Active Statistical Inference

Tijana Zrnic [1] **Emmanuel J. Candès** [2]

Abstract

Inspired by the concept of active learning, we propose *active inference*—a methodology for statistical inference with machine-learning-assisted data collection. Assuming a budget on the number of labels that can be collected, the methodology uses a machine learning model to identify which data points would be most beneficial to label, thus effectively utilizing the budget. It operates on a simple yet powerful intuition: prioritize the collection of labels for data points where the model exhibits uncertainty, and rely on the model's predictions where it is confident. Active inference constructs valid confidence intervals and hypothesis tests while leveraging any black-box machine learning model and handling any data distribution. The key point is that it achieves the same level of accuracy with far fewer samples than existing baselines relying on non-adaptively-collected data. This means that for the same number of collected samples, active inference enables smaller confidence intervals and more powerful tests. We evaluate active inference on datasets from public opinion research, census analysis, and proteomics.

1. Introduction

In the realm of data-driven research, collecting high-quality labeled data is a continuing impediment. The impediment is particularly acute when operating under stringent labeling budgets, where the cost and effort of obtaining each label can be substantial. Recognizing these limitations, many have turned to machine learning as a pragmatic solution, leveraging it to predict unobserved labels across various fields. In remote sensing, machine learning assists in annotating and interpreting satellite imagery (Jean et al., 2016; Xie et al., 2016; Rolf et al., 2021); in proteomics, tools like AlphaFold (Jumper et al., 2021) are revolutionizing our understanding of protein structures; even in the realm of elections—including most major US elections—technologies combining scanners and predictive models are used as efficient tools for vote counting (Zdun, 2022). These applications reflect a growing reliance on machine learning for extracting knowledge from unlabeled datasets.

However, this reliance on machine learning is not without its pitfalls. The core issue lies in the inherent biases of these models. No matter how sophisticated, predictions lead to dubious conclusions; as such, predictions cannot fully substitute for traditional data sources such as gold-standard experimental measurements, high-quality surveys, and expert annotations. This begs the question: is there a way to effectively leverage the predictive power of machine learning while still ensuring the integrity of our inferences?

Drawing inspiration from the concept of active learning, we propose *active inference*—a novel methodology for statistical inference that harnesses machine learning not as a replacement for data collection but as a strategic guide to it. The methodology uses a machine learning model to identify which data points would be most beneficial to label, thus effectively utilizing the labeling budget. It operates on a simple yet powerful intuition: prioritize the collection of labels for data points where the model exhibits uncertainty, and rely on the model's predictions where it is confident. Active inference constructs provably valid confidence intervals and hypothesis tests for any black-box machine learning model and any data distribution. The key takeaway is that it achieves the same level of accuracy with far fewer samples than existing baselines relying on non-adaptively-collected data. Put differently, this means that for the same number of collected samples, active inference enables smaller confidence intervals and more powerful p-values. We will show in our experiments that active inference can save over 80% of the sample budget required by classical methods.

Although quite different in scope, our work is inspired by the recent framework of *prediction-powered inference* (PPI) (Angelopoulos et al., 2023a). PPI assumes access to a small labeled dataset and a large unlabeled dataset, drawn i.i.d. from the population of interest. It then asks how one can use machine learning and the unlabeled dataset to sharpen inference about population parameters depending

[1]Department of Statistics and Stanford Data Science, Stanford University, USA [2]Department of Statistics and Department of Mathematics, Stanford University, USA. Correspondence to: Tijana Zrnic <tijana.zrnic@stanford.edu>, Emmanuel J. Candès <candes@stanford.edu>.

Proceedings of the 41^{st} International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024.

on the distribution of labels. Our objective in this paper is different since the core of our contribution is (1) designing *strategic* data collection approaches that enable more powerful inferences than collecting labels in an i.i.d. manner, and (2) showing how to perform inference with such strategically collected data. We will see that PPI can be seen as a special case of our methodology: while PPI ignores the issue of strategic data collection and instead uses a trivial, uniform data collection strategy, it leverages machine learning to enhance inference in a similar way to our method. We provide a further discussion of prior work in Section 3.

2. Problem Description

We introduce the formal problem setting. We observe unlabeled instances $X_1, \dots, X_n$, drawn i.i.d. from a distribution $\mathbb{P}_X$. The labels Y_i are unobserved, and we shall use $(X, Y) \sim \mathbb{P} = \mathbb{P}_X \times \mathbb{P}_{Y|X}$ to denote a generic feature–label pair drawn from the underlying data distribution. We are interested in performing inference—conducting a hypothesis test or forming a confidence interval—for a parameter θ^* that depends on the distribution of the unobserved labels; that is, the parameter is a functional of $\mathbb{P}_X \times \mathbb{P}_{Y|X}$. For example, we might be interested in forming a confidence interval for the mean label, $\theta^* = \mathbb{E}[Y_i]$, where Y_i is the label corresponding to X_i. Although we will primarily focus on forming confidence intervals, the standard duality between confidence intervals and hypothesis tests makes our results directly applicable to testing as well.

We have no collected labels a priori. Rather, the goal is to efficiently and strategically acquire labels for certain points X_i, so that inference is as powerful as possible for a given collection budget—more so than if labels were collected uniformly at random—while also remaining valid. We denote by n_{lab} the number of collected labels. We assume that we are constrained to collect, on average, $\mathbb{E}[n_{\mathrm{lab}}] \le n_b$ labels, for some budget n_b.[1] Typically, $n_b \ll n$.

To guide the choice of which instances to label, we will make use of a predictive model f. Typically this will be a black-box machine learning model, but it could also be a hand-designed decision rule based on expert knowledge. This is the key component that will enable us to get a significant boost in power. We do not assume any knowledge of the predictive performance of f, or any parametric form for it. Our key takeaway is that, if we have a reasonably good model for predicting the labels Y_i based on X_i, then we can achieve a significant boost in power compared to labeling a uniformly-at-random chosen set of instances.

We will consider two settings, depending on whether or not we update the predictive model f as we gather more labels.

- The first is a *batch* setting, where we simultaneously make decisions of whether or not to collect the corresponding label for all unlabeled points at once. In this setting, the model f is pre-trained and remains fixed during the label collection. The batch setting is simpler and arguably more practical if we already have a good off-the-shelf predictor.
- The second setting is *sequential*: we go through the unlabeled points one by one and update the predictive model as we collect more data. The benefit of the second approach is that it is applicable even when we do not have access to a pre-trained model, but we have to train a model from scratch.

Our proposed active inference strategy will be applicable to all convex *M-estimation* problems. This means that it handles all targets of inference θ^* that can be written as:

$$\theta^* = \arg\min_{\theta} \mathbb{E}[\ell_\theta(X, Y)], \text{ where } (X, Y) \sim \mathbb{P},$$

for a loss function ℓ_θ that is convex in θ. We denote $L(\theta) = \mathbb{E}[\ell_\theta(X, Y)]$ for brevity. M-estimation captures many relevant targets, such as the following.

Example 2.1 (Mean label)**.** *If $\ell_\theta(x, y) = \frac{1}{2}(y - \theta)^2$, then the target is the mean label, $\theta^* = \mathbb{E}[Y]$. Note that this loss has no dependence on the features.*

Example 2.2 (Linear regression)**.** *If $\ell_\theta(x, y) = \frac{1}{2}(y - x^\top \theta)^2$, then θ^* is the vector of linear regression coefficients obtained by regressing y on x, that is, the "effect" of x on y.*

Example 2.3 (Label quantile)**.** *For a given $q \in (0, 1)$, let $\ell_\theta(x, y) = q(y - \theta)\mathbf{1}\{y > \theta\} + (1 - q)(\theta - y)\mathbf{1}\{y \le \theta\}$ be the "pinball" loss. Then, θ^* is equal to the q-quantile of the label distribution: $\theta^* = \inf\{\theta : \mathbb{P}(Y \le \theta) \ge q\}$.*

3. Related Work

Our work is most closely related to prediction-powered inference (PPI) and other recent works on inference with machine learning predictions (Angelopoulos et al., 2023a;c; Zrnic & Candès, 2024; Motwani & Witten, 2023; Gan & Liang, 2023; Miao et al., 2023). This recent literature in turn relates to classical work on inference with missing data and semiparametric statistics (Rubin, 1976; 1987; 1996; Robins et al., 1994; Robins & Rotnitzky, 1995; Chernozhukov et al., 2018), as well as semi-supervised inference (Zhang et al., 2019; Azriel et al., 2022; Zhang & Bradic, 2022). We consider the same set of inferential targets as in (Angelopoulos et al., 2023a;c; Zrnic & Candès, 2024), building on classical M-estimation theory (Van der Vaart, 2000) to enable inference. While PPI assumes access to a small labeled dataset and a large unlabeled dataset, which are drawn i.i.d., our work is different in that it leverages machine learning in order to design *adaptive* label collection strategies, which breaks the i.i.d. structure between the labeled and the unlabeled data. We will see that our active inference estimator

[1] For simplicity we bound $\mathbb{E}[n_{\mathrm{lab}}]$, however the budget can be met with high probability, as n_{lab} has fast concentration around n_b.

reduces to the prediction-powered estimator when we apply a trivial, uniform label collection strategy. We will demonstrate empirically that the adaptivity in label collection enables significant improvements in statistical power.

There is a growing literature on inference from adaptively collected data (Kato et al., 2020; Zhang et al., 2021; Cook et al., 2023), often focusing on data collected via a bandit algorithm. These papers typically focus on average treatment effect estimation. In contrast to our work, these works generally do not focus on how to set the data-collection policy as to achieve good statistical power, but their main focus is on providing valid inferences given a fixed data-collection policy. Notably, Zhang et al. (2021) study inference for M-estimators from bandit data. However, their estimators do not leverage machine learning, which is central to our work.

A substantial line of work studies adaptive experiment design (Robbins, 1952; Lai & Robbins, 1985; Hu & Rosenberger, 2006; List et al., 2011; Hahn et al., 2011; Bhattacharya & Dupas, 2012; Kasy & Sautmann, 2021; Hadad et al., 2021; Chandak et al., 2023), often with the goal of maximizing welfare during the experiment or identifying the best treatment. Most related to our motivation, a subset of these works (List et al., 2011; Hahn et al., 2011; Chandak et al., 2023) study adaptive design with the goal of efficiently estimating average treatment effects. While our motivation is not necessarily treatment effect estimation, we continue in a similar vein—collecting data adaptively with the goal of improved efficiency—with a focus on using modern, black-box machine learning to produce uncertainty estimates that can be turned into efficient label collection methods. Related variance-reduction ideas appear in stratified survey sampling (Nassiuma, 2001; Särndal et al., 2003). Our proposal can be seen as stratifying the population of interest based on the certainty of a machine learning model.

Finally, our work draws inspiration from active learning, a subarea of machine learning centered around the observation that a predictive model can enhance its predictive capabilities if it is allowed to choose the data from which it learns. Our setup is analogous to pool-based active learning (Settles, 2009). Sampling according to a measure of predictive uncertainty is a central idea in active learning (Schohn & Cohn, 2000; Tong & Koller, 2001; Balcan et al., 2006; Joshi et al., 2009; Hanneke et al., 2014; Gal et al., 2017; Ash et al., 2019; Ren et al., 2021). Since our goal is statistical inference, rather than training a good predictor, our sampling rules are different and adapt to the inferential question.

4. Warm-up: Active Inference for the Mean

We first focus on the special case of estimating the mean label, $\theta^* = \mathbb{E}[Y]$, in the batch setting. The intuition derived from this example carries over to all other problems. Recall the setup: we observe n i.i.d. unlabeled instances $X_1, \dots, X_n$, and we can collect labels for at most n_b of them (on average). Consider first a "classical" solution, which does not leverage machine learning. Given a budget n_b, we can simply label any arbitrarily chosen n_b points. Since the instances are i.i.d., without loss of generality we can choose to label instances $\{1, \dots, n_b\}$ and compute $\hat{\theta}^{\texttt{noML}} = \frac{1}{n_b}\sum_{i=1}^{n_b} Y_i$. The estimator $\hat{\theta}^{\texttt{noML}}$ is clearly unbiased, and its variance is $\mathrm{Var}(\hat{\theta}^{\texttt{noML}}) = \frac{1}{n_b}\mathrm{Var}(Y)$.

Now, suppose that we are given a machine learning model $f(X)$, which predicts the label $Y \in \mathbb{R}$ from observed covariates $X \in \mathcal{X}$.[2] The idea behind our active inference strategy is to increase the effective sample size by using the model's predictions on points X where the model is confident and focusing the labeling budget on the points X where the model is uncertain. To implement this idea, we design a *sampling rule* $\pi : \mathcal{X} \to [0,1]$ and collect label Y_i with probability $\pi(X_i)$. The sampling rule is derived from f, by appropriately measuring its uncertainty. The hope is that $\pi(x) \approx 1$ signals that the model f is very uncertain about instance x, whereas $\pi(x) \approx 0$ indicates that the model f should be very certain about instance x. Let $\xi_i \sim \mathrm{Bern}(\pi(X_i))$ denote the indicator of whether we collect the label for point i. By definition, $n_{\mathrm{lab}} = \sum_{i=1}^n \xi_i$. The rule π will be carefully rescaled to meet the budget constraint: $\mathbb{E}[n_{\mathrm{lab}}] = \mathbb{E}[\pi(X)] \cdot n \le n_b$.

Our *active estimator* of the mean θ^* is given by:

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

This is essentially the augmented inverse propensity weighting (AIPW) estimator (Robins et al., 1994), with a particular choice of propensities $\pi(X_i)$ based on the certainty of the machine learning model that predicts the missing labels. When the sampling rule is uniform, i.e. $\pi(x) = n_b/n$ for all x, $\hat{\theta}^{\pi}$ is equal to the prediction-powered mean estimator (Angelopoulos et al., 2023a).

It is not hard to see that $\hat{\theta}^{\pi}$ is unbiased: $\mathbb{E}[\hat{\theta}^{\pi}] = \theta^*$. A short calculation shows that its variance equals

$$\mathrm{Var}(\hat{\theta}^{\pi}) = \frac{1}{n}\left(\mathrm{Var}(Y) + \mathbb{E}\left[(Y - f(X))^2\left(\pi(X)^{-1} - 1\right)\right]\right). \quad (2)$$

If the model is accurate for all x, i.e. $f(X) \approx Y$, then $\mathrm{Var}(\hat{\theta}^{\pi}) \approx \frac{1}{n}\mathrm{Var}(Y)$, which is far smaller than $\mathrm{Var}(\hat{\theta}^{\texttt{noML}})$ since $n_b \ll n$. Of course, f will never be accurate for all instances x. For this reason, we will aim to choose π such that π is small when $f(X) \approx Y$ and large otherwise, so that the relevant term $(Y - f(X))^2\left(\pi^{-1}(X) - 1\right)$ is always

[2] $\mathcal{X}$ is the set of values the covariates can take on, e.g. $\mathbb{R}^d$.

small (of course, subject to the sampling budget constraint). For example, for instances for which the predictor is correct, i.e. $f(X) = Y$, we would ideally like to set $\pi(X) = 0$ as this incurs no additional variance. We note that the variance reduction of active inference compared to the "classical" solution also implies that the resulting confidence intervals get smaller. This follows because interval width scales with the standard deviation for most standard intervals (e.g., those derived from the central limit theorem).

Finally, we explain how to set the sampling rule π. The rule will be derived from a measure of *model uncertainty* $u(x)$ and we shall provide different choices of $u(x)$ in the following paragraphs. At a high level, one can think of $u(X_i)$ as the model's best guess of $|Y_i - f(X_i)|$. We will choose $\pi(x)$ proportional to $u(x)$, that is, $\pi(x) \propto u(x)$, normalized to meet the budget constraint. Intuitively, this means that we want to focus our data collection budget on parts of the covariate space where the model is expected to make the largest errors. Roughly speaking, we will set $\pi(x) = \frac{u(x)}{\mathbb{E}[u(X)]} \cdot \frac{n_b}{n}$; this implies $\mathbb{E}[n_{\mathrm{lab}}] = \mathbb{E}[\pi(X)] \cdot n \leq n_b$. (This is an idealized form of $\pi(x)$ because $\mathbb{E}[u(X)]$ cannot be known exactly, though it can be estimated accurately from the unlabeled data; we will formalize this later on.)

We will take two different approaches for choosing the uncertainty $u(x)$, depending on whether we are in a regression or a classification setting.

Regression uncertainty In regression, we explicitly train a model $u(x)$ to predict $|f(X_i) - Y_i|$ from X_i. We note that we aim to predict only the magnitude of the error and not the directionality. In the batch setting, we typically have historical data of (X, Y) pairs that are used to train the model f. We thus train $u(x)$ on this historical data, by setting $|f(X) - Y|$ as the target label for instance X. The data used to train u should ideally be disjoint from the data used to train f to avoid overoptimistic estimates of uncertainty. We will typically use data splitting to avoid this issue, though there are more data efficient solutions such as cross-fitting. Notice that access to historical data will only be important in the batch setting; in the sequential setting we will be able to train $u(x)$ gradually on the collected data.

Classification uncertainty Next we look at classification, where Y is supported on a discrete set of values. Our main focus will be on binary classification, where $Y \in \{0, 1\}$. In such cases, our target is $\theta^* = \mathbb{P}(Y = 1)$.

In classification, $f(x)$ is usually obtained as the "most likely" class. If K is the number of classes, we have $f(x) = \arg\max_{i \in [K]} p_i(x)$, for some probabilistic output $p(x) = (p_1(x), \ldots, p_K(x))$ which satisfies $\sum_{i=1}^{K} p_i(x) = 1$. For example, $p(x)$ could be the softmax output of a neural network given input x. We will measure the uncertainty as $u(x) = \frac{K}{K-1} \cdot \left(1 - \max_{i \in [K]} p_i(x)\right)$. In binary classification, this reduces to

$$u(x) = 2 \min\{p(x), 1 - p(x)\}, \tag{3}$$

where we use $p(x)$ to denote the raw classifier output in $[0, 1]$. Therefore, $u(x)$ is large when $p(x)$ is close to uniform, i.e. $\max_i p_i(x) \approx 1/K$. On the other hand, if the model is confident, i.e. $\max_i p_i(x) \approx 1$, the uncertainty is close to zero.

5. Batch Active Inference

Building on the discussion from Section 4, we provide formal results for active inference in the batch setting. Recall that in the batch setting we observe i.i.d. unlabeled points $X_1, \ldots, X_n$, all at once. We consider a family of sampling rules $\pi_\eta(x) = \eta\, u(x)$, where $u(x)$ is the chosen uncertainty measure and $\eta \in \mathcal{H} \subseteq \mathbb{R}^+$ is a tuning parameter. We will discuss ways of choosing $u(x)$ in Section 7. The role of the tuning parameter is to scale the sampling rule to the sampling budget. We choose

$$\hat{\eta} = \max \left\{ \eta \in \mathcal{H} : \eta \sum_{i=1}^{n} u(X_i) \leq n_b \right\}, \tag{4}$$

and deploy $\pi_{\hat{\eta}}$ as the sampling rule. With this choice, we have $\mathbb{E}[n_{\mathrm{lab}}] = \mathbb{E}\left[\sum_{i=1}^{n} \hat{\eta}\, u(X_i)\right] \leq n_b$; therefore, $\pi_{\hat{\eta}}$ meets the label collection budget. We denote $\hat{\theta}^{\eta} \equiv \hat{\theta}^{\pi_\eta}$.

Mean estimation We first explain how to perform inference for mean estimation in Proposition 5.1. Recall the active mean estimator:

$$\hat{\theta}^{\hat{\eta}} = \frac{1}{n} \sum_{i=1}^{n} \left(f(X_i) + (Y_i - f(X_i)) \frac{\xi_i}{\pi_{\hat{\eta}}(X_i)} \right), \tag{5}$$

where $\xi_i \sim \mathrm{Bern}(\pi_{\hat{\eta}}(X_i))$. Following standard notation, z_q below denotes the qth quantile of the standard normal distribution.

Proposition 5.1. *Suppose that there exists $\eta^* \in \mathcal{H}$ such that $\mathbb{P}(\hat{\eta} \neq \eta^*) \to 0$. Then*

$$\sqrt{n}(\hat{\theta}^{\hat{\eta}} - \theta^*) \xrightarrow{d} \mathcal{N}(0, \sigma_*^2),$$

where $\sigma_*^2 = \mathrm{Var}(f(X) + (Y - f(X)) \frac{\xi^{\eta^*}}{\pi_{\eta^*}(X)})$ and $\xi^{\eta^*} \sim \mathrm{Bern}(\pi_{\eta^*}(X))$. Consequently, for any $\hat{\sigma}^2 \xrightarrow{p} \sigma_*^2$, $\mathcal{C}_\alpha = (\hat{\theta}^{\hat{\eta}} \pm z_{1-\alpha/2} \frac{\hat{\sigma}}{\sqrt{n}})$ is a valid $(1-\alpha)$-confidence interval: $\lim_{n \to \infty} \mathbb{P}(\theta^* \in \mathcal{C}_\alpha) = 1 - \alpha$.

A few remarks about Proposition 5.1 are in order: first, the consistency condition $\mathbb{P}(\hat{\eta} \neq \eta^*) \to 0$ is easily ensured if n_b/n has a limit $p \in (0, 1)$, that is, if n_b is asymptotically proportional to n. Then, as long as the space of tuning parameters $\mathcal{H}$ is discrete and there is no $\eta \in \mathcal{H}$ such that

$\eta\,\mathbb{E}[u(X)] = p$ exactly, the consistency condition is met (see Claim A.1). Second, obtaining a consistent variance estimate $\hat{\sigma}^2$ is straightforward, as one can simply take the empirical variance of the increments in the estimator (5).

We note that, while our main results will all focus on asymptotic confidence intervals, some of our results have direct non-asymptotic and time-uniform analogues; see Section C.

General M-estimation Next, we turn to general convex M-estimation. Recall this means that we can write $\theta^* = \arg\min_\theta L(\theta) = \arg\min_\theta \mathbb{E}[\ell_\theta(X,Y)]$, for a convex loss ℓ_θ. To simplify notation, let $\ell_{\theta,i} = \ell_\theta(X_i, Y_i)$, $\ell^f_{\theta,i} = \ell_\theta(X_i, f(X_i))$. We similarly use $\nabla\ell_{\theta,i}$ and $\nabla\ell^f_{\theta,i}$. For a general rule π, our *active estimator* is defined as

$$\hat{\theta}^\pi = \arg\min_\theta L^\pi(\theta), \text{ where} \tag{6}$$

$$L^\pi(\theta) = \frac{1}{n}\sum_{i=1}^n \left(\ell^f_{\theta,i} + (\ell_{\theta,i} - \ell^f_{\theta,i})\frac{\xi_i}{\pi(X_i)}\right).$$

As before, $\xi_i \sim \mathrm{Bern}(\pi(X_i))$. When π is the uniform rule, $\pi(x) = n_b/n$, the estimator (6) equals the general prediction-powered estimator (Angelopoulos et al., 2023c). Notice that the loss estimate $L^\pi(\theta)$ is unbiased: $\mathbb{E}[L^\pi(\theta)] = L(\theta)$. We again scale the sampling rule $\pi_\eta(x) = \eta\, u(x)$ according to the sampling budget, as in Eq. (4).

We next show asymptotic normality of $\hat{\theta}^{\hat{\eta}}$ for general targets θ^* which, in turn, enables inference. The result essentially follows from the usual asymptotic normality for M-estimators (Van der Vaart, 2000, Ch. 5), with some necessary modifications to account for the data-driven selection of $\hat{\eta}$. We require standard, mild smoothness assumptions on the loss ℓ_θ, formally stated in Ass. A.2 in the Appendix.

Theorem 5.2 (CLT for batch active inference)**.** *Assume the loss is smooth (Ass. A.2) and define the Hessian $H_{\theta^*} = \nabla^2\mathbb{E}[\ell_{\theta^*}(X,Y)]$. Suppose that there exists $\eta^* \in \mathcal{H}$ such that $\mathbb{P}(\hat{\eta} \neq \eta^*) \to 0$. Then, if $\hat{\theta}^{\eta^*} \xrightarrow{p} \theta^*$, we have*

$$\sqrt{n}(\hat{\theta}^{\hat{\eta}} - \theta^*) \xrightarrow{d} \mathcal{N}(0, \Sigma_*), \text{ where}$$

$$\Sigma_* = H_{\theta^*}^{-1}\mathrm{Var}\left(\nabla\ell^f_{\theta^*,i} + \left(\nabla\ell_{\theta^*,i} - \nabla\ell^f_{\theta^*,i}\right)\frac{\xi^{\eta^*}}{\pi_{\eta^*}(X)}\right)H_{\theta^*}^{-1}.$$

Consequently, for any $\hat{\Sigma} \xrightarrow{p} \Sigma_*$, $\mathcal{C}_\alpha = (\hat{\theta}^{\hat{\eta}}_j \pm z_{1-\alpha/2}\sqrt{\frac{\hat{\Sigma}_{jj}}{n}})$ is a valid $(1-\alpha)$-confidence interval for θ^*_j: $\lim_{n\to\infty} \mathbb{P}(\theta^*_j \in \mathcal{C}_\alpha) = 1-\alpha$.

The remarks following Proposition 5.1 again apply: the consistency condition on $\hat{\eta}$ is easily ensured if n_b/n has a limit, and $\hat{\Sigma}$ admits a simple plug-in estimate. The consistency condition on $\hat{\theta}^{\eta^*}$ is a standard requirement for analyzing M-estimators (see Van der Vaart, 2000, Ch. 5). It can be deduced if the empirical loss $L^\pi(\theta)$ is almost surely convex or if the parameter space is compact. The loss $L^\pi(\theta)$ is convex in a number of interesting cases, including means and GLMs (for the proof, see Angelopoulos et al., 2023c).

6. Sequential Active Inference

In the batch setting we observe all data points $X_1, \ldots, X_n$ at once and fix a predictive model f and sampling rule π that guide our choice of which labels to collect. An arguably more natural data collection strategy would operate in an online manner: as we collect more labels, we iteratively update the model and our strategy for which labels to collect next. This allows for further efficiency gains over using a fixed model throughout, as the latter ignores knowledge acquired during the data collection.

Formally, instead of having a fixed model f and rule π, we go through our data sequentially. At step $t \in \{1, \ldots, n\}$, we observe data point X_t and collect its label with probability $\pi_t(X_t)$, where $\pi_t(\cdot)$ is based on the uncertainty of model f_t. The model f_t can be fine-tuned on all information observed up to time t; formally, we require that $f_t, \pi_t \in \mathcal{F}_{t-1}$, where $\mathcal{F}_t$ is the σ-algebra generated by the first t points X_s, $1 \le s \le t$, their labeling decisions ξ_s, and their labels Y_s, if observed: $\mathcal{F}_t = \sigma((X_1, Y_1\xi_1, \xi_1), \ldots, (X_t, Y_t\xi_t, \xi_t))$. (Note that $Y_t\xi_t = Y_t$ if and only if $\xi_t = 1$; otherwise, $Y_t\xi_t = \xi_t = 0$.) We will again calibrate our decisions of whether to collect a label according to a budget on the sample size n_b. We denote by $n_{\mathrm{lab},t}$ the number of labels collected up to time t.

Inference in the sequential setting is more challenging than batch inference because the data points $(X_t, Y_t, \xi_t), t \in [n]$, are dependent; indeed, the purpose of the sequential setting is to leverage previous observations when deciding on future labeling decisions. We will construct estimators that respect a *martingale* structure, which will enable tractable inference via the martingale central limit theorem (Dvoretzky, 1972).

Mean estimation We begin by focusing on the mean. If we take ℓ_θ to be the squared loss as in Example 2.1, we obtain the sequential active mean estimator:

$$\hat{\theta}^{\vec{\pi}} = \frac{1}{n}\sum_{t=1}^n \Delta_t, \;\; \Delta_t = f_t(X_t) + (Y_t - f_t(X_t))\frac{\xi_t}{\pi_t(X_t)}.$$

Note that Δ_t are *martingale increments*; they share a common conditional mean $\mathbb{E}[\Delta_t|\mathcal{F}_{t-1}] = \theta^*$, and they are $\mathcal{F}_t$-measurable. We let $\sigma_t^2 = V(f_t, \pi_t) = \mathrm{Var}(\Delta_t|f_t, \pi_t)$ denote the conditional variance of the increments.

To show asymptotic normality of $\hat{\theta}^{\vec{\pi}}$, we require the Lindeberg condition, which is a standard assumption for proving central limit theorems when the increments are not i.i.d.: $\frac{1}{n}\sum_{t=1}^n \mathbb{E}[\bar{\Delta}_t^2 \mathbf{1}\{|\bar{\Delta}_t| > \epsilon\sqrt{n}\}|\mathcal{F}_{t-1}] \xrightarrow{p} 0$, for all $\epsilon > 0$, where $\bar{\Delta}_t = \Delta_t - \theta^*$. Roughly speaking, the Lindeberg condition requires that the increments do not have very heavy tails; it prevents any increment from having a disproportionate contribution to the overall variance.

Proposition 6.1. *Suppose $\frac{1}{n}\sum_{t=1}^n \sigma_t^2 \xrightarrow{p} \sigma_*^2 = V(f_*, \pi_*)$, for some fixed model–rule pair (f_*, π_*), and that the incre-*

ments Δ_t satisfy the Lindeberg condition (Ass. A.3). Then

$$\sqrt{n}(\hat{\theta}^{\vec{\pi}} - \theta^*) \xrightarrow{d} \mathcal{N}(0, \sigma_*^2).$$

Consequently, for any $\hat{\sigma}^2 \xrightarrow{p} \sigma_*^2$, $\mathcal{C}_\alpha = (\hat{\theta}^{\vec{\pi}} \pm z_{1-\alpha/2}\frac{\hat{\sigma}}{\sqrt{n}})$ is a valid $(1-\alpha)$-confidence interval: $\lim_{n\to\infty} \mathbb{P}(\theta^* \in \mathcal{C}_\alpha) = 1-\alpha$.

Intuitively, Proposition 6.1 requires that the model f_t and sampling rule π_t converge. For example, a sufficient condition for $\frac{1}{n}\sum_{t=1}^n \sigma_t^2 \xrightarrow{p} \sigma_*^2$ is $V(f_n, \pi_n) \xrightarrow{L^1} V(f_*, \pi_*)$. Since the sampling rule is typically based on the model, it makes sense that it would converge if f_t converges. At the same time, it makes sense for f_t to gradually stop updating after sufficient accuracy is achieved.

General M-estimation We generalize Proposition 6.1 to all convex M-estimation problems. The general version of our sequential active estimator takes the form

$$\hat{\theta}^{\vec{\pi}} = \arg\min_\theta L^{\vec{\pi}}(\theta), \text{ where } L^{\vec{\pi}}(\theta) = \frac{1}{n}\sum_{t=1}^n L_t(\theta), \quad (7)$$

$$L_t(\theta) = \ell_{\theta,t}^{f_t} + (\ell_{\theta,t} - \ell_{\theta,t}^{f_t})\frac{\xi_t}{\pi_t(X_t)}.$$

Let $V_{\theta,t} = V_\theta(f_t, \pi_t) = \mathrm{Var}(\nabla L_t(\theta)|f_t, \pi_t)$. We will again require that (f_t, π_t) converge in an analogous sense.

Theorem 6.2 (CLT for sequential active inference)**.** *Assume the loss is smooth (Ass. A.2) and define the Hessian $H_{\theta^*} = \nabla^2\mathbb{E}[\ell_{\theta^*}(X,Y)]$. Suppose also that $\frac{1}{n}\sum_{t=1}^n V_{\theta^*,t} \xrightarrow{p} V_* = V_{\theta^*}(f_*, \pi_*)$ entry-wise for some fixed model–rule pair (f_*, π_*), and that the increments $L_t(\theta)$ satisfy the Lindeberg condition (Ass. A.4). Then, if $\hat{\theta}^{\vec{\pi}} \xrightarrow{p} \theta^*$, we have*

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

where $\Sigma_* = H_{\theta^*}^{-1}V_*H_{\theta^*}^{-1}$. Consequently, for any $\hat{\Sigma} \xrightarrow{p} \Sigma_*$, $\mathcal{C}_\alpha = (\hat{\theta}_j^{\vec{\pi}} \pm z_{1-\alpha/2}\sqrt{\frac{\hat{\Sigma}_{jj}}{n}})$ is a valid $(1-\alpha)$-confidence interval for θ_j^*: $\lim_{n\to\infty} \mathbb{P}(\theta_j^* \in \mathcal{C}_\alpha) = 1-\alpha$.

The conditions of Theorem 6.2 are largely the same as in Theorem 5.2; the main difference is the requirement of convergence of the model–sampling rule pairs, which is similar to the analogous condition of Proposition 6.1.

Proposition 6.1 and Theorem 6.2 apply to any sampling rule π_t, as long as the variance convergence requirement is met. We discuss ways to set π_t so that the sampling budget n_b is met. Our default will be to "spread out" the budget n_b over the n observations. We will do so by having an "imaginary" budget for the expected number of collected labels by step t, equal to $n_{b,t} = tn_b/n$. Let $n_{\Delta,t} = n_{b,t} - n_{\mathrm{lab},t-1}$ denote the remaining budget at step t. We derive a measure of uncertainty u_t from model f_t, as before, and let

$$\pi_t(x) = \min\left\{\eta_t\, u_t(x), n_{\Delta,t}\right\}_{[0,1]}, \quad (8)$$

where η_t normalizes $u_t(x)$ and the subscript $[0,1]$ denotes clipping to $[0,1]$. The normalizing constant η_t can be arbitrary, but we find it helpful to set it roughly as $\eta_t = n_b/(n\,\mathbb{E}[u_t(X)])$. In words, the sampling probability is high if the uncertainty is high *and* we have not used up too much of the sampling budget thus far. Of course, if the model estimates low uncertainty $u_t(x)$ throughout, the budget will be underutilized. For this reason, to make sure we use up the budget in practice, we occasionally set $\pi_t(x) = (n_{\Delta,t})_{[0,1]}$ regardless of the uncertainty.

7. Choosing the Sampling Rule

We have seen how to perform inference given an abstract sampling rule, and argued that, intuitively, the sampling rule should be calibrated to the uncertainty of the model's predictions. Here we argue that this is in fact the *optimal* strategy. We derive an "oracle" rule, which optimally sets the sampling probabilities so that the variance of $\hat{\theta}^\pi$ is minimized. While the oracle rule cannot be implemented since it depends on unobserved information, it provides an ideal that our algorithms will try to approximate. We discuss in detail ways of tuning the approximations to make them practical and powerful in Section B.1 in the Appendix.

Mean estimation Recall the expression for $\mathrm{Var}(\hat{\theta}^\pi)$ (2). Given that $\mathbb{E}\left[\pi^{-1}(X)(Y - f(X))^2\right]$ is the only term that depends on π, we define the oracle rule as the solution to:

$$\min_\pi \mathbb{E}\left[\pi(X)^{-1}(Y - f(X))^2\right] \text{ s.t. } \mathbb{E}[\pi(X)] \le \frac{n_b}{n}. \quad (9)$$

The optimization problem (9) appears in importance sampling (Owen, 2013, Ch. 9), constrained utility optimization (Balcan et al., 2014), and, relatedly to our work, survey sampling (Särndal, 1980). The optimality conditions of (9) show that its solution π_{opt} satisfies:

$$\pi_{\mathrm{opt}}(X) \propto \sqrt{\mathbb{E}[(Y - f(X))^2|X]},$$

where $\propto$ ignores the normalizing constant required to make $\mathbb{E}[\pi_{\mathrm{opt}}(X)] \le n_b/n$. Therefore, the optimal sampling rule is one that samples data points according to the expected magnitude of the model error. Of course, $\mathbb{E}[(Y - f(X))^2|X]$ cannot be known since the label distribution is unknown, and that is why we call π_{opt} an oracle.

To develop intuition, it is instructive to consider an even more powerful oracle $\tilde{\pi}_{\mathrm{opt}}(X, Y)$ that is allowed to depend on Y. To be clear, we would commit to the same functional form as in (1) and would seek to minimize $\mathrm{Var}(\hat{\theta}^\pi)$ while allowing the sampling probabilities to depend on both X *and* Y. In this case, by the same argument we conclude that

$$\tilde{\pi}_{\mathrm{opt}}(X, Y) \propto |Y - f(X)|. \quad (10)$$

The perspective of allowing the oracle to depend on both X and Y is directly prescriptive: a natural way to approximate the rule $\tilde{\pi}_{\rm opt}$ is to train an arbitrary black-box model u on historical (X, Y) pairs to predict $|Y - f(X)|$ from X.

General M-estimation In general, we cannot hope to minimize the variance of $\hat{\theta}^{\pi}$ at a fixed sample size n since the finite-sample distribution of $\hat{\theta}^{\pi}$ is not tractable. However, we can minimize the *asymptotic* variance of $\hat{\theta}^{\pi}$. Since the estimator is potentially multi-dimensional, we assume that we want to minimize the asymptotic variance of a single coordinate $\hat{\theta}^{\pi}_j$ (for example, one regression coefficient). A short derivation similar to the one for mean estimation shows that

$$\pi_{\rm opt}(X) \propto \sqrt{\mathbb{E}[((\nabla \ell_{\theta^*}(X,Y) - \nabla \ell_{\theta^*}(X, f(X)))^\top h^{(j)})^2 | X]},$$

where $h^{(j)}$ is the j-th column of $H_{\theta^*}^{-1}$. This recovers $\pi_{\rm opt}$ for the mean, as the squared loss has $\nabla \ell_{\theta^*}(x,y) = \theta^* - y$.

Generalized linear models (GLMs) We simplify the general solution $\pi_{\rm opt}$ in the case of generalized linear models (GLMs). We define GLMs as M-estimators whose loss function takes the form $\ell_\theta(x,y) = -\log p_\theta(y|x) = -yx^\top\theta + \psi(x^\top\theta)$, for some convex log-partition function ψ. This definition recovers linear regression by taking $\psi(s) = \frac{1}{2}s^2$ and logistic regression by taking $\psi(s) = \log(1+e^s)$. By the definition of the GLM loss, we have $\nabla \ell_{\theta^*}(x,y) - \nabla \ell_{\theta^*}(x, f(x)) = (f(x) - y)x$ and, therefore,

$$\pi_{\rm opt}(X) \propto \sqrt{\mathbb{E}[(f(X) - Y)^2|X]} \cdot |X^\top h^{(j)}|,$$

where the Hessian is equal to $H_{\theta^*} = \mathbb{E}[\psi''(X^\top\theta_*)XX^\top]$ and $h^{(j)}$ is the j-th column of $H_{\theta^*}^{-1}$. In linear regression, for instance, $H_{\theta^*} = \mathbb{E}[XX^\top]$. Again, we see that the model errors play a role in determining the optimal sampling. In particular, again considering the more powerful oracle $\tilde{\pi}_{\rm opt}(X, Y)$ that is allowed to set the sampling probabilities according to both X and Y, we get

$$\tilde{\pi}_{\rm opt}(X,Y) \propto |f(X) - Y| \cdot |X^\top h^{(j)}|. \tag{11}$$

Therefore, as in the case of the mean, our measure of uncertainty will aim to predict $|f(X) - Y|$ from X and plug those predictions into the above rule.

8. Experiments

We evaluate active inference on several problems and compare it to two baselines. The first baseline replaces active sampling with the uniformly random sampling rule $\pi^{\rm unif}$. This baseline still uses machine learning predictions $f(X_i)$ and corresponds to prediction-powered inference (PPI) (Angelopoulos et al., 2023a). The purpose of this comparison is to quantify the benefits of machine-learning-driven data collection. In the rest of this section we refer to this baseline as the "uniform" baseline because the only difference from our estimator is that it replaces active sampling with uniform sampling. The second baseline removes machine learning altogether and computes the "classical" estimate based on uniformly random sampling, $\hat{\theta}^{\rm noML} = \arg\min_\theta \frac{1}{n_b}\sum_{i=1}^n \ell_\theta(X_i, Y_i)\xi_i$, where $\xi_i \sim {\rm Bern}(\frac{n_b}{n})$. This baseline serves to evaluate the cumulative benefits of machine learning for data collection *and* inference combined. We refer to this baseline as the "classical" baseline, or classical inference.

The target error level is $\alpha = 0.1$ throughout. We report the average interval width and coverage for varying sample sizes n_b, averaged over 1000 and 100 trials for the batch and sequential settings, respectively. We plot the interval width on a log–log scale. In Appendix B we also report the percentage of budget saved by active inference relative to the baselines when the methods are matched to be equally accurate. More precisely, for varying n_b we compute the average interval width achieved by the uniform and classical baselines; then, we look for the budget size $n_b^{\rm active}$ for which active inference achieves the same average interval width, and report $(n_b - n_b^{\rm active})/n_b \cdot 100\%$ as the percentage of budget saved. The batch and sequential active inference methods used in our experiments are outlined in Algorithm 1 and Algorithm 2 in the Appendix. We defer some experimental details to Appendix B. Code for reproducing the experiments is available at this link.

Post-election survey research We apply active inference to survey data collected by the Pew Research Center following the 2020 United States presidential election (Pew, 2020). We focus on one question in the survey, aimed at gauging people's approval of Joe Biden's (Donald Trump's, respectively) political messaging following the election. Approval is encoded as a binary response, $Y_i \in \{0, 1\}$. The respondents—a nationally representative pool of US adults—provide background information such as age, gender, education, political affiliation. We show that, by training an XGBoost model (Chen & Guestrin, 2016) to predict approval from background information and measuring the model's uncertainty via Eq. (3), we can allocate the per-question budget in a way that outperforms uniform allocation.

In Figure 1 (rows 1 and 2) we compare active inference to the uniform (PPI) and classical baselines. All methods meet the coverage requirement. Across different values of the budget n_b, active sampling reduces the confidence interval width of the uniform baseline (PPI) by a significant margin (at least $\sim 10\%$). Classical inference is highly suboptimal compared to both alternatives. In Figure 3 we report the percentage of budget saved due to active sampling. For estimating Biden's approval, we observe an over 85% save in budget over classical inference and around 25% save over the uniform baseline. For estimating Trump's approval, we observe an over 70% save in budget over classical inference and around 25% save over the uniform baseline.

Census data analysis Next, we study the American Community Survey (ACS) Public Use Microdata Sample (PUMS) collected by the US Census Bureau. ACS PUMS is an

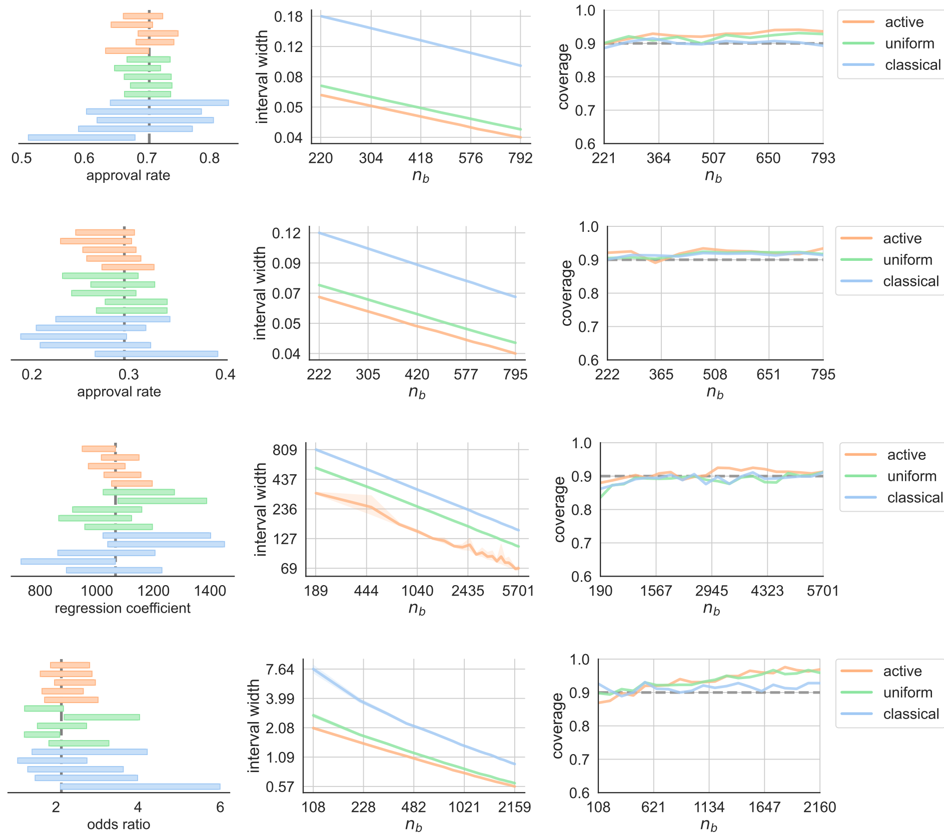

Figure 1. **Batch experiments.** Example intervals in five randomly chosen trials (left), average confidence interval width (middle), and coverage (right) in post-election survey research (rows 1 and 2), census analysis (row 3), and proteomics research (row 4).

annual survey that collects information about citizenship, education, income, employment, and more. We investigate the relationship between age and income in survey data collected in California in 2019, controlling for sex. Specifically, we target the linear regression coefficient when regressing income on age and sex (that is, its age coordinate). We train an XGBoost model (Chen & Guestrin, 2016) to predict a person's income from the available demographic covariates. To quantify the model's uncertainty, we use the strategy described in Section 4, training a separate XGBoost model $e(\cdot)$ to predict $|f(X) - Y|$ from X. We set the uncertainty $u(x)$ as prescribed in Eq. (11), replacing $|f(X) - Y|$ by $e(X)$.

The interval widths and coverage are shown in Figure 1 (row 3). As in the previous application, all methods approximately achieve the target coverage, however this time we observe more extreme gains over the uniform baseline (PPI): the interval widths almost double when going from active sampling to uniform sampling. Of course, the improvement of active inference over classical inference is even more substantial. The large gains of active sampling can also be seen in Figure 3: we save around 80% of the budget over classical inference and over 60% over the uniform baseline.

AlphaFold-assisted proteomics research Inspired by the findings of Bludau et al. (2022) and the subsequent analysis of Angelopoulos et al. (2023a), we study the odds ratio of a protein being phosphorylated, a functional property of a protein, and being part of an intrinsically disordered region (IDR), a structural property. Angelopoulos et al. (2023a) showed that forming a classical confidence interval around the odds ratio based on AlphaFold predictions is not valid given that the predictions are imperfect. They provide a valid alternative assuming access to a small subset of proteins with true structure measurements, uniformly sampled from the larger population of proteins of interest. We show that, by strategically choosing which protein structures to experimentally measure, active inference allows for intervals that retain validity and are tighter than intervals based on uniform sampling. Naturally, for the purpose of evaluating validity, we restrict the analysis to proteins where we have gold-standard structure measurements; we use the post-processed AlphaFold outputs made available by Angelopoulos et al. (2023a), which predict the IDR property based on the raw AlphaFold output. While the odds ratio is not a solution to an M-estimation problem, it is a func-

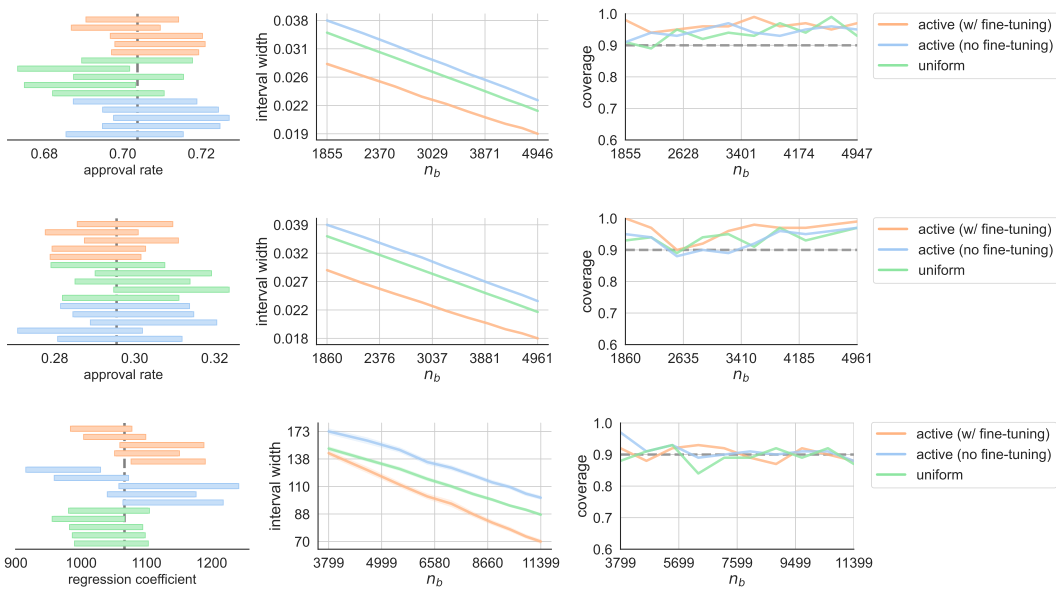

Figure 2. **Experiments with fine-tuning.** Example intervals in five randomly chosen trials (left), average confidence interval width (middle), and coverage (right) in post-election survey research (rows 1 and 2) and census analysis (row 3).

tion of two means (see also (Angelopoulos et al., 2023a;c)). Confidence intervals can thus be computed by applying the delta method to the asymptotic normality result for the mean. Since Y is binary, we measure uncertainty via Eq. (3).

Figure 1 (row 4) shows the interval widths and coverage for the three methods, and Figure 3 shows the percentage of budget saved due to adaptive data collection. The gains are substantial: over 75% of the budget is saved in comparison to classical inference, and around $20 - 25\%$ is saved in comparison to the uniform baseline (PPI).

Post-election survey research with fine-tuning We return to the first application, this time evaluating the benefits of sequential fine-tuning. We compare active inference, with and without fine-tuning, and PPI, which relies on uniform sampling. We show that active inference with no fine-tuning can hurt compared to PPI if the former uses a poorly trained model; fine-tuning, on the other hand, remedies this issue. We train an XGBoost model on only 10 labeled examples and use this model for active inference with no fine-tuning and PPI. Active inference with fine-tuning continues to fine-tune the model with every $B = 100$ new survey responses, also updating the sampling rule via update (8). The uncertainty measure $u_t(x)$ is given by Eq. (3), as before. As discussed in Section 6, we also periodically use up the remaining budget regardless of the uncertainty in order to avoid underutilizing the budget (every $100n/n_b$ steps).

The interval widths and coverage are reported in Figure 2 (rows 1 and 2). We find that fine-tuning greatly improves power and retains correct coverage. In Figure 4 we show the save in sample size budget over active inference with no fine-tuning and inference based on uniform sampling, i.e. PPI. For estimating Biden's approval, we observe a gain of around 40% and 30% relative to active inference without fine-tuning and PPI, respectively. For Trump's approval, we observe even larger gains, around 45% and 35%.

Census data analysis with fine-tuning We similarly evaluate the benefits of sequential fine-tuning in the census example. We again compare active inference, with and without fine-tuning, and PPI, i.e., active inference with a trivial, uniform sampling rule. Recall that we trained a separate model e to predict the prediction errors, which we in turn used to form the uncertainty $u(x)$ according to Eq. (11). This time we fine-tune both the prediction model, f_t, and the error model, e_t. We train initial XGBoost models f_1 and e_1 on 100 labeled examples. We use f_1 for PPI and both f_1 and e_1 for active inference with no fine-tuning. Active inference with fine-tuning continues to fine-tune the two models with every $B = 1000$ new survey responses, also updating the model uncertainty via update (8). We compute u_t from e_t based on Eq. (11). We again periodically use up the remaining budget regardless of the computed uncertainty in order to avoid underutilizing the budget (in particular, every $500n/n_b$ steps).

We show the interval widths and coverage in Figure 2 (row 3). The gains of fine-tuning are significant and increase as n_b increases. In Figure 4 we show the save in sample size budget: fine-tuning saves around $32-40\%$ over the baseline with no fine-tuning and around $20 - 30\%$ over PPI.

Acknowledgements

We thank Lihua Lei, Jann Spiess, and Stefan Wager for many insightful comments and pointers to relevant work. T.Z. was supported by Stanford Data Science through the Fellowship program. E.J.C. was supported by the Office of Naval Research grant N00014-20-1-2157, the National Science Foundation grant DMS-2032014, the Simons Foundation under award 814641, and the ARO grant 2003514594.

Impact Statement

Our proposal aims to provide a method for valid and more powerful statistical inference compared to current baselines. However, if applied poorly, our method could lead to less powerful inferences—and thus to worse downstream decisions—than classical inference or PPI. This could happen if the model's estimate of uncertainty is miscalibrated, for example. Another limitation of our work is its reliance on the assumption that the data points (X_i, Y_i) are i.i.d.; especially in the setting where we collect data sequentially, this assumption may be violated and our methods may fail to cover as a result.

References

Angelopoulos, A. N., Bates, S., Fannjiang, C., Jordan, M. I., and Zrnic, T. Prediction-powered inference. *Science*, 382 (6671):669–674, 2023a.

Angelopoulos, A. N., Bates, S., Fannjiang, C., Jordan, M. I., and Zrnic, T. Prediction-powered inference: Data sets, 2023b. URL `https://doi.org/10.5281/zenodo.8397451`.

Angelopoulos, A. N., Duchi, J. C., and Zrnic, T. PPI++: Efficient prediction-powered inference. *arXiv preprint arXiv:2311.01453*, 2023c.

Ash, J. T., Zhang, C., Krishnamurthy, A., Langford, J., and Agarwal, A. Deep batch active learning by diverse, uncertain gradient lower bounds. *arXiv preprint arXiv:1906.03671*, 2019.

Azriel, D., Brown, L. D., Sklar, M., Berk, R., Buja, A., and Zhao, L. Semi-supervised linear regression. *Journal of the American Statistical Association*, 117(540):2238–2251, 2022.

Balcan, M.-F., Beygelzimer, A., and Langford, J. Agnostic active learning. In *Proceedings of the 23rd international conference on Machine learning*, pp. 65–72, 2006.

Balcan, M.-F., Daniely, A., Mehta, R., Urner, R., and Vazirani, V. V. Learning economic parameters from revealed preferences. In *Web and Internet Economics: 10th International Conference, WINE 2014, Beijing, China, December 14-17, 2014. Proceedings 10*, pp. 338–353. Springer, 2014.

Bhattacharya, D. and Dupas, P. Inferring welfare maximizing treatment assignment under budget constraints. *Journal of Econometrics*, 167(1):168–196, 2012.

Bludau, I., Willems, S., Zeng, W.-F., Strauss, M. T., Hansen, F. M., Tanzer, M. C., Karayel, O., Schulman, B. A., and Mann, M. The structural context of posttranslational modifications at a proteome-wide scale. *PLoS biology*, 20(5):e3001636, 2022.

Chandak, Y., Shankar, S., Syrgkanis, V., and Brunskill, E. Adaptive instrument design for indirect experiments. *arXiv preprint arXiv:2312.02438*, 2023.

Chen, T. and Guestrin, C. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pp. 785–794, 2016.

Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., Newey, W., and Robins, J. Double/debiased machine learning for treatment and structural parameters, 2018.

Cook, T., Mishler, A., and Ramdas, A. Semiparametric efficient inference in adaptive experiments. *arXiv preprint arXiv:2311.18274*, 2023.

Ding, F., Hardt, M., Miller, J., and Schmidt, L. Retiring adult: New datasets for fair machine learning. *Advances in neural information processing systems*, 34:6478–6490, 2021.

Durrett, R. *Probability: theory and examples*, volume 49. Cambridge university press, 2019.

Dvoretzky, A. Asymptotic normality for sums of dependent random variables. In *Proceedings of the Sixth Berkeley Symposium on Mathematical Statistics and Probability, Volume 2: Probability Theory*, volume 6, pp. 513–536. University of California Press, 1972.

Gal, Y., Islam, R., and Ghahramani, Z. Deep Bayesian active learning with image data. In *International conference on machine learning*, pp. 1183–1192. PMLR, 2017.

Gan, F. and Liang, W. Prediction de-correlated inference. *arXiv preprint arXiv:2312.06478*, 2023.

Hadad, V., Hirshberg, D. A., Zhan, R., Wager, S., and Athey, S. Confidence intervals for policy evaluation in adaptive experiments. *Proceedings of the national academy of sciences*, 118(15):e2014602118, 2021.

Hahn, J., Hirano, K., and Karlan, D. Adaptive experimental design using the propensity score. *Journal of Business & Economic Statistics*, 29(1):96–108, 2011.

Hanneke, S. et al. Theory of disagreement-based active learning. *Foundations and Trends® in Machine Learning*, 7(2-3):131–309, 2014.

Hu, F. and Rosenberger, W. F. *The theory of response-adaptive randomization in clinical trials*. John Wiley & Sons, 2006.

Jean, N., Burke, M., Xie, M., Davis, W. M., Lobell, D. B., and Ermon, S. Combining satellite imagery and machine learning to predict poverty. *Science*, 353(6301):790–794, 2016.

Joshi, A. J., Porikli, F., and Papanikolopoulos, N. Multi-class active learning for image classification. In *2009 ieee conference on computer vision and pattern recognition*, pp. 2372–2379. IEEE, 2009.

Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Žídek, A., Potapenko, A., et al. Highly accurate protein structure prediction with alphafold. *Nature*, 596(7873):583–589, 2021.

Kasy, M. and Sautmann, A. Adaptive treatment assignment in experiments for policy choice. *Econometrica*, 89(1): 113–132, 2021.

Kato, M., Ishihara, T., Honda, J., and Narita, Y. Efficient adaptive experimental design for average treatment effect estimation. *arXiv preprint arXiv:2002.05308*, 2020.

Lai, T. L. and Robbins, H. Asymptotically efficient adaptive allocation rules. *Advances in applied mathematics*, 6(1): 4–22, 1985.

List, J. A., Sadoff, S., and Wagner, M. So you want to run an experiment, now what? some simple rules of thumb for optimal experimental design. *Experimental Economics*, 14:439–457, 2011.

Miao, J., Miao, X., Wu, Y., Zhao, J., and Lu, Q. Assumption-lean and data-adaptive post-prediction inference. *arXiv preprint arXiv:2311.14220*, 2023.

Motwani, K. and Witten, D. Valid inference after prediction. *arXiv preprint arXiv:2306.13746*, 2023.

Nassiuma, D. K. Survey sampling: Theory and methods, 2001.

Orabona, F. and Jun, K.-S. Tight concentrations and confidence sequences from the regret of universal portfolio. *IEEE Transactions on Information Theory*, 2023.

Owen, A. B. *Monte Carlo theory, methods and examples*. https://artowen.su.domains/mc/, 2013.

Pew. American trends panel (ATP) wave 79, 2020. URL https://www.pewresearch.org/science/dataset/american-trends-panel-wave-79/.

Ren, P., Xiao, Y., Chang, X., Huang, P.-Y., Li, Z., Gupta, B. B., Chen, X., and Wang, X. A survey of deep active learning. *ACM computing surveys (CSUR)*, 54(9):1–40, 2021.

Robbins, H. Some aspects of the sequential design of experiments. 1952.

Robins, J. M. and Rotnitzky, A. Semiparametric efficiency in multivariate regression models with missing data. *Journal of the American Statistical Association*, 90(429):122–129, 1995.

Robins, J. M., Rotnitzky, A., and Zhao, L. P. Estimation of regression coefficients when some regressors are not always observed. *Journal of the American statistical Association*, 89(427):846–866, 1994.

Rolf, E., Proctor, J., Carleton, T., Bolliger, I., Shankar, V., Ishihara, M., Recht, B., and Hsiang, S. A generalizable and accessible approach to machine learning with global satellite imagery. *Nature communications*, 12(1):4392, 2021.

Rubin, D. Multiple imputation for nonresponse in surveys. *Wiley Series in Probability and Statistics*, pp. 1, 1987.

Rubin, D. B. Inference and missing data. *Biometrika*, 63(3): 581–592, 1976.

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

Särndal, C. E. On π-inverse weighting versus best linear unbiased weighting in probability sampling. *Biometrika*, 67(3):639–650, 1980.

Särndal, C.-E., Swensson, B., and Wretman, J. *Model assisted survey sampling*. Springer Science & Business Media, 2003.

Schohn, G. and Cohn, D. Less is more: Active learning with support vector machines. In *ICML*, volume 2, pp. 6, 2000.

Settles, B. Active learning literature survey. *Department of Computer Sciences, University of Wisconsin-Madison*, 2009.

Tong, S. and Koller, D. Support vector machine active learning with applications to text classification. *Journal of machine learning research*, 2(Nov):45–66, 2001.

Van der Vaart, A. W. *Asymptotic statistics*, volume 3. Cambridge university press, 2000.

Waudby-Smith, I. and Ramdas, A. Estimating means of bounded random variables by betting. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 86(1):1–27, 2024.

Xie, M., Jean, N., Burke, M., Lobell, D., and Ermon, S. Transfer learning from deep features for remote sensing and poverty mapping. In *Proceedings of the AAAI conference on artificial intelligence*, volume 30, 2016.

Zdun, M. Machine politics: How America casts and counts its votes. *Reuters*, 2022.

Zhang, A., Brown, L. D., and Cai, T. T. Semi-supervised inference: General theory and estimation of means. *Annals of Statistics*, 47(5):2538–2566, 2019.

Zhang, K., Janson, L., and Murphy, S. Statistical inference with M-estimators on adaptively collected data. *Advances in neural information processing systems*, 34:7460–7471, 2021.

Zhang, Y. and Bradic, J. High-dimensional semi-supervised learning: in search of optimal inference of the mean. *Biometrika*, 109(2):387–403, 2022.

Zrnic, T. and Candès, E. J. Cross-prediction-powered inference. *Proceedings of the National Academy of Sciences*, 121(15):e2322083121, 2024.

A. Proofs

A.1. Proof of Proposition 5.1

Recall that $\xi_i \sim \mathrm{Bern}(\pi_{\hat{\eta}}(X_i))$. For any $\eta \in \mathcal{H}$, we define

$$\xi_i^\eta = \mathbf{1}\{\pi_\eta(X_i) \le \pi_{\hat{\eta}}(X_i)\}\xi_i(1-\xi_i^{\le}) + \mathbf{1}\{\pi_\eta(X_i) > \pi_{\hat{\eta}}(X_i)\}(\xi_i + (1-\xi_i)\xi_i^{>}), \tag{12}$$

where $\xi_i^{\le} \sim \mathrm{Bern}(\frac{\pi_{\hat{\eta}}(X_i)-\pi_\eta(X_i)}{\pi_{\hat{\eta}}(X_i)})$ and $\xi_i^{>} \sim \mathrm{Bern}(\frac{\pi_\eta(X_i)-\pi_{\hat{\eta}}(X_i)}{1-\pi_{\hat{\eta}}(X_i)})$ are drawn independently of ξ_i. This definition couples $\xi_i^{\eta^*}$ with ξ_i, while ensuring that $\xi_i^{\eta^*} \sim \mathrm{Bern}(\pi_{\eta^*}(X_i))$. Let

$$\hat{\theta}^{\eta^*} = \frac{1}{n}\sum_{i=1}^n \left(f(X_i) + (Y_i - f(X_i))\frac{\xi_i^{\eta^*}}{\pi_{\eta^*}(X_i)} \right).$$

By the central limit theorem, we know that

$$\sqrt{n}(\hat{\theta}^{\eta^*} - \theta^*) \xrightarrow{d} \mathcal{N}(0, \sigma_*^2), \tag{13}$$

where $\sigma_*^2 = \mathrm{Var}\left(f(X) + (Y - f(X))\frac{\xi^{\eta^*}}{\pi_{\eta^*}(X)} \right)$. On the other hand, we have

$$\sqrt{n}(\hat{\theta}^{\hat{\eta}} - \theta^*) = \sqrt{n}(\hat{\theta}^{\eta^*} - \theta^*) + \sqrt{n}(\hat{\theta}^{\hat{\eta}} - \hat{\theta}^{\eta^*}).$$

For any $\epsilon > 0$, we have $\mathbb{P}(|\sqrt{n}(\hat{\theta}^{\hat{\eta}} - \hat{\theta}^{\eta^*})| \ge \epsilon) \le \mathbb{P}(\hat{\eta} \ne \eta^*) \to 0$; therefore, $\sqrt{n}(\hat{\theta}^{\hat{\eta}} - \hat{\theta}^{\eta^*}) \xrightarrow{p} 0$. Putting this fact together with Eq. (13), we conclude that $\sqrt{n}(\hat{\theta}^{\hat{\eta}} - \theta^*) \xrightarrow{d} \mathcal{N}(0, \sigma_*^2)$ by Slutsky's theorem.

A.2. Sufficient Condition for Consistency of $\hat{\eta}$

Claim A.1. *Suppose that $n_b/n \to p \in (0,1)$. If $\mathcal{H}$ is discrete and there is no $\eta \in \mathcal{H}$ such that $\eta\mathbb{E}[u(X)] = p$ exactly, then there exists $\eta^* \in \mathcal{H}$ such that $\mathbb{P}(\hat{\eta} \ne \eta^*) \to 0$.*

Proof. We have that $\hat{\eta}$ is the maximum $\eta \in \mathcal{H}$ such that $\eta\frac{1}{n}\sum_{i=1}^n u(X_i) \le \frac{n_b}{n}$, or equivalently, $\eta \le (\frac{n_b}{n})/(\frac{1}{n}\sum_{i=1}^n u(X_i))$. The right-hand side converges in probability to $p/\mathbb{E}[u(X)]$. Now, take $\epsilon = \min_{\eta \in \mathcal{H}} |\eta - p/\mathbb{E}[u(X)]|$. Then, on the event that $(\frac{n_b}{n})/(\frac{1}{n}\sum_{i=1}^n u(X_i))$ is within ϵ of $p/\mathbb{E}[u(X)]$, we know $\hat{\eta}$ will be equal to $\eta^* = \max\{\eta \in \mathcal{H} : \eta\mathbb{E}[u(X)] \le p\}$. Therefore, $\mathbb{P}(\hat{\eta} \ne \eta^*) \le \mathbb{P}(|(\frac{n_b}{n})/(\frac{1}{n}\sum_{i=1}^n u(X_i)) - p/\mathbb{E}[u(X)]| \ge \epsilon) \to 0$, by the claimed convergence in probability. □

A.3. Proof of Theorem 5.2

The proof follows a similar argument as the classical proof of asymptotic normality for M-estimation; see (Van der Vaart, 2000, Thm. 5.23). A similar proof is also given for the prediction-powered estimator (Angelopoulos et al., 2023c), which is closely related to our active inference estimator. The main difference between our proof and the classical proof is that $\hat{\eta}$ is tuned in a data-adaptive fashion, so the increments in the empirical loss $L^{\pi_{\hat{\eta}}}(\theta)$ are not independent. We begin by formally stating the required smoothness assumption.

Assumption A.2 (Smoothness)**.** The loss ℓ is smooth if:

- $\ell_\theta(x,y)$ is differentiable at θ^* for all (x,y);
- ℓ_θ is locally Lipschitz around θ^*: there is a neighborhood of θ^* such that $\ell_\theta(x,y)$ is $C(x,y)$-Lipschitz and $\ell_\theta(x, f(x))$ is $C(x)$-Lipschitz in θ, where $\mathbb{E}[C(X,Y)^2] < \infty, \mathbb{E}[C(X)^2] < \infty$;
- $L(\theta) = \mathbb{E}[\ell_\theta(X,Y)]$ and $L^f(\theta) = \mathbb{E}[\ell_\theta(X, f(X))]$ have Hessians, and $H_{\theta^*} = \nabla^2 L(\theta^*) \succ 0$.

Using the same definition of ξ_i^η as in Eq. (12), let $L_{\theta,i}^\eta = \ell_\theta(X_i, f(X_i)) + (\ell_\theta(X_i, Y_i) - \ell_\theta(X_i, f(X_i)))\,\frac{\xi_i^\eta}{\pi_\eta(X_i)}$. We define $\nabla L_{\theta,i}^\eta$ analogously, replacing the losses with their gradients. Given a function g, let

$$\mathbb{G}_n[g(L_\theta^\eta)] := \frac{1}{\sqrt{n}}\sum_{i=1}^n \left(g(L_{\theta,i}^\eta) - \mathbb{E}[g(L_{\theta,i}^\eta)] \right); \quad \mathbb{E}_n[g(L_\theta^\eta)] := \frac{1}{n}\sum_{i=1}^n g(L_{\theta,i}^\eta).$$

We similarly use $\mathbb{G}_n[g(\nabla L^{\eta}_\theta)]$, $\mathbb{E}_n[g(\nabla L^{\eta}_\theta)]$, etc. Notice that $\mathbb{E}_n[L^{\hat\eta}_\theta] = L^{\pi_{\hat\eta}}(\theta)$.

By the differentiability and local Lipschitzness of the loss, for any $h_n = O_P(1)$ we have

$$\mathbb{G}_n[\sqrt{n}(L^{\eta^*}_{\theta^*+h_n/\sqrt{n}} - L^{\eta^*}_{\theta^*}) - h_n^\top \nabla L^{\eta^*}_{\theta^*}] \xrightarrow{p} 0.$$

By definition, this is equivalent to

$$n\mathbb{E}_n[L^{\eta^*}_{\theta^*+h_n/\sqrt{n}} - L^{\eta^*}_{\theta^*}] = n(L(\theta^* + h_n/\sqrt{n}) - L(\theta^*)) + h_n^\top \mathbb{G}_n[\nabla L^{\eta^*}_{\theta^*}] + o_P(1),$$

where $L(\theta) = \mathbb{E}[\ell_\theta(X,Y)]$ is the population loss. A second-order Taylor expansion now implies

$$n\mathbb{E}_n[L^{\eta^*}_{\theta^*+h_n/\sqrt{n}} - L^{\eta^*}_{\theta^*}] = \frac{1}{2}h_n^\top H_{\theta^*} h_n + h_n^\top \mathbb{G}_n[\nabla L^{\eta^*}_{\theta^*}] + o_P(1).$$

At the same time, since $\mathbb{P}(\hat\eta \neq \eta^*) \to 0$, we have

$$n\mathbb{E}_n[L^{\hat\eta}_{\theta^*+h_n/\sqrt{n}} - L^{\hat\eta}_{\theta^*}] = n\mathbb{E}_n[L^{\eta^*}_{\theta^*+h_n/\sqrt{n}} - L^{\eta^*}_{\theta^*}] + o_P(1).$$

Putting everything together, we have shown

$$n\mathbb{E}_n[L^{\hat\eta}_{\theta^*+h_n/\sqrt{n}} - L^{\hat\eta}_{\theta^*}] = \frac{1}{2}h_n^\top H_{\theta^*} h_n + h_n^\top \mathbb{G}_n[\nabla L^{\eta^*}_{\theta^*}] + o_P(1).$$

The rest of the proof is standard. We apply the previous display with $h_n = \hat h_n := \sqrt{n}(\hat\theta^{\hat\eta} - \theta^*)$ (which is $O_P(1)$ by the consistency of $\hat\theta^{\eta^*}$; see (Van der Vaart, 2000, Thm. 5.23)) and $h_n = \tilde h_n := -H_{\theta^*}^{-1}\mathbb{G}_n[\nabla L^{\eta^*}_{\theta^*}]$:

$$\begin{aligned} n\mathbb{E}_n[L^{\hat\eta}_{\hat\theta^{\hat\eta}} - L^{\hat\eta}_{\theta^*}] &= \frac{1}{2}\hat h_n^\top H_{\theta^*} \hat h_n + \hat h_n^\top \mathbb{G}_n[\nabla L^{\eta^*}_{\theta^*}] + o_P(1); \\ n\mathbb{E}_n[L^{\hat\eta}_{\theta^*+\tilde h_n/\sqrt{n}} - L^{\hat\eta}_{\theta^*}] &= \frac{1}{2}\tilde h_n^\top H_{\theta^*} \tilde h_n + \tilde h_n^\top \mathbb{G}_n[\nabla L^{\eta^*}_{\theta^*}] + o_P(1). \end{aligned}$$

By the definition of $\hat\theta^{\hat\eta}$, the left-hand side of the first equation is smaller than the left-hand side of the second equation. Therefore, the same must be true of the right-hand sides of the equations. If we take the difference between the equations and complete the square, we get

$$\frac{1}{2}\left(\sqrt{n}(\hat\theta^{\hat\eta} - \theta^*) - \tilde h_n\right)^\top H_{\theta^*}\left(\sqrt{n}(\hat\theta^{\hat\eta} - \theta^*) - \tilde h_n\right) + o_P(1) \leq 0.$$

Since the Hessian H_{θ^*} is positive-definite, it must be the case that $\sqrt{n}(\hat\theta^{\hat\eta} - \theta^*) - \tilde h_n \xrightarrow{p} 0$. By the central limit theorem, $\tilde h_n = -H_{\theta^*}^{-1}\mathbb{G}_n[\nabla L^{\eta^*}_{\theta^*}]$ converges to $\mathcal{N}(0, \Sigma_*)$ in distribution, where

$$\Sigma_* = H_{\theta^*}^{-1}\mathrm{Var}\left(\nabla\ell_{\theta^*}(X, f(X)) + (\nabla\ell_{\theta^*}(X,Y) - \nabla\ell_{\theta^*}(X, f(X)))\frac{\xi^{\eta^*}}{\pi_{\eta^*}(X)}\right)H_{\theta^*}^{-1}.$$

The final statement thus follows by Slutsky's theorem.

A.4. Proof of Proposition 6.1

We prove the result by an application of the martingale central limit theorem (see, for example, Theorem 8.2.4. in (Durrett, 2019)).

Let $\bar\Delta_t$ denote the increments Δ_t with their mean subtracted out, i.e. $\bar\Delta_t = \Delta_t - \theta^*$. To apply the theorem, we first need to verify that the increments $\bar\Delta_t = \Delta_t - \theta^*$ are martingale increments; this follows because

$$\mathbb{E}[\bar\Delta_t|\mathcal{F}_{t-1}] = \mathbb{E}[\bar\Delta_t|f_t, \pi_t] = \mathbb{E}[f_t(X_t)|f_t,\pi_t] + \mathbb{E}[Y_t - f_t(X_t)|f_t,\pi_t]\mathbb{E}\left[\frac{\xi_t}{\pi_t(X_t)}|f_t,\pi_t\right] - \theta^* = 0,$$

together with the fact that $\bar\Delta_t \in \mathcal{F}_t$.

The martingale central limit theorem is now applicable given two regularity conditions. The first is that $\frac{1}{n}\sum_{t=1}^n \sigma_t^2$ converges in probability, which holds by assumption. The second condition is the so-called Lindeberg condition, stated below.

Assumption A.3. We say that Δ_t satisfy the Lindeberg condition if

$$\frac{1}{n}\sum_{t=1}^{n}\mathbb{E}[\bar{\Delta}_t^2\mathbf{1}\{|\bar{\Delta}_t| > \epsilon\sqrt{n}\}|\mathcal{F}_{t-1}] \xrightarrow{p} 0$$

for all $\epsilon > 0$, where $\bar{\Delta}_t = \Delta_t - \theta^*$.

Since this condition holds by assumption, we can apply the central limit theorem to conclude $\sqrt{n}(\hat{\theta}^{\vec{\pi}} - \theta^*) = \frac{1}{\sqrt{n}}\sum_{t=1}^{n}\bar{\Delta}_t \xrightarrow{d} \mathcal{N}(0, \sigma_*^2)$.

A.5. Proof of Theorem 6.2

We follow a similar approach as in the proof of Theorem 5.2, which is in turn similar to the classical argument for M-estimation (Van der Vaart, 2000, Thm. 5.23). In this case, the main difference to the classical proof is that the empirical loss $L^{\vec{\pi}}(\theta)$ comprises martingale, rather than i.i.d. increments. We explain the differences relative to the proof of Theorem 5.2.

We define $L_{\theta,i} = \ell_\theta(X_i, f_i(X_i)) + (\ell_\theta(X_i, Y_i) - \ell_\theta(X_i, f_i(X_i)))\frac{\xi_i}{\pi_i(X_i)}$, and $\nabla L_{\theta,i}$ is defined analogously. We again use the notation $\mathbb{G}_n[g(L_\theta)], \mathbb{E}_n[g(L_\theta)], \mathbb{G}_n[g(\nabla L_\theta)], \mathbb{E}_n[g(\nabla L_\theta)]$, etc.

As in the classical argument, for any $h_n = O_P(1)$ we have $\mathbb{G}_n[\sqrt{n}(L_{\theta^*+h_n/\sqrt{n}} - L_{\theta^*}) - h_n^\top \nabla L_{\theta^*}] \xrightarrow{p} 0$. This can be concluded from the martingale central limit theorem, since the variance of the increments tends to zero. Specifically, define the triangular array $\mathcal{L}_{n,i} = \sqrt{n}(L_{\theta^*+h_n/\sqrt{n},i} - L_{\theta^*,i}) - h_n^\top \nabla L_{\theta^*,i}$, and let $\mathcal{V}_{n,i} = \mathrm{Var}(\mathcal{L}_{n,i}|\mathcal{F}_{i-1})$. We have $|\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\mathcal{L}_{n,i}| \leq \max_i \sqrt{\mathcal{V}_{n,i}}|\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\frac{\mathcal{L}_{n,i}}{\sqrt{\mathcal{V}_{n,i}}}|$. By the martingale central limit theorem, $\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\frac{\mathcal{L}_{n,i}}{\sqrt{\mathcal{V}_{n,i}}} \xrightarrow{d} \mathcal{N}(0,1)$ and, since $\max_i \sqrt{\mathcal{V}_{n,i}} \xrightarrow{p} 0$, we conclude by Slutsky's theorem that $\mathbb{G}_n[\sqrt{n}(L_{\theta^*+h_n/\sqrt{n}} - L_{\theta^*}) - h_n^\top \nabla L_{\theta^*}] \xrightarrow{p} 0$.

The following steps are the same as in the proof of Theorem 5.2; we conclude that $\sqrt{n}(\hat{\theta}^{\vec{\pi}} - \theta^*) - \tilde{h}_n \xrightarrow{p} 0$, where $\tilde{h}_n = -H_{\theta^*}^{-1}\mathbb{G}_n[\nabla L_{\theta^*}]$. Finally, we argue that $\tilde{h}_n$ converges to $\mathcal{N}(0, \Sigma_*)$ in distribution. To see this, first note that all one-dimensional projections $v^\top \tilde{h}_n$ converge to $v^\top Z$, $Z \sim \mathcal{N}(0, \Sigma_*)$, by the martingale central limit theorem, which is applicable because the Lindeberg condition holds by assumption (see below for statement) and the variance process $V_{\theta^*,n}$ converges to V_*. Once we have the convergence of all one-dimensional projections, convergence of $\tilde{h}_n$ follows by the Cramér-Wold theorem.

Assumption A.4. We say that the increments satisfy the Lindeberg condition if, for all $v \in \mathcal{S}^{d-1}$,

$$\frac{1}{n}\sum_{t=1}^{n}\mathbb{E}[(v^\top \nabla L_{\theta^*,t})^2\mathbf{1}\{|v^\top \nabla L_{\theta^*,t}| > \epsilon\sqrt{n}\}|\mathcal{F}_{t-1}] \xrightarrow{p} 0$$

for all $\epsilon > 0$.

B. Implementation Details

B.1. Practical Sampling Rules

As explained in Section 5 and Section 6, our sampling rule $\pi(x)$ is derived from a measure of uncertainty $u(x)$. As clear from Section 7, the right notion of uncertainty should measure a notion of error dependent on the estimation problem at hand. In particular, we hope to have $u(X) \approx |\left(\nabla\ell_{\theta^*}(X,Y) - \nabla\ell_{\theta^*}(X, f(X))\right)^\top h^{(j)}|$. For GLMs and means, in light of Eq. (10) and Eq. (11), this often boils down to training a predictor of $|f(X) - Y|$ from X and, in the case of GLMs, using a plug-in estimate of the Hessian. This is what we do in our experiments (except in the case of binary classification where we simply use the uncertainty from Eq. (3)).

Of course, the learned predictor of model errors cannot be perfect; as a result, $\pi(x) \propto u(x)$ cannot naively be treated as the oracle rule π_{opt}. For example, the model might mistakenly estimate (near-)zero uncertainty ($u(X) \approx 0$) when $|f(X) - Y|$ is large, which would blow up the estimator variance. To fix this issue, we find that it helps to stabilize the rule $\pi(x) \propto u(x)$ by mixing it with a uniform rule.

Denote the uniform rule by $\pi^{\mathrm{unif}}(x) = n_b/n$. Clearly the uniform rule meets the budget constraint, since $n\,\mathbb{E}[\pi^{\mathrm{unif}}(X)] =$

Algorithm 1 Batch active inference

Input: unlabeled data $X_1, \dots, X_n$, sampling budget n_b, predictive model f, error level $\alpha \in (0,1)$

Choose uncertainty measure $u(x)$ based on f
Let $\pi(x) = \hat{\eta}\, u(x)$, where $\hat{\eta} = \frac{n_b}{n\hat{\mathbb{E}}[u(X)]}$; let $\pi^{\text{unif}} = \frac{n_b}{n}$
Select $\tau \in (0,1)$ and choose sampling rule $\pi^{(\tau)}(x) = (1-\tau)\cdot\pi(x) + \tau \cdot \pi^{\text{unif}}$
Sample labeling decisions $\xi_i \sim \text{Bern}(\pi^{(\tau)}(X_i)), i \in [n]$
Collect labels $\{Y_i : \xi_i = 1\}$
Compute batch active estimator $\hat{\theta}^{\pi^{(\tau)}}$ (Eq. (6))

Algorithm 2 Sequential active inference

Input: unlabeled data $X_1, \dots, X_n$, sampling budget n_b, initial predictive model f_1, error level $\alpha \in (0,1)$, fine-tuning batch size B

```
Set D^tune ← ∅
for t = 1, ..., n do
  Choose uncertainty measure u_t(x) for f_t
  Set π_t(x) as in Eq. (8) with η_t = n_b / (n Ê[u_t(X)]); let π^unif = n_b / n
  Select τ ∈ (0,1) and choose sampling rule π_t^(τ)(x) = (1 − τ) · π_t(x) + τ · π^unif
  Sample labeling decision ξ_t ~ Bern(π_t^(τ)(X_t))
  if ξ_t = 1 then
    Collect label Y_t
    D^tune ← D^tune ∪ {(X_t, Y_t)}
    if |D^tune| = B then
      Fine-tune model on D^tune: f_{t+1} = finetune(f_t, D^tune)
      Set D^tune ← ∅
    else
      f_{t+1} ← f_t
    end if
  else
    f_{t+1} ← f_t
  end if
end for
Compute sequential active estimator θ̂^{π⃗^(τ)} (Eq. (7))
```

n_b. For a fixed $\tau \in [0,1]$ and $\pi(x) \propto u(x)$, we define the τ-mixed rule as

$$\pi^{(\tau)}(x) = (1-\tau)\cdot\pi(x) + \tau\cdot\pi^{\text{unif}}(x).$$

Any positive value of τ ensures that $\pi^{(\tau)}(x) > 0$ for all x, avoiding instability due to small uncertainty estimates $u(x)$. When historical data is available, one can tune τ by optimizing the empirical estimate of the (asymptotic) variance of $\hat{\theta}^{\pi^{(\tau)}}$ given by Theorem 5.2. For example, in the case of mean estimation, this would correspond to solving:

$$\hat{\tau} = \arg\min_{\tau\in[0,1]} \sum_{i=1}^{n_h} \frac{1}{\pi^{(\tau)}(X_i^h)}(Y_i^h - f(X_i^h))^2, \tag{14}$$

where $(X_i^h, Y_i^h), \dots, (X_{n_n}^h, Y_{n_h}^h)$ are the historical data points. Otherwise, one can set τ to be any user-specified constant. In our experiments, in the batch setting we tune τ on historical data when such data is available. In the sequential setting we simply set $\tau = 0.5$ as the default.

B.2. Experimental Details

The batch and sequential active inference methods used in our experiments are outlined in Algorithm 1 and Algorithm 2.

In Figure 3 and Figure 4 we report the percentage of budget saved by active inference relative to the baselines when the

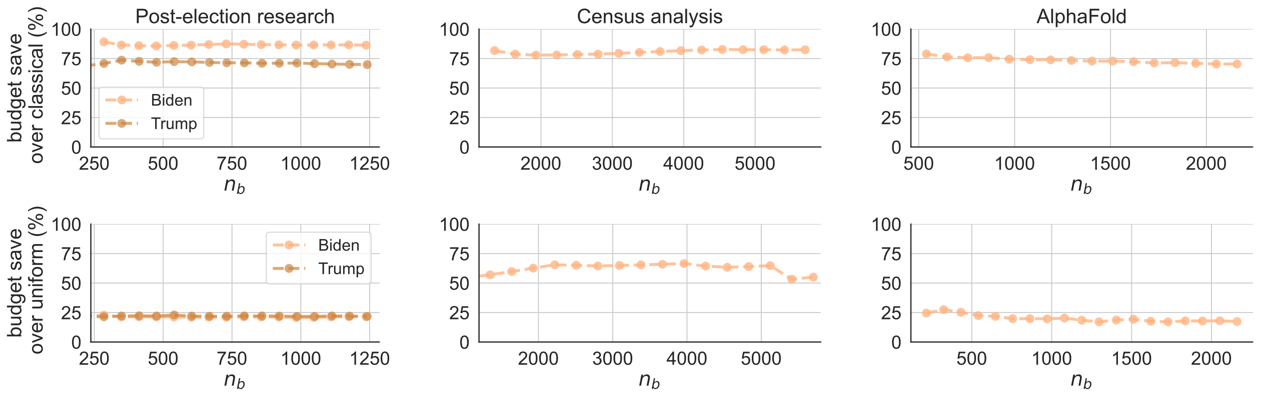

Figure 3. **Save in sample budget due to active inference.** Reduction in sample size required to achieve the same confidence interval width with active inference and (top) classical inference and (bottom) uniform sampling, respectively, across the applications shown in Figure 1.

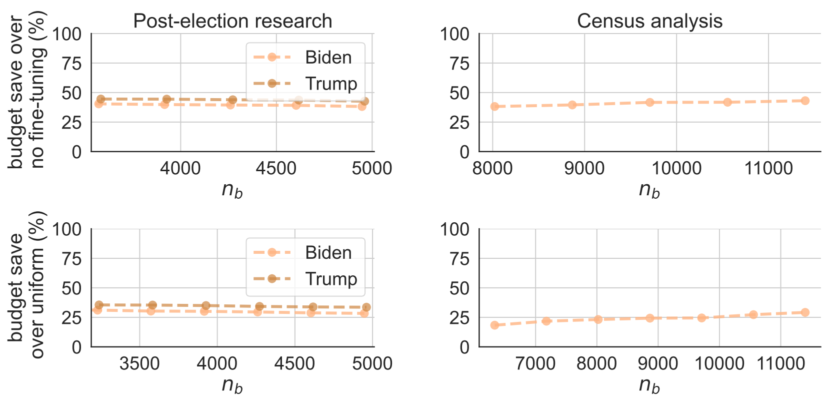

Figure 4. **Save in sample size budget due to fine-tuning.** Reduction in sample size required to achieve the same confidence interval width with active inference with fine-tuning and (top) active inference with no fine-tuning and (bottom) the uniform baseline (PPI), respectively, in the applications shown in Figure 2.

methods are matched to be equally accurate. More precisely, for varying n_b we compute the average interval width achieved by the uniform and classical baselines; then, we look for the budget size n_b^{active} for which active inference achieves the same average interval width, and report $(n_b - n_b^{\text{active}})/n_b \cdot 100\%$ as the percentage of budget saved.

In all our experiments, we have a labeled dataset of n examples. We treat the solution on the full dataset as the ground-truth θ^* for the purpose of evaluating coverage. In each trial, the underlying data points (X_i, Y_i) are fixed and the randomness comes from the labeling decisions ξ_i. In the sequential experiments, we additionally randomly permute the data points at the beginning of each trial. The experiments in the batch setting average the results over 1000 trials and the experiments in the sequential setting average the results over 100 trials. The Pew dataset is available at (Pew, 2020); the census dataset is available through Folktables (Ding et al., 2021); the Alphafold dataset is available at (Angelopoulos et al., 2023b).

As discussed in Section B.1, to avoid values of $\pi(x)$ that are very close to zero we mix the "standard" sampling rule based on the reported measure on uncertainty $u(x)$ with a uniform rule $\pi^{\text{unif}} = \frac{n_b}{n}$ according to a parameter $\tau \in (0, 1)$. In post-election survey research, we have training data for the prediction model and we use the same data to select τ so as to minimize an empirical approximation of the variance $\text{Var}(\hat{\theta}^{\pi^{(\tau)}})$, as in Eq. (14). In the AlphaFold example and both problems with model fine-tuning we set $\tau = 0.5$ for simplicity. In the census example, the trained predictor of model error $e(x)$ rarely gives very small values, and so we set $\tau = 0.001$.

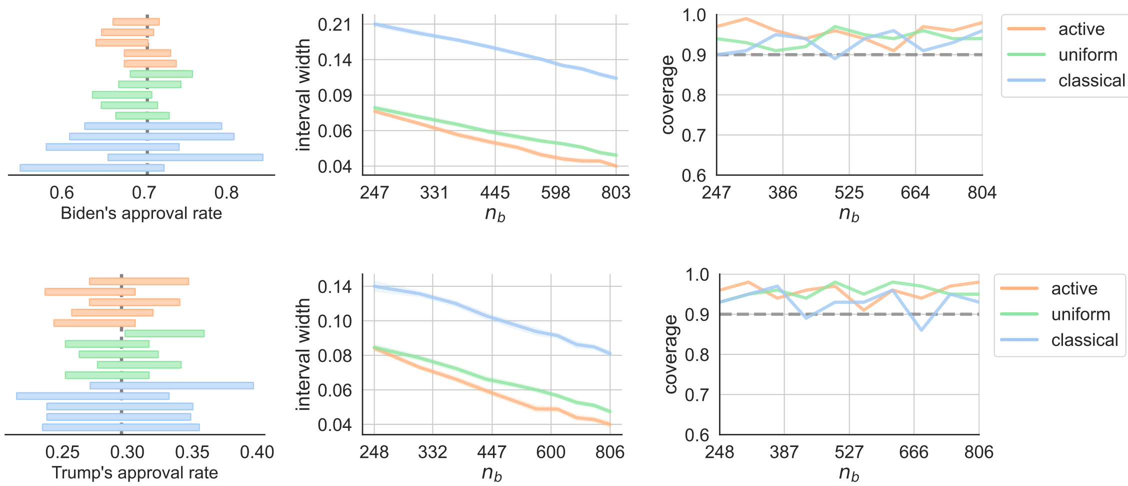

Figure 5. **Non-asymptotic experiments.** Example intervals in five randomly chosen trials (left), average confidence interval width (middle), and coverage (right) in post-election survey research with non-asymptotic confidence intervals.

In each experiment, we vary n_b over a grid of uniformly spaced values. We take 20 grid values for the batch experiments and 10 grid values for the sequential experiments. The plots of interval width and coverage linearly interpolate between the respective values obtained at the grid points. There linearly interpolated values are used to produce the plots of budget save: for all values of n_b from the grid, we look for n_b' such that the (linearly interpolated) width of active inference at sample size n_b' matches the interval width of classical (resp. uniform) inference at sample size n_b. For the leftmost plot in Figure 1 and Figure 2, we uniformly sample five trials for a fixed n_b and show the intervals for all methods in those same five trials. We arbitrarily select n_b to be the fourth largest value in the grid of budget sizes for all experiments.

C. Non-Asymptotic Results

While our main results focus on asymptotic confidence intervals based on the central limit theorem, some of our results—in particular, those for mean estimation—have direct non-asymptotic and time-uniform analogues.

We explain this extension for the sequential algorithm, as it subsumes the extension for the batch setting. Let $\Delta_t = f_t(X_t) + (Y_t - f_t(X_t))\frac{\xi_t}{\pi_t(X_t)}$. As explained in Section 6, Δ_t have a common conditional mean: $\mathbb{E}[\Delta_t|\Delta_1, \dots, \Delta_{t-1}] = \theta^*$. Moreover, if Y_t and $f_t(X_t)$ are almost surely bounded, and $\pi_t(X_t)$ is almost surely bounded from below, then Δ_t are bounded as well. (Given that we construct π_t by "τ-mixing" it with a uniform rule, as explained in Section B.1, in our applications $\pi_t(X_t)$ is always bounded from below since $\pi_t(x) \geq \tau\frac{n_b}{n}$.) Therefore, given that we have bounded observations (with a known bound) having a common conditional mean, we can apply the recent betting-based methods (Waudby-Smith & Ramdas, 2024; Orabona & Jun, 2023) for constructing non-asymptotic confidence intervals and time-uniform confidence sequences satisfying $\mathbb{P}(\theta^* \in C_t, \forall t) \geq 1 - \alpha$.

We demonstrate the non-asymptotic extension in the problem of post-election survey analysis from Section 8. Figure 5 provides a non-asymptotic analogue of the corresponding batch results from Figure 1, applying the method from Theorem 3 of Waudby-Smith & Ramdas (2024) to form a non-asymptotic confidence interval. Qualitatively we observe a similar comparison as before—active inference outperforms both uniform sampling and classical inference—though the methods naturally overcover as a result of using non-asymptotic intervals that do not have exact coverage.