Actively Testing Your Model While It Learns: Realizing Label-efficient Learning in Practice

Dayou Yu[1]
Weishi Shi[2]
Qi Yu[1]

[1]Rochester Institute of Technology

[2]University of North Texas

Active Learning / Active Testing

Active Learning

Low labeling budget for training, active training selection

Achieve better performance, with sampling bias?

Active Learning / Active Testing

Active Learning

Low labeling budget for training, active training selection

Achieve better performance, with sampling bias?

How to evaluate?

Active Learning / Active Testing

Active Learning

Low labeling budget for training, active training selection

Achieve better performance, with sampling bias?

Active Testing

Low labeling budget for testing (validation)

Active testing selection, unbiased loss (risk) estimation

Active risk estimation [1], active testing [2], active surrogate estimators [3]

[1] Christoph Sawade, et al. Active risk estimation. In ICML, 2010.

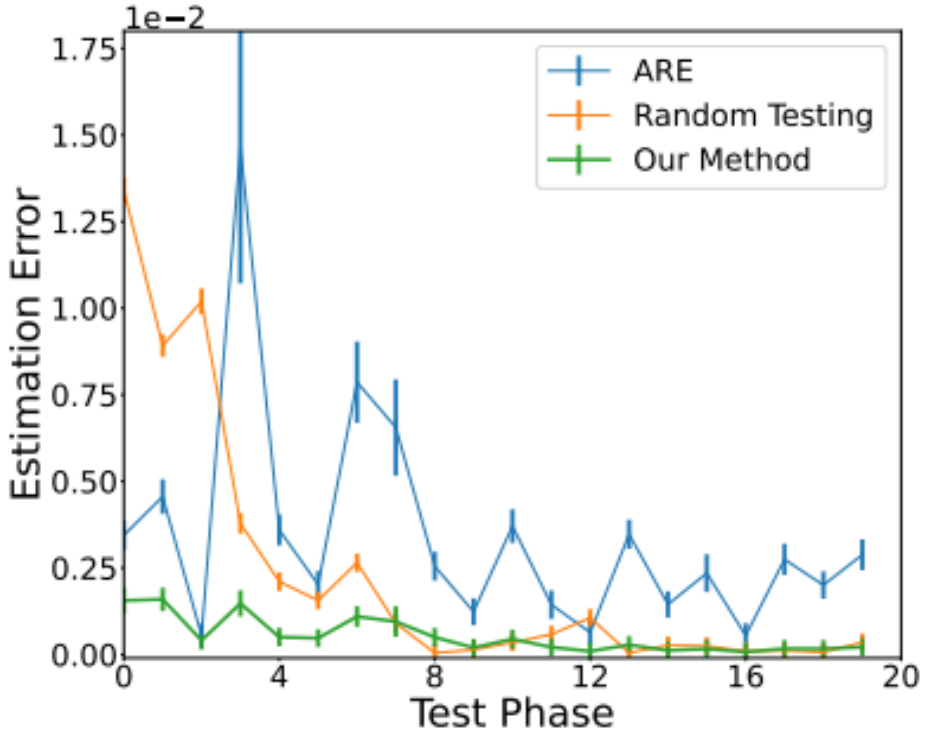
[2][3] Jannik Kossen, et al. Active testing: Sample efficient model evaluation. In International Conference on Machine Learning, pages 5753–5763. PMLR, 2021. / Active surrogate estimators: An active learning approach to label-efficient model evaluation, in Advances in Neural Information Processing Systems, 2022.

Active Testing while Learning - Challenges

Low labeling budget for both learning and testing

Model changes during active learning

Will we still have effective testing selection and unbiased loss (risk) estimation?



Active Quizzes

Selecting a quiz set after one active learning round

$$\mathcal{Q}_t = \{\mathbf{x}_t^{(1)}, \ldots, \mathbf{x}_t^{(n_t)}\}$$

A "locally" optimal test set – asymptotically converges to the true risk

$$\sqrt{n_t}(\widehat{R}_{\mathcal{Q}_t}(f_T) - R(f_T)) \xrightarrow{n_t \to \infty} \mathcal{N}(0, \sigma_t^2(f_T))$$

Selected by the current optimal selection proposal $q^*(\mathbf{x})$

Integrate $\tilde{R} = \sum_{t=1}^{T} v_t \widehat{R}_{\mathcal{Q}_t}$

Active Learning-Testing-Feedback Loop

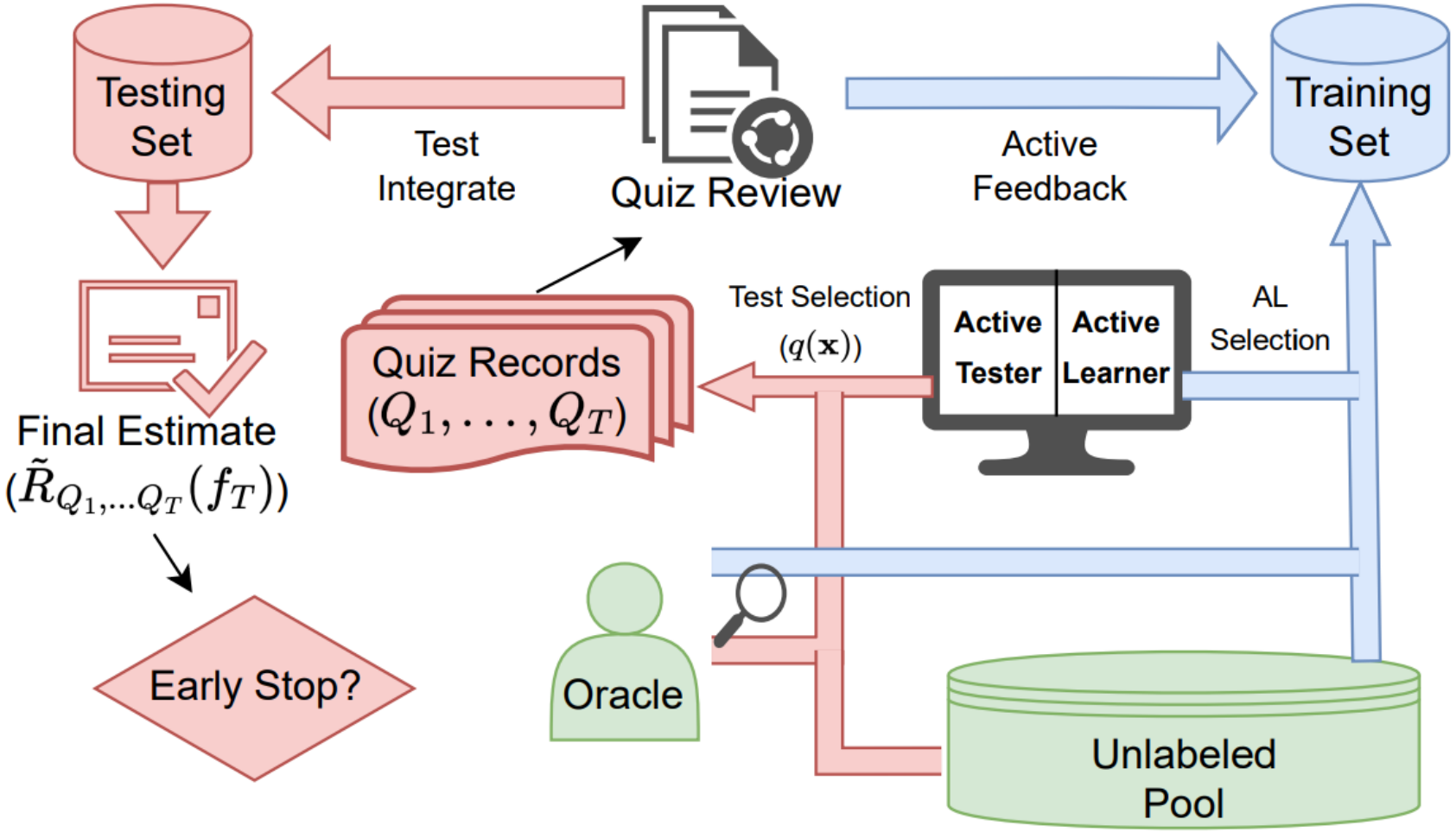
Combined learning-testing objective

$$\mathcal{S}^*_{\mathrm{FB}} = \arg\min_{\mathcal{S}_{\mathrm{FB}} \in \{Q_1,\ldots,Q_T\}} [R(f_{\theta|(\mathcal{S}_L \cup \mathcal{S}_{\mathrm{FB}})}) + C||R - \tilde{R}_{(\{Q_1,\ldots,Q_T\} \setminus \mathcal{S}_{\mathrm{FB}})}||]$$

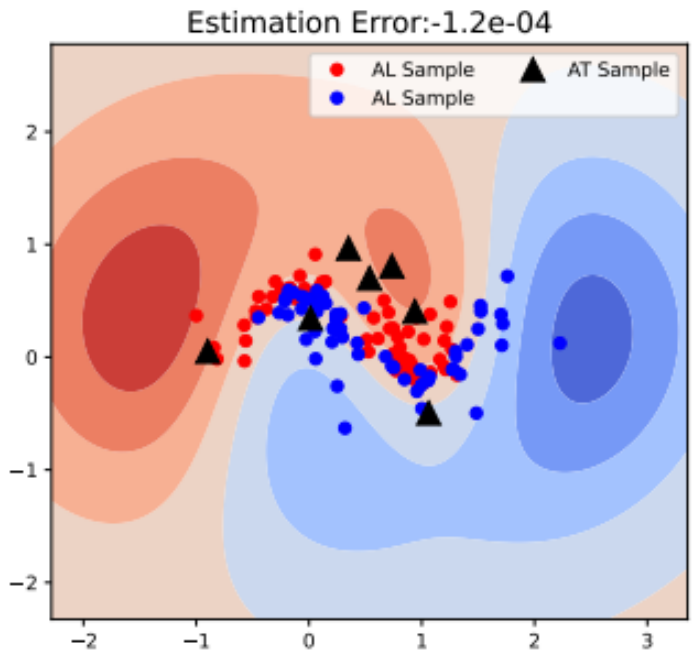
$$(\mathcal{O}(1/\sqrt{N_L + N_{FB}}) + \mathcal{O}(1/\sqrt{N_T - N_{FB}}))$$

Improving training performance while maintaining estimation error

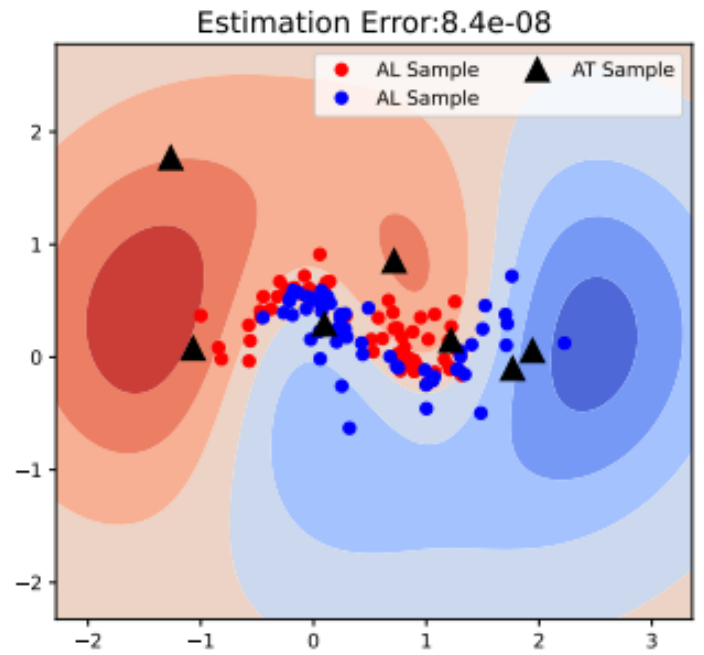
Active Testing while Learning Framework



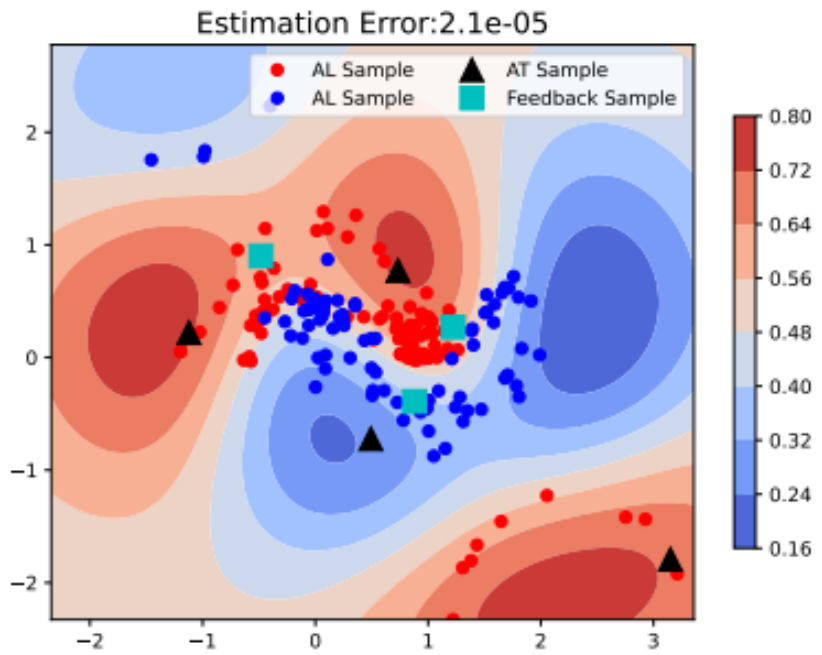
Active Learning-Testing-Feedback Loop



Random Test



ATL



ATL with Feedback

Visualization using synthetic dataset

Label-efficient estimation + Improved learning performance from feedback

Experiments

- Risk estimation error evaluation with active testing while learning

Table 1: Estimation error: squared difference between estimate and true risks ($\times 10^{-3}$)

Dataset	Method \ AL round	4	8	12	16	20
MNIST	ARE quiz	5.27 ± 5.42	6.39 ± 1.54	2.96 ± 3.45	8.85 ± 4.31	8.31 ± 3.96
	AT integrate	16.3 ± 24.5	32.8 ± 22.1	6.93 ± 18.0	8.72 ± 3.59	3.11 ± 2.98
	ASE integrate	3.45 ± 2.76	1.45 ± 1.00	2.17 ± 5.06	4.00 ± 2.37	5.88 ± 5.27
	ATL-NF	$\mathbf{2.57 \pm 1.17}$	$\mathbf{0.79 \pm 1.15}$	$\mathbf{0.17 \pm 0.15}$	$\mathbf{0.56 \pm 0.30}$	$\mathbf{1.32 \pm 0.37}$
Fashion MNIST	ARE quiz	4.24 ± 3.01	4.62 ± 7.77	8.63 ± 2.47	5.71 ± 1.87	23.78 ± 1.75
	AT integrate	11.9 ± 6.1	36.1 ± 30.7	34.1 ± 31.4	28.0 ± 36.9	22.5 ± 25.7
	ASE integrate	11.1 ± 3.63	3.72 ± 3.53	3.56 ± 8.78	5.29 ± 9.78	8.42 ± 6.72
	ATL-NF	$\mathbf{3.64 \pm 1.61}$	$\mathbf{0.67 \pm 0.38}$	$\mathbf{0.96 \pm 0.16}$	$\mathbf{0.98 \pm 0.43}$	$\mathbf{3.04 \pm 1.37}$
CIFAR10	ARE quiz	10.1 ± 8.79	13.8 ± 13.0	22.2 ± 14.7	21.9 ± 31.4	14.1 ± 13.4
	AT integrate	6.89 ± 6.98	12.0 ± 7.18	21.8 ± 5.73	12.9 ± 9.76	38.9 ± 25.6
	ASE integrate	10.9 ± 3.67	6.51 ± 2.87	7.53 ± 1.46	17.6 ± 2.66	23.2 ± 6.10
	ATL-NF	$\mathbf{8.83 \pm 7.79}$	$\mathbf{3.06 \pm 5.04}$	$\mathbf{4.95 \pm 7.12}$	$\mathbf{7.94 \pm 5.22}$	$\mathbf{6.20 \pm 5.79}$

Experiments

- Learning performance (hold-out test risk) with active feedback

Table 2: 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	$\mathbf{0.88 \pm 0.07}$	$\mathbf{0.53 \pm 0.04}$	$\mathbf{0.39 \pm 0.03}$	$\mathbf{0.26 \pm 0.01}$	$\mathbf{0.19 \pm 0.03}$
Fashion MNIST	ATL-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	$\mathbf{0.74 \pm 0.03}$	$\mathbf{0.65 \pm 0.04}$	$\mathbf{0.59 \pm 0.02}$	$\mathbf{0.56 \pm 0.03}$	$\mathbf{0.51 \pm 0.01}$
CIFAR10	ATL-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	$\mathbf{1.90 \pm 0.05}$	$\mathbf{1.76 \pm 0.02}$	$\mathbf{1.65 \pm 0.03}$	$\mathbf{1.58 \pm 0.02}$	$\mathbf{1.53 \pm 0.02}$

Extensions and Future Directions

Early Stopping

$$\Delta\tilde{R}_t = \frac{\sum_{i=t-w}^{t} v_i \tilde{R}_i}{\sum_{i=t-w}^{t} v_i} - \frac{\sum_{i=t-w-1}^{t-1} v_i \tilde{R}_i}{\sum_{i=t-w-1}^{t-1} v_i}$$

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%	$8.9e-5$

Future work: Principled feedback strategy, model-specific sampling proposals

Thank you