup\mathcal{S}_{\text{FB}})}[l_{f_{\theta|(\mathcal{S}_L\cup\mathcal{S}_{\text{FB}})}}(\cdot)] \\ &+ C'\left(\beta'\lambda'\sqrt{\frac{|\theta|}{N_L+N_{\text{FB}}}} + \sqrt{\frac{\log(1/\delta)}{N_L+N_{\text{FB}}}}\right) \end{aligned} \tag{29}$$

We notice that in both (28) and (29), we include the expected loss which is slightly different from the best possible AL model risks $R(f^*_{\theta|(\mathcal{S}_L\cup\mathcal{S}_{\text{FB}})})$ and $R(f^*_{\theta|\mathcal{S}_L})$. However, the difference is usually on a smaller scale than $(1/\sqrt{N_L} - 1/\sqrt{N_L + N_{\text{FB}}})$. In general, we assume that $R(f^*_{\theta|(\mathcal{S}_L \cup \mathcal{S}_T)}) \lesssim R(f^*_{\theta|(\mathcal{S}_L \cup \mathcal{S}_{\text{FB}})}) \lesssim R(f^*_{\theta|\mathcal{S}_L})$ since more labeled samples can benefit learning (we do not need to assume a strictly monotonic case for the sake of this analysis). For most AL strategies, the difference between the expected empirical risk and the optimal risks given the learning set size is on a higher order of dependency on n than the learning bound itself. If we ignore the higher order terms, we can simplify the results as shown in (26). Then, the term and more importantly the change in the learning objective that is related to the assumed feedback $\mathcal{S}_{\text{FB}}$ is only dependent on N_{FB} through $\mathcal{O}(1/\sqrt{N_L + N_{\text{FB}}})$.

The testing objective. The relation in (27) can be further analyzed by taking a probabilistic view. If we assume the risks are bounded in the third moment, w.l.o.g., the two risk-difference terms can both be generalized to a slightly more specific high-probability confidence interval [4, 9, 26] than the plain central limit theorem result itself: with probability of at least $1 - \alpha$, we have

$$\begin{aligned}
&||\tilde{R}_T(\{Q_1, ..., Q_T\}) - \tilde{R}_T(\{Q_1, ..., Q_T\} \setminus \mathcal{S}_{\text{FB}})|| \\
&\leq 2 \left[F^{-1}_{N_T} \left(1 - \frac{\alpha}{2}\right) \frac{\tilde{\sigma}_{N_T}}{\sqrt{N_T}} - F^{-1}_{N_T - N_{\text{FB}}} \left(1 - \frac{\alpha}{2}\right) \frac{\tilde{\sigma}_{N_T - N_{\text{FB}}}}{\sqrt{N_T - N_{\text{FB}}}} \right]
\end{aligned} \tag{30}$$

$$||\tilde{R}_T(\{Q_1, ..., Q_T\}) - R|| \leq 2 \left[F^{-1}_{N_T} \left(1 - \frac{\alpha}{2}\right) \frac{\tilde{\sigma}_{N_T}}{\sqrt{N_T}} \right] \tag{31}$$

where F^{-1} is the inverse cumulative distribution function of the Student-t distribution and $\tilde{\sigma}^2$ is the empirical variance. For the active feedback analysis, we only care about how N_{FB} affects the testing objective, thus also obtaining an $\mathcal{O}(1/\sqrt{N_T - N_{\text{FB}}})$ dependency. $\square$

The detailed balancing between the two objectives (I) and (II) requires specific knowledge about the constants involved in the bounds. However, if we only focus on terms involving N_{FB}, both dependencies on the sample numbers are on the $1/\sqrt{n}$ level, making it possible to be balanced by a constant factor C. Combining these results, we get the N_{FB} term as $\mathcal{O}(1/\sqrt{N_L + N_{\text{FB}}}) + \mathcal{O}(1/\sqrt{N_T - N_{\text{FB}}})$ (absorbing $\mathcal{O}(1)$ terms that do not depend on N_{FB}). The next key factor is that throughout the entire ATL process, we either keep N_{FB} fixed or only change it at a linear rate (flexible N_{FB} should be an interesting future direction). Combining with our previous assumption that N_L and N_T are of similar magnitudes, we know that an optimal balance could be achieved between (I) and (II) to minimize the joint learning-testing objective given in (8).

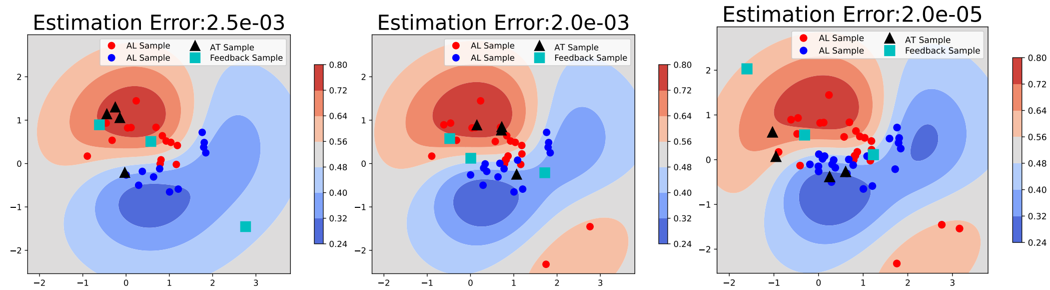
D Additional Experiment Results

In this section, we present the detailed experimental settings and additional experimental results.

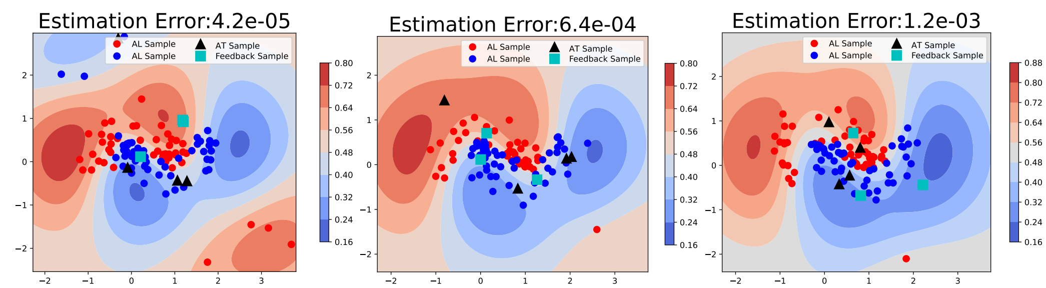
D.1 Synthetic Experiment

Figure 4 shows how the proposed feedback strategy helps to encourage exploration. The background color shows the model's predictive distribution. For each quiz, we display all the training samples obtained by an active learner (red and blue circles representing 2 classes) but only the current quiz (triangles) and feedback samples (squares, then added to circles in later AL rounds) from the active tester to make the visualization clear. Figure 4a shows that ATL selects a feedback sample in the bottom right corner because it is not included in the current knowledge base of the AL model. The AL model predicts it poorly in the quiz. In Figure 4b, we see that the AL model is guided by the feedback samples and starts to explore the bottom right corner. Once the AL model collects samples from the bottom right area, ATL stops to provide guidance for that region. In this way, the proposed feedback strategy manages to find the minority cluster at the other corner shortly as shown in Figure 4c.

In Figure 4d, to further demonstrate the effectiveness of the proposed feedback strategy, we compare it with the feedback samples selected using two other baselines: random feedback (in Figure 4f) and AL based feedback (in Figure 4e), when the samples at the bottom right corner are first discovered. First, we notice that those data samples are found by the AL model rather than through the feedback strategies. As a result, it happens at a much later quiz time compared with ATL. Therefore, they result in a less efficient learning process. Second, we observe that when an AL model discovers a



(a) ATL Feedback Quiz 2 (b) ATL Feedback Quiz 3 (c) ATL Feedback Quiz 4



(d) ATL Feedback Quiz 10 (e) AL Feedback Quiz 9 (f) Random Feedback Quiz 10

Figure 4: Effectiveness of active feedback for improving model training

Table 5: Test classification accuracy

Feedback Strategy	Quiz 2	Quiz 3	Quiz 4	Quiz 9	Quiz 10
ATL	0.75	0.87	0.90	0.96	0.97
AL	0.75	0.83	0.84	0.84	0.88
Random	0.74	0.78	0.81	0.83	0.91

new area to learn, both baseline feedback strategies fail to provide support even though they have some test samples (*i.e.,* the red point at the bottom right corner) available in the interesting region. Last, we can see that at around quiz 10, the AL model with the proposed ALT converges to a better decision boundary that captures the entire data distribution while the two other baselines both fail to correctly discover the predictive distribution at the two corners. As a result, ALT leads to a more accurate model (shown in Table 5) while maintaining lower estimation error in the end.

D.2 Real-world Experiments

In this section, we provide more results on the real-world datasets including MNIST, FashionMNIST and CIFAR10, mainly to demonstrate different feedback approaches and how we can implement early stopping in ATL.

D.2.1 Experimental Settings

In all experiments, we use a CNN model and standard data transformation for each dataset. In each AL training round, we run 10 epochs for MNIST and FashionMNIST and 50 epochs for CIFAR10. A threshold of 1×10^{-5} is used for probability outputs as required for the proposal $q(\mathbf{x})$ computation [13] to avoid 0 denominators.

An important detail to note is that for **ATL-NF** results, we **sample 50 additional data points during AL for fair comparison** (550 in each round), which is actually very similar to **ATL-RF**. The results with only 500 data points per round will be shown in the following section D.2.2. Another detail worth mentioning is that although we set the initial budget to be 500 labels and add 500 training samples plus 100 testing samples in each round, the final total budget is $12,450$ on average instead of $12,500$ because we allow replacement while sampling.

D.2.2 Random Hold-out Test and Random Feedback

In this section, we discuss the issues with traditional hold-out validation/testing procedures during AL and compare the results using random sampling for both test sample selection and feedback selection with the proposed learning-testing-feedback process. In Table 6, random test is referring

Table 6: Risk estimation error comparison with random methods

Dataset	Method \ AL round	4	8	12	16	20
MNIST	Random Test	7.80 ± 13.4	6.61 ± 3.72	5.94 ± 7.25	3.15 ± 5.83	6.18 ± 2.41
	Random Test & Feedback	30.4 ± 42.2	16.8 ± 16.1	5.34 ± 6.33	11.8 ± 10.6	5.62 ± 2.73
	Random Test & Weighted	71.3 ± 31.9	19.1 ± 16.5	12.3 ± 12.2	10.0 ± 11.7	5.64 ± 1.64
	ATL-NF	**2.57 ± 1.17**	**0.79 ± 1.15**	**0.17 ± 0.15**	**0.56 ± 0.30**	**1.32 ± 0.37**
Fashion MNIST	Random Test	6.97 ± 11.2	5.29 ± 3.56	10.9 ± 6.55	5.40 ± 2.76	8.56 ± 7.57
	Random Test & Feedback	12.7 ± 13.4	15.4 ± 16.0	12.7 ± 6.93	18.8 ± 20.9	32.7 ± 15.4
	Random Test & Weighted	6.85 ± 14.8	3.01 ± 11.3	2.80 ± 11.1	4.58 ± 29.0	9.51 ± 9.13
	ATL-NF	**3.64 ± 1.61**	**0.67 ± 0.38**	**0.96 ± 0.16**	**0.98 ± 0.43**	**3.04 ± 1.37**
CIFAR10	Random Test	20.5 ± 6.50	15.8 ± 10.3	13.0 ± 10.1	9.99 ± 6.58	9.89 ± 9.61
	Random Test & Feedback	44.7 ± 36.4	16.4 ± 15.4	31.0 ± 12.8	11.7 ± 8.65	55.3 ± 19.0
	Random Test & Weighted	43.5 ± 14.8	14.6 ± 11.3	15.1 ± 11.1	11.2 ± 29.0	48.7 ± 9.13
	ATL-NF	**8.83 ± 7.79**	**3.06 ± 5.04**	**4.95 ± 7.12**	**7.94 ± 5.22**	**6.20 ± 5.79**

to randomly sampling 100 test samples after each 550-sample (additional 50 for fair comparison, same as ATL-NF) AL round and simply averaging the loss over these test samples. Random test & feedback is referring to sampling 100 test samples after each 500-sample AL round and then randomly selecting 50 for feedback. Random test & weighted is referring to the same process but the quizzes are weighted by $1/R_t$. From Table 6, we can see that in the small-data regime, random sampling may not provide an accurate estimate of the true risk. However, in later AL rounds, the no feedback case (Random Test) can maintain an unbiased estimate, and we do see that some results are comparable with active risk estimation baselines without the ATL-integrate estimator. This is probably because existing active risk estimation baselines (ARE-quiz, AT-integrate, ASE-integrate) do not consider the biased selection and model change through the AL process. The methods that use surrogate models also suffers from the insufficient training of the surrogate model. However, random testing selection does not work well with the active feedback process. For Random Test & Feedback and Random Test & Weighted, we often see much worse estimation due to the feedback process involved.

D.2.3 Additional Active Feedback Comparisons

In this section, we show a more complete comparison between different feedback approaches. The feedback comparison consists of two parts: (1) baseline comparison including no feedback (ATL-NF), random feedback (ATL-RF), entropy-based feedback (ATL-EN) and (2) ablation study including loss-based feedback (ATL-LF), weighted loss-based feedback (ATL-WL) and the proposed weighted loss plus diversity feedback (ATL).

Table 7: Hold-out test risk using different feedback criteria over 20 AL rounds

Dataset	Method \ AL round	4	8	12	16	20
MNIST	ATL-NF	0.92 ± 0.06	0.55 ± 0.08	0.46 ± 0.06	0.32 ± 0.04	0.22 ± 0.02
	ATL-RF	0.92 ± 0.12	0.54 ± 0.02	0.41 ± 0.05	0.29 ± 0.03	0.21 ± 0.02
	ATL-EN	0.90 ± 0.12	0.55 ± 0.06	0.41 ± 0.02	0.34 ± 0.06	0.23 ± 0.03
	ATL-LF	0.89 ± 0.10	0.56 ± 0.04	0.41 ± 0.02	0.32 ± 0.07	0.20 ± 0.02
	ATL-WL	0.86 ± 0.06	0.53 ± 0.06	0.40 ± 0.05	0.32 ± 0.07	0.22 ± 0.03
	ATL	**0.88 ± 0.07**	**0.53 ± 0.04**	**0.39 ± 0.03**	**0.26 ± 0.01**	**0.19 ± 0.03**
Fashion MNIST	ATL-NF	0.75 ± 0.03	0.69 ± 0.02	0.61 ± 0.02	0.57 ± 0.04	0.56 ± 0.03
	ATL-RF	0.75 ± 0.04	0.68 ± 0.02	0.61 ± 0.01	0.58 ± 0.06	0.56 ± 0.04
	ATL-EN	0.76 ± 0.02	0.67 ± 0.05	0.58 ± 0.02	0.59 ± 0.03	0.56 ± 0.02
	ATL-LF	0.76 ± 0.04	0.65 ± 0.03	0.63 ± 0.01	0.56 ± 0.02	0.56 ± 0.04
	ATL-WL	0.76 ± 0.03	0.65 ± 0.02	0.62 ± 0.01	0.56 ± 0.02	0.53 ± 0.02
	ATL	**0.74 ± 0.03**	**0.65 ± 0.04**	**0.59 ± 0.02**	**0.56 ± 0.03**	**0.51 ± 0.01**
CIFAR10	ATL-NF	1.91 ± 0.04	1.76 ± 0.05	1.72 ± 0.01	1.66 ± 0.02	1.55 ± 0.03
	ATL-RF	1.91 ± 0.03	1.77 ± 0.04	1.69 ± 0.03	1.60 ± 0.04	1.54 ± 0.07
	ATL-EN	1.92 ± 0.09	1.76 ± 0.04	1.70 ± 0.03	1.66 ± 0.04	1.54 ± 0.02
	ATL-LF	1.94 ± 0.04	1.75 ± 0.03	1.65 ± 0.01	1.59 ± 0.03	1.54 ± 0.01
	ATL-WL	1.94 ± 0.04	1.75 ± 0.03	1.63 ± 0.01	1.63 ± 0.03	1.54 ± 0.01
	ATL	**1.90 ± 0.05**	**1.76 ± 0.02**	**1.65 ± 0.03**	**1.58 ± 0.02**	**1.53 ± 0.02**

First, we show the hold-out test risk of the AL model throughout AL using different active feedback approaches as the indicator of the model performance. From Table 7, we see that in most occasions, all active feedback approaches can reduce the test risk compared to ATL-NF.

Table 8: Squared difference between the estimate and the true risk over 20 AL rounds ($\times 10^{-3}$)

Dataset	Method \ AL round	4	8	12	16	20
MNIST	ATL-NF	2.57 ± 1.17	0.79 ± 1.15	0.17 ± 0.15	0.56 ± 0.30	1.32 ± 0.37
	ATL-RF	26.8 ± 21.4	21.4 ± 17.0	3.54 ± 4.01	5.54 ± 3.21	7.62 ± 4.41
	ATL-EN	23.6 ± 24.8	14.0 ± 15.8	13.8 ± 11.7	29.5 ± 21.7	21.8 ± 12.8
	ATL-LF	15.6 ± 12.6	42.4 ± 36.9	48.5 ± 25.8	15.7 ± 14.8	10.9 ± 7.44
	ATL-WL	16.5 ± 19.4	21.0 ± 24.3	7.36 ± 8.44	11.4 ± 12.7	7.59 ± 4.45
	ATL	$\mathbf{14.6 \pm 22.1}$	$\mathbf{16.9 \pm 13.7}$	$\mathbf{3.19 \pm 2.63}$	$\mathbf{4.15 \pm 3.20}$	$\mathbf{1.87 \pm 1.41}$
Fashion MNIST	ATL-NF	3.64 ± 1.61	0.67 ± 0.38	0.96 ± 0.16	0.98 ± 0.43	3.04 ± 1.37
	ATL-RF	10.2 ± 9.30	4.41 ± 3.77	2.19 ± 5.53	5.69 ± 4.52	11.6 ± 7.51
	ATL-EN	93.2 ± 23.4	50.2 ± 10.2	78.5 ± 32.4	76.2 ± 59.6	85.8 ± 25.9
	ATL-LF	9.36 ± 10.2	27.2 ± 26.0	22.6 ± 28.3	14.6 ± 12.0	11.0 ± 15.2
	ATL-WL	8.39 ± 8.97	7.52 ± 6.09	4.89 ± 6.50	7.29 ± 4.45	11.1 ± 7.02
	ATL	$\mathbf{2.50 \pm 2.93}$	$\mathbf{1.94 \pm 2.25}$	$\mathbf{1.78 \pm 1.07}$	$\mathbf{6.32 \pm 5.41}$	$\mathbf{5.03 \pm 4.41}$
CIFAR10	ATL-NF	8.83 ± 7.79	3.06 ± 5.04	4.95 ± 7.12	7.94 ± 5.22	6.20 ± 5.79
	ATL-RF	20.6 ± 17.6	19.1 ± 13.7	9.82 ± 8.03	33.6 ± 30.5	24.8 ± 32.4
	ATL-EN	30.3 ± 17.0	45.8 ± 24.4	20.3 ± 17.4	36.8 ± 31.7	27.0 ± 27.1
	ATL-LF	35.0 ± 27.9	45.8 ± 28.5	20.3 ± 10.1	57.2 ± 33.6	40.5 ± 34.0
	ATL-WL	22.7 ± 19.7	25.0 ± 13.2	12.9 ± 21.5	52.2 ± 45.9	28.7 ± 16.3
	ATL	$\mathbf{11.6 \pm 13.4}$	$\mathbf{5.11 \pm 3.45}$	$\mathbf{8.81 \pm 6.51}$	$\mathbf{11.9 \pm 16.7}$	$\mathbf{6.57 \pm 6.29}$

In Table 8, we show a full comparison of the squared error of risk estimation. All estimation results are based on the proposed ATL estimator $\tilde{R}$, where ATL-NF, ATL-RF, ATL-EN serve as baselines, meanwhile ATL-LF and ATL-WL serve as ablation studies since the proposed ATL utilizes the weighted loss as well. We see that all feedback approaches suffer from an increased estimation error, especially in the early stage when the number of test samples available is small. We see that the baseline methods suffer from increased estimation error. However, ATL can usually maintain a similar level of estimation error after 20 AL rounds. For ATL-LF, there is usually a larger variance of the estimation error. The potential reason for the unstable behavior of ATL-LF is that it only selects samples with larger losses in the feedback process. Although the importance mechanism can make up for some of the difference, there is still the potential risk of the estimate being biased. Further combining with the diversity metric, we achieve the best results with ATL.

Concluding from both the risk results and the estimation error results, we show that the proposed feedback approach achieves a good balance in the performance-estimation trade-off. This is because we consider both the loss L and the importance weight q in the selection criterion. Overall, ATL achieves a similar model test risk as ATL-LF/ATL-WL, both of which are much better than ATL-NF and ATL-RF. ATL also achieves a much lower estimation error than ATL-RF, ATL-EN, and ATL-LF.

D.2.4 Feedback Size Study

We also provide a study on the size of the feedback set. As mentioned in the proof for Theorem 2, we keep the size of feedback simple in this work. This is to be consistent with our theoretical analysis and the experiments show that the active feedback process is helpful in this generic setting. Further details about extending this will be mentioned in the future directions. However, even in the simple setting of fixed feedback size, we can see that the learning and testing performances do not consistently and monotonically change with respect to the feedback size. Although, from Table 9 and Table 10 below, we can see that in general, model risk (learning performance) is better when we use a larger feedback size, but at the same time the estimation error (testing performance) may become much worse. The model risk on CIFAR10 behaves differently with the feedback size, probably because the model performance is not good enough and adding difficult samples in this stage does not necessarily help with the generalization ability.

D.2.5 Single Feedback Round Comparison

In previous experiments, we add additional training points for the no feedback case (ATL-NF) to make fair comparison for the model risk. However, if we look at the risk change before and after a single feedback round, the difference is even more obvious.

Table 9: Hold-out test risk using different feedback criteria over 20 AL rounds

Dataset	AL round / Feedback size	4	8	12	16	20
MNIST	83%	0.86 ± 0.09	0.53 ± 0.04	0.40 ± 0.08	0.30 ± 0.02	0.20 ± 0.03
	67%	0.87 ± 0.08	0.52 ± 0.03	0.35 ± 0.03	0.30 ± 0.02	0.21 ± 0.02
	50%	0.88 ± 0.07	0.53 ± 0.04	0.39 ± 0.03	0.26 ± 0.01	0.19 ± 0.03
	25%	0.94 ± 0.06	0.54 ± 0.03	0.42 ± 0.08	0.35 ± 0.02	0.25 ± 0.02
	20%	0.99 ± 0.04	0.56 ± 0.08	0.43 ± 0.06	0.38 ± 0.01	0.24 ± 0.02
Fashion MNIST	83%	0.74 ± 0.02	0.67 ± 0.03	0.60 ± 0.03	0.54 ± 0.02	0.51 ± 0.03
	67%	0.77 ± 0.04	0.68 ± 0.03	0.59 ± 0.03	0.56 ± 0.03	0.52 ± 0.02
	50%	0.74 ± 0.03	0.65 ± 0.04	0.59 ± 0.02	0.56 ± 0.03	0.51 ± 0.01
	25%	0.76 ± 0.02	0.70 ± 0.01	0.62 ± 0.02	0.59 ± 0.05	0.53 ± 0.03
	20%	0.77 ± 0.02	0.71 ± 0.02	0.64 ± 0.02	0.61 ± 0.04	0.54 ± 0.04
CIFAR10	83%	1.92 ± 0.06	1.71 ± 0.02	1.67 ± 0.07	1.59 ± 0.04	1.57 ± 0.04
	67%	1.96 ± 0.05	1.75 ± 0.02	1.64 ± 0.04	1.58 ± 0.04	1.58 ± 0.06
	50%	1.90 ± 0.05	1.76 ± 0.02	1.65 ± 0.03	1.58 ± 0.02	1.53 ± 0.02
	25%	1.94 ± 0.08	1.76 ± 0.03	1.70 ± 0.03	1.64 ± 0.04	1.59 ± 0.02
	20%	1.91 ± 0.03	1.76 ± 0.02	1.73 ± 0.03	1.59 ± 0.02	1.63 ± 0.02

Table 10: Squared difference between the estimate and the true risk over 20 AL rounds ($\times 10^{-3}$)

Dataset	AL round / Feedback size	4	8	12	16	20
MNIST	83%	50.2 ± 39.8	21.0 ± 24.3	7.36 ± 8.44	11.4 ± 12.7	7.59 ± 4.45
	67%	25.6 ± 23.4	29.3 ± 29.7	6.90 ± 8.05	6.24 ± 6.71	7.50 ± 5.07
	50%	14.6 ± 22.1	16.9 ± 13.7	3.19 ± 2.63	4.15 ± 3.20	1.87 ± 1.41
	25%	11.7 ± 11.5	10.0 ± 7.98	9.73 ± 11.4	4.76 ± 5.25	1.59 ± 1.96
	20%	28.0 ± 24.4	11.8 ± 14.5	5.91 ± 3.82	4.31 ± 4.80	1.25 ± 1.36
Fashion MNIST	83%	8.39 ± 8.97	7.52 ± 10.4	2.77 ± 3.58	3.87 ± 4.45	11.1 ± 7.02
	67%	8.59 ± 8.77	8.60 ± 10.5	5.42 ± 5.96	4.05 ± 2.47	14.6 ± 13.8
	50%	2.50 ± 2.93	1.94 ± 2.25	1.78 ± 1.07	6.32 ± 5.41	5.03 ± 4.41
	25%	3.04 ± 4.00	2.38 ± 4.81	1.54 ± 1.18	6.40 ± 8.06	4.13 ± 3.99
	20%	2.62 ± 1.57	1.56 ± 1.77	2.42 ± 4.52	5.65 ± 4.33	5.22 ± 3.27
CIFAR10	83%	54.5 ± 54.1	14.3 ± 7.75	56.1 ± 17.0	47.2 ± 34.3	62.2 ± 43.3
	67%	24.6 ± 25.6	36.7 ± 20.5	24.1 ± 18.6	30.7 ± 40.8	36.2 ± 21.0
	50%	11.6 ± 13.4	5.11 ± 3.45	8.81 ± 6.51	11.9 ± 16.7	6.57 ± 6.29
	25%	4.88 ± 5.80	6.01 ± 8.22	6.80 ± 1.36	10.2 ± 13.4	4.48 ± 3.53
	20%	5.44 ± 6.65	3.65 ± 3.44	11.2 ± 11.0	4.21 ± 1.36	5.82 ± 3.34

Table 11: Hold-out test risk before and after a specific feedback round

Dataset	AL round / Method	4	8	12	16	20
MNIST	ATL-before	0.91 ± 0.09	0.54 ± 0.04	0.41 ± 0.08	0.29 ± 0.02	0.21 ± 0.03
	ATL-after	$\mathbf{0.88 \pm 0.07}$	$\mathbf{0.53 \pm 0.04}$	$\mathbf{0.39 \pm 0.03}$	$\mathbf{0.26 \pm 0.01}$	$\mathbf{0.19 \pm 0.03}$
Fashion MNIST	ATL-before	0.77 ± 0.03	0.66 ± 0.03	0.61 ± 0.02	0.57 ± 0.03	0.53 ± 0.03
	ATL-after	$\mathbf{0.74 \pm 0.03}$	$\mathbf{0.65 \pm 0.04}$	$\mathbf{0.59 \pm 0.02}$	$\mathbf{0.56 \pm 0.03}$	$\mathbf{0.51 \pm 0.01}$
CIFAR10	ATL-before	1.97 ± 0.07	1.82 ± 0.05	1.70 ± 0.03	1.67 ± 0.03	1.57 ± 0.04
	ATL-after	$\mathbf{1.90 \pm 0.05}$	$\mathbf{1.76 \pm 0.02}$	$\mathbf{1.65 \pm 0.03}$	$\mathbf{1.58 \pm 0.02}$	$\mathbf{1.53 \pm 0.02}$

D.2.6 Early Stopping in AL

In this section, we show how the ATL-based risk estimation can be readily used for early stopping in AL. In the above experiments, we observe a steady decrease of the estimated risk most of the times. However, we do find the decrease becomes more insignificant near the end of the 20 rounds of learning, especially for the MNIST and Fashion MNIST datasets. We observe that after a certain amount of AL rounds, the risk decrease is significantly small, and the corresponding test accuracy is also stabilized (MNIST around 94%, Fashion MNIST around 80%, CIFAR around 54%). This gives us the opportunity to apply early stopping in real-life AL applications. We here show the average stopping iteration and model performance (hold-out test accuracy) of the compared methods in Table 12. Following the same threshold value, by augmenting the moving average of active risk estimation given by (10) with stabilized prediction (SP), the combined method can stop at a similar testing accuracy as compared with the SP method, but with much lower variance in test accuracy. Based on the threshold setting, it is also possible to stop AL much earlier, saving the overall labeling budget.

Table 12: Average early stopping iteration and final test accuracy comparison (with variance)

Dataset	Method	Iteration	Variance	Test Accuracy	Variance
MNIST	SP	15	6.8	94.52%	$6.0e-5$
	Combined	11	1.2	94.08%	$3.1e-5$
Fashion MNIST	SP	16	4.4	81.32%	$3.7e-5$
	Combined	12.4	1.04	80.12%	$2.4e-5$
CIFAR10	SP	12	2.8	53.87%	$1.4e-4$
	Combined	12.8	0.16	54.43%	$8.9e-5$

E Details of Hardware for Experiments

All experiments were run on clusters with either NVIDIA A6000 or NVIDIA A100 graphic cards and Intel Xeon Gold 6150 CPU processors. The runtime of the experiments varies depending on the number of repeat runs, but is usually on the scale of a few hours. For example, to get the 5 runs results of one ATL setting for 20 AL rounds on MNIST or Fashion MNIST may take about 6 to 8 hours. The CIFAR10 experiments may take slightly longer.

F Limitation, Future work, and Social Impact

In this section, we first discuss some limitation of the proposed framework and identify some important future direction. We then discuss some potential social impact of our work.

F.1 Limitation and Future Directions

In this paper, we propose an integrated framework that combines active learning and testing. In the interactive framework, the exchange of training and testing information should be carefully guided. Although the proposed testing selection is statistically unbiased and the active feedback is backed by the high-level analysis, we still have room for improving the specification of methods in applicable settings, which we will introduce here as future directions:

- From the learning perspective, we can improve upon the general setting in this paper. In this paper, we focus on introducing a general framework and working under the agnostic setting. However, using specific AL strategies can potentially provide advantages in certain use cases. There have been works that analyze AL label complexity bounds using either importance weighting mechanism in stream-based settings [5, 8] or other methods in pool-based settings [11].
- Continuing on the results from D.2.4 and the discussion above, the feedback size is a very important factor in the process, especially if we allow the size to change during AL. Further investigating the relationship between the sample size and the combined learning-testing objective can potentially improve the framework.
- We also propose the ATL framework under an AL-agnostic assumption. Given specific AL strategies, we might be able to also incorporate the learning or testing proposal in the construction of feedback proposal.

F.2 Social Impact

The proposed ATL framework considers the practical challenges of applying active learning in real-world settings, where both model training and evaluation require labeled data. It is a critical step towards realizing label-efficient learning in practice, which can benefit many critical domains where data annotation is highly costly. To this end, the proposed ATL framework has the potential to fundamentally address the data annotation crisis and further broaden the usage of AI to benefit the entire society.

G Source Code

The data and source code for replicating the results are provided in this link:

`https://drive.google.com/drive/folders/10s9j2oUEuNCM0KjxDT852PtfvvsaoAtZ?usp=sharing`