Balanced Active Inference

Boyu Chen [1] Zhixiang Zhou [1,2] Liuhua Peng [3] Zhonglei Wang [1]

[1] Xiamen University [2] Shanghai Innovation Institute [3] The University of Melbourne

boyuchen@stu.xmu.edu.cn, zhixiangzhou.stat@outlook.com, liuhua.peng@unimelb.edu.au, wangzl@xmu.edu.cn

Neural Information Processing Systems 2025

Outline

The Labeling Bottleneck

- Limited labeling budget impedes data-driven researches:
  - Medical analysis (Kononenko, 2001)
  - Remote sensing (Jean et al., 2016)
  - Population census (Cochran, 1977)
- Core challenge: Abundance of unlabeled data vs. scarcity of annotations (Traugott, 2011)
- Conventional approaches lack statistical efficiency:
  - Random sampling: Ignores instance informativeness
  - Heuristic-based selection: Lacks systematic prioritization

Active Inference

- Active learning selects uncertain instances to maximize label efficiency (Settles, 2009; Gal et al., 2017; Li et al., 2024)
- **Active inference** extends active learning to **statistical inference** by querying labels where the model exhibits high uncertainty using independent sampling (Angelopoulos et al., 2023; Zrnic and Candes, 2024)
- Balancing ideas from survey sampling may help to improve statistical efficiency

Balanced sampling

Core Problem

How to achieve higher statistical efficincy given fixed labeling cost?

- Balanced sampling (Hájek, 1964; Royall and Herson, 1973) :
  - Enforce structural representativeness in selected instances through auxiliary variable alignment
  - Improve statistical efficiency without increasing labeling cost
- Cube method to implement balanced sampling (Deville and Tillé, 2004; Tillé, 2011) :
  - **Flight phase** adjusts inclusion probabilities to approximate target distributions
  - **Landing phase** resolves residual imbalances via constrained optimization

Intuition

- **Active inference** leverages both model-derived uncertainty estimates and strategically acquired sample labels to **maximize statistical information**.
- The **cube method** optimizes sampling through balancing constraints, reconciling **variance reduction** with asymptotic normality.

Outline

Problem Setup

- Datasets:
  - Labeled: $\mathcal{D}_l = \{(X_j, Y_j)\}_{j=1}^m$
  - Unlabeled: $\mathcal{D}_u = \{X_i\}_{i=1}^n$ (labels unobserved)

- Target parameter: Population mean of unobserved labels in $\mathcal{D}_u$

$$\theta^* = \frac{1}{n}\sum_{i=1}^{n} Y_i$$

Active Inference Framework

- Resources:
  - Labeling budget: Acquire n_b labels from $\mathcal{D}_u$

$$\xi_i = \mathbb{I}(\text{label } Y_i \text{ acquired}), \quad \mathbb{E}\left[\sum_{i=1}^{n} \xi_i\right] = n_b$$

  - $\hat{f}(\cdot)$: predictive model of label trained on labeled $\mathcal{D}_l$
  - $\hat{u}(X_i)$: the model's estimated uncertainty measure for instance i, typically approximating $|Y_i - \hat{f}(X_i)|$
- Label strategy: Normalized sampling probabilities for each instance

$$\pi(X_i) = \frac{n_b}{n} \cdot \frac{\hat{u}(X_i)}{\frac{1}{n}\sum_{j=1}^{n} \hat{u}(X_j)}$$

Balanced Active Inference

- **Balanced constraint**: Enforce uncertainty balancing

$$\sum_{i=1}^{n} \frac{\hat{u}(X_i)\xi_i}{\pi(X_i)} = \sum_{i=1}^{n} \hat{u}(X_i)$$

- Employ the cube method for sampling to satisfy the balancing constraint

- Estimate the mean of unobserved labels through a Generalized Difference estimator (Cassel et al., 1976):

$$\tilde{\theta} = \frac{1}{n} \sum_{i=1}^{n} \left[\hat{f}(X_i) + \{Y_i - \hat{f}(X_i)\} \frac{\xi_i}{\pi(X_i)} \right]$$

Outline

1. Introduction
2. Methodology
3. Theoretical Results
4. Experiments
5. Conclusion

Asymptotic Normality

- The theorem establishes the asymptotic normality of the balanced active inference estimator

Theorem

Under regularity conditions, for the balanced active sampling scheme, the estimator $\tilde{\theta}$ satisfies:

$$\sqrt{n}(\tilde{\theta} - \theta^*) \xrightarrow{d} \mathcal{N}(0, V_0)$$

where $V_0 = \mathbb{E}(\varepsilon_1^2/\pi_1) + \mathrm{Var}[f(X_1)]$.

- The asymptotic variance includes the scaled noise variance and the intrinsic variance of $f(X_1)$

Statistical Efficiency

Theorem

Under regularity conditions, for traditional active inference with independent $\{\xi_i\}_{i=1}^n$, the estimator $\hat{\theta}$ satisfies:

$$\sqrt{n}(\hat{\theta} - \theta^*) \xrightarrow{d} \mathcal{N}\left(0, V_0 + \mathbb{E}\left[\left(f(X_1) - \hat{f}(X_1)\right)^2 \left(\frac{1}{\pi(X_1)} - 1\right)\right]\right).$$

- We can show that the asymptotic variance V_1 of the traditional active inference estimator with $\{\xi_i\}_{i=1}^n$ being independent (Poisson sampling in survey sampling) satisfies

$$V_1 - V_0 \succcurlyeq 0$$

.

Outline

Stabilization via mixed sampling

- Direct implementation of $\pi(X_i) \propto \hat{u}(X_i)$ may be statistically inefficent due to variance inflation from near-zero $\pi(X_i)$'s.
- A τ-mixed rule is proposed (Zrnic and Candes, 2024):

$$\pi^{(\tau)}(X_i) = \tau \cdot \frac{n_b \hat{u}(X_i)}{\sum_{j=1}^{n} \hat{u}(X_j)} + (1-\tau) \cdot \frac{n_b}{n}$$

  - $\tau \in [0, 1]$ controls the trade-off between uncertainty prioritization and robustness.
  - τ defaults to 0.5 in implementation.

Algorithm

1. Split data into training/test sets
2. Train two XGBoost models:
   - $\hat{f}(\cdot)$: Label predictor
   - $\hat{u}(X_i)$: Uncertainty estimator ($|\ \hat{f}(X_i) - Y_i\ |$)
3. Compute blended sampling probabilities:

$$\pi^{(\tau)}(X_i) = \tau \cdot \frac{n_b \hat{u}(X_i)}{\sum_{j=1}^{n} \hat{u}(X_j)} + (1-\tau) \cdot \frac{n_b}{n}$$

4. Apply the cube method to obtain $\{\xi_i\}$ with balancing:

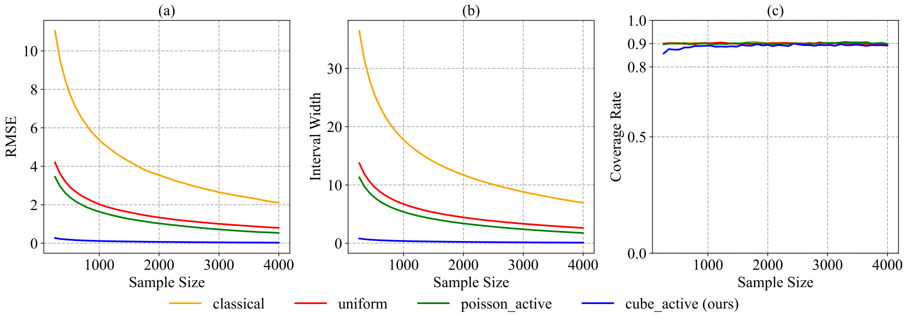
$$\sum_{i=1}^{n} \frac{\hat{u}(X_i)\xi_i}{\pi^{(\tau)}(X_i)} = \sum_{i=1}^{n} \hat{u}(X_i)$$

5. Compute balanced GD estimator $\tilde{\theta}$

Experimental Setup

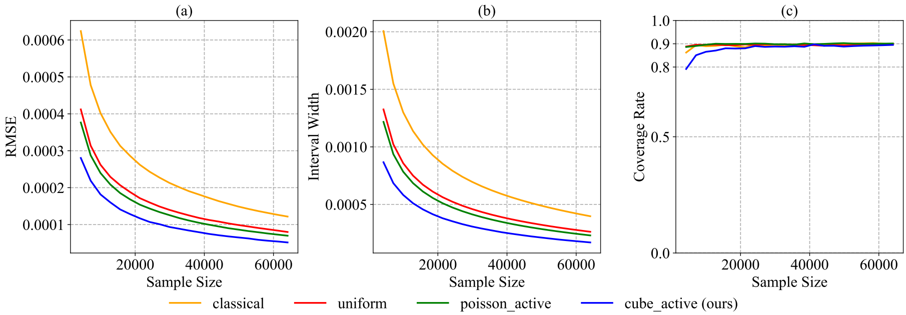
- **Datasets**:
  - Synthetic: Linear, Nonlinear, Friedman
  - Real: 6 regression (Bike, Communities, Concrete, Energy, Life, Superconductor)
    2 classification (Credit Fraud Detection, Post-election)
- **Baselines**:
  - Classical: Simple random sampling
  - Uniform: Active inference with uniform sampling
  - Traditional-active: Active inference with independent sampling
- **Evaluation**:
  - RMSE, CI width, coverage rate (90% nominal)
  - $T = 10,000$ Monte Carlo repetitions

Performance: Bike Sharing Dataset



- 17k hourly rentals (weather/temporal features)
- ↓30-50% RMSE vs alternatives
- **Tighter inference**: Narrowest intervals achieved
- **Reliable coverage**: Coverage aligns with nominal 90% target

Performance: Credit Fraud Dataset



- 492 frauds / 284k transactions (PCA features)
- **Best precision**: Tightest CIs + accurate coverage
- **Superior RMSE**: Outperforms all benchmarks
- Robust in high-imbalance scenarios and Classification problems

Accuracy Improvement (Label Budget = 0.1)

Table 1: Comparison of Confidence Interval Width with 0.1 Label Budget across Methods (Partial Results)

Dataset	Classical	Uniform	Traditional-active	**Cube-active**
Linear	0.3171	0.1690	0.1443	**0.0722**
Nonlinear	0.7786	0.3626	0.2963	**0.1370**
Communities	0.0722	0.0516	0.0509	**0.0466**
Concrete	7.1217	3.9367	3.7415	**2.6406**
Life	3.0828	1.4509	1.3050	**0.7265**
Post-election	0.0571	0.0426	0.0403	**0.0381**
Superconductor	3.2664	1.6711	1.5950	**1.1680**

Outline

Conclusion

- We propose balanced active inference, integrating balanced sampling with active inference to address variance inflation and dataset imbalance in label-scarce settings.
- Theoretical guarantees establish the estimator's asymptotic normality and statistical efficiency.
- Experiments across datasets demonstrate RMSE reduction, label savings, and reliable coverage versus traditional active inference.

Thank you for your attention!

References I

Anastasios N. Angelopoulos, Stephen Bates, Clara Fannjiang, Michael I. Jordan, and Tijana Zrnic. Prediction-powered inference. *Science*, 382(6671):669–674, 2023. doi: 10.1126/science.adi6000. URL https://www.science.org/doi/abs/10.1126/science.adi6000.

Claes M Cassel, Carl E Särndal, and Jan H Wretman. Some results on generalized difference estimation and generalized regression estimation for finite populations. *Biometrika*, 63(3): 615–620, 12 1976. ISSN 0006-3444. doi: 10.1093/biomet/63.3.615. URL https://doi.org/10.1093/biomet/63.3.615.

William G Cochran. *Sampling Techniques*. John Wiley & Sons, 1977.

Jean-Claude Deville and Yves Tillé. Efficient balanced sampling: The cube method. *Biometrika*, 91(4):893–912, 2004. ISSN 0006-3444. doi: 10.1093/biomet/91.4.893. URL https://doi.org/10.1093/biomet/91.4.893.

Yarin Gal, Riashat Islam, and Zoubin Ghahramani. Deep Bayesian active learning with image data. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 1183–1192. PMLR, 06–11 Aug 2017. URL https://proceedings.mlr.press/v70/gal17a.html.

Jaroslav Hájek. Asymptotic theory of rejective sampling with varying probabilities from a finite population. *Annals of Mathematical Statistics*, 35(4):1491–1523, 1964.

Neal Jean, Marshall Burke, Michael Xie, William Matthew Davis, David B Lobell, and Stefano Ermon. Combining satellite imagery and machine learning to predict poverty. *Science*, 353 (6301):790–794, 2016.

References II

Igor Kononenko. Machine learning for medical diagnosis: History, state of the art and perspective. *Artificial Intelligence in Medicine*, 23(1):89–109, 2001.

Dongyuan Li, Zhen Wang, Yankai Chen, Renhe Jiang, Weiping Ding, and Manabu Okumura. A survey on deep active learning: Recent advances and new frontiers. *IEEE Transactions on Neural Networks and Learning Systems*, 2024.

Richard M Royall and Jay Herson. Robust estimation in finite populations I. *Journal of the American Statistical Association*, 68(344):880–889, 1973.

Burr Settles. Active learning literature survey. *Technical Report*, 2009.

Yves Tillé. Ten years of balanced sampling with the cube method: An appraisal. *Survey methodology*, 37(2):215–226, 2011.

Michael Traugott. The accuracy of opinion polling and its relation to its future. In *The Oxford Handbook of American Public Opinion and the Media*, pages 316–331. Oxford University Press, 2011.

Tijana Zrnic and Emmanuel Candes. Active statistical inference. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 62993–63010. PMLR, 21–27 Jul 2024. URL `https://proceedings.mlr.press/v235/zrnic24a.html`.