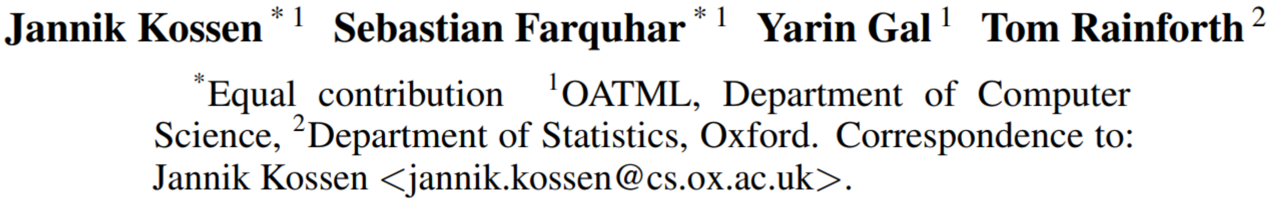
Active Testing:

Sample–Efficient Model Evaluation



Jannik Kossen [*] [1] **Sebastian Farquhar** [*] [1] **Yarin Gal** [1] **Tom Rainforth** [2]

[*] Equal contribution [1] OATML, Department of Computer Science, [2] Department of Statistics, Oxford. Correspondence to: Jannik Kossen <jannik.kossen@cs.ox.ac.uk>.

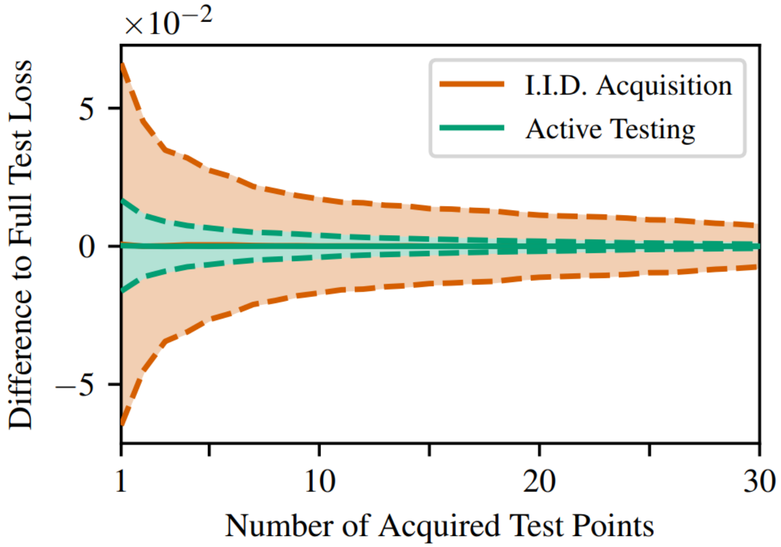
ICML 2021

Motivation

- Accurately evaluate model performance **need a large set of test data**, which is expensive in real tasks.

- **Actively** sample M examples from N unlabeled data pool ($M \ll N$) to evaluate model, so that:

$$\frac{1}{M}\sum_{i_m\in\mathcal{D}_{\text{test}}^{\text{observed}}}\mathcal{L}(f(\mathbf{x}_{i_m}),y_{i_m}) \xrightarrow{\text{close}} \frac{1}{N}\sum_{i_n\in\mathcal{D}_{\text{test}}}\mathcal{L}(f(\mathbf{x}_{i_n}),y_{i_n})$$

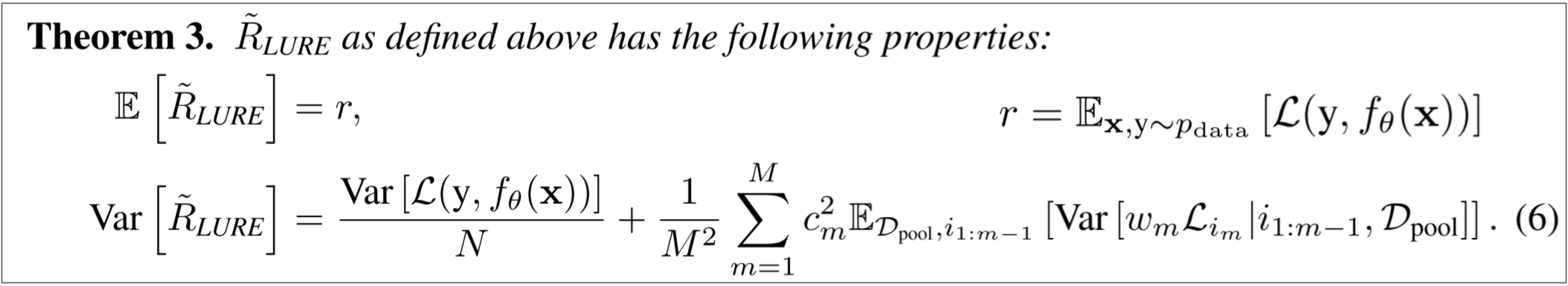


A weighting approach

- introduce an acquisition distribution $q(i_m)$ that denotes the probability of selecting index i_m to be labeled.

$$\hat{R}_{\text{LURE}} = \frac{1}{M}\sum\nolimits_{m=1}^{M} v_m \mathcal{L}\left(f\left(\mathbf{x}_{i_m}\right), y_{i_m}\right)$$

$$v_m = 1 + \frac{N-M}{N-m}\left(\frac{1}{(N-m+1)q(i_m)} - 1\right)$$



Theorem 3. $\tilde{R}_{LURE}$ *as defined above has the following properties:*

$$\mathbb{E}\left[\tilde{R}_{LURE}\right] = r, \qquad r = \mathbb{E}_{\mathbf{x},\mathrm{y}\sim p_{\mathrm{data}}}\left[\mathcal{L}(\mathrm{y}, f_\theta(\mathbf{x}))\right]$$

$$\mathrm{Var}\left[\tilde{R}_{LURE}\right] = \frac{\mathrm{Var}\left[\mathcal{L}(\mathrm{y}, f_\theta(\mathbf{x}))\right]}{N} + \frac{1}{M^2}\sum_{m=1}^{M} c_m^2 \mathbb{E}_{\mathcal{D}_{\mathrm{pool}}, i_{1:m-1}}\left[\mathrm{Var}\left[w_m \mathcal{L}_{i_m} | i_{1:m-1}, \mathcal{D}_{\mathrm{pool}}\right]\right]. \quad (6)$$

Optimal Distribution

Theorem 7. *Given a non-negative loss, the optimal proposal distribution*

$$q^*(i_m; i_{1:m-1}, \mathcal{D}_{\text{pool}}) = \mathcal{L}_{i_m} / \Sigma_{n \notin i_{1:m-1}} \mathcal{L}_n$$

yields estimators exactly equal to the pool risk, that is $\tilde{R}_{\text{LURE}} = \hat{R}$ *almost surely* $\forall M$.

Where: $\hat{R} = \frac{1}{N} \sum_{n=1}^{N} \mathcal{L}(\mathrm{y}_n, f_\theta(\mathbf{x}_n))$.

Proof:

$$\tilde{R}_{\text{LURE}} = \frac{1}{M} \sum_{m=1}^{M} v_m \mathcal{L}_{i_m}$$

$$= \frac{1}{M} \sum_{m=1}^{M} \mathcal{L}_{i_m} + \frac{N-M}{N-m} \left(\frac{\mathcal{L}_{i_m}}{(N-m+1) q^*(i_m; i_{1:m-1}, f_{\theta_{m-1}}, \mathcal{D}_{\text{pool}})} - \mathcal{L}_{i_m} \right)$$

$$= \frac{1}{M} \sum_{m=1}^{M} \mathcal{L}_{i_m} + \frac{N-M}{N-m} \left(\frac{\sum_{t=m}^{N} \mathcal{L}_{i_t}}{(N-m+1)} - \mathcal{L}_{i_m} \right),$$

Optimal Distribution

Proof: pulling out the loss

$$=\frac{1}{M}\sum_{m=1}^{M}\mathcal{L}_{i_m}\left(1-\frac{N-M}{N-m}+(N-M)\underbrace{\sum_{k=1}^{m}\frac{1}{(N-k)(N-k+1)}}_{=m/(N(N-m))}\right)$$

$$+\frac{1}{M}\sum_{t=M+1}^{N}\mathcal{L}_{i_t}(N-M)\underbrace{\sum_{k=1}^{M}\frac{1}{(N-k)(N-k+1)}}_{=M/(N(N-M))},$$

simplifying and rearranging

$$=\frac{1}{M}\sum_{m=1}^{M}\mathcal{L}_{i_m}\left(1-\frac{N-M}{N-m}+\frac{N-M}{N-m}\frac{m}{N}\right)+\frac{1}{N}\sum_{t=M+1}^{N}\mathcal{L}_{i_t}$$

$$=\frac{1}{M}\sum_{m=1}^{M}\mathcal{L}_{i_m}\left(1-\frac{N-M}{N-m}\left(\frac{N-m}{N}\right)\right)+\frac{1}{N}\sum_{t=M+1}^{N}\mathcal{L}_{i_t}$$

$$=\frac{1}{N}\sum_{m=1}^{M}\mathcal{L}_{i_m}+\frac{1}{N}\sum_{t=M+1}^{N}\mathcal{L}_{i_t}$$

$$=\frac{1}{N}\sum_{n=1}^{N}\mathcal{L}_n$$

$= \hat{R}$ as required.

Implementation

- **In practice, we cannot know the true loss. Use expectation.**

$$q^*(i_m) \propto \mathbb{E}_{p(y|\mathbf{x}_{i_m})}\left[\mathcal{L}(f(\mathbf{x}_{i_m}), y)\right]$$

- **Train a surrogate model**

$$q(i_m) \propto \mathbb{E}_{\pi(\theta)\pi(y|\mathbf{x}_{i_m},\theta)}\left[\mathcal{L}(f(\mathbf{x}_{i_m}), y)\right]$$

- **Why not use the original model for the surrogate?**
  - the surrogate can never disagree with *f*
  - May not calibrate well

Choose surrogate

- **Uncertainty**

  Use surrogates that incorporate both epistemic and aleatoric uncertainty effectively. For example, Bayesian neural networks, deep ensembles, and Gaussian processes.

- **Ensemble (QBC)**

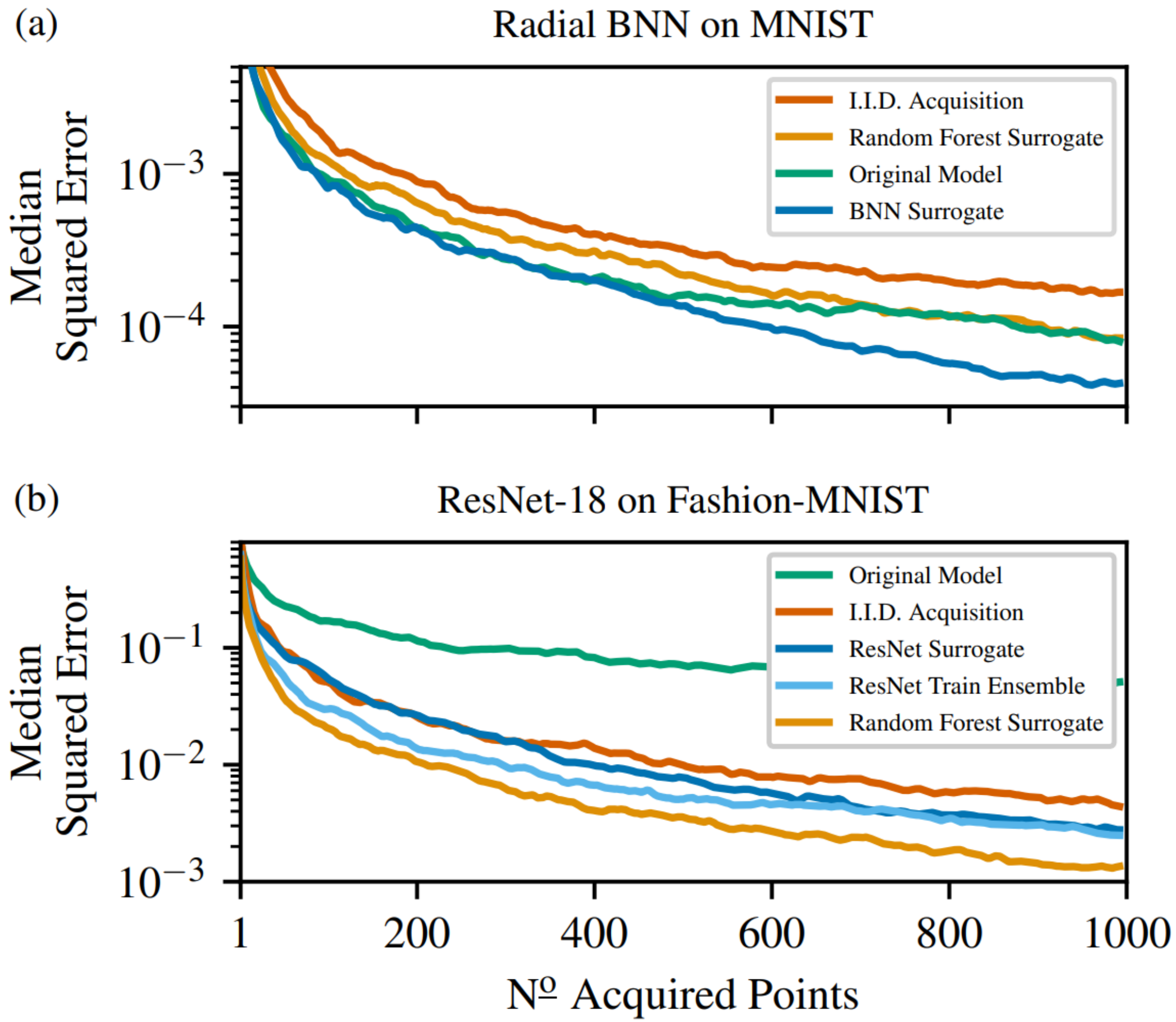
- **Diversity**

  Choosing the surrogate from a different model family or adjusting its hyperparameters

- **Extra data**

  Retrain the surrogate on $\mathcal{D}_{\text{test}}^{\text{observed}} \cup \mathcal{D}_{\text{train}}$ after each step

Surrogate Choice Case Study: Image Classification



- 250 training and 5000 test points
- BNN surrogate & ResNet Surrogate: Use queried test points to retrain the model

Experiments

Validate the effectiveness of ensemble strategy

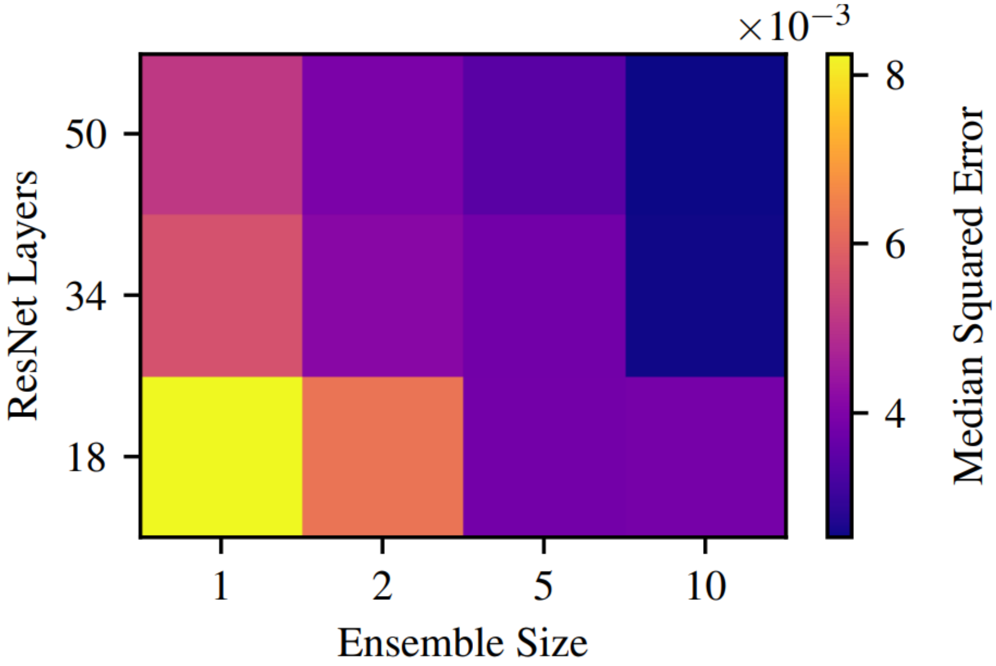


Figure 7. Both diversity and fidelity of the surrogate contribute to sample-efficient active testing. However, the effect of increasing diversity seems larger than that of increased fidelity. We vary the layers (fidelity) and ensemble size (diversity) of the surrogate for active evaluation of a ResNet-18 trained on CIFAR-10. Experiments are repeated for 1000 randomly drawn test sets and we report average values over acquisition steps 100–200.

- ResNet-18 on CIFAR-10
- Using different ResNet ensembles as surrogates.

Experiments

Resnet-18 on CIFAR-10 and Fashion-MNIST & WideResNet on CIFAR-100

Using ensemble as the surrogate

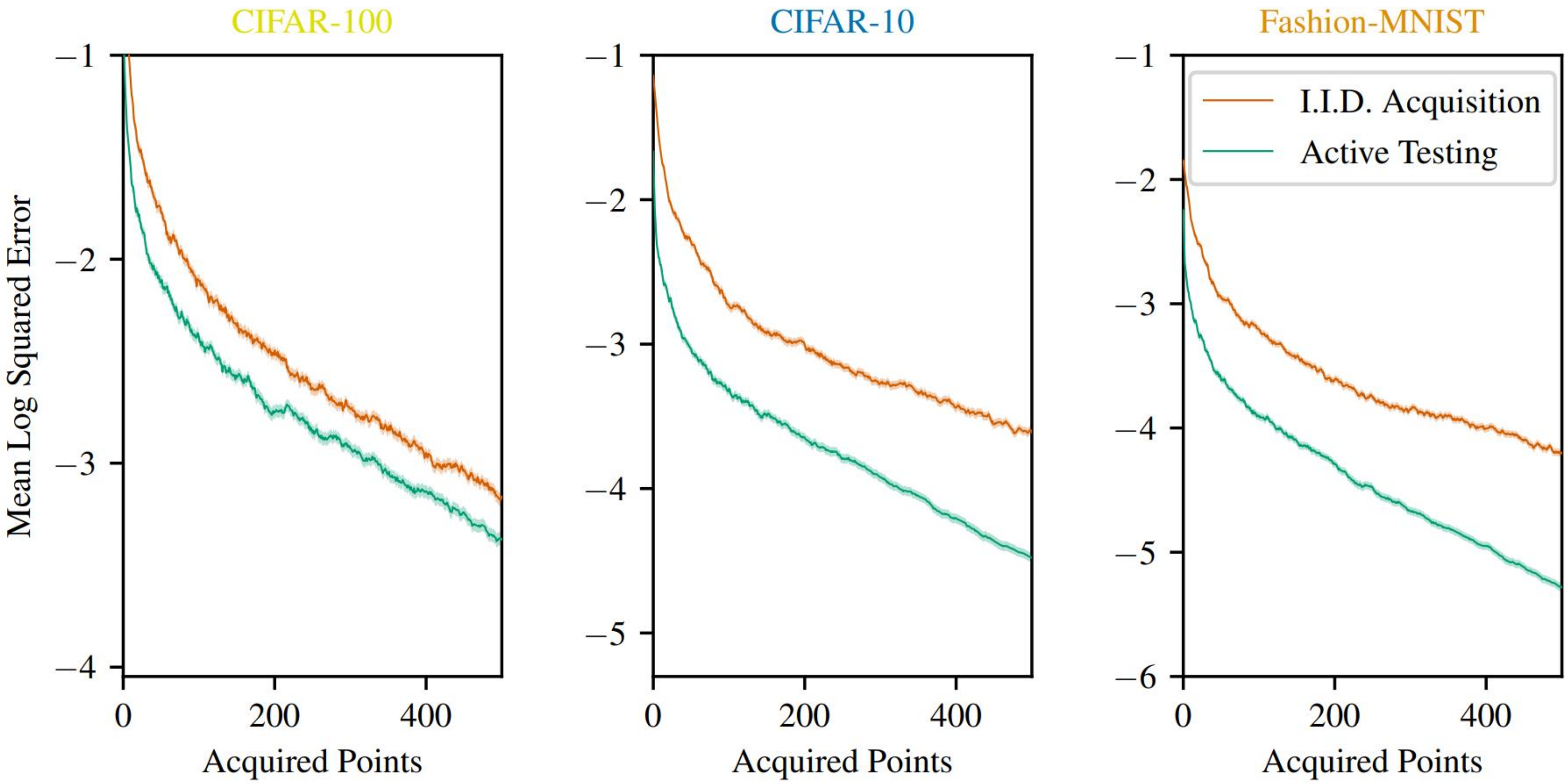


Figure A.4. Figure 6 (a) from the main paper but now showing the mean of the log squared difference instead of the median. Additionally, shading indicates the standard error of the log squared difference. Averages over 1000 random test set draws. See text and main paper for details.

Experiments

Optimal Proposals and Unbiasedness （ResNet-18 trained on CIFAR-10）

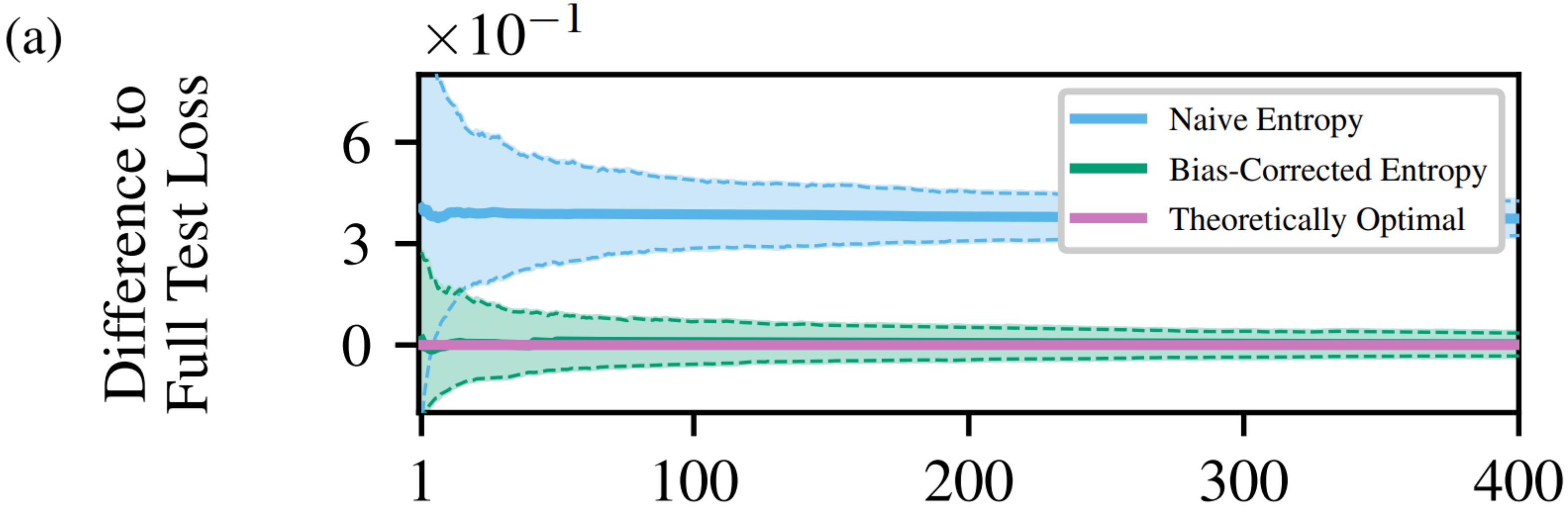


Figure 8. (a) Naively acquiring proportional to the predictive entropy and using the unweighted estimator $\hat{R}_{\text{iid}}$ leads to biased estimates with high variance compared to active testing with $\hat{R}_{\text{LURE}}$. Sampling from the unknown true loss distribution would yield unbiased, zero-variance estimates. While this is in practice impossible, the result validates a main theoretical assumption.

Conclusion

- Active testing allows much more precise estimates of test loss and accuracy using fewer data labels.

- The optimal sampling strategy for active testing is to sample in proportion to the true loss.

- Ensemble is an effective surrogate method for loss approximation.

THANKS