Balanced Active Inference

Boyu Chen[1]*, **Zhixiang Zhou**[1,2]*, **Liuhua Peng**[3]†, **Zhonglei Wang**[4]†
[1]School of Economics, Xiamen University, Xiamen, China
[2]Shanghai Innovation Institute, Shanghai, China
[3]School of Mathematics and Statistics, The University of Melbourne, Melbourne, Australia
[4]Wang Yanan Institute for Studies in Economics, Xiamen University, Xiamen, China
boyuchen@stu.xmu.edu.cn, zhixiangzhou.stat@outlook.com,
liuhua.peng@unimelb.edu.au, wangzl@xmu.edu.cn

Abstract

Limited labeling budget severely impedes data-driven research, such as medical analysis, remote sensing and population census, and active inference is a solution to this problem. Prior works utilizing independent sampling have achieved improvements over uniform sampling, but its insufficient usage of available information undermines its statistical efficiency. In this paper, we propose balanced active inference, a novel algorithm that incorporates balancing constraints based on model uncertainty utilizing the cube method for label selection. Under regularity conditions, we establish its asymptotic properties and also prove that the statistical efficiency of the proposed algorithm is higher than its alternatives. Various numerical experiments, including regression and classification in both synthetic setups and real data analysis, demonstrate that the proposed algorithm outperforms its alternatives while guaranteeing nominal coverage. Our code is available at: https://github.com/Uninfty/Balanced_Active_Inference

1 Introduction

Machine learning has revolutionized data-driven fields, yet its success still hinges on access to high-quality labeled data, which is a critical component for reliable inference. This dependency on labeled data is particularly pronounced in precision-sensitive fields such as medical diagnostics [1], financial risk assessment [2], and remote sensing [3], where the accuracy of predictions directly affects decision making. However, labeling remains a costly and time-consuming process, resulting in a persistent gap between the abundance of unlabeled data and the limitation of annotated resources [4, 5]. Conventional methods, including random sampling and heuristic-based selection, lack systematic prioritization of informative instances, leading to inefficient labeling [6, 7].

Active learning addresses the labeling bottleneck by iteratively selecting uncertain instances to maximize label efficiency [7, 8, 9]. Extending it to statistical inference, active inference [10] strategically queries labels where the model exhibits high uncertainty using independent sampling, and makes statistical inference, including confidence intervals and hypothesis tests, based on the acquired labels. Combined with prediction-powered inference [11], integrating model predictions with limited labeled instances, active inference outperforms random sampling in terms of statistical efficiency. Nonetheless, its reliance on independent sampling induces variance inflation and often yields imbalanced datasets. These issues may degrade estimation performance [12], especially under systematic bias or distribution shifts, ultimately limiting the practical utility of existing methods.

*Equal contribution.
†Corresponding author.

39th Conference on Neural Information Processing Systems (NeurIPS 2025).

Therefore, a critical challenge in active inference is to further improve statistical efficiency under a constrained labeling budget. To address the inefficiency introduced by traditional active inference strategies, we propose balanced active inference. This method strategically selects informative instances while maintaining statistical representativeness of the population. The key idea is to enforce structural balance in the selected samples so that they preserve key characteristics of the overall data distribution, thereby improving statistical efficiency without increasing the labeling budget. Our method builds upon the principle of covariate balancing, a classical technique in survey sampling and causal inference that aligns the distribution of sampled instances with population-level summaries of auxiliary variables [13, 14].

To enforce covariate balance in active instance selection, we implement balanced active inference using the cube method [15, 16], a sampling algorithm that by consecutively updating selection probabilities to satisfy certain balancing constraints. The cube method operates in two phases: a *flight phase*, which iteratively adjusts inclusion probabilities to approximate target distributions, and a *landing phase*, which finalizes the selection by resolving residual imbalances through constrained optimization.

Compared with existing works, the proposed balanced active inference framework offers three advantages. First, we reconceptualize model uncertainty estimates as dynamic auxiliary variables, enabling simultaneous optimization for informativeness and representativeness during instance selection. Second, we introduce a balancing condition that constrains the weighted sum of uncertainties among the labeled instances to match the corresponding population total, effectively preventing oversampling from specific uncertainty regions and promoting more stable estimates. Third, we provide a theoretical guarantee that he proposed balanced active inference framework yields lower asymptotic variance compared to conventional active inference methods.

Our contributions can be summarized as follows.

- Innovatively increase the statistical efficiency of active inference through balanced sampling.
- Incorporate model uncertainty as an auxiliary covariate in balanced sampling using a cube method.
- Provide closed-form expressions for the asymptotic variance, illustrating the variance reduction property of the proposed method.
- Demonstrate the broad applicability and superiority of our method through extensive experiments on diverse real-world and synthetic datasets.

Related work. *(1) Label Inference.* The challenge of drawing valid inferences from partially labeled data has inspired diverse methodological developments. Traditional methods for missing data, such as inverse probability weighting and multiple imputation [17, 18], established foundational principles for handling label scarcity. Semi-supervised inference methods [19, 20] demonstrated how unlabeled data could improve efficiency in parameter estimation, particularly under smoothness assumptions. A pivotal advancement emerged with prediction-powered inference (PPI) [11], which integrates machine learning predictions with a small labeled dataset to estimate population quantities. By treating predictions as noisy proxies for missing labels, PPI constructs debiased estimators while maintaining statistical validity. However, its reliance on uniform random sampling undermines its statistical efficiency, as it fails to prioritize informative instances for labeling.

(2) Active Learning. Active learning addresses label efficiency by adaptively selecting instances for annotation based on model uncertainty [7, 8, 9, 21]. While classical active learning focuses on optimizing model training [22], recent work extends these principles to statistical inference. Active inference [10] formalizes this paradigm, employing unequal-probability sampling to prioritize uncertain instances. By combining actively acquired labels with model predictions via a GD (general difference) estimator [23], it incorporates auxiliary information to correct for sampling bias and improve estimation efficiency. Despite its advantages, the independent sampling mechanism inherent to active inference introduces variance inflation, as independent label selections may yield imbalanced instances that poorly represent critical regions of the data distribution.

(3) Balanced Sampling. The use of auxiliary information in finite population sampling is widely recognized for enhancing estimation precision. Classical methods, such as stratification [12, 24] and probability proportional-to-size sampling [25, 26], exploit known auxiliary variables to reduce variance. More advanced balanced sampling techniques impose equality constraints, ensuring that

weighted sample summations match population totals, guaranteeing significant variance reduction [27, 28].

2 Problem setup

Suppose we have access to two datasets: a small labeled dataset $\mathcal{D}_l = \{(X_j, Y_j)\}_{j=1}^m$ with m instances, independently and identically distributed (i.i.d.) from a distribution $\mathbb{P} = \mathbb{P}_X \times \mathbb{P}_{Y|X}$, and a large unlabeled dataset $\mathcal{D}_u = \{X_i\}_{i=1}^n$ with n instances drawn i.i.d. from $\mathbb{P}_X$, where the corresponding labels $\{Y_i\}_{i=1}^n$ are unobserved. Denote $\mathcal{X} \subseteq \mathbb{R}^d$ as the feature space, and $\mathcal{Y} \subseteq \mathbb{R}$ as the label space. The primary goal is to estimate the population mean of the unobserved labels in $\mathcal{D}_u$, defined as

$$\theta^* = \mathbb{E}(Y_1). \tag{1}$$

To leverage feature-label relationships, a predictive model $\hat{f} : \mathcal{X} \to \mathcal{Y}$ is trained on $\mathcal{D}_l$. Additionally, a labeling budget of n_b allows the query of labels for a subset of $\mathcal{D}_u$. Let $\xi_i \in \{0, 1\}$ denote an indicator for whether the label Y_i is acquired for the instance $X_i \in \mathcal{D}_u$ with $\mathbb{E}[\sum_{i=1}^n \xi_i] = n_b$. The challenge lies in designing an estimator that optimally combines the predictive power of $\hat{f}(\cdot)$ with strategically sampled labels to reduce the variance of an estimator of (1).

Active inference employs a machine learning model $\hat{f}(\cdot)$ to predict labels for unobserved instances, coupled with an adaptive sampling strategy that corrects potential prediction biases. The sampling policy $\pi : \mathcal{X} \to [0, 1]$ determines label acquisition probabilities through uncertainty quantification. Specifically, let $\hat{u}(X_i)$ represent the model's estimated uncertainty measure for instance i, typically equal to $|Y_i - \hat{f}(X_i)|$. The sampling probabilities are then normalized as

$$\pi(X_i) = \frac{n_b}{n} \cdot \frac{\hat{u}(X_i)}{\frac{1}{n}\sum_{j=1}^n \hat{u}(X_j)}, \tag{2}$$

ensuring $\mathbb{E}(\sum_{i=1}^n \xi_i) = n_b$ through scaling. This allocation prioritizes regions where the model exhibits higher prediction uncertainty.

A GD estimator [23] of (1) for the inference,

$$\hat{\theta} = \frac{1}{n}\sum_{i=1}^n \left[\hat{f}(X_i) + (Y_i - \hat{f}(X_i))\frac{\xi_i}{\pi(X_i)}\right], \tag{3}$$

which is unbiased regardless of the form of the predictive model $\hat{f}(\cdot)$. Specifically, prediction-powered inference [11] emerges as a special case when $\pi(X_i) = n_b/n$, corresponding to uniform random sampling.

3 Balanced active inference

Our framework advances active inference by integrating balanced sampling through the cube method, which enforces structural constraints on auxiliary variables to improve statistical efficiency. Specifically, we impose the balancing condition

$$\sum_{i=1}^n \frac{\hat{u}(X_i)\xi_i}{\pi(X_i)} = \sum_{i=1}^n \hat{u}(X_i), \tag{4}$$

where $\hat{u}(X_i)$ quantifies the uncertainty of the predictive model $\hat{f}(X_i)$. The balancing constraint (4) ensures that the selected instances preserve the population structure of model uncertainties, mitigating selection bias. To operationalize this, we employ the cube method, a two-phase sampling algorithm that first iteratively adjusts inclusion probabilities to satisfy balancing constraints (flight phase), and then resolves residual imbalances via a landing phase when exact balancing is infeasible [16]; see Section S1 of Supplementary Material for an introduction to the cube method.

Once we obtain sampling indicators $\{\xi_i\}_{i=1}^n$ through the proposed balanced sampling strategy, we still consider a GD estimator

$$\tilde{\theta} = \frac{1}{n}\sum_{i=1}^n \left[\hat{f}(X_i) + (Y_i - \hat{f}(X_i))\frac{\xi_i}{\pi(X_i)}\right]. \tag{5}$$

Different from (3), the sampling indicators associated with (5) satisfy the balancing constraint (4).

Remark 1. *It may seem preferable to balance $\hat{f}(X_i)$ as well, since $\hat{f}(X_i)$ directly captures predictive information about Y_i. However, enforcing the balancing constraint on $\hat{f}(X_i)$ leads the estimator in* (5) *to degenerate into the form $n^{-1}\sum_{i=1}^{n}(Y_i\xi_i/\pi_i)$, thereby forfeiting any benefits from active sampling. In contrast, balancing on $\hat{u}(X_i)$ preserves the correction term in* (5)*. If $\hat{u}(X_i)$ accurately approximates the residual error, i.e., $\hat{u}(X_i) \approx Y_i - \hat{f}(X_i)$, then the proposed estimator $\tilde{\theta} \approx n^{-1}\sum_{i=1}^{n} Y_i$, thereby improving the statistical efficiency of the GD estimator.*

Uncertainty measures To effectively quantify prediction uncertainty for different task types, we define specific uncertainty measures tailored to regression and classification settings. For regression problems, the uncertainty is captured by the absolute residual $|Y_i - f(X_i;\hat{\theta})|$. For classification tasks, let $p(X_i) = (p_1(X_i), \ldots, p_K(X_i))$ represent the predicted class probabilities. The uncertainty measure is defined as $u(X_i) = \frac{K}{K-1}\left(1 - \max_{j\in[K]} p_j(X_i)\right)$, which attains its maximum when the model is maximally uncertain and decreases to zero when the model exhibits high confidence in a single class. The results of using other uncertainty quantification are provided in Section S6 in the Supplementary Material.

Stabilization via mixed sampling As suggested by the traditional active inference literature [10], direct implementation of $\pi(X_i) \propto \hat{u}(X_i)$ risks instability when $\hat{u}(X_i)$ is misspecified. To safeguard against variance inflation from near-zero $\pi(X_i)$, we consider the following τ-mixed rule:

$$\pi^{(\tau)}(X_i) = \tau \cdot \frac{n_b \hat{u}(X_i)}{\sum_{j=1}^{n} \hat{u}(X_j)} + (1-\tau)\cdot \frac{n_b}{n}, \tag{6}$$

where $\tau \in [0,1]$ controls the trade-off between uncertainty prioritization and robustness. Empirical analysis demonstrates that $\tau = 0.5$ achieves favorable bias-variance trade-offs across diverse scenarios, aligning with findings in [10]. This mixture ensures $\pi^{(\tau)}(X_i) > 0$ universally while retaining adaptivity to $\hat{u}(X_i)$. The relevant sensitivity analysis of τ is discussed in Section S5 in the Supplementary Material.

Implementation Algorithm 1 outlines our cube-based balanced active inference procedure. The flight phase enforces the balancing constraint via iterative geometric projections, followed by a landing phase that minimizes deviation from target inclusion probabilities. The integration of the cube method with uncertainty-aware sampling distinguishes it from conventional balanced sampling, as the auxiliary variable $\hat{u}(x)$ directly links to the statistical efficiency of the GD estimator. The computational complexity of the cube method is $\mathcal{O}(n \times p^2)$ [16], where n is the population size and p the number of balancing covariates. In our implementation, the uncertainty measure u is used as the only auxiliary covariate, and the complexity of balanced sampling reduces to $\mathcal{O}(n)$.

This synthesis of balanced sampling and active inference provides a unified framework for semi-supervised mean estimation, where model predictions guide sampling, while balancing constraints safeguard against distributional shifts—a key advancement over existing works.

Extension to M-estimation Consider a general M-estimation problem that, given a class of functions $f(X_i;\theta)$,

$$\theta^* = \arg\min_{\theta} \mathbb{E}\left[L(X_1, Y_1; \theta)\right],$$

where $L(X_1, Y_1; \theta)$ is a loss function measuring the discrepancy between the true label Y_1 and the predicted label $f(X_1;\theta)$, and θ is the parameter of interest. The goal is to estimate θ^* using the labeled data $\mathcal{D}_l$ and the unlabeled data $\mathcal{D}_u$. Suppose that there have been an estimation $f(X_i;\hat{\theta})$ trained on the labeled data $\mathcal{D}_l$ and a uncertainty estimator $\hat{u}(X_i)$ trained on $L(X_i, Y_i; \hat{\theta}) - L(X_i, f(X_i;\hat{\theta}); \hat{\theta})$, a sampling scheme $\{\pi(X_i)\}_{i=1}^{n}$ is derived following (2) given a budget n_b. We use the cube method to generate an assignment $\{\xi_i\}_{i=1}^{n}$ such that the balancing constraint in (4) holds. The proposed estimator for M-estimation is then defined as

$$\tilde{\theta} = \arg\min_{\theta} \frac{1}{n}\sum_{i=1}^{n}\left\{ L(X_i, f(X_i;\hat{\theta});\theta) + \left[L(X_i, Y_i;\theta) - L(X_i, f(X_i;\hat{\theta});\theta)\right]\frac{\xi_i}{\pi(X_i)}\right\}. \tag{7}$$

Algorithm 1 Balanced active inference

1: Train a prediction model $\hat{f}(\cdot)$ on labeled training data $\mathcal{D}_l = \{(X_j, Y_j)\}_{j=1}^m$.
2: Compute residuals $e_j = \hat{f}(X_j) - Y_j$ and train an uncertainty model $\hat{u}(X_j) \approx |e_j|$.
3: Sample n_b instances for labeling given the label budget $b = \mathbb{E}[n_b/n]$.
4: **for** $X_i \in \mathcal{D}_u$ **do**
5: Predict label $\hat{Y}_i = \hat{f}(X_i)$.
6: Predict uncertainty $\hat{u}_i = \hat{u}(X_i)$.
7: Compute active probability $p_{a,i} = \frac{b \cdot \hat{u}_i}{\bar{u}}$, where $\bar{u} = \frac{1}{n}\sum \hat{u}_i$ and uniform probability $p_{e,i} = b$.
8: Blend with uniform sampling: $\pi_i^{(\tau)} = \tau p_{a,i} + (1-\tau) p_{e,i}$.
9: **end for**
10: Apply the cube method on $\{\pi_i^{(\tau)}\}_{i=1}^n$ to get $\{\xi_i\}_{i=1}^n$ satisfying the balancing constraint

$$\sum_{i=1}^{n} \frac{\hat{u}_i \xi_i}{\pi_i^{(\tau)}} = \sum_{i=1}^{n} \hat{u}_i.$$

11: Compute $\tilde{\theta}$ via (5).

The first term in the curly braces is the loss function evaluated at the predicted label $f(X_i; \hat{\theta})$, while the second term is the correction term that accounts for the difference between the loss function evaluated at the true label Y_i and the predicted label $f(X_i; \hat{\theta})$. When the prediction $f(X_i; \hat{\theta})$ is poor, the correction helps to reduce the variance of the estimator by adaptively adjusting the inclusion probabilities based on the uncertainty of the predictions. By leveraging the cube method, we ensure that the selected instances are balanced with respect to the uncertainty estimates, leading to a more efficient estimation process.

4 Theoretical properties

Before presenting the main theoretical results, we introduce some generality assumptions under which our analysis is conducted.

Assumption 1. *There exists a function f, such that*

$$Y_i = f(X_i) + \varepsilon_i,$$

where ε_i satisfies $\mathbb{E}(\varepsilon_i | X_i) = 0$.

Assumption 2. *Assume $\mathbb{E}(Y_1^2 + f(X_1)^2 + \hat{f}(X_1)^2 + \hat{u}(X_1)^2) < \infty$.*

Assumption 3. *There exists a positive constant $c \in (0, 1)$, such that*

$$\mathbb{P}\left(\xi_i = 1 \mid X_1, \ldots, X_n\right) \in [c, 1-c].$$

Assumption 4. *For any $k \in \mathbb{N}$, we have with probability one,*

$$\lim_{n\to\infty} \sup_{i_1,\ldots,i_k} \left| \mathbb{E}\left(\prod_{j=1}^{k} \left(\xi_{i_j} - \pi_{i_j}\right) \mid X_1, \ldots, X_n \right) \right| = 0.$$

Assumption 5. *The estimator $\hat{f}(X_i)$ and $\hat{u}(X_i)$ statisfy*

$$\mathbb{E}\left\{ \left[f(X_1) - \hat{f}(X_1) - \operatorname{sgn}\left(f(X_1) - \hat{f}(X_1) \right) \hat{u}(X_1) \right]^2 \right\} = o(1),$$

where $\operatorname{sgn}(x) = I\{x > 0\} - I\{x < 0\}$ *with* $\operatorname{sgn}(0) = 0$.

Assumption 1 states that there exists an underlying regression function f such that the observed outcomes Y_i can be decomposed into a systematic component $f(X_i)$ and a zero-mean noise term ε_i, conditional on X_i. This is a standard assumption in supervised learning and nonparametric regression, ensuring the model is well-specified in the conditional expectation sense. Assumption 2 imposes a moment condition that ensures the second moments of the response variable Y_i, the true

regression function $f(X_i)$, and its estimator $\hat{f}(X_i)$ are all finite almost surely. Assumption 3 is standard in the semi-supervised inference literature and has been adopted in works such as [29]. Assumption 4 follows a conjecture from [30], asserting that as $n \to \infty$, the dependence among inclusion indicators for any fixed subset of instances vanishes asymptotically. Assumption 5 states that $\mathrm{sgn}(f(X_i) - \hat{f}(X_i))\hat{u}(X_i)$ is a good estimator of $f(X_i) - \hat{f}(X_i)$, and we impose this to facilitate analytical tractability.

Theorem 1 (Asymptotic normality for mean estimation)**.** *Suppose Assumptions 1–5 hold,*

(a) For the balanced active sampling scheme, the estimator $\tilde{\theta}$ defined in (5) *satisfies*

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

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

(b) For the traditional active inference with $\{\xi_i\}_{i=1}^n$ being independent, if $\mathbb{E}[\hat{f}(X_1)] = \theta^*$, the estimator $\hat{\theta}$ in (3) *satisfies:*

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

The proof of Theorem 1 is provided in Appendix A. Theorem 1 establishes the asymptotic normality of the balanced active inference estimator. Specifically, the asymptotic variance of the proposed method consists of two components, including the scaled noise variance and the intrinsic variance of $f(X_1)$. Notably, this variance is reduced compared to that of the classical active inference methods, as the additional variability from estimation error in $\hat{f}(X_1)$ and non-uniform sampling is mitigated through the balancing constraint. Furthermore, the balanced active inference does not require the predictive model $\hat{f}(\cdot)$ to be unbiased, so it is more robust than existing active inference methods.

Remark 2. *Assumption 5 is introduced to enable the derivation of an explicit expression for the asymptotic variance of the proposed balanced active inference method, but it is not strictly necessary in practice. We conjecture that the proposed estimator retains its variance reduction benefits even when Assumption 5 is mildly violated. Empirical evidence in Section 5 supports our conjecture.*

5 Experiments

In this section, we evaluate the performance of our proposed method through both numerical simulations and real data applications. More numerical results about the real data analysis are shown in Section S3 of the Supplementary Material.

Datasets We consider three synthetic experiments, including a linear setup, a nonlinear setup and a Friedman setup [31]. Besides, we also consider real data applications, including six regression datasets and two classification datasets. Regression datasets include Bike Sharing [32], Communities and Crime [33], Concrete Compressive Strength [34], Energy Efficiency [35], Life Expectancy [36], Superconductivity Data [37], and binary classification datasets including Credit Fraud Detection [38] and Post-election Survey Research [39].

Baselines Our method (*cube-active*) is compared with three baseline methods across all experiments. As introduced in Section 2 and Section 3, the baselines include: (i) a simple random sampling labeling strategy using sample mean estimator $\hat{\theta} = \frac{1}{n_b}\sum_{j=1}^{n_b} Y_j$ with no involvement of machine learning models (*classical*); (ii) prediction-powered inference with GD estimator $\hat{\theta} = \frac{1}{n}\sum_{i=1}^{n}\left[\hat{f}(X_i) + (Y_i - \hat{f}(X_i))\frac{\xi_i}{\pi}\right]$ using a uniform random labeling strategy (*uniform*); and (iii) active inference based on independent sampling strategies designed using machine learning model predictions with GD estimator (*traditional-active*).

Evaluation metrics For the four methods, we first report their Root Mean Squared Error (RMSE), which directly quantifies the deviation between point estimates and the true population mean. Additionally, leveraging the asymptotic normality of each method and its respective variance estimator, we compute the confidence intervals at a fixed confidence level 0.9, and compare their empirical coverage rates with respect to the true population mean. The variance estimators of each method are shown in Section S2 of the Supplementary Material.

Experiment setup For all the datasets, the reported results are based on $T = 10\,000$ Monte Carlo simulations. Following recommendations from the traditional active inference literature [10], we set $\tau = 0.5$ in (6) across all experiments. All predictive models are obtained by XGBoost [40]. All experiments were conducted on a machine equipped with an Intel® Xeon® Gold 5118 CPU @ 2.30GHz, featuring 12 cores and 24 threads.

Protocol The numerical analysis proceeds as follows. First, we generate (X, Y) pairs and randomly split them into training/test sets. Then, an XGBoost regressor $\hat{f}(\cdot)$ is trained on the training set, and an uncertainty model $\hat{u}(\cdot)$ is fitted using XGBoost on $|\hat{f}(X_i) - Y_i|$ in $\mathcal{D}_l$. The specific model hyperparameters used for each dataset are detailed in Section S4 of the Supplementary Material. The details of computational resources and efficiency are provided in Section S3 of the Supplementary Material for completeness. The cube method is implemented by the R package `balancesampling` via Python's `rpy2` interface [41]. Other experimental details are provided in Section S4 in Supplementary Material.

5.1 Performance across diverse budgets

In this subsection, we evaluate the proposed method on three representative datasets under different labeling budgets, covering both synthetic and real-world scenarios. Specifically, we consider (i) a synthetic regression dataset generated by the nonlinear model

$$y = 10\sin(\pi x_1 x_2) + 20(x_3 - 0.5)^2 + 10x_4 + 5x_5 + \varepsilon,$$

where the predictors $x_1, \ldots, x_{10}$ are uniformly distributed on $[0, 1]$ and ε follows a standard normal distribution; (ii) a real regression dataset (UCI Bike Sharing); and (iii) a real classification dataset (Credit Fraud Detection) with severe class imbalance. These datasets together provide a comprehensive test bed to evaluate the method's performance across different task types and data characteristics.

The results in Figure 1 reveal that the proposed cube-active method consistently outperforms alternatives across all labeling budgets. It achieves the lowest RMSE and the narrowest 90% confidence intervals, demonstrating superior predictive performance and sharper uncertainty quantification. Moreover, the empirical coverage rates closely match the nominal level in all cases, highlighting the validity and robustness of the inference procedure. Notably, even in the imbalanced classification setting, our method maintains stable coverage and tight intervals, highlighting its effectiveness across varying scenarios and labeling budgets.

5.2 Efficiency improvement

Table 1 presents the confidence interval widths (with empirical coverage rates in parentheses) for all methods under a labeling budget of 0.1. Across the majority of datasets, the empirical coverage rates remain closely aligned with the nominal confidence level of 0.9, confirming the validity of the asymptotic normal property and the statistical reliability of the proposed estimator.

Our proposed cube-active method achieves substantial and consistent efficiency gains over all baselines. Compared to traditional active inference based on independent sampling, cube-active reduces confidence interval widths by approximately 25%–85% across both synthetic and real-world datasets. When benchmarked against classical methods including uniform and simple random sampling, the improvement remains significant, with at least 30% narrower intervals on all tasks.

These gains can be attributed to the enhanced covariate balancing induced by the cube sampling design, which effectively controls estimator variance by aligning the labeled set with the underlying feature distribution. As a result, cube-active achieves sharper inference and more precise uncertainty quantification without sacrificing coverage validity. This improvement is particularly valuable in low-budget regimes, where efficient use of labeled data is crucial for reliable statistical inference.

5.3 Label budget saving

Label budget saving refers to the percentage reduction in the required sample size by our method compared to raditional-active inference under a given estimation accuracy. We establish a precision benchmark using the confidence interval width of traditional-active inference at label budget 0.2.

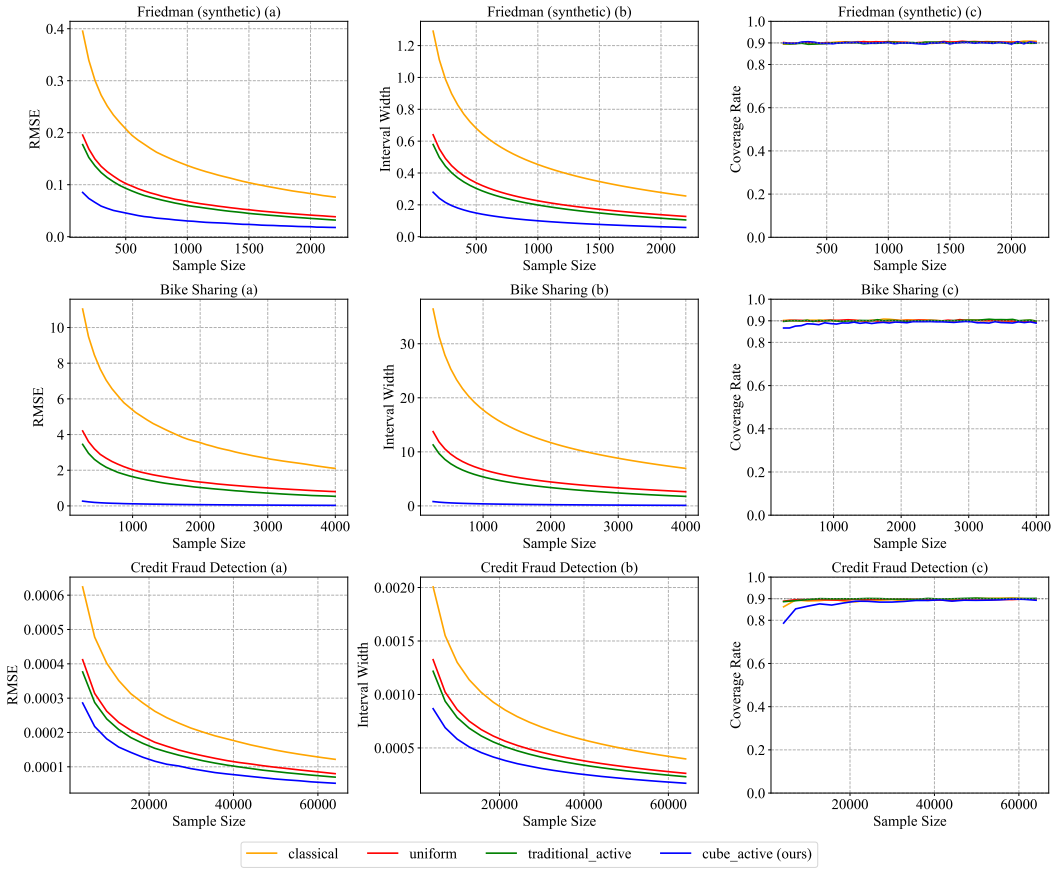

Figure 1: Performance comparison across three datasets with varying sample sizes. The top row (synthetic Friedman dataset), middle row (UCI Bike Sharing dataset), and bottom row (Credit Card Fraud Detection dataset) each display: (a) Root mean squared error (RMSE); (b) Average width of the 90% confidence intervals, reflecting inference precision; and (c) Empirical coverage rate, confirming interval estimation validity. The proposed cube-active sampling method consistently achieves superior performance across these datasets.

Table 1: Comparison of confidence interval width with 0.1 label budget across methods on 11 datasets. The bold values indicate the narrowest confidence intervals under valid coverage.

Dataset	**classical**	**uniform**	**traditional-active**	**cube-active**
Linear (synthetic)	0.3171 (0.8998)	0.1690 (0.9053)	0.1443 (0.8973)	**0.0722** (0.8998)
Nonlinear (synthetic)	0.7786 (0.8931)	0.3626 (0.8923)	0.2963 (0.8967)	**0.1370** (0.9063)
Friedman (synthetic)	0.6803 (0.8975)	0.3798 (0.9040)	0.3305 (0.8979)	**0.1052** (0.9056)
Bike	19.2033 (0.9028)	2.8213 (0.8987)	2.2052 (0.8977)	**0.3316** (0.8969)
Communities	0.0722 (0.8912)	0.0516 (0.8942)	0.0509 (0.8900)	**0.0466** (0.8885)
Concrete	7.1217 (0.8970)	3.9367 (0.9024)	3.7415 (0.8999)	**2.6406** (0.8943)
Credit-fraud-detection	0.0011 (0.8981)	0.0007 (0.8930)	0.0006 (0.8985)	**0.0005** (0.8700)
Energy	5.1520 (0.8938)	2.0364 (0.8954)	**1.8944** (0.9055)	0.4136 (0.8025)
Life	3.0828 (0.8967)	1.4509 (0.8963)	1.3050 (0.9036)	**0.7265** (0.8959)
Post-election	0.0571 (0.8949)	0.0426 (0.9019)	0.0403 (0.8998)	**0.0381** (0.8980)
Superconductor	3.2664 (0.9040)	1.6711 (0.9067)	1.5950 (0.9002)	**1.1680** (0.9003)

For each dataset, we determine the minimal label budget required by alternative methods to achieve the most similar confidence interval width of the benchmark. Given our experimental grid of label budgets (0.03˘0.45 with 0.01 increments), we employ linear interpolation between adjacent budget points for precision matching. When a method's precision at the minimal tested budget (0.03) exceeds the benchmark, its required budget is conservatively denoted as "$< 10\%$". Conversely, if precision remains below benchmark at the maximal budget (0.45), the required budget is cautiously reported as "$> 150\%$".

Table 2 quantifies the label budget efficiency required by different methods to match the precision benchmark of traditional-active at 0.2 label budget. Our method achieves substantial budget savings across all datasets, requiring less label budget of the benchmark to attain equivalent precision. In several cases, our method attains superior precision with less than 40% of the benchmark budget, demonstrating exceptional statistical efficiency. Compared to uniform sampling, our method only needs 5% or less labeled instances to achieve higher precision in real-world applications. These savings stem from our method's optimal balanced sampling that simultaneously maximizes information gain and minimizes distributional discrepancy. The method's adaptive balancing mechanism proves particularly effective in high-dimensional settings where conventional active learning methods exhibit diminishing returns due to covariate mismatch. This systematic budget reduction, coupled with maintained statistical validity, establishes cube-active as a practical solution for label-constrained inference scenarios.

Table 2: Comparison of budget saving with the precision benchmarks across methods on 11 datasets.

Dataset	**classical**	**uniform**	**traditional-active**	**cube-active**
Linear(synthetic)	>150%	127%	100%	37%
Nonlinear(synthetic)	>150%	117%	100%	70%
Friedman(synthetic)	>150%	123%	100%	33%
Bike	>150%	147%	100%	63%
Communities	150%	103%	100%	90%
Concrete	>150%	110%	100%	67%
Credit-fraud-detection	>150%	110%	100%	63%
Energy	>150%	113%	100%	<10%
Life	>150%	123%	100%	43%
Post-election	>150%	107%	100%	93%
Superconductor	>150%	110%	100%	63%

5.4 Performance of M-estimation problems

Figure 2 summarizes the statistical performance of our method in estimating linear regression coefficients for selected variables across the Linear and Bike Sharing datasets. Owing to the complexity of the underlying estimation procedures, deriving closed-form variance expressions for the estimators is intractable. Instead, we report the RMSE under different settings, which serves as a robust indicator of statistical efficiency and allows for clear performance comparisons across methods.

The results in Figure 2 consistently highlight the advantages of the proposed cube-active sampling strategy. For both the synthetic linear dataset and the real-world Bike Sharing dataset, cube-active sampling yields substantially lower RMSE values compared to all baseline methods, including uniform sampling, classical inference, and traditional active inference. These improvements illustrate the method's ability to provide more accurate coefficient estimation, thereby enhancing the overall statistical efficiency of the inference procedure.

6 Discussion

In this paper, we propose balanced active inference and demonstrate its superior statistical efficiency theoretically and numerically in various synthetic setups and real data analysis; see Section S8 and S9 of Supplementary Material for the limitations and societal impacts. A natural direction for future research is to explore the extension of this method into a sequential active inference framework.

Transitioning balanced active inference to a sequential setting introduces intriguing yet challenging methodological considerations. The principal challenge lies in the dynamic balancing requirement.

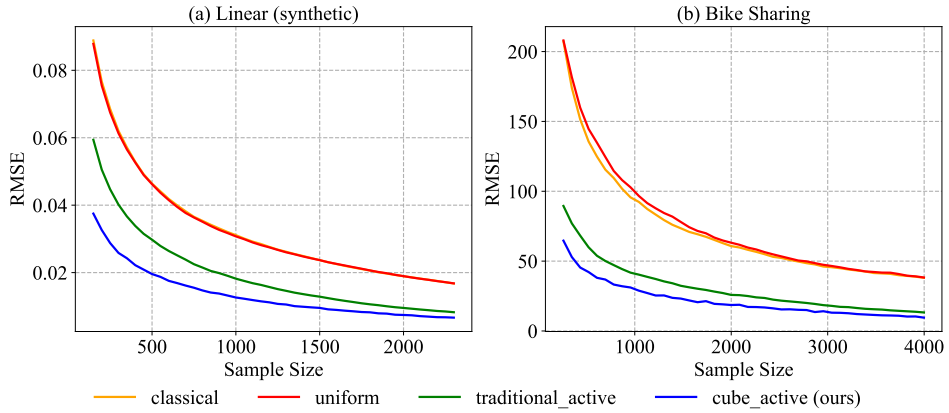

Figure 2: Root mean squared error (RMSE) of the least squares estimator in linear regression across two datasets with varying sample sizes. (a) the parameter of x_1 in synthetic Linear dataset ; (b) the parameter of *temperature* variable in Bike Sharing dataset. The proposed cube-active sampling method consistently achieves superior performance across these datasets.

In contrast to the batch scenario, where auxiliary variables are fixed, sequential balanced sampling involves continuously updating auxiliary variables when new instances become available. An adaptive implementation of the cube method, or a similar balanced sampling procedure, needs to integrate seamlessly with evolving model predictions and uncertainties in a computationally efficient manner. This dynamical adjustment could further reduce variance and improve representativeness, allowing for more precise allocation of labeling resources.

However, theoretical justification for such sequentially balanced sampling remains open. Extending existing results, such as martingale-based analyses that underpin sequential active inference, to balanced sampling contexts is not straightforward due to dependencies and complexity introduced by continually updated balancing constraints. Establishing rigorous statistical properties, such as unbiasedness, variance reduction, and asymptotic normality, within this sequential balanced framework will likely require novel analytical techniques or approximations.

Empirically, preliminary exploration through synthetic experiments and real data applications would be a valuable first step toward understanding sequential balanced active inference. Future studies should investigate various update strategies, examining how frequently and substantially the balancing conditions should be adjusted. Developing heuristics and computationally efficient algorithms capable of handling online updates of the balancing constraints would substantially advance the feasibility of sequential balanced active inference. Such advancements have great potential to further enhance labeling efficiency in practical applications, where data arrives continuously, and timely decision-making is critical.

Acknowledgments and Disclosure of Funding

The authors would like to thank anonymous reviewers for their valuable comments. Wang's research is supported in part by Humanities and Social Sciences Foundation of the Ministry of Education of China Grant (No. 23YJA910005), NSFC (No. 12571291, 72533007, 71988101, 72033002). Peng was supported by ARC (Grant No. LP240100101).

References

[1] Andre Esteva, Brett Kuprel, Roberto A Novoa, Justin Ko, Susan M Swetter, Helen M Blau, and Sebastian Thrun. Dermatologist-level classification of skin cancer with deep neural networks. *nature*, 542(7639):115–118, 2017.

[2] Wei Bao, Jun Yue, and Yulei Rao. A deep learning framework for financial time series using stacked autoencoders and long-short term memory. *PloS one*, 12(7):e0180944, 2017.

[3] 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.

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

[5] 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.

[6] Burr Settles and Mark Craven. An analysis of active learning strategies for sequence labeling tasks. In *proceedings of the 2008 conference on empirical methods in natural language processing*, pages 1070–1079, 2008.

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

[8] 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.

[9] 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.

[10] 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.

[11] Anastasios N. Angelopoulos, Stephen Bates, Clara Fannjiang, Michael I. Jordan, and Tijana Zrnic. Prediction-powered inference. *Science*, 382(6671):669–674, 2023.

[12] Jerzy Neyman. *On the Two Different Aspects of the Representative Method: the Method of Stratified Sampling and the Method of Purposive Selection*, pages 123–150. Springer, 1992.

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

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

[15] Jean-Claude Deville and Yves Tillé. Efficient balanced sampling: The cube method. *Biometrika*, 91(4):893–912, 2004.

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

[17] Donald B Rubin. Inference and missing data. *Biometrika*, 63(3):581–592, 1976.

[18] Donald B Rubin. Multiple imputation after 18+ years. *Journal of the American Statistical Association*, 91(434):473–489, 1996.

[19] Anru Zhang, Lawrence D Brown, and T Tony Cai. Semi-supervised inference: General theory and estimation of means. *Annals of Statistics*, 47(5):2538–2566, 2019.

[20] David Azriel, Lawrence D Brown, Michael Sklar, Richard Berk, Andreas Buja, and Linda Zhao. Semi-supervised linear regression. *Journal of the American Statistical Association*, 117(540):2238–2251, 2022.

[21] Greg Schohn and David Cohn. Less is more: Active learning with support vector machines. In *Proceedings of the Seventeenth International Conference on Machine Learning*, volume 2, page 839–846, 2000.

[22] Burr Settles. From theories to queries: Active learning in practice. In *Active Learning and Experimental Design Workshop in Conjunction with AISTATS 2010*, pages 1–18. JMLR Workshop and Conference Proceedings, 2011.

[23] 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.

[24] Al. A. Tchouproff. On the mathematical expectation of the moments of frequency distributions. *Biometrika*, 12(1/2):140–169, 1918.

[25] Morris H Hansen and William N Hurwitz. On the theory of sampling from finite populations. *Annals of Mathematical Statistics*, 14(4):333–362, 1943.

[26] William G Madow. On the theory of systematic sampling, II. *Annals of Mathematical Statistics*, 20(3):333–354, 1949.

[27] Frank Yates. A review of recent statistical developments in sampling and sampling surveys. *Journal of the Royal Statistical Society*, 109(1):12–43, 1946.

[28] Frank Yates. *Sampling methods for censuses and surveys*. Charles Griffin & Co., Ltd., 1949.

[29] Paul R Rosenbaum and Donald B Rubin. The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1):41–55, 1983.

[30] Laurent Davezies, Guillaume Hollard, and Pedro Vergara Merino. Revisiting randomization with the cube method. *arXiv preprint arXiv:2407.13613*, 2024.

[31] Jerome H Friedman. Multivariate adaptive regression splines. *Annals of Statistics*, 19(1):1–67, 1991.

[32] Hadi Fanaee-T and Joao Gama. Event labeling combining ensemble detectors and background knowledge. *Progress in Artificial Intelligence*, 2:113–127, 2014.

[33] Michael Redmond and Alok Baveja. A data-driven software tool for enabling cooperative information sharing among police departments. *European Journal of Operational Research*, 141(3):660–678, 2002.

[34] I-C Yeh. Modeling of strength of high-performance concrete using artificial neural networks. *Cement and Concrete Research*, 28(12):1797–1808, 1998.

[35] Athanasios Tsanas and Angeliki Xifara. Accurate quantitative estimation of energy performance of residential buildings using statistical machine learning tools. *Energy and Buildings*, 49:560–567, 2012.

[36] Abhinav Kumar. Life expectancy (WHO). Kaggle, 2020. https://www.kaggle.com/datasets/kumarajarshi/life-expectancy-who.

[37] Kam Hamidieh. A data-driven statistical model for predicting the critical temperature of a superconductor. *Computational Materials Science*, 154:346–354, 2018.

[38] Emmanuel Lopez. Credit card fraud detection. Kaggle, 2019. https://www.kaggle.com/datasets/mlg-ulb/creditcardfraud.

[39] Pew. American trends panel (atp) wave 79, 2020. https://www.pewresearch.org/science/dataset/american-trends-panel-wave-79/.

[40] Tianqi Chen and Carlos Guestrin. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, page 785–794, New York, NY, USA, 2016. Association for Computing Machinery.

[41] Don L Stevens Jr and Anthony R Olsen. Spatially balanced sampling of natural resources. *Journal of the American Statistical Association*, 99(465):262–278, 2004.

[42] Jiahua Chen and J. N. K. Rao. Asymptotic normality under two-phase sampling designs. *Statistica Sinica*, 17(3):1047–1064, 2007.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The formulation of the proposed balanced active inference can be found in Section 3. Theoretical properties can be found in Section 4, and numerical results can be found in Section 5.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: The limitations of our proposed method are discussed in Section S8 of the Supplementary material.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [Yes]

Justification: All the assumptions can be found in Section 4, and all the detailed proofs can be found in Appendix A.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: We provide detailed descriptions and formulations required to reproduce the main experimental results, including information on the model, datasets, baselines, and evaluation metrics. Additionally, we present the implementation details in Section S4 of Supplementary Material to offer clearer implementation guidance.

Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: The datasets utilized in this study are publicly accessible via the cited references, and the implementation code is available upon request.

   Guidelines:

   - The answer NA means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: The main settings can be found in Section 5, and detailed settings and additional experiments are presented in Section S4 and S3 of Supplementary Material to provide further insights into the results.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: The coverage rates of the proposed estimator and its alternatives are reported in Section 5.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.

- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: Compute resources are clearly stated in Section 5.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: We have carefully reviewed the NeurIPS Code of Ethics and have ensured that all aspects of our research fully comply with its guidelines.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: The societal impacts of our proposed method are discussed in Section S9 of the supplementary material.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.

- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: We believe that our work poses no foreseeable risk of misuse.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: All datasets used in our experiments are properly cited, and source links are provided for reference.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.

- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [NA]

    Justification: Our work does not release new assets.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: Our work does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: Our work does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.

- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

    Answer:[NA]

    Justification: The core method development in this work does not involve LLMs as any important, original, or non-standard components.

    Guidelines:

    - The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy (`https://neurips.cc/Conferences/2025/LLM`) for what should or should not be described.

Appendix

A Proof

In this section, we provide the detailed proof of Theorem 1. Before the formal proof, some technical lemmas are presented.

A.1 Technical Lemmas

Lemma 1 (Proposition 1 in [30]). *Suppose Assumptions 2 and 3 hold, and a cube method is conducted to balance auxiliary information* $\{X_i : i = 1, \ldots, n\}$ *under a sampling scheme* $\{\pi_i : 1 = 1, \ldots, n\}$*, then*

$$\frac{1}{n}\sum_{i=1}^{n}\frac{X_i\xi_i}{\pi_i} - \frac{1}{n}\sum_{i=1}^{n}X_i = o_p\left(\frac{q}{\sqrt{n}}\right),$$

where $\xi_i \in \{0, 1\}$ *indicates if the* i*-th instance is selected.*

Lemma 2 (Lemma C.2 in [30]). *Let* f *and* g *be two functions such that for* $f_i = f\left(\delta_i(1), \delta_i(0), X_i\right)$ *and* $g_i = g\left(\delta_i(1), \delta_i(0), X_i\right)$ *we have* $\mathbb{E}\left(f_i^2 + g_i^2\right) < \infty$ *and* $\mathbb{E}\left[f_i \mid X_i\right] = \mathbb{E}\left[g_i \mid X_i\right] = 0$*.*

If Assumptions 2, 3 and 4 hold. Then, conditional on $(X_i)_{i\geq 1}$*,*

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n} f_i + g_i D_i \xrightarrow{d} \mathcal{N}\left(0, V_0\right)$$

with $V_0 = \mathbb{E}\left[f_1^2 + \left(2g_1f_1 + g_1^2\right)\pi_1\right]$*.*

Lemma 3 (Theorem 2 in [42]). *Let* U_n, V_n *be two sequences of random variables and* $\mathcal{B}_n$ *be a* σ*-algebra. Assume that*
(1) there exists $\sigma_{1n} > 0$ *such that* $\sigma_{1n}^{-1}V_n \to N(0, 1)$ *in distribution as* $n \to \infty$*, and* V_n *is* $\mathcal{B}_n$ *measurable;*
(2) $E\left\{U_n \mid \mathcal{B}_n\right\} = 0$ *and* $\mathrm{Var}\left(U_n \mid \mathcal{B}_n\right) = \sigma_{2n}^2$ *such that*

$$\sup_t \left|P\left(\sigma_{2n}^{-1}U_n \leq t \mid \mathcal{B}_n\right) - \Phi(t)\right| = o_p(1),$$

where $\Phi(t)$ *is the cumulative distribution function of the standard normal random variable;*
(3) $\gamma_n^2 = \sigma_{1n}^2/\sigma_{2n}^2 \to \gamma^2$ *in probability as* $n \to \infty$*. Then*

$$\frac{U_n + V_n}{\sqrt{\sigma_{1n}^2 + \sigma_{2n}^2}} \to \mathcal{N}(0, 1)$$

in distribution as $n \to \infty$*.*

A.2 Proof of Theorem 1

By definition, we have

$$\begin{aligned}\widetilde{\theta} - \theta^* &= \frac{1}{n}\sum_{i=1}^{n}\left[\hat{f}(X_i) + \left(Y_i - \hat{f}(X_i)\right)\frac{\xi_i}{\pi(X_i)}\right] - \mathbb{E}Y_1 \\ &= \frac{1}{n}\sum_{i=1}^{n}\left[\hat{f}(X_i) - f(X_i) + \left(Y_i - \hat{f}(X_i)\right)\frac{\xi_i}{\pi(X_i)}\right] + \frac{1}{n}\sum_{i=1}^{n}\left(f(X_i) - \mathbb{E}Y\right). \qquad (8)\end{aligned}$$

For the first term in (8), we have

$$
\begin{aligned}
&\frac{1}{n}\sum_{i=1}^{n}\left[\hat{f}(X_i)-f(X_i)+\left(Y_i-\hat{f}(X_i)\right)\frac{\xi_i}{\pi(X_i)}\right]\\
=&\frac{1}{n}\sum_{i=1}^{n}\left[\hat{f}(X_i)-f(X_i)+\operatorname{sgn}\left(\hat{f}(X_i)-f(X_i)\right)\hat{u}(X_i)\right.\\
&\left.+\left(f_i-\hat{f}(X_i)+\varepsilon_i-\operatorname{sgn}\left(\hat{f}(X_i)-f(X_i)\right)\hat{u}(X_i)\right)\frac{\xi_i}{\pi(X_i)}\right]\\
&-\frac{1}{n}\sum_{i=1}^{n}\hat{u}(X_i)+\frac{1}{n}\sum_{i=1}^{n}\hat{u}(X_i)\frac{\xi_i}{\pi(X_i)}\\
=&\frac{1}{n}\sum_{i=1}^{n}\left\{\left[\hat{f}(X_i)-f(X_i)+\operatorname{sgn}\left(\hat{f}(X_i)-f(X_i)\right)\hat{u}(X_i)\right]\left(1-\frac{\xi_i}{\pi(X_i)}\right)\right\} \qquad (9)\\
&+\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i\frac{\xi_i}{\pi(X_i)}+o_p\left(\frac{1}{\sqrt{n}}\right).
\end{aligned}
$$

Denoting $\left[\hat{f}(X_i)-f(X_i)+\operatorname{sgn}\left(\hat{f}(X_i)-f(X_i)\right)\hat{u}(X_i)\right]\left(1-\frac{\xi_i}{\pi(X_i)}\right)$ by A_i and applying Chebyshev's inequality on (9), we have

$$
\mathbb{P}\left(\left|\frac{1}{n}\sum_{i=1}^{n}A_i-\mathbb{E}A_i\right|\geq\frac{1}{\sqrt{n}}\right)\leq\frac{1}{n}\operatorname{Var}\left(\sum_{i=1}^{n}A_i\right). \qquad (10)
$$

Further, we have $\operatorname{Var}\left(\sum_{i=1}^{n}A_i\right)=\sum_{i=1}^{n}\operatorname{Var}(A_i)+\sum_{i\neq j}\operatorname{Cov}(A_i,A_j)$. For $\operatorname{Var}(A_i)$, we have

$$
\begin{aligned}
\operatorname{Var}(A_i)=&\mathbb{E}\left[\left(\hat{f}(X_i)-f(X_i)+\operatorname{sgn}\left(\hat{f}(X_i)-f(X_i)\right)\hat{u}(X_i)\right)^2\left(1-\frac{\xi_i}{\pi(X_i)}\right)^2\right]\\
&-\mathbb{E}\left\{\left[\hat{f}(X_i)-f(X_i)+\operatorname{sgn}\left(\hat{f}(X_i)-f(X_i)\right)\hat{u}(X_i)\right]\left(1-\frac{\xi_i}{\pi(X_i)}\right)\right\}^2.
\end{aligned}
$$

Since $\pi(X_i)$ is bounded from c to $1-c$, there exists a positive constant C such that

$$
\begin{aligned}
&\mathbb{E}\left[\left(\hat{f}(X_i)-f(X_i)+\operatorname{sgn}\left(\hat{f}(X_i)-f(X_i)\right)\hat{u}(X_i)\right)^2\left(1-\frac{\xi_i}{\pi(X_i)}\right)^2\right]\\
\leq&C\cdot\mathbb{E}\left[\left(\hat{f}(X_i)-f(X_i)+\operatorname{sgn}\left(\hat{f}(X_i)-f(X_i)\right)\hat{u}(X_i)\right)^2\right].
\end{aligned}
$$

Further, by the assumption that $\mathbb{E}\left[f(X_i)-\hat{f}(X_i)-\operatorname{sgn}\left(\hat{f}(X_i)-f(X_i)\right)\hat{u}(X_i)\right]^2=o_p(1)$, we have $\operatorname{Var}(A_i)=o_p(1)$. For $\operatorname{Cov}(A_i,A_j)=\mathbb{E}\left[A_iA_j\right]-\mathbb{E}\left[A_i\right]\mathbb{E}\left[A_j\right]$, first,

$$
\begin{aligned}
&\mathbb{E}\left[A_iA_j\right]\\
=&\mathbb{E}\Bigg[\left(\hat{f}(X_i)-f(X_i)+\operatorname{sgn}\left(\hat{f}(X_i)-f(X_i)\right)\hat{u}(X_i)\right)\frac{\xi_i-\pi(X_i)}{\pi(X_i)}\\
&\qquad\left(\hat{f}(X_j)-f(X_j)+\operatorname{sgn}\left(\hat{f}(X_i)-f(X_i)\right)\hat{u}(X_j)\right)\frac{\xi_j-\pi(X_j)}{\pi(X_j)}\Bigg]\\
=&\mathbb{E}\Bigg\{\frac{\hat{f}(X_i)-f(X_i)+\operatorname{sgn}\left(\hat{f}(X_i)-f(X_i)\right)\hat{u}(X_i)}{\pi(X_i)}\\
&\qquad\cdot\frac{\hat{f}(X_j)-f(X_j)+\operatorname{sgn}\left(\hat{f}(X_i)-f(X_i)\right)\hat{u}(X_j)}{\pi(X_j)}\\
&\qquad\cdot\mathbb{E}\left[\left(\xi_i-\pi(X_i)\right)\left(\xi_j-\pi(X_j)\right)\Big|X_i,X_j\right]\Bigg\}.
\end{aligned}
$$

By Assumption 4, we have with probability 1, $\lim_{n\to\infty}\mathbb{E}\left[\left(\xi_i-\pi(X_i)\right)\left(\xi_j-\pi(X_j)\right)\Big|X_i,X_j\right]=0$. Thus, with probability 1,$\lim_{n\to\infty}\mathbb{E}\left[A_iA_j\right]=0$. And by a similar argument, we have $\mathbb{E}\left[A_i\right]=\mathbb{E}\left[\left(\hat{f}(X_i)-f(X_i)+\operatorname{sgn}\left(\hat{f}(X_i)-f(X_i)\right)\hat{u}(X_i)\right)\frac{\xi_i-\pi(X_i)}{\pi(X_i)}\right]=0$ as $n\to\infty$. Thus, it can be concluded that, with probability 1,$\lim_{n\to\infty}\sum_{i\neq j}\operatorname{Cov}(A_i,A_j)=0$. Therefore, we have $\operatorname{Var}\left(\sum_{i=1}^n A_i\right)=o_p(1)$, then (10) implies that

$$\mathbb{P}\left(\left|\frac{1}{n}\sum_{i=1}^n A_i-\mathbb{E}A_i\right|\geq\frac{1}{\sqrt{n}}\right)=o_p(1).$$

Thus we have

$$\frac{1}{n}\sum_{i=1}^n\left\{\left[\hat{f}(X_i)-f(X_i)+\operatorname{sgn}\left(\hat{f}(X_i)-f(X_i)\right)\hat{u}(X_i)\right]\left(1-\frac{\xi_i}{\pi(X_i)}\right)\right\}=o_p(\frac{1}{\sqrt{n}}).$$

By Lemma 2, we have conditional on $X_1,X_2,\ldots,X_n$

$$\frac{1}{\sqrt{n}}\sum_{i=1}^n\varepsilon_i\frac{\xi_i}{\pi(X_i)}\xrightarrow{d}\mathcal{N}\left(0,\mathbb{E}\left(\frac{\varepsilon_i^2}{\pi_i}\right)\right).$$

Thus, applying Lemma 3, we have

$$\sqrt{n}\left(\widetilde{Y}-\mathbb{E}Y\right)\xrightarrow{d}\mathcal{N}\left(0,\mathbb{E}\left(\frac{\varepsilon_1^2}{\pi_1}\right)+\operatorname{Var}\left[f(X_1)\right]\right).$$

If a Poisson sampling is used, by a standard CLT, we have

$$\sqrt{n}(\hat{\theta}-\theta^*)\xrightarrow{d}\mathcal{N}(0,V_1),$$

where

$$\begin{aligned}
V_1&=\operatorname{Var}\left(\hat{f}\left(X_1\right)+\left(Y_1-\hat{f}\left(X_1\right)\right)\frac{\xi_1}{\pi\left(X_1\right)}\right)\\
&=\operatorname{Var}\left[\mathbb{E}\left(\hat{f}\left(X_1\right)+\left(Y_1-\hat{f}\left(X_1\right)\right)\frac{\xi_1}{\pi\left(X_1\right)}|X_1\right)\right]\\
&\quad+\mathbb{E}\left[\operatorname{Var}\left(\hat{f}\left(X_1\right)+\left(Y_1-\hat{f}\left(X_1\right)\right)\frac{\xi_1}{\pi\left(X_1\right)}|X_1\right)\right]\\
&=\operatorname{Var}(Y_1)+\mathbb{E}\left[\left(Y_1-\hat{f}\left(X_1\right)\right)^2\frac{1-\pi(X_1)}{\pi\left(X_1\right)}\right]\\
&=\operatorname{Var}(f_1)+\mathbb{E}\left(\frac{\varepsilon_1^2}{\pi_1}\right)+\mathbb{E}\left[\left(f(x)-\hat{f}(x)\right)^2\left(\frac{1}{\pi(x)}-1\right)\right].
\end{aligned}$$

Then, the proof is completed.