Enhancing Deep Batch Active Learning for Regression with Imperfect Data Guided Selection

Yinjie Min* [1], Furong Xu* [1], Xinyao Li[2], Changliang Zou†[1], Yongdao Zhou†[1]

1 School of Statistics and Data Science, Nankai University
2 School of Computer Science and Engineering, University of Electronic Science and Technology of China

NeurIPS 2025

Outline

1. Introduction & Motivation
2. Methodology: AGBAL Framework
3. Experimental Results
4. Conclusion & Impact

Active Learning Challenge in Regression

- **Active Learning (AL)** reduces annotation costs by selecting informative samples.
- **Informativeness** consists of two components:
  - Model sensitivity (measured via parameter gradients).
  - Predictive uncertainty (hard to estimate without labels).
- Regression tasks face fundamental challenge: no direct uncertainty estimation.

Key Problem

How to estimate predictive uncertainty for regression when true labels are unavailable?

Auxiliary Data: An Underutilized Resource

- Real-world scenarios often have **imperfect auxiliary data**:
  - Medical images with varying symptom manifestations.
  - Autonomous vehicle data from varied environments.
  - Industrial sensor logs with recording inaccuracies.
- These data are typically discarded due to distribution shifts.
- Our insight: Auxiliary data can provide **reliable uncertainty estimation** when properly weighted.

AGBAL: Core Idea

- **Key innovation**: Weighted loss approximation based on density ratio.
- Auxiliary data guides uncertainty estimation despite distribution shifts.
- Three-step process: density ratio estimation $\rightarrow$ auxiliary loss computation $\rightarrow$ weighting.

Mathematical Formulation

The Decompose of the Loss Gradient $\partial R(\theta, P)/\partial\theta$

$$
\begin{aligned}
\frac{\partial R(\theta; P)}{\partial \theta} &= \mathbb{E}_{X,Y\sim P} \frac{\partial l(Y, f(X;\theta))}{\partial \theta} \\
&= \mathbb{E}_{X,Y\sim P} \frac{\partial l(Y, f(X;\theta))}{\partial f(X;\theta)} \frac{\partial f(X;\theta)}{\partial \theta} \\
&= \mathbb{E}_{X\sim P_X} \phi_1(\theta; X) \left\{ \mathbb{E}_{Y\sim P_{Y|X}} \frac{\partial l(Y, f(X;\theta))}{\partial f(X;\theta)} \right\} \\
&= \mathbb{E}_{X\sim P_X} \phi_1(\theta; X) \phi_2(\theta; X).
\end{aligned}
$$

Mathematical Formulation

Density Ratio Weighting

We estimate the expected loss gradient using auxiliary data:

$$\widehat{\phi}_2(\theta; x) = \frac{1}{n'} \sum_{i=1}^{n'} \widehat{r}(X_i', Y_i') \cdot \frac{\partial l(Y_i', f(X_i'; \theta))}{\partial f(X_i'; \theta)},$$

where $\widehat{r}(x, y)$ is the density ratio estimator.

Auxiliary-Guided Gradient Kernel

$$K_{\text{grad-aux}}(x, x'; \theta) = \{\widehat{\phi}_2(\theta; x)\phi_1(\theta; x)\}^\top \{\widehat{\phi}_2(\theta; x')\phi_1(\theta; x')\}.$$

Theoretical Guarantees

Theorem 1 (Uncertainty Estimation Consistency)

Under Neural Tangent Kernel (NTK) theory, our auxiliary data guided estimator $\widehat{\phi}_2(\theta; x)$ provides a consistent surrogate for the true expected loss gradient, with variance proportional to the ridge estimator variance.

- Provides theoretical foundation for uncertainty estimation.
- Justifies use of distributionally shifted auxiliary data.
- Ensures reliability of the selection process.

Comprehensive Evaluation Setup

- **Datasets**: 2 synthetic (S1, S2) + 5 real-world (BIO, BIKE, DIAMOND, CT, STOCK).
- **Comparison**: 8 selection methods + random baseline.
- **Metrics**: Area Under Curve (AUC) of MSE learning curves, RMSE at step 10.
- **Settings**: $|\mathcal{L}_0| = 200$, batch size $N = 200$, 15 active learning steps.

AUC Performance Across Datasets (AGBAL vs BMDAL)

Table 1: Comparison of 8 selection methods across synthetic and real-world datasets in terms of AUC, where Avg Impro represents improvement over BMDAL averaged across 7 experiments.

Method	S1	S2	BIO	BIKE	DIAMOND	CT	STOCK	Avg Impro
random	0.928	1.421	0.451	0.459	21.009	0.380	0.392	–
lcmd	1.011	1.517	0.417	0.394	19.714	0.255	0.370	–
lcmd (AGBAL)	0.846	1.279	0.420	0.435	20.687	0.291	0.363	+0.6%
maxdist	0.863	1.310	0.428	0.439	20.643	0.270	0.389	–
maxdist (AGBAL)	0.834	1.266	0.417	0.401	21.179	0.298	0.361	+1.8%
kmeanspp	0.894	1.378	0.414	0.404	19.691	0.271	0.372	–
kmeanspp (AGBAL)	0.842	1.294	**0.406**	0.364	**19.613**	0.264	**0.355**	+4.5%
fw	0.953	1.448	0.434	0.455	21.115	0.347	0.388	–
fw (AGBAL)	0.899	1.341	0.418	0.404	21.840	0.310	0.377	+5.5%
bait	0.853	1.340	0.441	0.481	22.057	0.435	0.392	–
bait (AGBAL)	0.835	1.293	0.431	0.398	22.134	0.374	0.371	+6.3%
maxdet	0.876	1.318	0.418	0.486	19.702	0.320	0.378	–
maxdet (AGBAL)	**0.833**	**1.254**	0.409	**0.362**	19.846	**0.254**	0.359	+8.9%
maxdiag	0.903	1.401	0.451	0.597	24.594	0.526	0.415	–
maxdiag (AGBAL)	0.836	1.270	0.420	0.410	20.745	0.304	0.361	+18.0%

AGBAL Outperforms Across Datasets

- AGBAL consistently outperforms BMDAL (no auxiliary data).
- Significant improvements in both synthetic and real-world datasets.
- Best performance with maxdet and kmeanspp selection methods.

Visualization: Better Uncertainty Estimation

- AGBAL identifies truly high-uncertainty points.
- BMDAL selects well-trained points.

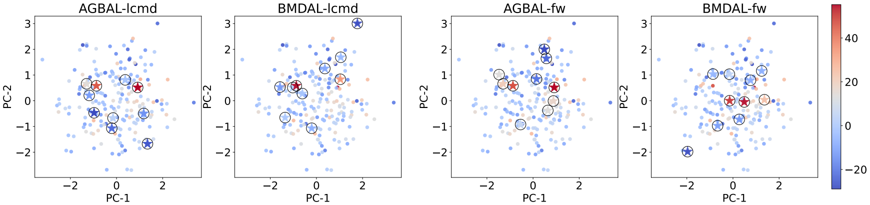

Figure 1: Visualization of the loss of selected points across four AL configurations. Left, right panels display lcmd, fw results of AGBAL and BMDAL, respectively.

Auxiliary Data Quality Analysis

Table 2: Worst case AUC comparison between AGBAL and BMDAL with distributional shift parameter $\zeta = 64$.

		Selection Methods						
Dataset	Method	maxdiag	maxdet	bait	fw	maxdist	kmeanspp	lcmd
S1	BMDAL	0.956	0.952	0.914	1.038	0.933	0.975	1.131
	AGBAL	0.942	0.918	0.904	1.038	0.947	0.940	0.975
	Improv.	1.5%	3.6%	1.1%	0.0%	-1.5%	3.6%	13.8%
S2	BMDAL	1.501	1.430	1.417	1.583	1.406	1.479	1.647
	AGBAL	1.437	1.398	1.390	1.543	1.436	1.426	1.468
	Improv.	4.3%	2.2%	1.9%	2.5%	-2.1%	3.6%	10.9%

AUC with Varying N_{aux}

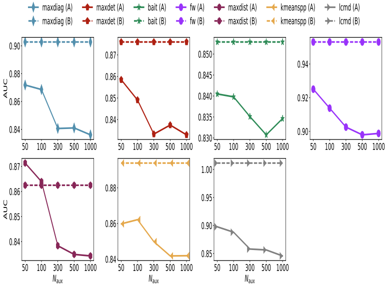

Figure 2: AUC for S1 (N_{aux} variation).

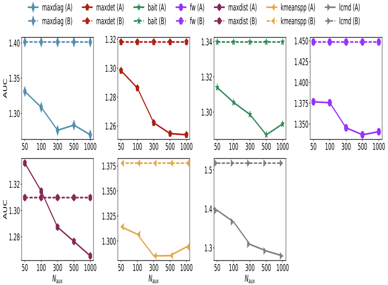

Figure 3: AUC for S2 (N_{aux} variation).

Contributions Summary

- **Theoretical**: Formal decomposition of informativeness into sensitivity and uncertainty.
- **Methodological**: AGBAL framework for leveraging imperfect auxiliary data.
- **Empirical**: Consistent improvements across diverse datasets and selection strategies.
- **Practical**: Lightweight implementation with minimal computational overhead.

Broader Impact & Future Directions

Positive Impacts

- Reduces annotation costs in resource-constrained domains.
- Enables use of otherwise discarded imperfect data.
- Applicable to healthcare, autonomous driving, industrial monitoring.

Future Work

- Extend to high-dimensional structured data (images, time series).
- Investigate privacy-preserving variants.
- Explore cross-modal auxiliary data utilization.

Thank You!