Curvature-Aware Active Statistical Inference : Reducing Labeling via Data Coherence

Pinaki Mohanty[1], Rajiv Khanna[1]
[1]Department of Computer Science, College of Science & College of Engineering,
Purdue University, West Lafayette, IN, USA
{pmohanty, rajivak}@purdue.edu

Abstract

Inspired by Active Learning, Active Statistical Inference (ASI) is an inference framework that leverages machine learning predictions to guide data-label acquisition, efficiently utilizing the labeling budget. However, relying solely on model-output uncertainty can lead to labeling redundant instances with diminishing informational returns. To address this, we propose Curvature-Aware Active Statistical Inference (CA-ASI), which prioritizes points of high model-output uncertainty while penalizing redundant points based on their *structural* similarity to the inference target. The structural similarity itself is evaluated by incorporating second-order information into the sampling rule, ensuring diverse and informative points are selected to be labeled. Further, we show that CA-ASI constructs provably valid confidence intervals and hypothesis tests for any black-box model. Under the same budget, CA-ASI enables smaller confidence intervals and more powerful statistical tests than ASI. We demonstrate CA-ASI's effectiveness on real-world datasets across standard baselines.

1 Introduction

In modern data-driven science, high-quality labeled data is often the key bottleneck, especially in settings with limited labeling budgets. Annotating each instance requires substantial human effort, expert knowledge, or time. To mitigate this, machine learning (ML) model-generated predictions on unlabeled data are used in various domains, such as remote sensing [31, 74, 60], proteomics [32], and electoral systems [76]. These applications reflect a growing dependence on predictive models as substitutes for manual annotation. However, simulated predictions can only assist with scalability and cannot replace the statistical guarantees of labeled data.

To reconcile ML efficiency with statistical inference rigor, Zrnić and Candès [80] propose Active Statistical Inference (ASI). ASI strategically guides data labeling based on model-output uncertainty, enabling statistically valid confidence intervals and hypothesis tests for any black-box model and data distribution, even under tight labeling budgets. Extensions incorporate weak supervision from LLMs, using model-generated annotations to reduce human burden while maintaining inference guarantees [25].

While ASI offers substantial gains over uniform labeling, its acquisition rule is inherently *uncertainty-only*: labels are spent where the predictive model is most uncertain, irrespective of whether those points contribute *new* information for the inferential target. This becomes a bottleneck in modern, high-dimensional datasets that can exhibit strong local structure by correlations, data-generation artifacts, and representation learning effects [8, 35, 70]. In such regimes, predictive uncertainty often concentrates on dense, near-duplicate regions of the pool, so uncertainty sampling repeatedly queries

Accepted at ICML 2026 Workshop on Decision-Making from Offline Datasets to Online Adaptation: Black-Box Optimization to Reinforcement Learning

points that are different in input space yet effectively *redundant* for inference: they induce similar gradients/curvatures and hence yield diminishing returns per label.

This redundancy is not merely an efficiency nuisance; it directly limits statistical precision. For M-estimation, the asymptotic covariance of the parameter estimator is governed by second-order information - through the (generalized) Fisher/Hessian geometry of the loss [71]. Intuitively, repeatedly labeling points whose local curvature aligns with what is already well-covered by previously labeled data does not expand the estimator's effective information in new directions, and can therefore waste budget without shrinking confidence intervals commensurately. Moreover, repeatedly querying "more of the same" can degrade human-in-the-loop engagement and annotation quality [64].

A natural response in the *active learning* literature is to enforce diversity—e.g., via Bayesian batch acquisition [35], core-set style representativeness [62], or gradient-space diversity heuristics [6]. However, these mechanisms are designed to improve predictive performance, not to minimize inferential variance, and they do not come with the central guarantee that makes ASI compelling: *valid* confidence intervals and hypothesis tests under adaptive data collection. This raises a fundamental question:

Can we endow active statistical inference with a notion of geometric non-redundancy—aligned with Fisher/Hessian information—while preserving ASI's distribution-free inferential validity?

In this paper we answer in the affirmative by introducing *Curvature-Aware Active Statistical Inference (CA-ASI)*, a strict generalization of ASI that couples uncertainty with a curvature-based redundancy penalty. The key idea is to treat the statistical problem's local curvature as the right proxy for "information contribution," and to explicitly downweigh points whose curvature is highly overlapping with the rest of the pool. Concretely, CA-ASI augments uncertainty with an instance-level redundancy score (derived from curvature overlap), and samples according to a *geometry-pruned utility*

$$\tilde{u}_\beta(x) = \tilde{u}(x) - \beta\,\tilde{\rho}(x),$$

where $\tilde{u}(x)$ is the normalized ASI uncertainty and $\tilde{\rho}(x)$ quantifies curvature redundancy. This simple modification has two decisive consequences: (i) it preserves the inferential backbone of ASI—yielding valid confidence intervals and tests under the same black-box flexibility—and (ii) it enables a principled efficiency theory: we provide an interpretable sufficient condition under which turning on the curvature penalty provably reduces the asymptotic covariance (in trace) relative to ASI, formalizing when and why redundancy-pruning improves statistical precision.

We summarize our contributions as follows:

- **Curvature-aware acquisition for active inference.** We introduce *Curvature-Aware Active Statistical Inference (CA-ASI)*, a strict generalization of ASI that augments uncertainty-based acquisition with an *instance-level curvature-redundancy penalty* based on Hessian overlap. This yields a simple, actionable sampling rule that explicitly avoids spending labels on geometrically redundant points, in both *batch* and *sequential* settings.
- **Distribution-free inferential validity under geometry-aware sampling.** We show that CA-ASI preserves the inferential backbone of ASI: under standard smoothness conditions for M-estimation [71], our curvature-aware active estimator is asymptotically normal and yields *valid* confidence intervals and hypothesis tests despite adaptive, non-uniform label acquisition.
- **A provable efficiency theorem for redundancy-pruned active inference.** Beyond validity, we provide a principled efficiency theory: we derive an interpretable *variance-spotlight* condition under which *turning on* the curvature penalty (i.e. $\beta > 0$) provably *reduces* the asymptotic covariance trace relative to uncertainty-only ASI. This formalizes when and why geometric non-redundancy translates into strictly tighter inference, rather than serving as a heuristic diversity add-on.
- **Empirical gains in precision, power, and label utilization.** On real datasets in both batch and sequential regimes, CA-ASI consistently delivers *narrower* confidence intervals at *matched* coverage, higher label diversity (lower redundancy), and often *fewer* labels used than the allocated budget—demonstrating that curvature-aware redundancy pruning yields practical, measurable gains in statistical precision under realistic labeling constraints. CA-ASI reduces label usage by up to 45% while simultaneously shrinking confidence interval widths upto by 25.4%, all while maintaining valid coverage-against baselines on real-world datasets.

2 Related Works

ML-enabled Data Annotation and Inference. A growing literature exists on inference from adaptively collected data, including [33, 77, 16], and adaptive experimental design, such as [27, 13]. The work most closely related to our pursuits is ASI as it positions itself uniquely by adaptively leveraging the uncertainties in modern, black-box ML for data collection, all while adhering to budget constraints for statistical inference. ASI shares similarities with prediction-powered inference (PPI) and other recent works on inference with ML predictions [4, 5]. More recently, Puheng et al. [56] introduce robust sampling to ensure active inference performance never drops below uniform sampling when model uncertainty is misspecified. While their work focuses on safety (mitigating bad uncertainty estimates), our method (CA-ASI) focuses on informative data-collection (mitigating data redundancy), making these approaches complementary.

Data Coherence. Data coherence, examined through loss landscape geometry, reveals how individual data points shape learning and model behavior. Dexter et al. [17], Chang and Khanna [14] introduce a novel coherence measure to characterize point-wise loss Hessian alignment and understand Stochastic Gradient Descent stability. Further, Agiollo et al. [2] explores how loss function curvature around samples indicates memorization, enabling data coherence for tasks like dataset pruning and summarization.

3 Preliminaries

Data Coherence for Active Learning. Importance of diversity in labeled data to maximize informational gain and avoid costly repetition is given key consideration [35]. To gauge data quality, for a labeled dataset $\mathcal{X}_{\text{label}} \subset \mathbb{R}^d$, we define label density, $\nu(.)$, with radius $r > 0$ as

$$\nu(r) \;=\; \frac{1}{|\mathcal{X}_{\text{label}}|} \sum_{i=1}^{|\mathcal{X}_{\text{label}}|} \Big| \big\{ x_j \in \mathcal{X}_{\text{label}} \setminus \{x_i\} : \; \|x_j - x_i\| \le r \big\} \Big| \tag{1}$$

$\nu(r)$ captures on average for each labeled point x_i, how many other labeled points x_j fall within a radius r. A higher value means more redundancy (less diversity in labeled data), whereas a lower value means labels are more spread out, which is desirable from an experimental design and expenditure perspective [23, 41]. Uncertainty sampling can over-query near-duplicate points from dense regions of the pool, yielding diminishing returns [63]. While active learning has long incorporated diversity via feature- or gradient-space criteria [62, 6], these objectives are primarily designed to improve predictive accuracy.

In contrast, our goal is inference efficiency: for M-estimation, statistical precision is governed by the Fisher/Hessian geometry of the loss [71], motivating a redundancy notion defined directly in curvature space. Inspired by Dexter et al. [17], we construct a coherence matrix $\mathbf{S} \in \mathbb{R}^{n \times n}$ where each entry $S_{ij} = \|\hat{H}_i^{1/2} \hat{H}_j^{1/2}\|_F$ measures the curvature overlap between data points i and j, where $\hat{H}_i = \nabla^2_\theta \ell_\theta(x_i, y_i)$. While they used $\mathbf{S}$ to derive a single, global diagnostic for a dataset, a key challenge lies in transforming this concept into an actionable, instance-level score. For this, we define for each row i, the sum of S_{ij} over all $j \neq i$ (across the $n-1$ other points). We then divide each row sum by the total sum across all rows, resulting in scores in $[0, 1]$. This yields the coherence score:

$$\tilde{\rho}_i \;=\; \frac{\sum_{j \neq i} S_{ij}}{\sum_{k=1}^{n} \sum_{j \neq k} S_{kj}}, \tag{2}$$

A large $\tilde{\rho}_i$ indicates that the curvature information of point x_i is highly redundant with many other points in the pool, while a small score indicates it provides more complementary geometric information.

Importantly, the curvature-redundancy score $\tilde{\rho}(x)$ is often *label-free* (or nearly so) for common statistical models: for squared-loss linear regression, the curvature overlap reduces to $S_{ij} = \|\hat{H}_i^{1/2} \hat{H}_j^{1/2}\|_F = |x_i^\top x_j|$ (Appendix D.3). More broadly, for many GLM negative log-likelihoods (e.g., logistic regression), the Hessian depends on (x, θ) through the predicted mean/variance and not on the realized label, so $\tilde{\rho}$ can be computed from unlabeled covariates once a provisional θ is fixed. When the Hessian *does* depend on y, we adopt the standard pseudo-labeling strategy [38, 69], using

$\hat{y}_i = f(x_i)$ to form a proxy curvature $\hat{H}_i = \nabla^2_\theta \ell_\theta(x_i, \hat{y}_i)$. Like ASI [80], the effectiveness of this approach depends on the assumption that $f(x)$ provides well-calibrated and generalizable predictions for y.

Problem Setup. We observe unlabeled samples $X_1, \ldots, X_n \sim \mathbb{P}_X$, with corresponding labels Y_i initially unobserved, and aim to estimate a population parameter θ^* defined as the minimizer of an expected loss:

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

for a convex loss function ℓ_θ. Here, $(X, Y) \sim \mathbb{P} = \mathbb{P}_X \times \mathbb{P}_{Y|X}$ denotes a generic feature-label pair drawn from the underlying data distribution. This formulation captures a broad class of M-estimation problems, including mean estimation, quantiles, and regression coefficients.

Similar to ASI, our challenge is to perform inference under a strict labeling budget ($n_b \ll n$). To overcome this, we use an off-the-shelf model as a strategic guide to select the most informative points to label. We denote the number of empirically collected labels as n_{lab}. We consider this problem in both batch and sequential settings.

Curvature-Aware Active Estimator. We assume a predictive model $f(X)$ for labels $Y \in \mathbb{R}$ given covariates $X \in \mathcal{X}$. The idea behind our curvature-aware active inference strategy is to increase the effective sample size by focusing the labeling budget on *structurally informative* points where the model is uncertain . To implement this idea, we employ a *sampling rule*, denoted generally as $\pi : \mathcal{X} \to [0, 1]$, and collect label Y_i with probability $\xi_i \sim \text{Bernoulli}(\pi(X_i))$. The intuition is to model $\pi(.)$ to have a high value for instances where f is uncertain and provides novel curvature information, crucial for shaping the loss landscape. Similarly, $\pi(.)$ will be low for redundant instances where the model f should be very certain. Concretely, we propose a specific family of sampling rules, denoted as π_β, derived from a measure of model-output uncertainty $u(x)$ and data coherence $\rho(x)$. We first standardize uncertainty via $\tilde{u}(x) = \frac{u(x)}{\sum_x u(x)} \in [0, 1]$. We choose $\pi(x) \propto \tilde{u}_\beta(x)$, where $\tilde{u}_\beta(x) = \tilde{u}(x) - \beta \cdot \tilde{\rho}(x)$ for some $\beta > 0$ (see Eq. (2) for definition of $\tilde{\rho}$). In other words, β acts as a regularizer that counterbalances uncertainty with redundancy, with the goal of reducing expected label usage $\mathbb{E}[n_{\text{lab}}]$. Let $\xi_i \sim \text{Bernoulli}(\pi_\beta(X_i))$ indicate whether point i is labeled, so that $n_{\text{lab}} = \sum_{i=1}^n \xi_i$. To stay under a budget n_b, we define our specific policy π_β as: $\pi_\beta(x) = \frac{\tilde{u}_\beta(x)}{\mathbb{E}[\tilde{u}_\beta(X)]} \cdot \frac{n_b}{n}$, which yields $\mathbb{E}[n_{\text{lab}}] = \mathbb{E}[\pi_\beta(X)] \cdot n \leq n_b$. Thus, our *curvature-aware active estimator* is defined as $\hat{\theta}^{\pi_\beta} = \arg\min_\theta L^{\pi_\beta}(\theta)$, where our *curvature-aware active empirical risk* is as follows,

$$L^{\pi_\beta}(\theta) := \frac{1}{n} \sum_{i=1}^n \left[\ell^f_{\theta,i} + \left(\ell_{\theta,i} - \ell^f_{\theta,i} \right) \frac{\xi_i}{\pi_\beta(X_i)} \right] \tag{3}$$

where $L(\theta) = \mathbb{E}[\ell_\theta(X, Y)]$, $\ell_{\theta,i} = \ell_\theta(X_i, Y_i)$, and $\ell^f_{\theta,i} = \ell_\theta(X_i, f(X_i))$. If the risk in Eq. 3, did not leverage f, our method would reduce to the *classical* baseline; we simply label any arbitrarily chosen n_b points such that, $\hat{\theta}^{\texttt{noML}} := \arg\min_\theta \frac{1}{n_b} \sum_{i=1}^{n_b} \ell_{\theta,i}$.

Distinction vs ASI/PPI. It is evident that our estimator, akin to the *active estimator* (Equation 1 from Zrnić and Candès [80]), exemplifies augmented inverse propensity weighting (AIPW) estimation [59]. The key distinction lies in using π_β versus π, when $\beta = 0$ our estimator recovers the ASI. In other words, the use of π_β invokes *curvature awareness* in our estimator. Further, when the sampling rule is uniform, i.e., $\pi_\beta(x) = \frac{n_b}{n}$ for all x, the estimator imitates the prediction-powered estimator [4]. Replacing π by π_β introduces new algorithmic and analytical challenges absent in ASI. First, $\tilde{\rho}(x)$ is defined through *pool-level* curvature interactions (Hessian overlap), so the sampling rule is no longer a pointwise function of uncertainty alone; it depends on global geometric structure in the unlabeled pool. Second, in many losses the Hessian may depend on the unknown label, requiring a proxy curvature (via pseudo-labels) and careful normalization to keep π_β feasible under a strict budget. Third, the practical policy involves clipping $[\tilde{u}(x) - \beta\tilde{\rho}(x)]_+$ and a data-dependent normalizer, making π_β *non-smooth* and *strongly data-adaptive*. Establishing that the resulting risk is unbiased ($\mathbb{E}[L^{\pi_\beta}(\theta)] = L(\theta)$) and AIPW estimator remains consistent and asymptotically normal requires showing that these empirical policy parameters stabilize, and that the influence-function/martingale arguments underlying ASI continue to hold under geometry-aware, pool-dependent sampling.

Variance mechanism and what β buys beyond ASI. As in ASI, the policy-dependent contribution to variance is $\frac{1}{n}\mathbb{E}\left[\delta_{(f,\theta)}(X)\left(\pi_\beta(X)^{-1} - 1\right)\right]$ which can be expressed via the model-induced loss

error $\delta_{(f,\theta)}(X) := \mathbb{E}\Big[\big(\ell_\theta(X,Y) - \ell_\theta(X, f(X))\big)^2 \mid X\Big] \geq 0$. Uncertainty-based sampling already reduces this term relative to uniform baselines by assigning larger π where the model is unreliable (large δ). CA-ASI adds a second control knob: it further downweighs *redundant* points whose curvature contribution is highly overlapping (large $\tilde{\rho}$), reallocating label probability from correlated, diminishing-return regions toward geometrically complementary ones. Our analysis formalizes when this reallocation yields a strict statistical gain: under an interpretable variance-spotlight condition, there exists $\beta > 0$ for which the asymptotic covariance trace is provably smaller than that of uncertainty-only ASI (see Theorems 5.4, 5.6).

4 Main Algorithm

Our algorithm constructions closely resemble those presented in Sections 5 and 6 in [80]. However, our primary challenge lies in incorporating *curvature-awareness* into the sampling procedure.

Batch Setting. We consider the batch setting with i.i.d. unlabeled covariates $X_{1:n}$, observed at once. Recall the redundancy–pruned uncertainty score $\tilde{u}_\beta(x) = \tilde{u}(x) - \beta\,\tilde{\rho}(x) \in [-\beta, 1]$, where $\tilde{u}$ and $\tilde{\rho}$ are the normalized uncertainty and coherence scores, and $\beta > 0$ controls pruning. To avoid negative scores, we clip to $[0,1]$. $\bar{u}_\beta(x) := [\tilde{u}_\beta(x)]_{[0,1]} = \min\{\max\{\tilde{u}_\beta(x), 0\}, 1\}$. In practice, we compute $\mathbb{E}[\tilde{u}_\beta(X)]$ as $\hat{\mu}_\beta := \max\Big\{\frac{1}{n}\sum_{j=1}^n \bar{u}_\beta(X_j),\, \varepsilon\Big\}$ for some $\epsilon > 0$.

We consider a budget–scaled family of sampling rules $\pi_\gamma(x) = \eta\,\bar{u}_\beta(x), \eta \in \mathcal{H} \subseteq \mathbb{R}^+, \gamma = (\eta, \beta)$; η is data-adaptive scaling parameter chosen to match the label budget and β is the fixed pruning hyperparameter. We set $\hat{\eta} = \max\Big\{\eta \in \mathcal{H} : \eta\sum_{i=1}^n \bar{u}_\beta(X_i) \leq n_b\Big\}$, and deploy $\pi_{\hat{\gamma}}$, where $\hat{\gamma} = (\hat{\eta}, \beta)$ is the empirically determined parameter vector to ensure the expected number of collected labels does not exceed the budget. We emphasize $\hat{\theta}^\eta \equiv \hat{\theta}^{\pi_\gamma}$.

Optionally, to guarantee stability and some level of acceptance, i.e. to lower bound our acceptance probabilities, we rely on τ-mixing, i.e. essentially interpolation of our sampling rule and uniform sampling. Thus, $\pi_\beta^{(\tau)}(x) := \frac{n_b}{n}\left[(1-\tau)\,\frac{\bar{u}_\beta(x)}{\hat{\mu}_\beta} + \tau\right]$, where $\tau \in [0,1]$.

Sequential Setting. In the sequential setting, we observe a stream $(X_t, Y_t)_{t=1}^n$ and decide online whether to acquire Y_t at each step. At round t, we choose to acquire the label with probability $\pi_t(X_t)$, where π_t is built from the model-output's uncertainty at time $t-1$. Formally, let $\mathcal{F}_t = \sigma((X_1, Y_1\xi_1, \xi_1), \ldots, (X_t, Y_t\xi_t, \xi_t))$ and require *predictability*: $f_t,\, \pi_t \in \mathcal{F}_{t-1}$. The predictor f_t may be updated using all information observed up to time t. Let $n_{\mathrm{lab},t}$ denote the number of labels collected up to time t.

The batch empirical risk in Eq. (3) specializes to a sequential setting at time t to,

$$L_t^{\overrightarrow{\pi_\beta}}(\theta) = \frac{1}{t}\sum_{i=1}^t \left[\ell_\theta^{f_i}(X_i) + \big(\ell_\theta(X_i,Y_i) - \ell_\theta^{f_i}(X_i)\big)\frac{\xi_i}{\pi_{(\beta,i)}(X_i)}\right] \tag{4}$$

with $\xi_i \sim \mathrm{Bernoulli}(\pi_{(\beta,i)}(X_i))$. We note that unlike Zrnić and Candès [80], our sequential estimator is not time agnostic. Thus, our sequential estimator at t is $\hat{\theta}_t^{\overrightarrow{\pi}_\beta} = \arg\min_\theta L_t^{\overrightarrow{\pi_\beta}}(\theta)$. We define the per-timestep term: $\Delta_i(\theta) := \ell_\theta^{f_i}(X_i) + \Big(\ell_\theta(X_i,Y_i) - \ell_\theta^{f_i}(X_i)\Big)\cdot\frac{\xi_i}{\pi_{(\beta,i)}(X_i)}$. Then $\{\Delta_i(\theta) - L(\theta), \mathcal{F}_i\}_{i\geq 1}$ is a martingale difference sequence: $\mathbb{E}[\Delta_i(\theta) \mid \mathcal{F}_{i-1}] = L(\theta)$ and $\Delta_i(\theta) \in \mathcal{F}_i$(We assume a positivity floor for $\pi_{(\beta,i)}(X_i)$). Under this setting, let $\mathcal{L}_t = \{(X_j, Y_j) : \xi_j = 1,\, j < t\}$ denote the labeled set before round t, with size $n_{\mathrm{lab,t}} = |\mathcal{L}_t|$. For incoming candidate X_t, we compute an uncertainty score $\tilde{u}_t := u(X_t)$(using f_t). However, the main challenge in the sequential setting is pointwise incorporation of redundancy without access to the full pool. Redundancy is inactive at initialization and turns on only once at least two data points are labeled i.e. $n_{\mathrm{lab,t}} \geq 2$. The idea is to gauge, *"What is this candidate's proportional informativeness within the small world of data I labeled so far?"* [39, 12]. Based on this idea, we systematically compute data coherence for candidate datapoint using $n_{(\mathrm{lab,t})} + 1$ data points, using label proxies from f_t. Finally, we compute the modified local score with clipping, exactly as in batch: $\bar{u}_{\beta,t} = \big[\tilde{u}_t - \beta\,\tilde{\rho}_t\big]_{[0,1]}$ for some $\beta > 0$. So, early rounds behave like pure uncertainty (i.e. ASI like), while later rounds down-weight candidates that are redundant with the growing labeled set.

We also use the concept of *spreading out* an *imaginary* budget as a function of time. Let *imaginary* budget for the expected number of collected labels by step t, equal to $n_{(b,t)} = t\frac{n_b}{n}$. Let $n_{(\Delta,t)} = n_{(b,t)} - n_{\text{lab,t}}$ denote the remaining budget at step t. Thus, the final sampling probability becomes, $\pi_{\beta,t}(x) = \min\{\eta_t\, \bar{u}_{\beta,t}, n_{\Delta,t}\}_{[0,1]}$. In practice, we set the scaling factor $\eta_t = \frac{n_b}{n\,\mathbb{E}[\bar{u}_{\beta,t}]}$. This construction ensures we are on track to spend the entire budget. Both our algorithms are presented as 1, 2 in the Appendix. We provide runtime analysis and highlight tuning strategies for β in the Appendix D.5 and D.1.2 respectively.

5 Theoretical Analysis

Our analysis first focuses on the batch setting (Section 5), where all unlabeled data is available at once and model f is pre-trained and fixed during the label collection phase, but the principles extend naturally to the sequential setting (Section 5), where the model f is updated as new labels are collected. We begin by proving that our estimator, with its geometry-aware sampling policy, produces valid confidence intervals and is asymptotically normal. We then prove our main result: under a mild and interpretable condition, our CA-ASI estimator is guaranteed to be more statistically efficient than the standard ASI estimator, resulting in a smaller asymptotic covariance. Like ASI, our setup also requires standard, mild Smoothness and Lindeberg condition assumptions on the loss ℓ_θ, articulated as follows,

Assumption 5.1 (Smoothness)**.** Loss ℓ is smooth if:

1. $\ell_\theta(x, y)$ is differentiable at θ^* for all (x, y);
2. 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 $\mathbb{E}[C(X, Y)^2] < \infty, \mathbb{E}[C(X)^2] < \infty$;
3. $L(\theta) = \mathbb{E}[\ell_\theta(X, Y)]$ and $L^f(\theta) = \mathbb{E}[\ell_\theta(X, f(X))]$ have Hessians, and $H_{\theta^*} = \nabla^2 L(\theta^*) \succ 0$

Assumption 5.2 (Lindeberg Condition)**.** Increments satisfy the Lindeberg condition if, for all $v \in \mathcal{S}^{d-1}$ for all $\epsilon > 0$, $\frac{1}{n}\sum_{t=1}^{n} \mathbb{E}\left[(v^\top \nabla L_{\theta^*,t})^2 \mathbf{1}\{|v^\top \nabla L_{\theta^*,t}| > \epsilon\sqrt{n}\} \,\middle|\, \mathcal{F}_{t-1}\right] \xrightarrow{p} 0$. Here, $\mathcal{S}^{d-1}$ is the set of all vectors $v \in \mathbb{R}^d$ with unit Euclidean norm ($\|v\|_2 = 1$).

Batch Setting. Our key insight is that replacing the uncertainty-based sampling policy of ASI with our geometry-aware policy, $\pi_\beta(x) \propto u(x) - \beta\rho(x)$, does not alter the fundamental structure of the M-estimation problem.

Theorem 5.3. *Assume the loss is smooth(Ass. 5.1) and define the Hessian $H_{\theta^*} = \nabla^2\mathbb{E}[\ell_{\theta^*}(X, Y)]$. Suppose that there exists a deterministic limiting parameter $\gamma^* = (\eta^*, \beta)$ such that $\mathbb{P}(\hat{\gamma} \neq \gamma^*) \to 0$. Then, if $\hat{\theta}^{\gamma^*} \xrightarrow{p} \theta^*$, we have:*

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

$\Sigma_{\gamma^*} = H_{\theta^*}^{-1}\text{Var}\left(\nabla\ell_{\theta^*}^f + (\nabla\ell_{\theta^*} - \nabla\ell_{\theta^*}^f)\frac{\xi^{\gamma^*}}{\pi_{\gamma^*}(X)}\right)H_{\theta^*}^{-1}$ *and* $\xi^{\gamma^*} \sim \text{Bernoulli}(\pi_{\gamma^*}(\text{X}))$. *Consequently, for any consistent plug-in estimate $\hat{\Sigma} \xrightarrow{p} \Sigma_{\gamma^*}$, the confidence interval $\mathcal{C}_\alpha = (\hat{\theta}_j^{\hat{\gamma}} \pm z_{1-\alpha/2}\sqrt{\frac{\hat{\Sigma}_{jj}}{n}})$ is a valid $(1-\alpha)$-asymptotic confidence interval for θ_j^*:*

$$\lim_{n\to\infty} \mathbb{P}(\theta_j^* \in \mathcal{C}_\alpha) = 1 - \alpha.$$

The consistency condition on $\hat{\gamma}$ is easily satisfied if n_b/n converges to a limit(see formal proof in Appendix C.1), and $\hat{\Sigma}$ allows for a straightforward plug-in estimate by replacing all quantities with their empirical counterparts. The consistency condition on $\hat{\theta}^{\gamma^*}$ is a standard requirement for analyzing M-estimators [see 71, Ch. 5].

$\hat{\theta}^{\pi_\beta}$'s statistical efficiency is captured by its asymptotic covariance matrix. A sampling policy based on model-output uncertainty $u(x)$ reduces this covariance by focusing the labeling budget on informative points. However, a policy based on model-output uncertainty alone can be inefficient if it repeatedly samples from dense clusters of similar, uncertain points. Such points often yield highly correlated gradients, leading to an ill-conditioned covariance matrix V_β and consequently, high parameter variance.

Our curvature-aware modification provides its benefit by identifying points whose informational contribution is already well-represented by other points in the dataset. Through the judicious selection of β, CA-ASI strategically prunes these redundant points, forcing the algorithm to collect uncertain points with de-tangled gradients. This leads to a better-conditioned and "smaller" covariance matrix V_β, ensuring the collected information is not repetitive and controlling the parameter variance from inflating. To formalize this, we define scalar variance pressure as $w_H(x) := \mathrm{Tr}(H^{-1}W(x)H^{-1})$, where $H = \nabla^2\mathbb{E}[\ell_{\theta^*}(X,Y)]$and $W(x) := \mathbb{E}_{Y|X}[(\nabla\ell_{\theta^*}(X,Y) - \nabla\ell^f_{\theta^*}(X))(.)^T]$ is the oracle gradient residual covariance matrix. We use this to define the variance-spotlight measure,

$$\omega(x) := \frac{w_H(x)}{\pi_0^2(x)} \tag{5}$$

, where π_0 is the sampling rule under ASI. This measure helps quantify the weight given to points that contribute most to the ASI estimator's variance. We demonstrate that CA-ASI is guaranteed to be more efficient than ASI if on average across the batch the points that contribute most to the variance (Eq. 5) are less redundant than they are uncertain.

Theorem 5.4. *Assume the conditions for asymptotic normality of the active M-estimator hold(Theorem 5.3). Let Σ_0 and Σ_β be the asymptotic covariance matrices for the ASI and CA-ASI estimators, respectively.*[1] *If the following condition holds true,*

$$\mathbb{E}_\omega[\tilde{\rho}(X)] \leq \mathbb{E}_\omega[\tilde{u}(X)]$$

then there exists a $\beta_0 > 0$ such that for all $0 < \beta < \beta_0 : \mathrm{Tr}(\Sigma_\beta) \leq \mathrm{Tr}(\Sigma_0)$

Sequential Setting. The analysis of the sequential setting directly extends the principles established in the batch case. The fundamental structure of our estimator remains unchanged. The primary distinction arises from the adaptive nature of the model f_t, which introduces dependencies across observations in a non-i.i.d setting. Let $V_{\theta,t}(\beta) = V_\theta(f_t, \pi_{(\beta,t)}) = \mathrm{Var}\big(\nabla L_t(\theta) \mid f_t, \pi_{(\beta,t)}\big)$.

In ASI, validity of the Martingale CLT hinges on convergence of the conditional variance process i.e. when the predictive model f_t stabilizes, stabilizing sampling rule π_t. In our curvature-aware setting, the policy $\pi_{(\beta,t)}$ depends on both f_t and the streaming coherence score $\tilde{\rho}_t$. Thus, we make the same stability assumption regarding the data geometry. As the history of labeled points grows, the coherence matrix $\mathbf{S}_t$ provides an increasingly accurate approximation of the true population manifold, and the empirical redundancy scores $\tilde{\rho}_t(x)$ settle to their deterministic population values. This convergence, combined with the stabilization of f_t, ensures that the sampling policy $\pi_{\beta,t}$ converges to a fixed oracle rule, stabilizing the resulting variance process.

Theorem 5.5. *Assume the loss is smooth(Ass. 5.1) and define the Hessian $H_{\theta^*} = \nabla^2\mathbb{E}[\ell_{\theta^*}(X,Y)]$. Let $\hat{\theta}_n^{\overrightarrow{\pi_\beta}}$ be the final CA-ASI estimator after n steps, obtained using the rule $\pi_{(\beta,t)}$. Suppose also that $\frac{1}{n}\sum_{t=1}^n V_{\theta^*,t}(\beta) \xrightarrow{p} V^*_\beta = V_{\theta^*}(f_*, \pi_{\beta*})$ entry-wise for some fixed model-rule pair $(f_*, \pi_{\beta*})$ and that increments $\Delta_i(\theta)$ satisfy the Lindeberg condition(Ass. 5.2). Then, if the estimator is consistent, $\hat{\theta}_n^{\overrightarrow{\pi_\beta}} \xrightarrow{p} \theta^*$, we have:*

$$\sqrt{n}(\hat{\theta}_n^{\overrightarrow{\pi_\beta}} - \theta^*) \xrightarrow{d} N(0, \Sigma_\beta)$$

where $\Sigma_\beta = H^{-1}_{\theta^*} V^*_\beta H^{-1}_{\theta^*}$. Consequently, for any consistent plug-in estimate $\hat{\Sigma}_\beta \xrightarrow{p} \Sigma_\beta$, the confidence interval $\mathcal{C}_\alpha = (\hat{\theta}_{n_j}^{\overrightarrow{\pi_\beta}} \pm z_{1-\alpha/2}\sqrt{\frac{\hat{\Sigma}_{\beta,jj}}{n}})$ is a valid $(1-\alpha)$-asymptotic confidence interval for θ^*_j:

$$\lim_{n\to\infty} \mathbb{P}(\theta^*_j \in \mathcal{C}_\alpha) = 1 - \alpha.$$

At each step t, we define the conditional scalar variance pressure $w_{H,t}(x) := \mathrm{Tr}(H_t^{-1}W_t(x)H_t^{-1})$, where $W_t(x) = \mathbb{E}_{Y|X}[(\nabla\ell_{\theta^*_t} - \nabla\ell^f_{\theta^*_t})(\cdot)^T | X_t = x, \mathcal{F}_{t-1}]$ and $H_t = \nabla^2\mathbb{E}[\ell_{\theta^*}(X,Y)|X_t = x, \mathcal{F}_{t-1}]$ is the true Hessian conditional on the past. Naturally, Eq.5 can be extended to sequential setting, to define conditional sequential variance-spotlight measure,

$$\omega_t := \frac{w_{H,t}(x)}{\pi^2_{(0,t)}}$$

[1]For full rigor, the covariance depends on the full limiting parameter $\gamma^* = (\eta^*, \beta)$. Since β is the parameter of interest that distinguishes the methods, we denote the covariance Σ_β for clarity.

Theorem 5.6. *Assume the conditions for the Martingale Central Limit Theorem hold (Theorem 5.5). Let Σ_0 and Σ_β be the asymptotic covariance matrices for the sequential ASI and CA-ASI estimators, respectively. Then, if the following condition holds in expectation over time,*

$$\mathbb{E}_t[\mathbb{E}_{\omega_t}[\tilde{\rho}_t(X)|\mathcal{F}_{t-1}]] \leq \mathbb{E}_t[\mathbb{E}_{\omega_t}[\tilde{u}_t(X)|\mathcal{F}_{t-1}]]$$

then there exists a $\beta_0 > 0$ such that for all $0 < \beta < \beta_0 : \mathrm{Tr}(\Sigma_\beta) \leq \mathrm{Tr}(\Sigma_0)$

It is worth noting, our theoretical guarantees established in Theorems 5.3 and 5.5 can be construed as a general and unified framework for analyzing active inference methods. This strict generalization becomes clear when the redundancy penalty is set to zero($\beta = 0$). In this special case, our geometry-aware policies π_β (batch) and $\pi_{\beta,t}$ (sequential) reduce to their unpenalized, uncertainty-only counterparts used in ASI (Theorems 5.2 and 6.2 from Zrnić and Candès [80]). As a direct consequence, the asymptotic covariance matrices Σ_β derived in our theorems simplify to the corresponding covariance matrices for the standard ASI estimators. Thus, our analysis not only validates our new method but also formally subsumes the original ASI results, providing a single, coherent framework for exploring the entire spectrum of strategies from pure uncertainty-based sampling to fully geometry-aware inference.

Furthermore, while the ASI framework is empirically successful, it does not provide a formal theorem establishing the conditions under which its sampling strategy is guaranteed to reduce parameter variance. Our analysis addresses this critical gap directly. Theorems 5.4 and 5.6 establish a clear, interpretable condition under which our geometry-aware policy is provably more efficient, guaranteeing a smaller asymptotic covariance. We also show finite sample gains of CA-ASI under the sequential setting in the Appendix C.

6 Experiments

Motivational Synthetic Example . To gauge the impact of redundancy awareness on the sampling rule, we construct a controlled 1D logistic setting. We draw $x \sim \mathcal{U}_{[-6,6]}$ and generate labels from a ground-truth model $y \mid x \sim \mathrm{Bernoulli}(\sigma(\theta_1 \cdot x + \theta_0))$, where $(\theta_1, \theta_0) = (1, 0)$. A logistic regression model is fit on 5000 such samples, yielding an acquisition model $\hat{f}(x) = \sigma(\hat{\theta}_1 \cdot x + \hat{\theta}_0)$. For the test pool, we generate two Gaussian clusters, each of size 250 points, symmetrically placed around the decision boundary $x = 0$, namely $\mathcal{C}_1 : x \sim \mathcal{N}(1, 0.1^2)$ and $\mathcal{C}_2 : x \sim \mathcal{N}(-1, 0.05^2)$. Both clusters lie at comparable margin from the boundary (hence similar uncertainty under $\hat{f}$) (See Figure 1). For this setting, we minimize over logistic loss.

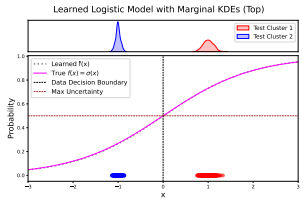

Figure 1: 2 Gaussian Clusters equidistant from $f(.)$

With a budget of 50 points, we visualize the labeling behavior of CA-ASI under curvature awareness. As shown in Figure 2, when $\beta = 0$, CA-ASI (acting similarly to ASI) is insensitive to redundancy and solely relies on uncertainty as its signal, resulting in roughly equal sampling from both clusters. However, as β increases, CA-ASI exhibits a preference for sampling from $\mathcal{C}_1$, which contains unique and non-redundant points. The proportion of samples from $\mathcal{C}_1$ gradually increases from approximately 45% to 67%. We provide additional experimental insights in the Appendix D.

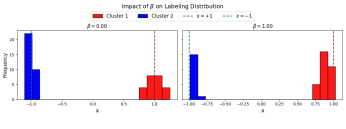

Figure 2: Targeted sampling due to curvature awareness

Real-world Datasets. The training details are provided in the Appendix D.2.1.

HTRU2: We use the HTRU2 dataset (17898 instances) from the UCI repository [42] for a binary classification task (pulsar(+1) vs. non-pulsar(-1)). In our setup, we optimize over exponential loss with the aim to determine the nature of the signal based on the characteristics of its profile and DM-SNR curve.

From Figure 3, Classical method produces narrow confidence intervals, however it suffers from catastrophic undercoverage, with empirical coverage below 40% at low budgets. This makes the CIs statistically invalid and misleadingly overconfident due to the severe class imbalance (10:1 non-pulsars to pulsars). In both settings (Figures 3 and 4), Uniform has the widest CIs, explaining its

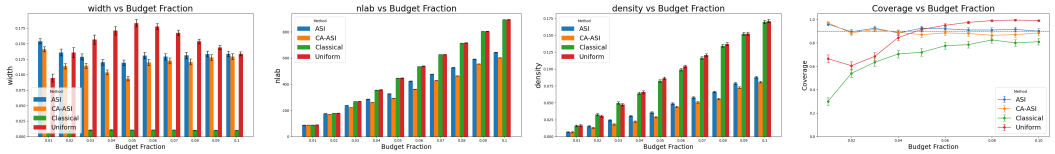

Figure 3: Batch Setting in HTRU2 Dataset

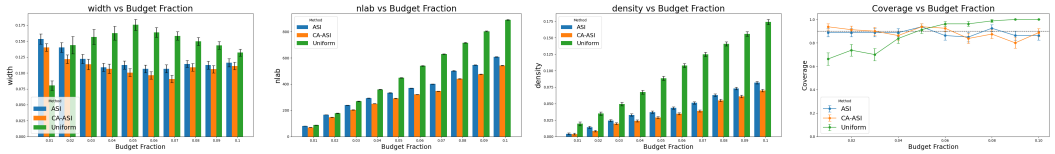

Figure 4: Sequential Setting in HTRU2 Dataset

overly optimistic coverage and making it unreliable. Both ASI and CA-ASI successfully maintain fidelity with respect to coverage across all budget fractions across both settings; CA-ASI consistently produces narrower confidence intervals, against other ML-assisted methods. Roughly, the average width decrease in batch settings is 10% (up to 21.7%) and in sequential settings is 9% (up to 15.1%). Most impressively, this superior precision is achieved while using significantly fewer and diverse labels. As shown in Figures 3 and 4, CA-ASI on average approximately uses 25% (up to 34.3%) and 32.4% (up to 44.4%) less labels in batch and sequential settings, respectively, compared to the Uniform baseline.

California Housing: For this setup, we use the California Housing dataset (20640 instances) [50] with the goal to predict the median house value for California districts(expressed in hundreds of thousands of dollars), using demography, location, and general information, while minimizing over squared loss.

From Figures 5, 6 we notice CA-ASI uses the least amount of labels, across all the methods, while maintaining the impressive coverage. It uses about 15.6% (up to 25.5%) less labels in batch settings and 24.1% (up to 45%) less labels in sequential settings compared to the Uniform baseline. Additionally, CA-ASI attains better label density and smaller CI Width compared to ASI; the average width decrease in batch setting is about 15.1% (up to 18.8%), while in sequential setting, it is about 16.6% (up to 25.4%). In the batch and sequential setting, while the Uniform method achieves similar or smaller CIs respectively, it suffers from undercoverage while utilizing entire budget, making them biased, unreliable and expensive.

In summary, when evaluated from a comprehensive standpoint, considering CI widths, label density, coverage, and label utilization, CA-ASI demonstrates superior performance compared to all other methods. Additional real-world data experiment in Appendix D.4.

7 Conclusion

Limitations. CA-ASI inherits the calibration and i.i.d. assumptions that define seminal ML-powered inference frameworks (ASI, PPI); validity is preserved regardless of calibration quality, while efficiency gains depend on it. On datasets with little to no hessian overlap, CA-ASI may not be able prevent wasteful labeling. Computational overhead from the coherence computation is characterized in Appendix D.5 and discussed as a label-versus-compute trade-off favorable in the regimes active inference targets.

Discussion. We propose Curvature-Aware Active Statistical Inference (CA-ASI), a novel inference framework that directs labeling effort toward points with the greatest impact on the statistical estimate. By leveraging both model uncertainty and the local curvature of the loss function, CA-ASI achieves more powerful inferences with significantly fewer labels. We provide theoretical guarantees for its asymptotic validity and demonstrate its practical effectiveness across datasets. We believe this work establishes a new principle for efficient data collection, showing that the geometry of the statistical problem is as crucial as the uncertainty of the predictive model.

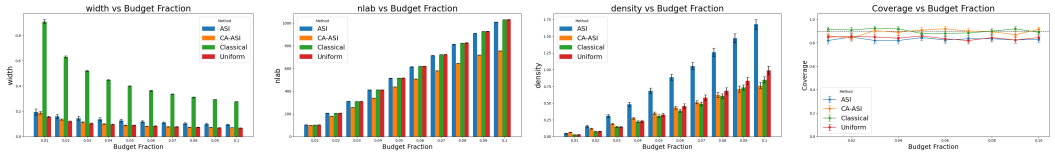

Figure 5: Batch Setting in Housing Dataset

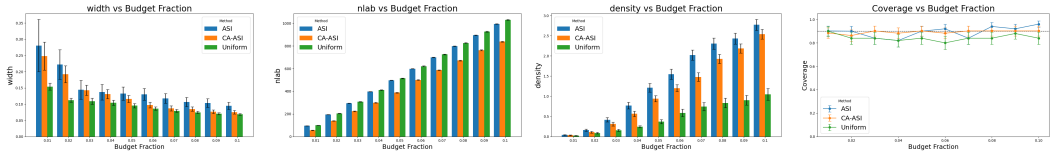

Figure 6: Sequential Setting in Housing Dataset

Acknowledgments

We thank the Central Indiana Corporate Partnership AnalytiXIN Initiative for their support.

References

[1] Y. Abbasi-yadkori, D. Pál, and C. Szepesvári. Improved algorithms for linear stochastic bandits. In J. Shawe-Taylor, R. Zemel, P. Bartlett, F. Pereira, and K. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 24. Curran Associates, Inc., 2011. URL https://proceedings.neurips.cc/paper_files/paper/2011/file/e1d5be1c7f2f456670de3d53c7b54f4a-Paper.pdf.

[2] A. Agiollo, Y. I. Kim, and R. Khanna. Approximating memorization using loss surface geometry for dataset pruning and summarization. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, KDD '24, page 17–28, New York, NY, USA, 2024. Association for Computing Machinery. ISBN 9798400704901. doi: 10.1145/3637528.3671985. URL https://doi.org/10.1145/3637528.3671985.

[3] A. Agresti. *Categorical Data Analysis*. Wiley Series in Probability and Statistics. Wiley, 2003. ISBN 9780471458760. URL https://books.google.com/books?id=hpEzw4T0sPUC.

[4] A. N. Angelopoulos, S. Bates, C. Fannjiang, M. I. Jordan, and T. Zrnic. Prediction-powered inference, 2023. URL https://arxiv.org/abs/2301.09633.

[5] A. N. Angelopoulos, J. C. Duchi, and T. Zrnic. Ppi++: Efficient prediction-powered inference, 2024. URL https://arxiv.org/abs/2311.01453.

[6] J. T. Ash, C. Zhang, A. Krishnamurthy, J. Langford, and A. Agarwal. Deep batch active learning by diverse, uncertain gradient lower bounds. *arXiv preprint arXiv:1906.03671*, 2019. URL https://arxiv.org/abs/1906.03671.

[7] R. Bardenet, A. Doucet, and C. Holmes. On markov chain monte carlo methods for tall data. *Journal of Machine Learning Research*, 18(47):1–43, 2017. URL http://jmlr.org/papers/v18/15-205.html.

[8] Y. Bengio, A. Courville, and P. Vincent. Representation learning: A review and new perspectives, 2014. URL https://arxiv.org/abs/1206.5538.

[9] M. Bernhardt, D. C. Castro, R. Tanno, A. Schwaighofer, K. C. Tezcan, M. Monteiro, S. Bannur, M. P. Lungren, A. Nori, B. Glocker, J. Alvarez-Valle, and O. Oktay. Active label cleaning for improved dataset quality under resource constraints. *Nature Communications*, 13(1):1161, 2022. ISSN 2041-1723. doi: 10.1038/s41467-022-28818-3. URL https://doi.org/10.1038/s41467-022-28818-3.

[10] S. C. Bourassa, E. Cantoni, and M. Hoesli. Spatial dependence, housing submarkets, and house price prediction. *Journal of Real Estate Finance and Economics*, 35(2):143–160, 2007. URL https://doi.org/10.1007/s11146-007-9036-8.

[11] P. Calem, J. Kenney, L. Lambie-Hanson, and L. Nakamura. Appraising home purchase appraisals. *Real Estate Economics*, 49(S1):134–168, 2021. doi: https://doi.org/10.1111/1540-6229.12326. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/1540-6229.12326.

[12] N. Cesa-Bianchi, C. Gentile, and L. Zaniboni. Worst-case analysis of selective sampling for linear classification. *Journal of Machine Learning Research*, 7(44):1205–1230, 2006. URL http://jmlr.org/papers/v7/cesa-bianchi06b.html.

[13] Y. Chandak, S. Shankar, V. Syrgkanis, and E. Brunskill. Adaptive instrument design for indirect experiments, 2023. URL https://arxiv.org/abs/2312.02438.

[14] W.-K. Chang and R. Khanna. A unified stability analysis of sam vs sgd: Role of data coherence and emergence of simplicity bias, 2025. URL https://arxiv.org/abs/2511.17378.

[15] T. Chen and C. Guestrin. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, page 785–794. ACM, Aug. 2016. doi: 10.1145/2939672.2939785. URL http://dx.doi.org/10.1145/2939672.2939785.

[16] L. N. Cook, A. Mishler, A. Ramdas, and T. Zrnić. Semiparametric efficient inference from adaptively collected data. In *Proceedings of the Conference on Learning Theory and Representations*, 2024. URL https://openreview.net/forum?id=mfSgIZpwi0.

[17] G. Dexter, B. Ocejo, S. Keerthi, A. Gupta, A. Acharya, and R. Khanna. A precise characterization of sgd stability using loss surface geometry, 2024. URL https://arxiv.org/abs/2401.12332.

[18] A. Dieuleveut, G. Fort, E. Moulines, and H.-T. Wai. Stochastic approximation beyond gradient for signal processing and machine learning. *IEEE Transactions on Signal Processing*, 71:3117–3148, 2023. doi: 10.1109/TSP.2023.3301121. URL https://ieeexplore.ieee.org/document/10202577.

[19] A. I. Dumachi and F. Negoita. Hardware acceleration for machine learning in medical imaging: Case study on colorectal polyp segmentation. In *2024 International Semiconductor Conference (CAS)*, pages 321–324, 2024. doi: 10.1109/CAS62834.2024.10736775. URL https://ieeexplore.ieee.org/document/10736775.

[20] R. P. Eatough, N. Molkenthin, M. Kramer, A. Noutsos, M. J. Keith, B. W. Stappers, and A. G. Lyne. Selection of radio pulsar candidates using artificial neural networks: Selection of radio pulsar candidates. *Monthly Notices of the Royal Astronomical Society*, 407(4):2443–2450, 2010. ISSN 0035-8711. doi: 10.1111/j.1365-2966.2010.17082.x. URL http://dx.doi.org/10.1111/j.1365-2966.2010.17082.x.

[21] A. L. Ferguson, M. LaFleur, L. Ruthotto, J. Thaler, Y.-S. Ting, P. Tiwary, and S. Villar. The future of artificial intelligence and the mathematical and physical sciences (ai+mps). *Machine Learning: Science and Technology*, Jan. 2026. ISSN 2632-2153. doi: 10.1088/2632-2153/ae3e4e. URL http://dx.doi.org/10.1088/2632-2153/ae3e4e.

[22] W. A. Fischel. *The Homevoter Hypothesis: How Home Values Influence Local Government Taxation, School Finance, and Land-Use Policies*. Harvard University Press, Cambridge, MA, 2001. URL https://books.google.com/books?id=kVgwEAAAQBAJ.

[23] Y. Fu, X. Zhu, and B. Li. A survey on instance selection for active learning. *Knowledge and Information Systems*, 35(2):249–283, 2013. doi: 10.1007/s10115-012-0507-8. URL https://link.springer.com/article/10.1007/s10115-012-0507-8.

[24] D. Fudenberg and J. Tirole. *Game Theory*. Mit Press. MIT Press, 1991. ISBN 9780262061414. URL https://books.google.com/books?id=pFPHKwXro3QC.

[25] K. Gligoric, T. Zrnic, C. Lee, E. Candes, and D. Jurafsky. Can unconfident LLM annotations be used for confident conclusions? In L. Chiruzzo, A. Ritter, and L. Wang, editors,

Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), pages 3514–3533, Albuquerque, New Mexico, Apr. 2025. Association for Computational Linguistics. ISBN 979-8-89176-189-6. doi: 10.18653/v1/2025.naacl-long.179. URL https://aclanthology.org/2025.naacl-long.179/.

[26] R. Greer, B. Antoniussen, M. V. Andersen, A. Møgelmose, and M. M. Trivedi. The why, when, and how to use active learning in large-data-driven 3d object detection for safe autonomous driving: An empirical exploration, 2024. URL https://arxiv.org/abs/2401.16634.

[27] J. Hahn, K. Hirano, and D. Karlan. Adaptive experimental design using the propensity score. *Journal of Business & Economic Statistics*, 29(1):96–108, 2011. doi: 10.1198/jbes.2009.08161. URL https://doi.org/10.1198/jbes.2009.08161.

[28] A. M. A. Heikal, S. H. E. A. Aleem, R. A. El-Sehiemy, and A. Y. Abdelaziz. Robust techno-economic optimization of energy hubs under uncertainty using active learning with artificial neural networks. *Scientific Reports*, 15, 2025. URL https://doi.org/10.1038/s41598-025-12358-z.

[29] W. Hoeffding. A class of statistics with asymptotically normal distribution. *The Annals of Mathematical Statistics*, 19(3):293–325, 1948. doi: 10.1214/aoms/1177730196. URL https://doi.org/10.1214/aoms/1177730196.

[30] T. Jaakkola and D. Haussler. Exploiting generative models in discriminative classifiers. In M. Kearns, S. Solla, and D. Cohn, editors, *Advances in Neural Information Processing Systems*, volume 11. MIT Press, 1998. URL https://proceedings.neurips.cc/paper_files/paper/1998/file/db1915052d15f7815c8b88e879465a1e-Paper.pdf.

[31] N. Jean, M. Burke, M. Xie, W. M. Davis, D. B. Lobell, and S. Ermon. Combining satellite imagery and machine learning to predict poverty. *Science*, 353(6301):790–794, 2016. URL https://www.science.org/doi/abs/10.1126/science.aaf7894.

[32] J. Jumper, R. Evans, A. Pritzel, T. Green, M. Figurnov, O. Ronneberger, K. Tunyasuvunakool, R. Bates, A. Zídek, A. Potapenko, et al. Highly accurate protein structure prediction with alphafold. *Nature*, 596(7873):583–589, 2021. URL https://doi.org/10.1038/s41586-021-03819-2.

[33] M. Kato, T. Ishihara, J. Honda, and Y. Narita. Efficient adaptive experimental design for average treatment effect estimation, 2025. URL https://arxiv.org/abs/2002.05308.

[34] G. I. Kim, S. Hwang, and B. Jang. Efficient compressing and tuning methods for large language models: A systematic literature review. *ACM Comput. Surv.*, 57(10), May 2025. ISSN 0360-0300. doi: 10.1145/3728636. URL https://doi.org/10.1145/3728636.

[35] A. Kirsch, J. van Amersfoort, and Y. Gal. Batchbald: Efficient and diverse batch acquisition for deep bayesian active learning, 2019. URL https://arxiv.org/abs/1906.08158.

[36] B. Krishnapuram, S. Yu, and R. Rao. *Cost-Sensitive Machine Learning*. Chapman & Hall/CRC machine learning & pattern recognition series. CRC Press, 2011. ISBN 9781439839287. URL https://books.google.com/books?id=8TrNBQAAQBAJ.

[37] F. Kunstner, L. Balles, and P. Hennig. Limitations of the empirical fisher approximation for natural gradient descent, 2020. URL https://arxiv.org/abs/1905.12558.

[38] D.-H. Lee. Pseudo-label : The simple and efficient semi-supervised learning method for deep neural networks. 2013. URL https://api.semanticscholar.org/CorpusID:18507866.

[39] D. D. Lewis and W. A. Gale. A sequential algorithm for training text classifiers, 1994. URL https://arxiv.org/abs/cmp-lg/9407020.

[40] D. Li, Y. Ma, N. Wang, Z. Ye, Z. Cheng, Y. Tang, Y. Zhang, L. Duan, J. Zuo, C. Yang, and M. Tang. Mixlora: Enhancing large language models fine-tuning with lora-based mixture of experts, 2024. URL https://arxiv.org/abs/2404.15159.

[41] B. Liu and V. Ferrari. Active learning for human pose estimation. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 4373–4382, 2017. doi: 10.1109/ICCV.2017.468. URL https://ieeexplore.ieee.org/document/8237730.

[42] R. Lyon. HTRU2. UCI Machine Learning Repository, 2015. URL https://doi.org/10.24432/C5DK6R.

[43] R. J. Lyon, B. W. Stappers, S. Cooper, J. M. Brooke, and J. D. Knowles. Fifty years of pulsar candidate selection: from simple filters to a new principled real-time classification approach. *Monthly Notices of the Royal Astronomical Society*, 459(1):1104–1123, 06 2016. ISSN 0035-8711. doi: 10.1093/mnras/stw656. URL https://doi.org/10.1093/mnras/stw656.

[44] J. Martens. Deep learning via hessian-free optimization. In *Proceedings of the 27th International Conference on International Conference on Machine Learning*, ICML'10, page 735–742, Madison, WI, USA, 2010. Omnipress. ISBN 9781605589077.

[45] J. Martens and R. Grosse. Optimizing neural networks with kronecker-factored approximate curvature. In F. Bach and D. Blei, editors, *Proceedings of the 32nd International Conference on Machine Learning (ICML)*, volume 37 of *Proceedings of Machine Learning Research*, pages 2408–2417, Lille, France, July 2015. PMLR. URL https://proceedings.mlr.press/v37/martens15.html.

[46] J. Martens, I. Sutskever, and K. Swersky. Estimating the hessian by back-propagating curvature. In *Proceedings of the 29th International Conference on Machine Learning (ICML)*. PMLR, 2012. URL https://arxiv.org/abs/1206.6464.

[47] P. McCullagh and J. Nelder. *Generalized Linear Models*. Springer, 2014. ISBN 9781489932457. URL https://books.google.com/books?id=oGwMswEACAAJ.

[48] A. A. Melnikov, H. P. Nautrup, M. Krenn, V. Dunjko, M. Tiersch, A. Zeilinger, and H. J. Briegel. Active learning machine learns to create new quantum experiments. *Proceedings of the National Academy of Sciences*, 115(6):1221–1226, 2018. doi: 10.1073/pnas.1714936115. URL https://www.pnas.org/doi/abs/10.1073/pnas.1714936115.

[49] A. O'Sullivan, T. Sexton, and S. Sheffrin. *Property Taxes and Tax Revolts: The Legacy of Proposition 13*. Cambridge University Press, 1995. ISBN 9780521461597. URL https://books.google.com/books?id=sEnB7PrE7lsC.

[50] R. K. Pace and R. Barry. Sparse spatial autoregressions. *Statistics & Probability Letters*, 33(3):291–297, 1997. URL https://doi.org/10.1016/S0167-7152(96)00140-X.

[51] V. Pagel-Theisen. *Diamond Grading ABC: The Manual*. Rubin & Son n.v., 2001. ISBN 9783980043465. URL https://books.google.com/books?id=M3pUPgAACAAJ.

[52] E. Pagourtzi, V. Assimakopoulos, T. Hatzichristos, and N. French. Real estate appraisal: a review of valuation methods. *Journal of Property Investment Finance*, 21(4):383–401, 08 2003. ISSN 1463-578X. doi: 10.1108/14635780310483656. URL https://doi.org/10.1108/14635780310483656.

[53] M. Pilanci and M. J. Wainwright. Newton sketch: A linear-time optimization algorithm with linear-quadratic convergence, 2015. URL https://arxiv.org/abs/1505.02250.

[54] M. Porter. *Competitive Strategy: Techniques for Analyzing Industries and Competitors*. Business book summary. Free Press, 1980. ISBN 9780029253601. URL https://books.google.com/books?id=Nl21AAAAIAAJ.

[55] D. Price. Reduced-resolution beamforming: Lowering the computational cost for pulsar and technosignature surveys. *Publications of the Astronomical Society of Australia*, 41:e037, 2024. doi: 10.1017/pasa.2024.35. URL https://doi.org/10.1017/pasa.2024.35.

[56] L. Puheng, Z. Tijana, and C. Emmanuel. Robust sampling for active statistical inference, 2025. URL https://arxiv.org/abs/2511.08991.

[57] P. Ren, Y. Xiao, X. Chang, P.-Y. Huang, Z. Li, B. B. Gupta, X. Chen, and X. Wang. A survey of deep active learning. *ACM Comput. Surv.*, 54(9), Oct. 2021. ISSN 0360-0300. doi: 10.1145/3472291. URL https://doi.org/10.1145/3472291.

[58] L. Renneboog and C. Spaenjers. Hard assets: The returns on rare diamonds and gems. *Finance Research Letters*, 9(4):220–230, 2012. URL https://papers.ssrn.com/sol3/papers.cfm?abstract_id=1845446.

[59] J. M. Robins, A. Rotnitzky, and L. P. Zhao. Estimation of regression coefficients when some regressors are not always observed. *Journal of the American Statistical Association*, 89(427): 846–866, 1994. doi: 10.1080/01621459.1994.10476818. URL https://doi.org/10.1080/01621459.1994.10476818.

[60] E. Rolf, J. Proctor, T. Carleton, I. Bolliger, V. Shankar, M. Ishihara, B. Recht, and S. Hsiang. A generalizable and accessible approach to machine learning with global satellite imagery. *Nature Communications*, 12(1):4392, 2021. URL https://doi.org/10.1038/s41467-021-24638-z.

[61] S. Rosen. Hedonic prices and implicit markets: Product differentiation in pure competition. *Journal of Political Economy*, 82(1):34–55, 1974. ISSN 00223808, 1537534X. URL http://www.jstor.org/stable/1830899.

[62] O. Sener and S. Savarese. Active learning for convolutional neural networks: A core-set approach. In *International Conference on Learning Representations (ICLR)*, 2018. URL https://openreview.net/forum?id=H1aIuk-RW.

[63] B. Settles. Active learning literature survey. *University of Wisconsin-Madison Department of Computer Sciences*, 52(55-66), 2009. URL http://digital.library.wisc.edu/1793/60660.

[64] B. Settles. Closing the loop: Fast, interactive semi-supervised annotation with queries on features and instances. In R. Barzilay and M. Johnson, editors, *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 1467–1478, Edinburgh, Scotland, UK., July 2011. Association for Computational Linguistics. URL https://aclanthology.org/D11-1136/.

[65] V. S. Sheng, F. Provost, and P. G. Ipeirotis. Get another label? improving data quality and data mining using multiple, noisy labelers. In *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '08, page 614–622, New York, NY, USA, 2008. Association for Computing Machinery. ISBN 9781605581934. doi: 10.1145/1401890.1401965. URL https://doi.org/10.1145/1401890.1401965.

[66] M. D. Shields, K. R. Gurley, R. A. Catarelli, M. Chauhan, M. Ojeda-Tuz, and F. J. Masters. Active learning applied to automated physical systems increases the rate of discovery. *Scientific Reports*, 13, 2023. URL https://api.semanticscholar.org/CorpusID:258890103.

[67] D. Slabbert, M. Lourens, and F. Petruccione. Pulsar classification: comparing quantum convolutional neural networks and quantum support vector machines. *Quantum Machine Intelligence*, 6(56), 2024. doi: 10.1007/s42484-024-00194-9. URL https://doi.org/10.1007/s42484-024-00194-9.

[68] A. Soen and K. Sun. Trade-offs of diagonal fisher information matrix estimators. In *NeurIPS 2024 Poster Track*, 2024. URL https://openreview.net/forum?id=TVbCKAqoD8.

[69] K. Sohn, D. Berthelot, C.-L. Li, Z. Zhang, N. Carlini, E. D. Cubuk, A. Kurakin, H. Zhang, and C. Raffel. Fixmatch: Simplifying semi-supervised learning with consistency and confidence, 2020. URL https://arxiv.org/abs/2001.07685.

[70] B. Sorscher, R. Geirhos, S. Shekhar, S. Ganguli, and A. S. Morcos. Beyond neural scaling laws: beating power law scaling via data pruning, 2023. URL https://arxiv.org/abs/2206.14486.

[71] A. van der Vaart. *Asymptotic Statistics*. Asymptotic Statistics. Cambridge University Press, 2000. ISBN 9780521784504. URL https://books.google.com/books?id=UEuQEM5RjWgC.

[72] H. Wickham. *ggplot2: Elegant Graphics for Data Analysis*. Use R! Springer International Publishing, 2016. ISBN 9783319242750. URL https://books.google.com/books?id=RTMFswEACAAJ.

[73] H. Wickham, M. Çetinkaya Rundel, and G. Grolemund. *R for Data Science*. O'Reilly Media, 2nd edition, 2023. URL https://r4ds.hadley.nz.

[74] M. Xie, N. Jean, M. Burke, D. Lobell, and S. Ermon. Transfer learning from deep features for remote sensing and poverty mapping. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 30, 2016. URL https://dl.acm.org/doi/10.5555/3016387.3016457.

[75] J. Yu. A multivariate regression analysis of factors influencing california housing prices. In *Proceedings of the International Conference on Mathematics and Machine Learning*, ICMML '23, page 165–169, New York, NY, USA, 2024. Association for Computing Machinery. ISBN 9798400716973. doi: 10.1145/3653724.3653753. URL https://doi.org/10.1145/3653724.3653753.

[76] M. Zdun. Machine politics: How america casts and counts its votes. *Reuters*, 2022. URL https://www.reuters.com/graphics/USA-ELECTION/VOTING/mypmnewdlvr/.

[77] K. W. Zhang, L. Janson, and S. Murphy. Statistical inference with m-estimators on adaptively collected data. In A. Beygelzimer, Y. Dauphin, P. Liang, and J. W. Vaughan, editors, *Advances in Neural Information Processing Systems*, 2021. URL https://openreview.net/forum?id=TJOQw_vMlAj.

[78] C. Zhu, R. H. Byrd, P. Lu, and J. Nocedal. Algorithm 778: L-bfgs-b: Fortran subroutines for large-scale bound-constrained optimization. *ACM Trans. Math. Softw.*, 23(4):550–560, Dec. 1997. ISSN 0098-3500. doi: 10.1145/279232.279236. URL https://doi.org/10.1145/279232.279236.

[79] L. Zhu, W. Wen, J. Li, C. Zhang, B. Zhou, and Z. Shuai. Deep active learning-enabled cost-effective electricity theft detection in smart grids. *IEEE Transactions on Industrial Informatics*, 20(1):256–268, 2024. doi: 10.1109/TII.2023.3249212. URL https://ieeexplore.ieee.org/document/10080868.

[80] T. Zrnić and E. J. Candès. Active statistical inference. In *Proceedings of the 41st International Conference on Machine Learning (ICML)*, Proceedings of Machine Learning Research, 2024. URL https://proceedings.mlr.press/v235/zrnic24a.html. PMLR v235.

[81] T. Zrnic and E. J. Candès. Cross-prediction-powered inference. *Proceedings of the National Academy of Sciences*, 121(15):e2322083121, 2024. doi: 10.1073/pnas.2322083121. URL https://www.pnas.org/doi/abs/10.1073/pnas.2322083121.

A Impact Statement

This work aims to further rigorous statistical inference by significantly reducing the cost of data labeling. By integrating curvature-aware sampling with Active Statistical Inference (ASI), our method enables smaller research teams, non-profits, and organizations with limited budgets to perform high-quality scientific inquiry in domains where data annotation is expensive and time intensive or requires expert knowledge. Positively, CA-ASI can improve the human-in-the-loop experience by reducing annotator fatigue and redirect human expertise toward higher-value tasks. By pruning geometrically redundant instances, our method ensures annotators focus on statistically critical, diverse cases requiring expert judgment—effectively upskilling the annotation pipeline from mechanical repetition to quality assurance and edge-case curation. Beyond this, we do not foresee direct negative societal consequences, though we acknowledge standard machine learning risks regarding the reliance on the quality of the underlying predictive models.

B Algorithm

Algorithm 1 Batch curvature-aware active inference

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

1. Choose uncertainty measure $\tilde{u}(x)$ based on f
2. Compute data coherence $\tilde{\rho}(x)$ based on f(Eq. (2))
3. Compute $\tilde{u}_\beta(x) = \tilde{u}(x) - \beta\,\tilde{\rho}(x)$
4. Let $\pi_\beta(x) = \hat{\eta}\ \bar{u}_\beta(x)$, where $\bar{u}_\beta(x) := [\tilde{u}_\beta(x)]_{[0,1]} = \min\{\max\{\tilde{u}_\beta(x), 0\}, 1\}$., $\hat{\eta} = \frac{n_b}{n\hat{\mathbb{E}}[\tilde{u}_\beta(X)]}$; let $\pi^{\text{unif}} = \frac{n_b}{n}$
5. Optionally: Select $\tau \in (0, 1)$ and choose sampling rule $\pi_\beta^{(\tau)}(x) = (1-\tau)\cdot\pi_\beta(x) + \tau\cdot\pi^{\text{unif}}$
6. Sample labeling decisions $\xi_i \sim \text{Bern}(\pi_\beta^{(\tau)}(X_i)), i \in [n]$
7. Collect labels $\{Y_i : \xi_i = 1\}$
8. Compute batch active estimator $(\hat{\theta}^{\pi_\beta^{(\tau)}})$ (minimizer of Eq. (3))

C Proofs

Claim C.1 (Consistency of $\hat{\gamma}$ for CA-ASI)**.** *Suppose that the budget fraction converges, i.e., $n_b/n \to p \in (0, 1)$. Let the curvature-aware utility be denoted by $\bar{u}_\beta(X)$, defined as: $\bar{u}_\beta(X) := \min(\max(\tilde{u}(X) - \beta\tilde{\rho}(X), 0), 1)$, where $\tilde{u}(X)$ and $\tilde{\rho}(X)$ are the normalized uncertainty and redundancy scores, respectively. Assuming the set of possible budget parameters $\mathcal{H}$ is discrete. If there is no $\eta \in \mathcal{H}$ such that $\eta\mathbb{E}[\bar{u}_\beta(X)] = p$ exactly, then there exists a unique deterministic $\eta^* \in \mathcal{H}$ such that: $\mathbb{P}(\hat{\eta} \neq \eta^*) \to 0$*

Proof. Proof of Claim C.1. By the definition of our sampling rule (Algorithm 1), the empirical parameter $\hat{\eta}$ is chosen as the maximum value in $\mathcal{H}$ that satisfies the budget constraint on the batch: $\hat{\eta} \cdot \frac{1}{n}\sum_{i=1}^n \bar{u}_\beta(X_i) \leq \frac{n_b}{n}$. Rearranging this inequality for $\hat{\eta}$, we have: $\hat{\eta} \leq \frac{n_b/n}{\frac{1}{n}\sum_{i=1}^n \bar{u}_\beta(X_i)}$

We now analyze the convergence of the terms on the right-hand side:

In the numerator, by assumption, $n_b/n \to p$.

In the denominator, the term $\frac{1}{n}\sum_{i=1}^n \bar{u}_\beta(X_i)$ is the empirical average of the curvature-aware utilities. In standard ASI, the analogous denominator is the empirical mean of a pointwise uncertainty score, which converges directly by the Law of Large Numbers (LLN). Here the situation is more subtle: $\bar{u}_\beta(X_i) = [\tilde{u}(X_i) - \beta\tilde{\rho}(X_i)]_{[0,1]}$ involves the difference of two pool-normalized quantities. Specifically, $\tilde{u}(X_i)$ is a ratio involving a degree-1 U-statistic in the denominator, and $\tilde{\rho}(X_i)$ is a ratio of two degree-2 U-statistics with bounded symmetric kernel S_{ij}. By the U-statistic LLN [29] and the Continuous Mapping Theorem (CMT) applied to each, both $\tilde{u}(X_i)$ and $\tilde{\rho}(X_i)$ converge in probability

Algorithm 2 Sequential curvature-aware active inference

Input: unlabeled data $X_1, \ldots, X_n$, sampling budget n_b, initial predictive model f_1, pruning parameter β, error level $\alpha \in (0,1)$, fine-tuning batch size B
Initialize: $\mathcal{D}^{\text{tune}} \leftarrow \emptyset$, $\mathbf{S}_1 \leftarrow \mathbf{0}$, $n_{\text{lab},1=0}$
for $t = 1, \ldots, n$ **do**
1. Choose uncertainty measure $\tilde{u}_t(x)$ for f_t
2.
if $n_{\text{lab,t}} \geq 2$ **then**
2.1 Compute $\tilde{\rho}_t(x)$ using f_t and $\mathbf{S}_t$.
else
2.2 Set $\tilde{\rho}_t(x) = 0$
end if
3. Set $\pi_{\beta,t}(x)$(as discussed in Section 4) with $\eta_t = \frac{n_b}{n\hat{\mathbb{E}}[u_{\beta,t}(X)]}$; let $\pi^{\text{unif}} = \frac{n_b}{n}$
4. Optionally: Select $\tau \in (0,1)$ and choose sampling rule $\pi^{(\tau)}_{\beta,t}(x) = (1-\tau)\cdot\pi_{\beta,t}(x) + \tau\cdot\pi^{\text{unif}}$
5. Sample labeling decision $\xi_t \sim \text{Bern}(\pi^{(\tau)}_{\beta,t}(X_t))$
6.
if $\xi_t = 1$ **then**
6.1.1 Collect label Y_t
6.1.2 $n_{\text{lab,t+1}} \leftarrow n_{\text{lab,t}} + 1$
6.1.3 Grow Coherence Matrix $\mathbf{S}_{t+1}$ by incorporating $\{(X_t, Y_t)\}$.
6.1.4 $\mathcal{D}^{\text{tune}} \leftarrow \mathcal{D}^{\text{tune}} \cup \{(X_t, Y_t)\}$
6.1.5
if $|\mathcal{D}^{\text{tune}}| = B$ **then**
6.1.5.1.1 Fine-tune model on $\mathcal{D}^{\text{tune}}$: $f_{t+1} = \texttt{finetune}(f_t, \mathcal{D}^{\text{tune}})$
6.1.5.1.2 Set $\mathcal{D}^{\text{tune}} \leftarrow \emptyset$
else
6.1.5.2 $f_{t+1} \leftarrow f_t$
end if
else
6.2 $f_{t+1} \leftarrow f_t$
end if
end for
7. Compute sequential active estimator $\hat{\theta}^{\overrightarrow{\pi}^{\tau}_{\beta}}$ (minimzer of Eq. (4))

to their respective population limits, and hence so does their difference. Thus, the empirical average converges in probability to the population expectation[2]: $\frac{1}{n}\sum_{i=1}^{n} \bar{u}_\beta(X_i) \xrightarrow{p} \mathbb{E}[\bar{u}_\beta(X)]$

Applying CMT, the entire ratio converges in probability,

$$R_n := \frac{n_b/n}{\frac{1}{n}\sum_{i=1}^{n} \bar{u}_\beta(X_i)} \xrightarrow{p} \frac{p}{\mathbb{E}[\bar{u}_\beta(X)]} := R^*$$

Per definition, population optimal parameter $\eta^* = \max\{\eta \in \mathcal{H} : \eta \leq R^*\}$. Since $\mathcal{H}$ is discrete and no η exactly equals R^* (by assumption), there exists a positive gap $\epsilon > 0$ between R^* and the nearest incorrect parameter value. Specifically, let $\epsilon = \min_{\eta\in\mathcal{H}} |\eta - R^*|$. On the event that the empirical ratio R_n is within ϵ of the population limit R^* (i.e., $|R_n - R^*| < \epsilon$), the discrete maximization will select exactly the same value η^*.

Therefore,

$$\mathbb{P}(\hat{\eta} \neq \eta^*) \leq \mathbb{P}(|R_n - R^*| \geq \epsilon)$$

.

Since $R_n \xrightarrow{p} R^*$, the probability on the right-hand side goes to 0 as $n \to \infty$. Thus, $\mathbb{P}(\hat{\eta} \neq \eta^*) \to 0$. Further, since the pruning parameter β is a fixed hyperparameter independent of n, the convergence

[2]The clipping operation $[\cdot]_{[0,1]}$ makes the limiting value of the empirical average non-trivial to characterize directly. We therefore rely on the ε regularization in $\hat{\mu}_\beta := \max\left\{\frac{1}{n}\sum_j \bar{u}_\beta(X_j), \varepsilon\right\}$, which ensures the denominator is bounded away from zero by construction.

of the scalar $\hat{\eta} \xrightarrow{p} \eta^*$ implies the convergence of the parameter vector $\hat{\gamma} = (\hat{\eta}, \beta)$ to the deterministic limit $\gamma^* = (\eta^*, \beta)$. Thus, the condition $\mathbb{P}(\hat{\gamma} \neq \gamma^*) \to 0$ holds. □

□

Proof for Theorem 5.3. The proof follows the argument for Theorem 5.2 in Zrnić and Candès [80]. Their proof relies on the consistency of the data-adaptive parameter $\hat{\eta}$ that defines the sampling policy. Our argument is a direct generalization. We introduce a composite parameter $\gamma = (\eta, \beta)$ to define the CA-ASI sampling policy π_γ. The key assumption for our theorem is the consistency of its empirical estimate, $\hat{\gamma} = (\hat{\eta}, \beta)$, to a deterministic limit, $\gamma^* = (\eta^*, \beta)$(from Claim C.1). This assumption is the direct analogue of their condition on η and ensures that the empirical sampling rule converges to the limiting rule. Since the core condition of a consistent, data-adaptive sampling policy is preserved, the original proof's machinery i.e. use of Slutsky's theorem to establish equivalence with an oracle estimator and then application a Central Limit Theorem hold without modification. The only change is the resulting asymptotic covariance matrix, which is now Σ_{γ^*} instead of Σ_*. The validity of the confidence intervals follows directly.

□

Lemma C.2. *Let $w(x) \geq 0$ be an arbitrary non-negative weight function. Let $J(\beta) = \mathbb{E}[w(X)/\pi_\beta(X)]$, where π_β is the CA-ASI sampling policy. Let π_0 be the standard ASI policy. The derivative of $J(\beta)$ evaluated at $\beta = 0$ is given by:*

$$\left.\frac{dJ}{d\beta}\right|_{\beta=0} \propto \mathbb{E}\left[\frac{w(X)}{\pi_0(X)^2}\left(\tilde{\rho}(X) - \frac{\tilde{u}(X)}{\mathbb{E}[\tilde{u}(X)]}\mathbb{E}[\tilde{\rho}(X)]\right)\right]$$

where $\tilde{u}(x)$ and $\tilde{\rho}(x)$ are the normalized [0,1] uncertainty and redundancy scores.

Proof for C.2. We aim to show that for a sufficiently small $\beta > 0$, the variance is a decreasing function of the penalty, formally proving that "turning on" the geometry-aware pruning is beneficial.

Recall, $\tilde{u}(x), \tilde{\rho}(x) \in [0, 1]$ denote the normalized uncertainty and redundancy scores, respectively. For $\beta > 0$, define the β-suppressed utility

$$\tilde{u}_\beta(x) := \tilde{u}(x) - \beta\,\tilde{\rho}(x), \quad \bar{u}_\beta(x) := [\,\tilde{u}_\beta(x)\,]_{[0,1]} := \min\{\max\{\tilde{u}_\beta(x), 0\}, 1\}.$$

Let

$$\mu_\beta = \mathbb{E}[\bar{u}_\beta(X)] = \mathbb{E}\left[\max\{\min\{\tilde{u}(X) - \beta\tilde{\rho}(X), 1\}, 0\}\right], \quad \hat{\mu}_\beta := \max\left\{\frac{1}{n}\sum_{j=1}^{n}\bar{u}_\beta(x_j),\ \varepsilon\right\}$$

The above is estimated in practice using,

$$\hat{\mu}_\beta := \max\left\{\frac{1}{n}\sum_{j=1}^{n}\bar{u}_\beta(x_j),\ \varepsilon\right\}$$

for some $\varepsilon > 0$ and define the τ–mixed sampling probability

$$\pi_\beta^{(\tau)}(x) := \frac{n_b}{n}\left[(1-\tau)\,\frac{\bar{u}_\beta(x)}{\mu_\beta} + \tau\right], \quad \tau \in [0, 1].$$

We define clipped set as

$$\mathcal{Z}_\beta := \{\, x \in \mathcal{D} : \bar{u}_\beta(x) = 0 \,\} = \left\{x \in \mathcal{D} \;\middle|\; \beta > \frac{\tilde{u}(x)}{\tilde{\rho}(x)}\right\} \quad \text{(assuming } \tilde{\rho}(x) > 0\text{)}.$$

Similarly , we define active set as, $\mathcal{Z}_\beta^{\mathsf{c}} := \mathcal{D} \setminus \mathcal{Z}_\beta$

Let, $s_\beta(x) := \frac{\bar{u}_\beta(x)}{\mu_\beta}, \quad g_\beta(x) := \tau + (1-\tau)s_\beta(x), \quad \pi_\beta^{(\tau)}(x) = \frac{n_b}{n}\,g_\beta(x)$

The proof follows by direct computation. The policy-dependent component of the objective is $J(\beta) \propto \mathbb{E}[w(X)/g_\beta(X)]$. Differentiating with respect to β and applying Leibniz's rule to move the derivative inside the expectation gives:

$$\frac{dJ}{d\beta} \propto -\mathbb{E}\left[\frac{w(X)g'_\beta(X)}{g_\beta(X)^2}\right]$$

From the definition of $g_\beta(x)$, its derivative is $g'_\beta(x) = (1-\tau)s'_\beta(x)$. We must therefore compute the derivative of the normalized score, $s_\beta(x) = \bar{u}_\beta(x)/\mu_\beta$. Using the quotient rule,

$$s'_\beta(x) = \frac{(\partial_\beta \bar{u}_\beta(x))\mu_\beta - \bar{u}_\beta(x)(\partial_\beta \mu_\beta)}{\mu_\beta^2}$$

We need the derivatives of the clipped utility, $\bar{u}_\beta(x)$, and the mean utility, μ_β. Due to the clipping operation $\bar{u}_\beta(x) = [\tilde{u}(x) - \beta\tilde{\rho}(x)]_{[0,1]}$, the derivative with respect to β is non-zero only on the active set $\mathcal{Z}_\beta^c$:

$$\partial_\beta \bar{u}_\beta(x) = -\tilde{\rho}(x)\,\mathbf{1}_{\mathcal{Z}_\beta^c}(x)$$

Consequently, the derivative of the mean utility is:

$$\partial_\beta \mu_\beta = \partial_\beta \mathbb{E}[\bar{u}_\beta(X)] = \mathbb{E}[\partial_\beta \bar{u}_\beta(X)] = -\mathbb{E}[\tilde{\rho}(X)\,\mathbf{1}_{\mathcal{Z}_\beta^c}(X)] =: -\tilde{\rho}_{\mathcal{Z}^c}(\beta)$$

At $\beta = 0$, several simplifications occur. The clipped set $\mathcal{Z}_0$ is empty, and the active set $\mathcal{Z}_0^c$ is the entire domain $\mathcal{D}$. The utility becomes $\bar{u}_0(x) = u(x)$. The score becomes $s_0(x) = u(x)/\mu_0$. The derivatives simplify to $\partial_\beta \bar{u}_\beta(x)|_{\beta=0} = -\tilde{\rho}(x)$ and $\partial_\beta \mu_\beta|_{\beta=0} = -\mathbb{E}[\tilde{\rho}(X)]$. This makes,

$$s'_0(x) = \frac{(-\tilde{\rho}(x))\mu_0 - u(x)(-\mathbb{E}[\tilde{\rho}(X)])}{\mu_0^2} = \frac{1}{\mu_0}\left(-\tilde{\rho}(x) + \frac{u(x)}{\mu_0}\mathbb{E}[\tilde{\rho}(X)]\right) = \frac{1}{\mu_0}\left(-\tilde{\rho}(x) + s_0(x)\mathbb{E}[\tilde{\rho}(X)]\right)$$

We now finally substitute this result back into the expression for the derivative of $J(\beta)$ at $\beta = 0$:

$$\left.\frac{dJ}{d\beta}\right|_{\beta=0} \propto -\mathbb{E}\left[\frac{w(X)(1-\tau)s'_0(X)}{g_0(X)^2}\right]$$

Plugging in $s'_0(x)$, $g_0(x)$ we get,

$$\left.\frac{dJ}{d\beta}\right|_{\beta=0} \propto -\frac{(1-\tau)}{\mathbb{E}[\tilde{u}(X)]}\mathbb{E}\left[\frac{w(X)}{(\frac{n_b}{n}\pi_0)^2}\left(-\tilde{\rho}(X) + \frac{\tilde{u}(X)}{\mathbb{E}[\tilde{u}(X)]}\mathbb{E}[\tilde{\rho}(X)]\right)\right]$$

Absorbing the positive constants into the proportionality sign and distributing the negative sign gives the final result: $\left.\frac{dJ}{d\beta}\right|_{\beta=0} \propto \mathbb{E}\left[\frac{w(X)}{\pi_0(X)^2}\left(\tilde{\rho}(X) - \frac{\tilde{u}(X)}{\mathbb{E}[\tilde{u}(X)]}\mathbb{E}[\tilde{\rho}(X)]\right)\right]$ This completes the proof of the lemma. □

Proof for 5.4. We begin with the asymptotic covariance of the curvature aware active M-estimator(Theorem 5.3), which is given by the formula:

$$\Sigma_\beta = H^{-1}V_\beta H^{-1}$$

Here, H is the Hessian of the true expected loss, which is independent of the sampling policy. Our analysis can therefore focus exclusively on the policy-dependent part, V_β, which is the covariance of the AIPW influence function. The influence function for a single data point i at the true parameter θ^* is:

$$g_i(\theta^*) := \nabla \ell^f_{\theta^*}(X_i) + \frac{\xi_i}{\pi_\beta(X_i)}\left(\nabla \ell_{\theta^*}(X_i, Y_i) - \nabla \ell^f_{\theta^*}(X_i)\right)$$

The covariance matrix is defined as $V_\beta = \mathbb{E}[g_i(\theta^*)g_i(\theta^*)^T]$. For simplicity, we define the gradient difference vector as $\tilde{\Delta}_i := \nabla \ell_{\theta^*}(X_i, Y_i) - \nabla \ell^f_{\theta^*}(X_i)$. Our influence function is now $g_i = \nabla \ell^f_{\theta^*} + \frac{\xi_i}{\pi_\beta}\tilde{\Delta}_i$.

We will use the Law of Total Expectation: $\mathbb{E}[Z] = \mathbb{E}_X[\mathbb{E}_{Y|X}[\mathbb{E}_{\xi|X,Y}[Z|\xi, X, Y]]]$. We take the expectation in three stages: first, over the randomness of sampling ξ, and second, over the randomness of the data (X, Y). We expand the outer product inside the expectation:

$$g_i g_i^T = (\nabla \ell^f_{\theta^*})(\nabla \ell^f_{\theta^*})^T + \frac{\xi_i}{\pi_\beta}(\nabla \ell^f_{\theta^*}\tilde{\Delta}_i^T + \tilde{\Delta}_i(\nabla \ell^f_{\theta^*})^T) + \frac{\xi_i^2}{\pi_\beta^2}(\tilde{\Delta}_i\tilde{\Delta}_i^T)$$

Now we take the first inner expectation over ξ. We know that for Bernoulli variable ξ: $\mathbb{E}[\xi_i|X_i, Y_i] = \pi_\beta(X_i)$ and $\mathbb{E}[\xi_i^2|X_i, Y_i] = \pi_\beta(X_i)$. Applying this:

$$\mathbb{E}[g_i g_i^T|X_i, Y_i] = (\nabla \ell^f_{\theta^*})(\nabla \ell^f_{\theta^*})^T + (\nabla \ell^f_{\theta^*}\tilde{\Delta}_i^T + \tilde{\Delta}_i(\nabla \ell^f_{\theta^*})^T) + \frac{\tilde{\Delta}_i\tilde{\Delta}_i^T}{\pi_\beta(X_i)}$$

The first two terms do not depend on our policy π. The only part that does is the last term. This gives us the policy-dependent part of V_β:

$$V_\beta^{\text{dep}} = \mathbb{E}_{X,Y}\left[\frac{\tilde{\Delta}_i\tilde{\Delta}_i^T}{\pi_\beta(X_i)}\right]$$

Next, we focus on taking expectation along $Y|X$ and then X

$$V_\beta^{\text{dep}} = \mathbb{E}_X\left[\mathbb{E}_{Y|X}\left[\frac{\tilde{\Delta}_i\tilde{\Delta}_i^T}{\pi_\beta(X_i)}\right]\right] = \mathbb{E}_X\left[\frac{\mathbb{E}_{Y|X}[\tilde{\Delta}_i\tilde{\Delta}_i^T]}{\pi_\beta(X_i)}\right]$$

Now, we connect this back to the trace of the full parameter covariance matrix. By the linearity of the trace, the policy-dependent part of $\text{Tr}(\Sigma_\beta)$ is:

$$\text{Tr}(\Sigma_\beta)^{\text{dep}} = \text{Tr}(H^{-1}V_\beta^{\text{dep}}H^{-1}) = \text{Tr}\left(H^{-1}\mathbb{E}\left[\frac{W(X)}{\pi_\beta(X)}\right]H^{-1}\right)$$

As expectation and trace are linear operators, they can be exchanged. Since $\pi_\beta(X)$ is a scalar, it can be factored out of the trace operation. Finally, by the cyclic property of the trace, we arrive at:

$$\text{Tr}(\Sigma_\beta)^{\text{dep}} = \mathbb{E}\left[\frac{1}{\pi_\beta(X)}\text{Tr}\left(H^{-1}W(X)H^{-1}\right)\right]$$

This successfully reduces the complex matrix problem to a scalar one. The policy-dependent part of our objective is $J(\beta) = \mathbb{E}[w_H(X)/\pi_\beta(X)]$, where $w_H(x) := \text{Tr}(H^{-1}W(x)H^{-1})$ is the scalar variance pressure.

With the reduction established, we can now analyze the derivative of $J(\beta)$ with respect to β. We apply Lemma C.2 with the scalar weight function $w(x) = w_H(x)$. The lemma states that the sign of $dJ/d\beta$ evaluated at $\beta = 0$ is determined by the sign of the following expectation:

$$\mathbb{E}\left[\frac{w_H(X)}{\pi_0(X)^2}\left(\tilde{\rho}(X) - \tilde{u}(X)\frac{\mathbb{E}[\tilde{\rho}(X)]}{\mathbb{E}[\tilde{u}(X)]}\right)\right]$$

To make this expression interpretable, we use the variance-spotlight measure $\omega(x)$:

$$\mathbb{E}[w_H(X)/\pi_0(X)^2] \cdot \mathbb{E}_\omega\left[\tilde{\rho}(X) - \tilde{u}(X)\frac{\mathbb{E}[\tilde{\rho}(X)]}{\mathbb{E}[\tilde{u}(X)]}\right]$$

The pre-factor, $\mathbb{E}[w_H(X)/\pi_0(X)^2]$, is an expectation of a non-negative quantity and is therefore positive. Thus, the derivative $dJ/d\beta$ is non-positive if and only if:

$$\mathbb{E}_\omega\left[\tilde{\rho}(X) - \tilde{u}(X)\frac{\mathbb{E}[\tilde{\rho}(X)]}{\mathbb{E}[\tilde{u}(X)]}\right] \leq 0$$

By linearity of expectation, this is equivalent to the condition stated in the theorem:

$$\frac{\mathbb{E}_\omega[\tilde{\rho}(X)]}{\mathbb{E}_\omega[\tilde{u}(X)]} \leq \frac{\mathbb{E}[\tilde{\rho}(X)]}{\mathbb{E}[\tilde{u}(X)]} \tag{6}$$

But, by construction of CA-ASI(Section 3), the scores $\tilde{u}(X)$ and $\tilde{\rho}(X)$ are normalized such that

$$\sum_{i=1}^n \tilde{u}(X_i) = \sum_{i=1}^n \tilde{\rho}(X_i) = 1.$$

Therefore, taking the expectation with respect to the empirical measure $\mathbb{P}_n$ (uniform over the pool), we have

$$\mathbb{E}_{\mathbb{P}_n}[\tilde{u}(X)] = \frac{1}{n}\sum_{i=1}^n \tilde{u}(X_i) = \frac{1}{n}, \qquad \mathbb{E}_{\mathbb{P}_n}[\tilde{\rho}(X)] = \frac{1}{n}\sum_{i=1}^n \tilde{\rho}(X_i) = \frac{1}{n}.$$

Consequently, $\frac{\mathbb{E}[\tilde{\rho}(X)]}{\mathbb{E}[\tilde{u}(X)]}$ becomes unity. This makes Equation (6),

$$\mathbb{E}_\omega[\tilde{\rho}(X)] \leq \mathbb{E}_\omega[\tilde{u}(X)] \tag{7}$$

If this condition holds, then $d(\text{Tr}(\Sigma_\beta))/d\beta \leq 0$ at $\beta = 0$. By the definition of the derivative, this implies that there exists some $\beta_0 > 0$ for which $\text{Tr}(\Sigma_\beta) \leq \text{Tr}(\Sigma_0)$ for all $0 < \beta < \beta_0$. This completes the proof.

□

Proposition C.3 (Upper Bound on Label Usage in CA-ASI)**.** *Let $\tilde{u}(x_i), \tilde{\rho}(x_i) \in [0,1]$ be normalized uncertainty and redundancy scores for each $x_i \in \mathcal{D}$. Define the geometry-corrected utility and ϵ clipped score:*

$$\tilde{u}_\beta(x_i) := \tilde{u}(x_i) - \beta \cdot \tilde{\rho}(x_i)$$

$$\bar{u}_\beta(x_i) := [\tilde{u}_\beta(x_i)]_{[0,1]} := \min\left\{\max\left\{\tilde{u}_\beta(x_i), 0\right\}, 1\right\}$$

Let $\hat{\mu}_\beta := \min\left\{\max\left\{\frac{1}{n}\sum_{j=1}^n \bar{u}_\beta(x_j), \epsilon\right\}, 1\right\}$ for some $\epsilon > 0$, $\mu_\beta = \mathbb{E}[\bar{u}_\beta(X)] = \mathbb{E}\left[\max\{\min\{\tilde{u}(X) - \beta\tilde{\rho}(X), 1\}, 0\}\right]$, and define the τ-mixed sampling probability:

$$\pi_\beta^{(\tau)}(x_i) := (1-\tau) \cdot \left[\frac{\bar{u}_\beta(x_i)}{\mu_\beta} \cdot \frac{n_b}{n}\right] + \tau \cdot \frac{n_b}{n}$$

Then:

1. *Let the clipped region be defined as*

$$\mathcal{Z}_\beta := \{x_i \in \mathcal{D} \mid \bar{u}_\beta(x_i) = 0\} = \left\{x_i \in \mathcal{D} \mid \beta > \frac{\tilde{u}(x_i)}{\tilde{\rho}(x_i)}\right\} \quad \textit{(when } \tilde{\rho}(x_i) > 0\textit{)}$$

   Then the expected number of labeled points is upper bounded as:

$$\mathbb{E}[n_{\text{lab}}] \leq \frac{\tau n_b}{n} \cdot |\mathcal{Z}_\beta| + \frac{n_b}{n}(\frac{1-\tau}{\epsilon} + \tau) \cdot (n - |\mathcal{Z}_\beta|)$$

2. *Let $\psi(x_i) := \frac{\tilde{u}(x_i)}{\tilde{\rho}(x_i)}$ for all x_i with $\tilde{\rho}(x_i) > 0$, and define the empirical CDF:*

$$\hat{F}_\psi(\beta) := \frac{1}{n}\sum_{i=1}^n \mathbb{I}\{\psi(x_i) \leq \beta\}$$

   Then the clipped region size satisfies: $\frac{|\mathcal{Z}_\beta|}{n} = \hat{F}_\psi(\beta)$, and the bound becomes:

$$\mathbb{E}[n_{\text{lab}}] \leq n_b \cdot \left((\frac{1-\tau}{\epsilon}) \cdot (1 - \hat{F}_\psi(\beta)) + \tau\right)$$

Proof. We begin by writing the expected label count:

$$\mathbb{E}[n_{\text{lab}}] = \sum_{i=1}^n \pi_\beta^{(\tau)}(x_i) = \sum_{i \in \mathcal{Z}_\beta} \pi_\beta^{(\tau)}(x_i) + \sum_{i \notin \mathcal{Z}_\beta} \pi_\beta^{(\tau)}(x_i)$$

From the sampling rule, if $x_i \in \mathcal{Z}_\beta$, then $\bar{u}_\beta(x_i) = 0$, so:

$$\pi_\beta^{(\tau)}(x_i) = \tau \cdot \frac{n_b}{n}, \quad \forall i \in \mathcal{Z}_\beta$$

Else, for $x_i \notin \mathcal{Z}_\beta$, the clipping and normalization ensure:

$$\pi_\beta^{(\tau)}(x_i) \leq (1-\tau) \cdot \frac{1}{\epsilon} \cdot \frac{n_b}{n} + \tau \cdot \frac{n_b}{n} = \frac{n_b}{n}(\frac{1-\tau}{\epsilon} + \tau)$$

Hence:

$$\mathbb{E}[n_{\text{lab}}] \leq \frac{\tau n_b}{n} \cdot |\mathcal{Z}_\beta| + \frac{n_b}{n}(\frac{1-\tau}{\epsilon} + \tau) \cdot (n - |\mathcal{Z}_\beta|)$$

Finally, we note that the expression

$$\mathcal{Z}_\beta = \{x_i \mid \beta > \psi(x_i)\} \quad \text{implies} \quad \frac{|\mathcal{Z}_\beta|}{n} = \hat{F}_\psi(\beta)$$

by definition of the empirical CDF. □

Geometry-Aware Suppression via β:

Let $C_\beta := \left(\frac{1-\tau}{\epsilon} \cdot (1 - \hat{F}_\psi(\beta)) + \tau\right)$ denote the label usage multiplier in the upper bound

$$\mathbb{E}[n_{\text{lab}}] \leq n_b \cdot C_\beta.$$

Then:

1. Monotonic Suppression: Since $\hat{F}_\psi(\beta)$ is the empirical CDF of $\psi(x_i) := \tilde{u}(x_i)/\tilde{\rho}(x_i)$, it is monotonically increasing in β. As β increases, more points have $\tilde{u}_\beta(x_i) = \tilde{u}(x_i) - \beta\tilde{\rho}(x_i) < 0$ and are clipped to zero. Hence, $\hat{F}_\psi(\beta) \to 1$, and

   $$\lim_{\beta \to \infty} C_\beta = \tau.$$

   The bound tightens to $n_b \cdot \tau$ in the limit, as most points are suppressed deterministically, and only a uniform τ-fraction are queried.
2. Threshold for Strict Suppression: To ensure strict suppression, i.e. $C_\beta < 1$, it suffices that

   $$\hat{F}_\psi(\beta) > 1 - \epsilon.$$

Higher Coherence Reduces Label Usage: Let $\sigma_\mathcal{D} := \min_i \tilde{\rho}(x_i)$ denote the dataset-level coherence. Define the score function:

$$\psi(x_i) := \frac{\tilde{u}(x_i)}{\tilde{\rho}(x_i)}$$

Then for all $x_i \in \mathcal{D}$, we have

$$\psi(x_i) \leq \frac{1}{\sigma_\mathcal{D}}$$

Letting $\beta = \frac{1}{\sigma_\mathcal{D}}$, we get

$$\hat{F}_\psi(\beta) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{I}\{\psi(x_i) \leq \beta\} = 1$$

and hence the expected number of labeled points under CA-ASI satisfies:

$$\mathbb{E}[n_{\text{lab}}] \leq n_b \cdot \left(\frac{1-\tau}{\epsilon} \cdot (1 - \hat{F}_\psi(\beta)) + \tau\right) = n_b \cdot \tau$$

Thus, higher coherence $\sigma_\mathcal{D}$ implies a smaller threshold for a larger clipping set $\mathcal{Z}_\beta$, and hence fewer labels used in expectation. Consider dataset $\mathcal{D}$ having n identical points. A short calculation shows us $\tilde{\rho}(x_i) = \frac{1}{n} \forall i \in [n]$. This implies, $\sigma_\mathcal{D} = \frac{1}{n}$. Further, from our discussion above, on setting $\beta = n$, we notice as dataset size increases i.e. as $n \to \infty$, pruning becomes increasingly aggressive, leading to fewer points being labelled in expectation.

Proof for 5.5. The proof directly parallels that of Theorem 6.2 in Zrnić and Candès [80]. The increments of our estimator's objective function form a martingale difference sequence. The introduction of our geometry-aware policy $\pi_{\beta,t}$ only changes the definition of the conditional covariance $V_{\theta^*,t}(\beta)$ at each step. The core conditions required for the Martingale Central Limit Theorem to hold are the convergence of the average conditional covariance and the Lindeberg condition. By assuming these conditions are met, the result follows directly from the application of the MCLT. □

Proof for 5.6. The proof extends the logic of the batch setting (Theorem 5.4) to the sequential case by analyzing the effect of the penalty parameter β on the conditional variance at each time step. The total asymptotic variance is determined by the limit of the time-average of these conditional variances.

Recall, from Theorem 5.5, the asymptotic covariance of the sequential estimator is $\Sigma_\beta = H^{-1}V_\beta^* H^{-1}$, where V_β^* is the limit of the time-average of the conditional covariance matrices:

$$V_\beta^* = \lim_{n\to\infty} \frac{1}{n}\sum_{t=1}^{n} V_t(\beta), \quad \text{where} \quad V_t(\beta) := \mathrm{Var}(\nabla \Delta_t(\theta^*)|\mathcal{F}_{t-1})$$

We are interested in the trace of this matrix. By the linearity of the trace, limit, and summation operators, we have:

$$\mathrm{Tr}(\Sigma_\beta) = \lim_{n\to\infty} \frac{1}{n}\sum_{t=1}^{n} \mathrm{Tr}(H_t^{-1}V_t(\beta)H_t^{-1})$$

where we allow for a time-varying Hessian H_t that is predictable with respect to the filtration $\mathcal{F}_{t-1}$. This reduces our problem to analyzing the expected value of the per-step objective function, $J_t(\beta) := \mathrm{Tr}(H_t^{-1}V_t(\beta)H_t^{-1})$.

At any given time step t, all quantities related to the model (f_t) and policy ($\pi_{\beta,t}$) are fixed, as they are measurable with respect to the history $\mathcal{F}_{t-1}$. The structure of the conditional variance $V_t(\beta)$ is therefore identical to that of the batch variance, with all expectations replaced by conditional expectations given $\mathcal{F}_{t-1}$. Following the same derivation as in the proof of Theorem 5.4, we can express the policy-dependent part of our per-step objective as:

$$J_t(\beta) \propto \mathbb{E}\left[\frac{w_{H,t}(X)}{\pi_{\beta,t}(X)} \;\middle|\; \mathcal{F}_{t-1}\right]$$

where $w_{H,t}(x)$ is the conditional scalar variance pressure defined in the theorem statement.

Lemma C.2 holds equally well when all expectations are made conditional on a filtration. Applying a conditional version of Lemma C.2 to the objective $J_t(\beta)$, we find that the derivative with respect to β, evaluated at $\beta = 0$, is proportional to:

$$\left.\frac{dJ_t}{d\beta}\right|_{\beta=0} \propto \mathbb{E}\left[\frac{w_{H,t}(X)}{\pi_{0,t}(X)^2}\left(\tilde{\rho}_t(X) - \frac{\tilde{u}_t(X)}{\mathbb{E}[\tilde{u}_t(X)|\mathcal{F}_{t-1}]}\mathbb{E}[\tilde{\rho}_t(X)|\mathcal{F}_{t-1}]\right) \;\middle|\; \mathcal{F}_{t-1}\right] \tag{8}$$

The derivative is non-positive if the conditional expectation is non-positive. Following the same rearrangement as in the proof of Theorem 5.4, this occurs if the condition stated in the theorem holds for the specific step t.

Asymptotically, the running normalized statistics stabilize and converge to their population counterparts, implying,

$$\frac{\mathbb{E}[\tilde{\rho}_t(X) \mid \mathcal{F}_{t-1}]}{\mathbb{E}[\tilde{u}_t(X) \mid \mathcal{F}_{t-1}]} \xrightarrow{p} \frac{1}{1} = 1$$

.

This is similar to the observation in batch setting(Equation 7 in Theorem 5.4).

The theorem assumes that the condition for variance reduction holds in expectation over time. This implies that, on average, the derivative of the per-step objective from Eq. 8 is non-positive:

$$\mathbb{E}_t\left[\left.\frac{dJ_t}{d\beta}\right|_{\beta=0}\right] \leq 0$$

Since $\mathrm{Tr}(\Sigma_\beta)$ is the limit of the average of $J_t(\beta)$, a non-positive average derivative at $\beta = 0$ implies that there must exist some $\beta_0 > 0$ such that for all $0 < \beta < \beta_0$, the total objective is reduced. Thus, $\mathrm{Tr}(\Sigma_\beta) \leq \mathrm{Tr}(\Sigma_0)$. This completes the proof.

□

Finite-Sample Time-Uniform Confidence Ellipsoids for CA-ASI

While asymptotic results establish long-term validity and efficiency(Theorem 5.5), a key advantage of sequential methods is the ability to provide guarantees that hold at any finite time t. In this section, we develop a non-asymptotic, time-uniform confidence ellipsoid for the parameter θ^*.

Theorem C.4 (Finite-Sample Time-Uniform Confidence Ellipsoids)**.** *Let $\ell_\theta(x, y)$ be a convex, twice-differentiable loss function. Define the true parameter $\theta^* = \arg\min_\theta \mathbb{E}[\ell_\theta(X, Y)]$. Assume:*

1. *The stochastic gradient process $g_s := \nabla_\theta \ell_{\theta^*}(X_s, Y_s)$ forms a martingale difference sequence with respect to the filtration $\mathcal{F}_s := \sigma((X_i, Y_i, \xi_i)_{i=1}^s)$, and $\|g_s\|_2 \le c$ almost surely.*
2. *Let A_t be the notation for the unregularized sum of gradient covariance matrix: $A_t = \sum_{s=1}^t \xi_s g_s g_s^\top$. Consequently we define the regularized gradient covariance matrix $V_t := \lambda I + A_t$ for some $\lambda > 0$.*
3. *Information Matrix Equivalence: In a neighborhood $\mathcal{N}$ around θ^*, the empirical Hessian is lower-bounded by the gradient outer products such that $\sum_{s=1}^t \xi_s \nabla^2 \ell_{\tilde\theta}(X_s, Y_s) \succeq \mu V_t$ for some constant $\mu > 0$ and all $\tilde\theta \in \mathcal{N}$.*

Then, for any $\alpha \in (0, 1)$, with probability at least $1 - \alpha$, the CA-ASI estimator $\hat\theta_t$ satisfies the following uniformly for all $t \ge 1$:

$$(\hat\theta_t - \theta^*)^\top V_t (\hat\theta_t - \theta^*) \le \frac{1}{\mu^2} \left(\sqrt{\lambda}\|\theta^*\|_2 + c\sqrt{2\log\left(\frac{\det(V_t)^{1/2}\det(\lambda I)^{-1/2}}{\alpha}\right)} \right)^2.$$

Proof. By the first-order optimality condition of the empirical risk, $\sum_{s=1}^t \xi_s \nabla_\theta \ell_{\hat\theta_t}(X_s, Y_s) = 0$. Using a first-order Taylor expansion of the gradient around θ^*, we have:

$$\sum_{s=1}^t \xi_s \Big(g_s + H_s(\hat\theta_t - \theta^*) \Big) = 0,$$

where $H_s = \nabla^2 \ell_{\tilde\theta_s}(X_s, Y_s)$ for some $\tilde\theta_s$ on the line segment between $\hat\theta_t$ and θ^*. Rearranging this yields:

$$\left(\sum_{s=1}^t \xi_s H_s \right) (\hat\theta_t - \theta^*) = -\sum_{s=1}^t \xi_s g_s.$$

By Assumption 3, $\sum \xi_s H_s \succeq \mu V_t$. Multiplying both sides by $V_t^{-1/2}$ and taking the ℓ_2 norm, we obtain:

$$\mu \|V_t^{1/2}(\hat\theta_t - \theta^*)\|_2 \le \left\| \sum_{s=1}^t \xi_s g_s \right\|_{V_t^{-1}}.$$

Assuming the gradient process is σ-sub-Gaussian with $\sigma = c$, with probability at least $1 - \alpha$, we can bound this self-normalized sum uniformly [1] over all $t \ge 1$ by:

$$\left\| \sum_{s=1}^t \xi_s g_s \right\|_{V_t^{-1}} \le \sqrt{\lambda}\|\theta^*\|_2 + c\sqrt{2\log\left(\frac{\det(V_t)^{1/2}\det(\lambda I)^{-1/2}}{\alpha}\right)} =: R_t(\alpha).$$

Squaring both sides and substituting $\|V_t^{1/2}(\hat\theta_t - \theta^*)\|_2^2 = (\hat\theta_t - \theta^*)^\top V_t (\hat\theta_t - \theta^*)$ yields:

$$(\hat\theta_t - \theta^*)^\top V_t (\hat\theta_t - \theta^*) \le \frac{R_t(\alpha)^2}{\mu^2}.$$

This defines a valid time-uniform confidence ellipsoid where the geometry is strictly governed by the accumulated gradient covariance A_t. □

Corollary C.5 (Time-Uniform ℓ_1 Confidence Bound and Coordinate Intervals)**.** *Under the assumptions of Theorem C, with probability at least $1 - \alpha$, the following hold for all $t \ge 1$:*

$$\left\| \hat\theta_t - \theta^* \right\|_1 \le \frac{\sqrt{R_t(\alpha)}}{\mu} \cdot \left\| \mathrm{diag}(V_t^{-1}) \right\|_1^{1/2},$$

and for each coordinate $j \in [d]$,

$$\theta_j^* \in \left[\hat{\theta}_{t,j} \pm \frac{\sqrt{R_t(\alpha)}}{\mu} \cdot \sqrt{(V_t^{-1})_{jj}}\right].$$

Claim C.6. *Spectral Truncation via Coherence Penalty Lemma under Same Label Usage. (Spectral Truncation via Coherence Penalty). Let V_{t-1} be the accumulated gradient covariance matrix at time $t-1$, with eigendecomposition $V_{t-1} = \sum_{i=1}^{d} \lambda_i v_i v_i^\top$, where $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_d$. Assume the coherence penalty $\tilde{\rho}(x_t)$ strictly penalizes alignment with the dominant eigenspace of V_{t-1}. Under the CA-ASI policy, a candidate point is queried if the penalized utility $\tilde{u}_\beta(x_t) = \tilde{u}(x_t) - \beta\tilde{\rho}(x_t) > 0$. If a candidate gradient g_t is highly collinear with the top eigenvector v_1, its coherence penalty is maximized. For any fixed utility $\tilde{u}(x_t)$, there exists a threshold β_0 such that for all $\beta > \beta_0$:*

$$\beta\tilde{\rho}(x_t) \geq \tilde{u}(x_t)$$

Consequently, the point is deterministically suppressed ($\xi_t = 0$), preventing any rank-1 update in the direction of v_1. For a fixed sampling budget B, this truncation strictly upper-bounds the growth of $\lambda_1(V_t)$, forcing the reallocation of the remaining trace budget to strictly increase the smaller eigenvalues $\lambda_d(V_t)$

Proof. When both policies used the same number of labeled samples, B, both generated covariance matrices exhibit same expected trace: $\mathbb{E}[\mathrm{Tr}(V_t)] = d\lambda + cB$. The proof follows directly from the definition of the CA-ASI suppression threshold; deterministically CA-ASI suppresses highly coherent updates. Because the rank-1 update to V_t is entirely dictated by the binary inclusion variable $\xi_t \in \{0, 1\}$, by Weyl's Inequality bounding $\xi_t = 0$ in the direction of v_1 ensures $\lambda_1(V_t) = \lambda_1(V_{t-1})$. Consequently, to reach the same label budget B, CA-ASI must select labels from a more diverse set of directions, strictly bounding the growth of λ_1 and forcing the inflation of the smaller eigenvalues $\lambda_2, \ldots, \lambda_d$. □

Proposition C.7 (Volume Reduction via Curvature Penalty Under Same Label Usage)**.** *Let $\mathcal{E}_t(\beta)$ denote the confidence ellipsoid from Theorem C.4 generated by the CA-ASI sampling policy $\pi_{\beta,t}$ with a fixed sampling budget $\mathbb{E}[\sum_{s=1}^{t} \xi_s] = B$.*

$$\mathcal{E}_t(\beta) = \{\theta \in \mathbb{R}^d : (\hat{\theta}_t - \theta)^\top V_t(\hat{\theta}_t - \theta) \leq R_t(\alpha)^2/\mu^2\}$$

Setting the curvature penalty $\beta > 0$ strictly reduces the expected volume of $\mathcal{E}_t(\beta)$ compared to standard active sampling ($\beta = 0$).

Proof. The volume of the d-dimensional ellipsoid $\mathcal{E}_t$ is given by:

$$\mathrm{Vol}(\mathcal{E}_t) = \frac{\pi^{d/2}}{\Gamma(d/2+1)} \det(V_t)^{-1/2} \left(\frac{R_t(\alpha)}{\mu}\right)^d.$$

Note that $R_t(\alpha) = \mathcal{O}(\sqrt{\log\det(V_t)})$, implying the radius term $R_t(\alpha)^d$ grows only polylogarithmically with respect to $\det(V_t)$, while the term $\det(V_t)^{-1/2}$ shrinks polynomially. Consequently, for all t such that the confidence bound is non-vacuous, the volume $\mathrm{Vol}(\mathcal{E}_t)$ is monotonically decreasing with respect to $\det(V_t)$. Thus, minimizing the volume is equivalent to maximizing $\det(V_t)$.

By the inequality of arithmetic and geometric means (AM-GM), for a fixed trace, the product of the eigenvalues (the determinant) is maximized when the eigenvalues are equal:

$$\det(V_t) = \prod_{i=1}^{d} \lambda_i \leq \left(\frac{1}{d}\sum_{i=1}^{d} \lambda_i\right)^d = \left(\frac{\mathrm{Tr}(V_t)}{d}\right)^d.$$

Under standard uncertainty sampling ($\beta = 0$), the policy repeatedly samples points near the decision boundary. These points exhibit high gradient coherence (redundancy), causing V_t to become ill-conditioned (for eg., one strictly dominant eigenvalue and $d-1$ small eigenvalues).

From Claim C.6, setting $\beta > 0$ deterministically suppresses updates along the dominant eigenvectors when coherence is high, strictly capping $\lambda_1(V_t^{(\beta>0)}) < \lambda_1(V_t^{(\beta=0)})$. Under the trace conservation

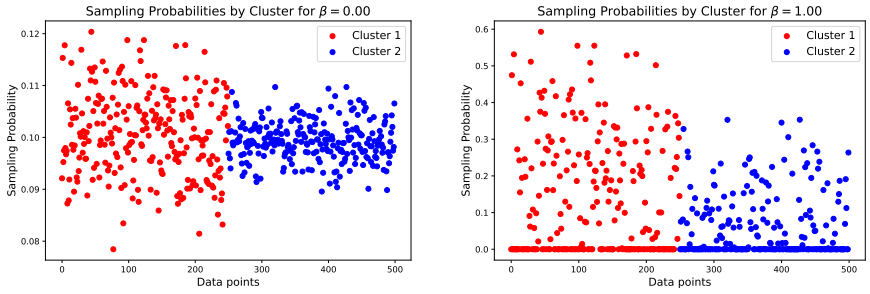

Figure 7: Distinguishable Sampling using Curvature Information.

property, this truncation forces the reallocation of the sampling budget to the orthogonal directions, strictly increasing the smaller eigenvalues $\lambda_i(V_t^{(\beta>0)}) > \lambda_i(V_t^{(\beta=0)})$ for $i \in [2, \cdots, d]$.

Because the eigenvalue spectrum under CA-ASI is strictly more uniform (closer to the AM-GM maximization condition) than the spectrum under standard sampling, it strictly follows that $\mathbb{E}[\det(V_t^{(\beta>0)})] > \mathbb{E}[\det(V_t^{(\beta=0)})]$. Consequently, the expected volume of the confidence ellipsoid strictly decreases. □

D Additional Experimental Results

D.1 Synthetic Example

D.1.1 Impact of β on Labelling Probabilities

Under the sample experimental settings as in 6 under Section 6, we examine every candidate's sampling probabilities. From Figure 7, when redundancy awareness is absent, both clusters exhibit a remarkably similar sampling probability distribution, indicating no favoritism. However, when redundancy awareness becomes effective, selection becomes *targeted*. Under ASI, the sampling–probability gap between $\mathcal{C}_1$ and $\mathcal{C}_2$ is small (max: 0.010, mean: 0.001). With CA-ASI, these gaps increase substantially (max: 0.240, mean: 0.085)—a $24\times$ rise in the peak separation and an $85\times$ rise in the average separation. This widened separation shows that CA-ASI leverages data coherence to concentrate labels on the more informative cluster under the black-box model f.

D.1.2 Parameter Tuning of β

The pruning parameter β is arguably the most important parameter for CA-ASI. We suggest two approaches for tuning β.

Automatic Schedule: Mathematically model β to be equal to some monotonic function of the budget fraction i.e $c = \frac{n_b}{n}$. [3] For practical purposes, one could use a power-odds schedule[3]. For some $a > 0$, $\beta = \frac{c^a}{1-c^a}$, which is strictly increasing in c, with $\lim_{c\to 0^+} \beta = 0$ and $\lim_{c\to 1^-} \beta = \infty$. a essentially works as a knob controlling the decay; for a fixed c, values close to 0 make β large($\lim_{a\to 0^+} \beta = \infty$), whereas on increasing a, β becomes smaller, making pruning less aggressive.

Data-driven Approach: Recall,

$$\psi(x) \;=\; \frac{\tilde{u}(x)}{\tilde{\rho}(x)}.$$

Let $\widehat{F}_\psi$ be the empirical CDF of $\psi(x)$ over the pool. Since a point survives iff $\psi(x) > \beta$, the resulting keep rate is

$$\text{keep}(\beta) \;\approx\; 1 - \widehat{F}_\psi(\beta).$$

[3] In many systems, having a larger budget (in terms of data, computational power, or financial resources) grants the luxury of imposing higher standards. When we have more redundancy, we can be less tolerant of marginal or noisy components, choosing to prune them aggressively to maintain overall quality or efficiency. We can downvote more readily because we trust that sufficient high-quality alternatives remain within the larger budget. Applications of this intuition are seen in quality control and resource allocation [24, 54].

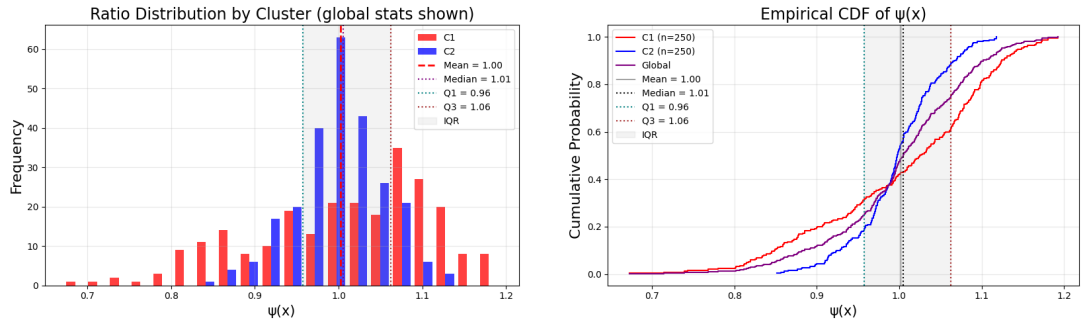

Figure 8: Data-driven tuning of β

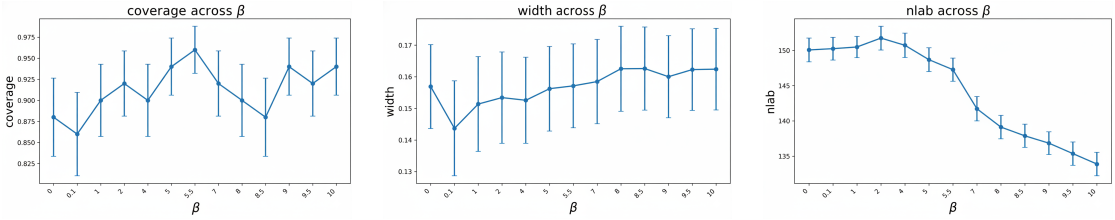

Figure 9: Grid Tuning of β in Batch Setting

We can therefore target a desired survivor fraction $s \in (0, 1]$ by choosing

$$\boxed{\beta \;=\; \widehat{F}_{\psi}^{-1}(1-s).} \tag{9}$$

We notice from Figure 8, the mass of $\psi(x)$ concentrates near 1. Hence $\beta = \text{median}(\psi)$ prunes $\approx 50\%$ of the pool; moving β toward the right tail (e.g., $Q_{0.75}$ or $Q_{0.9}$) prunes $\approx 75\%$ or 90%, respectively. Conversely, setting β below the first quartile produces light pruning. The CDF panel gives an immediate, data-driven way to pick β for any desired keep rate at the current pool, without use of any labeled data. For all our experiments in Section 6, we use grid-search combined with data-driven approach over validation data to judiciously set the value of β.

Sensitivity Analysis and the *Sweet Spot* for β: Figures 9 and 10 present a comprehensive grid search on the Housing Dataset over the curvature penalty parameter β in both batch and sequential settings. In the batch setting, we observe that for moderate pruning levels ($\beta \lesssim 7$), CA-ASI maintains coverage and confidence interval widths comparable to the ASI baseline ($\beta = 0$) while achieving progressive label reduction. However, beyond this threshold, the variance reduction guarantee breaks down: as β drives sampling probabilities $\pi_\beta(x) = [\tilde{u}(x) - \beta\tilde{\rho}(x)]_{[0,1]}$ toward zero for an increasing fraction of the pool, the inverse propensity weights $1/\pi_\beta(x)$ in our estimator (Eq. (3)) explode, inflating the asymptotic covariance Σ_β and widening confidence intervals. A similar pattern emerges in the sequential setting, where aggressive pruning beyond the optimal range degrades statistical efficiency despite reducing annotation volume. These empirical results precisely confirm the existence of the theoretical boundary β_0 predicted in Theorems 5.4 and 5.6—there exists a finite range $0 < \beta < \beta_0$ where curvature-aware pruning strictly improves upon uncertainty-only sampling, but beyond which the algorithm becomes overly aggressive, over-pruning informative points and destabilizing the propensity-weighted estimator.

D.2 Experimental Design

D.2.1 Training Details

For both our real-world dataset setups, we train XGBoost [15] as our black box model on 50% of the dataset and perform inference on the remaining half [80]. We keep our label budget values between 1% to 10%. To ensure the density(Eq. (1)) reflects structural redundancy rather than feature scaling artifacts, we compute it on the standardized features (zero mean, unit variance) of the labeled set.

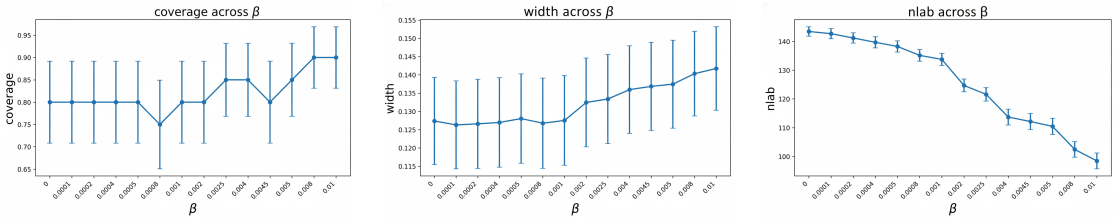

Figure 10: Grid Tuning of β in Sequential Setting

We set r to 0.2 and 1 for HTRU2 and Housing datasets respectively. We optimize over chosen loss using linear predictors that include main effects and an intercept($d = 9$). For simplicity sake, we report confidence-intervals(CIs) and coverage for the first parameter with $\alpha = 0.1$. We repeat our analysis for 500 and 80 trials for batch and sequential settings respectively. We use `L-BFGS-B` [78] to calculate $\hat{\theta}$ where closed form solution does not exist. All experiments were run on a single RTX A6000.

D.2.2 Robustness of Experimental Setup

Evaluation: Seminal ML-powered inference works like PPI and ASI [4, 81] evaluate estimator quality with standard metrics like coverage and width. For the latter, since the focus is much more on adaptive data collection for experimental design, assessing spatial quality of the acquired labels is imperative. In fact, no prior work diagnoses the geometric structure of the acquired dataset itself. In our setup, we categorically diagnose the labeled set directly through the density (Eq. (1)) of selected points in standardized feature space. Density is not a formal inference objective, but it is a revealing proxy for data-collection health: in our standardized features, even modest density reductions reflect large non-redundant geometric expansions in the input manifold. We report it as a diagnostic companion to our statistical results. We also demonstrate the effectiveness of curvature-aware redundancy pruning beyond previous simple setups like linear regression, simple logistic regression; our experimental setup also validates across a broader and more challenging set of GLMs and loss surfaces.

Dataset Choices: The datasets we consider reflect settings where obtaining high-quality labels is non-trivial. First, median house prices for atypical properties (coastal cliffs, fire-prone hillsides, historic designations) require licensed appraisers conducting physical inspections, with fees that scale with property complexity [10]. Moreover, valuations/predictions in this domain inform high-stakes decisions such as mortgage lending, taxation, investment, and urban planning [11, 49, 22]. Automated valuation models (AVMs) are well-documented to perform reliably on standard properties but degrade precisely in the high-leverage tails of the distribution, where expert appraisal is both most needed and most expensive [61, 52]. This motivates selective label acquisition that concentrates budget on geometrically isolated, high-uncertainty census tracts.

Second, pulsars are valuable to cosmic laboratories for detecting gravitational waves and galaxy mapping [55]. Real pulsars are rare and labeling them is bottlenecked by telescope follow-up time and time for manual classification of ambiguous signals [43, 20].

Third, for rare large diamonds (high carat, unusual cuts, exceptional clarity grades), precise valuation requires time consuming and costly GIA-certified gemologist appraisal or auction-house assessment [51]. Economically, the cost of label acquisition thus scales directly with diamond rarity, which coincides precisely with the geometric structure of the dataset [58]: rare large diamonds occupy sparse, high-leverage regions of the carat–dimension manifold making them exactly the instances CA-ASI preferentially targets via its uncertainty score while avoiding redundant sampling of the densely clustered common-stone region via its coherence penalty.

Dimensionality: Our experimental evaluation represents a deliberate escalation over the established literature, where $d \leq 3$. Our challenge of evaluating on $d = 9$ evaluation does not lie in its raw data size, but in the M-estimation stability and geometric conditioning under extreme budget constraints.

As d increases, the number of support points required to make an estimator stable grows linearly, but the volume of the search space grows exponentially(R^d). Further, the risk of inverting rank-

deficient or ill-conditioned global Hessian, composed of highly redundant data, becomes pronounced as dimensionality increases. To verify this, we measure the Condition Number of the empirical variance covariance matrix $\hat{\Sigma}$, (κ) for $d = 9$ vs $d = 2$ for the California Housing Dataset at 1% budget. $\kappa_{d=9}(121 \times 10^{10})$ is 6×10^9 times larger than $\kappa_{d=2}(202.726)$. A large condition number in $d = 9$ indicates that the M-estimator is under much higher 'geometric stress'. The matrix is nearing rank-deficiency, making the resulting confidence intervals far more sensitive to noise and sparse sampling. By evaluating on $d = 9$, we are specifically testing the robustness of statistical inference in a regime where matrix conditioning and redundancy are primary bottlenecks—challenges that simply do not emerge in $d = 2$ settings.

To further test our setup, we compute the Range and Coefficient of Variance(CV)% for $\tilde{\rho}$ with $d = 2(0.0003, 49.39\%)$ vs $d = 9(0.0019, 78.93\%)$ for Housing Dataset. For $d = 9$ range is $5.3\times$ larger and CV% is 60% higher. This indicates at $d = 9$ redundancy scores capture heterogeneous, complex, and heavy-tailed behavior, eventually benefiting selective pruning. This is true because in datasets like California Housing, many of the features are pairwise highly correlated (e.g., `total_rooms`, `total_bedrooms`, `population`, `households`) [75]. This makes it ideal for our cause: CA-ASI is required here to enforce Hessian penalty to avoid sampling data in redundant manifold and find the rare points that provide information about the other dimensions of θ. HTRU2(0.007, 139%) also has multiple highly correlated features [67].

D.3 Data Coherence

In the section, we assess Coherence Matrix $\mathbf{S}$ for settings discussed in Section 6.

1. **Linear Regression (Squared Loss):** The loss is $\ell(\theta) = \frac{1}{2}(y - x^T\theta)^2$. The Hessian becomes $\nabla^2\ell(\theta) = xx^T$ (Constant, independent of θ and y). Since xx^T is rank-1, its square root is just x (up to scale). Therefore, $\hat{H}_i^{1/2} = x_i$
   Coherence Matrix Entry:
   $$S_{ij} = \|\hat{H}_i^{1/2}\hat{H}_j^{1/2}\|_F = |x_i^T x_j|$$
   Interpretation: For Linear Regression, Curvature Coherence is simply the magnitude of the dot product (cosine similarity if normalized). Highly correlated features imply High Redundancy.

2. **Binary Classification (Exponential Loss)**: The loss is $\ell(\theta) = e^{-yx^T\theta}$ (where $y \in \{-1, 1\}$). The Hessian becomes $\nabla^2\ell(\theta) = e^{-yx^T\theta}xx^T$ (because $y^2 = 1$). Hence, Square Root of Hessian is, $\hat{H}_i^{1/2} = \sqrt{e^{-\hat{y}_i \cdot x_i^T\hat{\theta}}} \cdot x_i = e^{-\frac{1}{2}\hat{y}_i \cdot {x_i}^T\hat{\theta}}x_i$, where $\hat{\theta} := \arg\min_\theta \frac{1}{n}\sum_{i=1}^n \ell_\theta(X_i, f(X_i))$.
   Coherence Matrix Entry:
   $$\begin{aligned} S_{ij} = \|\hat{H}_i^{1/2}\hat{H}_j^{1/2}\|_F &= e^{-\frac{1}{2}\hat{y}_i({x_i}^T\hat{\theta})}e^{-\frac{1}{2}\hat{y}_j({x_j}^T\hat{\theta})}|x_i^T x_j| \\ &= \exp\left(-\frac{1}{2}(\hat{y}_i{x_i}^T\hat{\theta} + \hat{y}_j{x_j}^T\hat{\theta})\right) \cdot |x_i^T x_j| \end{aligned}$$
   Interpretation: Redundancy is the dot product $|x_i^T x_j|$ weighted by the Margin. If points are far from the decision boundary (high margin $yf(x)$), the exponential term is tiny $\rightarrow$ $S_{ij} \approx 0$. If points are near the boundary or misclassified (low margin), the weight is large.

D.3.1 Design Choice: Why Hessian-style penalty?

Generic diversity regularizers (e.g., Euclidean distance) are designed to improve predictive accuracy by uniformly covering the input space. However, our goal is performing statistical inference using a combination of human labels and pseudolabels. Because parameter variance in M-estimation is governed by the generalized Fisher Information (the Hessian, H), *any generic feature-space metric is geometrically blind to the loss surface and misalignned with the the goal of pruning information-wise redundant points while constructing valid CIs*. In fact, our core theoretical results (Theorems 5.4 and 5.6) which prove the exact conditions for variance reduction are mathematically dependent on the trace of the inverse Hessian.

If two points are far apart in standard Euclidean space (high generic diversity) but are both deeply entrenched in a highly confident region, their Hessian overlap is near zero because they offer no new gradient information. A generic distance metric would waste the labeling budget sampling these points simply to “cover the space," whereas CA-ASI correctly ignores them.

To experimentally verify this, we construct $\mathbf{S}$ st. S_{ij} only looks at the features x, completely ignoring the model's loss gradients or predictions. We define:

$$S_{ij}^{div} = \exp\left(-\frac{\|x_i - x_j\|_2^2}{2\sigma^2}\right)$$

Consequently,

$$\tilde{\rho}_i^{div} = \frac{\sum_{j \neq i} S_{ij}^{div}}{\sum_{k=1}^{n} \sum_{j \neq k} S_{kj}^{div}}$$

Our sampling rule now becomes,

$$\tilde{u}_\beta^{div}(x) = \tilde{u}(x) - \beta \tilde{\rho}^{div}(x)$$

We test this new sampling policy under HTRU2 Dataset.

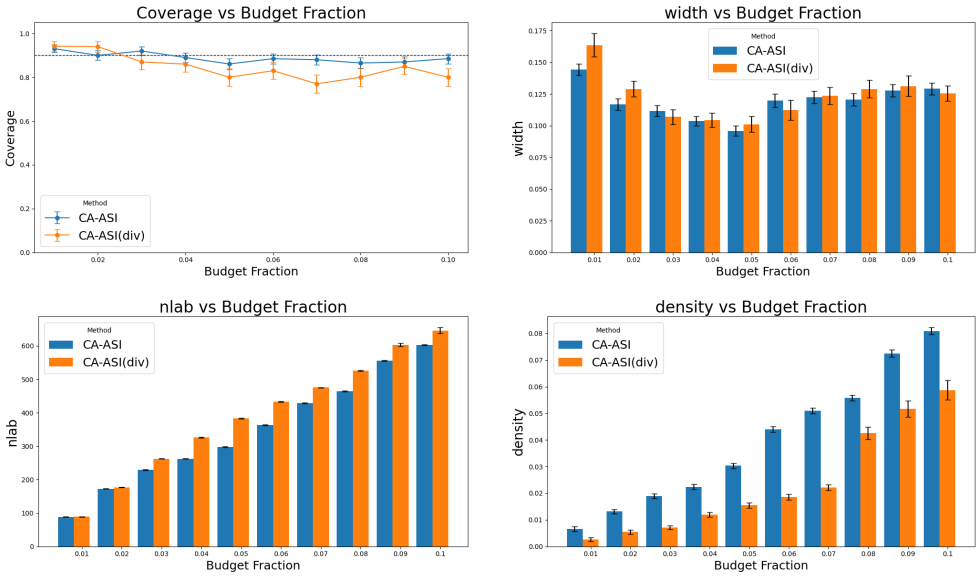

Figure 11: Inferior performance of diversity-based pruning on HTRU2 Dataset(Batch)

From Figures 11, 12 diversity-based penalization falls short in terms of coverage, employs substantially more labels, and has a wider CI width compared to redundancy-based penalization. However, it's worth noting that diversity-based penalization improves density, which aligns with its intended design.

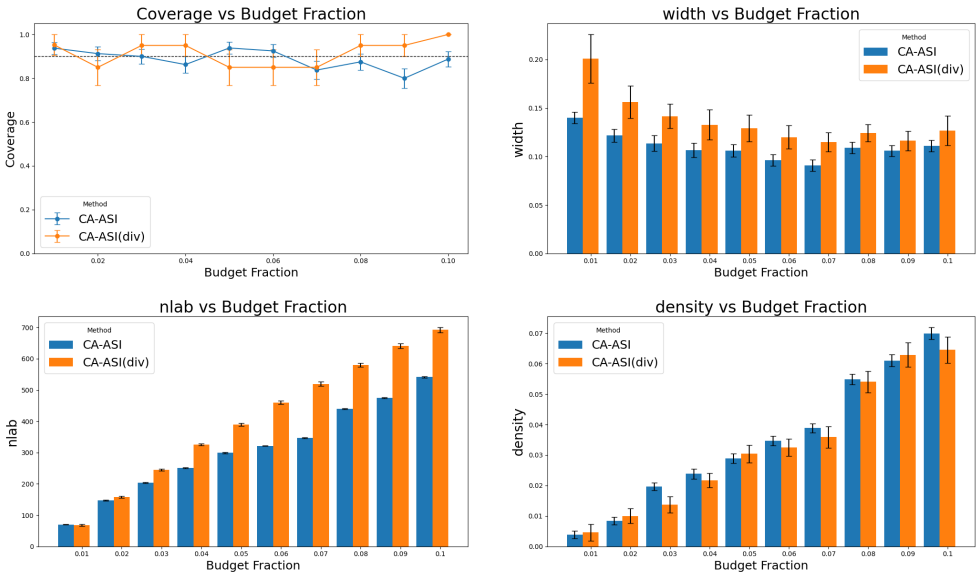

Figure 12: Inferior performance of diversity-based pruning on HTRU2 Dataset(Sequential)

D.3.2 Why does CA-ASI work ?

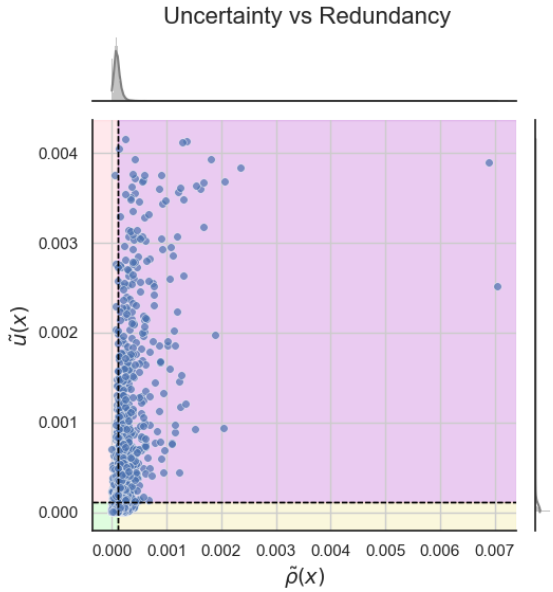

Figure 13: Mean Quadrant-wise Splitting of Inference Data

Figure 13 provides a joint scatterplot of normalized uncertainty $\tilde{u}(x)$ versus normalized redundancy $\tilde{\rho}(x)$, computed on the inference split of the HTRU2 dataset (see Section 6). The horizontal and vertical axes represent predictive uncertainty and redundancy respectively—both derived from the black-box model.

To enhance interpretability, we overlay kernel density estimates (KDEs) of the marginal distributions and segment the plane into four shaded quadrants using the empirical means of $\tilde{u}(x)$ and $\tilde{\rho}(x)$. These quadrants serve to characterize the data landscape and illustrate the key differences between ASI and CA-ASI sampling behavior:

1. **Quadrant I(Pink) – Top-Left (High Uncertainty, Low Redundancy)**
    - **Description:** Points that are informative (high uncertainty) yet non-redundant (low similarity with others). Close to 3% of the data.
    - **Interpretation:** Most valuable points for learning. These are prioritized by both ASI and CA-ASI, CA-ASI. This region is the core of active learning, representing optimal points for labeling.
2. **Quadrant II(Purple) – Top-Right (High Uncertainty, High Redundancy)**
    - **Description:** Points are uncertain, but also highly redundant (many similar neighbors). Close to 11% of the data.
    - **Interpretation:** ASI may still assign probability mass here, but CA-ASI heavily penalizes such points due to redundancy. Main target of pruning as redundancy increases—CA-ASI aims to save budget by avoiding repeated labeling of redundant information.
3. **Quadrant III(Green) – Bottom-Left (Low Uncertainty, Low Redundancy)**
    - **Description:** These points($\sim 63.67\%$) are both confidently predicted and non-redundant.
    - **Interpretation:** Typically not sampled under any method, as uncertainty is too low to begin with.
4. **Quadrant IV(Yellow) – Bottom-Right (Low Uncertainty, High Redundancy)**
    - **Description:** Confident predictions, but also highly redundant. They make up for 22.42% of the data.
    - **Interpretation:** CA-ASI double-downvotes this region: (i) Low uncertainty, (ii) High redundancy. CA-ASI significantly prunes this region even when ASI might still assign some (non-zero) probability. These points offer minimal marginal gain.

From this, we can infer, if there is more redundancy to exploit in the data , CA-ASI will exploit sampling in regions Quandrant II and Quandrant IV, especially the later which makes up a good 22.42% of the data .

We can say, *ASI applies a unidimensional filter(along X axis), whereas CA-ASI enforces a multi-dimensional pruning criterion(joint axes), leading to better selection of informative and diverse examples.*

D.3.3 Data Coherence and Fischer Kernel

Fisher Kernel is the canonical way to measure the distance between data points in the parameter space of a statistical model. The Fisher Kernel [30] is defined as:

$$K(x_i, x_j) = \nabla_\theta \log p(x_i|\theta)^\top \mathcal{I}(\theta)^{-1} \nabla_\theta \log p(x_j|\theta)$$

Where $\nabla_\theta \log p(x_i|\theta)$ is known as the Fisher Score (or simply the score function) and $\mathcal{I}(\theta)$ is the Fisher Information Matrix (FIM).

Our Coherence matrix entry S_{ij} is defined as:

$$S_{ij} = \|\hat{H}_i^{1/2} \hat{H}_j^{1/2}\|_F$$

If we use the Generalized Fisher Approximation [37], we can see that:

$$\hat{H}_i^{1/2} \approx \nabla_\theta \ell_\theta(x_i, y_i)$$

Substituting this into the definition:

$$S_{ij} \approx \|\nabla_\theta \ell_i \cdot \nabla_\theta \ell_j^\top\|_F \approx \nabla_\theta \ell_i^\top \nabla_\theta \ell_j$$

Thus, if we assume the FIM is approximately Identity or a known pre-conditioner M, coherence is effectively:

$$S_{ij} \approx K_{Fisher}(x_i, x_j) \text{ under the metric } M$$

So, while the standard Fisher Kernel $K(x_i, x_j) = \nabla_\theta \ell_i^\top \mathcal{I}(\theta)^{-1} \nabla_\theta \ell_j$ measures global similarity in parameter space, our measure $S_{ij} = \|\hat{H}_i^{1/2} \hat{H}_j^{1/2}\|_F$ provides an instance-specific, loss-surface-aware kernel. By utilizing the individual Hessian factors $\hat{H}^{1/2}$, we measure overlap in the gradients at the

current estimate, making it directly responsive to the model's local loss geometry, rather than the global average information.

Data Coherence from an Information-Gain perspective. Every time we label a point (x, y), we add a rank-1 contribution to the FIM $\mathcal{I}(\theta)$. The Fisher information of a new sample is:

$$\mathcal{I}_{new} = \mathcal{I}_{old} + \underbrace{\nabla_\theta \ell_\theta(x, y) \nabla_\theta \ell_\theta(x, y)^\top}_{\text{Information Gain } \Delta\mathcal{I}}$$

ASI uses uncertainty $u(x)$ alone to select samples. As a consequence, if we query x_i, we add $\Delta\mathcal{I}_i$ to the FIM. But ASI doesn't care in which direction this information is added. If the current FIM $\mathcal{I}_{old}$ is already highly "stretched" (ill-conditioned) in a certain direction, adding more samples that provide information in that same direction does not improve the precision of the estimator.

Our coherence score $\tilde{\rho}_i$ is an empirical proxy for the overlap of the new information $\Delta\mathcal{I}_i$ with the existing Fisher Information $\mathcal{I}_{old}$. From standard A-optimality criterion in Experimental Design the 'Information Contribution' of a point x_i as:

$$\text{Gain}(x_i) = \text{Tr}(\mathcal{I}_{old}^{-1} \Delta\mathcal{I}_i)$$

Because $S_{ij} \approx \nabla_\theta \ell_i^\top \nabla_\theta \ell_j$, the sum $\sum_j S_{ij}$ is essentially the projection of point i's gradient onto the gradients of the points already in the pool:

$$\tilde{\rho}_i \approx \nabla \ell_i^\top \underbrace{\left(\sum_{j \in \text{Labeled}} \nabla \ell_j \right)}_{\text{Existing Geometry}}$$

Thus we are not calculating $\mathcal{I}_{old}^{-1}$ (the inverse FIM) for every point because that is $O(d^3)$ and expensive. Our approach avoids this inversion by utilizing the Nyström-like property of the Fisher Kernel: $S_{ij} \approx \nabla \ell_i^\top \nabla \ell_j$. By summing these pairwise similarities, $\tilde{\rho}_i = \sum_j S_{ij}$ provides an explicit, non-aggregated proxy for how much the gradient $\nabla \ell_i$ aligns with the span of the existing labeled set. Thus, while standard leverage methods attempt to approximate the inverse $\mathcal{I}_t^{-1}$ via inversion, *CA-ASI approximates the information saturation directly in the original gradient space using explicit pairwise kernels. This preserves the individual identity of each labeled point, allowing us to detect and prune near-duplicate information that matrix inversion 'smears' out*.

D.3.4 Validity of Curvature Proxy

While our redundancy measure $\tilde{\rho}(x)$ is inspired by recent advances in data coherence [2, 17, 14], to analyze loss curvature on true data, we are the first to repurpose its use: extending it from a dataset-level, global loss curvature diagnostic to principally capturing redundancy in the active setting where only pseudo-labels $\hat{y} = f(x)$ are available.

We emphasize that the objective of CA-ASI is not the explicit estimation of oracle curvature quantities, but the construction of a computable sampling signal that captures sufficient second-order geometric information to drive sampling probability of uncertain and redundant points to zero. In particular, both $\tilde{u}(x)$ and $\tilde{\rho}(x)$ are model-derived signals, not oracle quantities. $\tilde{u}(x)$ simply represents the model's internal state of predictive ambiguity. Similarly, $\tilde{\rho}(x)$ represents the model's internal geometric assessment of redundancy based on its own loss surface Hessian $\hat{H}_i = \nabla_\theta^2 \ell_\theta(x_i, f(x_i))$. In accordance with ASI, the sampling policy $\pi_\beta(.)$ still remains a deterministic function of $f()$ and X. This combination $\tilde{u}_\beta(x) = \tilde{u}(x) - \beta\tilde{\rho}(x)$ creates a composite model-derived score.

Yet, to characterize the representational validity of our model-derived signal, we analyze the gap between the practical proxy Hessian and the oracle Hessian H_i^* (computed at θ^* with true labels Y), under some additional assumptions. The gap decomposes as ,

$$\|\nabla_\theta^2 \ell_{\hat{\theta}}(x, \hat{y}) - \nabla_\theta^2 \ell_{\theta^*}(x, y)\| \leq \underbrace{\|\nabla_\theta^2 \ell_{\hat{\theta}}(x, \hat{y}) - \nabla_\theta^2 \ell_{\theta^*}(x, \hat{y})\|}_{\text{Parameter error}} + \underbrace{\|\nabla_\theta^2 \ell_{\theta^*}(x, \hat{y}) - \nabla_\theta^2 \ell_{\theta^*}(x, y)\|}_{\text{Label error}}$$

- Parameter error: Recall that $\hat{\theta} := \arg\min_\theta \frac{1}{n} \sum_{i=1}^n \ell_\theta(X_i, f(X_i))$ minimizes the empirical proxy risk using pseudo-labels. Let $\theta_f^* := \arg\min_\theta \mathbb{E}[\ell_\theta(X, f(X))]$ denote the population minimizer of

this proxy risk. The empirical minimizer converges to the population proxy minimizer: $\|\hat{\theta} - \theta_f^*\| \xrightarrow{p} 0$ (Estimation error vanishes)

However, the proxy minimizer θ_f^* differs from the true oracle θ^* due to the calibration error of f. Since our loss is strictly convex:

$$\|\theta_f^* - \theta^*\| \leq \frac{1}{\mu}\|\nabla L(\theta_f^*)\|$$

Since θ_f^* minimizes the proxy risk, $\nabla L_f(\theta_f^*) = 0$, so:

$$\nabla L(\theta_f^*) = \nabla L(\theta_f^*) - \nabla L_f(\theta_f^*) = \mathbb{E}[\nabla_\theta \ell_{\theta_f^*}(X, Y) - \nabla_\theta \ell_{\theta_f^*}(X, f(X))]$$

Assuming Lipschitz continuity of the gradient:

$$\|\nabla L(\theta_f^*)\| \leq L_{grad} \cdot \mathbb{E}[|Y - f(X)|] = L_{grad} \cdot \epsilon$$

Therefore:

$$\|\theta_f^* - \theta^*\| \leq \frac{L_{grad}}{\mu} \cdot \epsilon = O(\epsilon)$$

By the continuity of the Hessian (implied by its existence in Assumption 5.1, applying the mean value theorem to the mapping $\theta \mapsto \nabla_\theta^2 \ell_\theta$), we obtain the parameter error bound:

$$\|\nabla_\theta^2 \ell_{\hat{\theta}}(x, \hat{y}) - \nabla_\theta^2 \ell_{\theta^*}(x, \hat{y})\| \leq L_\theta \|\hat{\theta} - \theta^*\| \leq L_\theta(\|\hat{\theta} - \theta_f^*\| + \|\theta_f^* - \theta^*\|) = \underbrace{o_p(1)}_{\text{Estimation Error}} + \underbrace{O(\epsilon)}_{\text{Calibration bias}}$$

- Label error: Here we need Lipschitz continuity of the Hessian with respect to its label argument, which is true for GLMS and Exponential Loss . If $\epsilon = |f(x) - \mathbb{E}[Y|X]|$ as the model calibration error, then, $\|\nabla_\theta^2 \ell_{\theta^*}(x_i, f(x_i)) - \nabla_\theta^2 \ell_{\theta^*}(x_i, \mathbb{E}[Y|x_i])\|_F \leq L_y \cdot |f(x_i) - \mathbb{E}[Y|x_i]| = L_y \cdot \epsilon$

Finally combining both we get,

$$\|\hat{H}_i - H_i\|_F = O(\epsilon) + o_p(1)$$

The redundancy score $\tilde{\rho}(x_i)$ is a normalized row-sum of the coherence matrix $\mathbf{S}$ (or equivalently, depends linearly on the entries of $\hat{H}$). Specifically:

$$\tilde{\rho}(x_i) \propto \sum_{j \neq i} \|\hat{H}_i^{1/2} \hat{H}_j^{1/2}\|_F$$

By the triangle inequality and submultiplicativity of the Frobenius norm:

$$|\tilde{\rho}(x_i; \hat{H}) - \tilde{\rho}(x_i; H)| \leq C \cdot \max_j \|\hat{H}_j - H_j\|_F = O(\epsilon)$$

where C depends on the boundedness of $\mathbf{S}$. This result implies that as the black-box surrogate f becomes better calibrated ($\epsilon \to 0$), the practical proxy converges to the oracle curvature.

Crucially, an active sampling strategy does not require the proxy to match the oracle in absolute value; it only requires the preservation of the relative ranking of points. In other words, for selection, if $\|\hat{H}_A - H_A\| < \epsilon$ and $\|\hat{H}_B - H_B\| < \epsilon$, the error propagates to the scalar scores via a local Lipschitz constant K. If the true redundancy gap satisfies $|\rho_A^* - \rho_B^*| > 2K\epsilon$, then the proxy preserves the ranking:

$$\text{sign}(\tilde{\rho}_A - \tilde{\rho}_B) = \text{sign}(\rho_A^* - \rho_B^*) \tag{10}$$

Thus, the $\mathcal{O}(\epsilon)$ bound on individual Hessians ensures that the redundancy scores (and their relative ordering) are preserved up to the same order of error. Practically, in high-dimensional manifolds (like our $d = 9$ setups), the redundancy scores exhibit high variance (as shown by our CV=79%, 139% diagnostic), meaning the vast majority of pairs have large gaps that are robust to $O(\epsilon)$ noise.

D.4 Additional Real World Example

We evaluate CA-ASI on the *Diamonds* dataset [72], a well-known benchmark comprising 53,940 round-cut diamonds with recorded sale prices and physical measurements. Following standard preprocessing [73], we remove entries with zero or physically implausible dimensions (i.e., length $x = 0$, width $y = 0$, depth $z = 0$, or $y > 20\,\text{mm}$, $z > 20\,\text{mm}$), retaining $n = 53{,}917$ observations. The response variable is *price* (USD), which is strictly positive ($\min = \$326$) and right-skewed (skewness ≈ 1.62), making it well-suited for Gamma generalised linear model (GLM) with a log link([47], Figure 14)

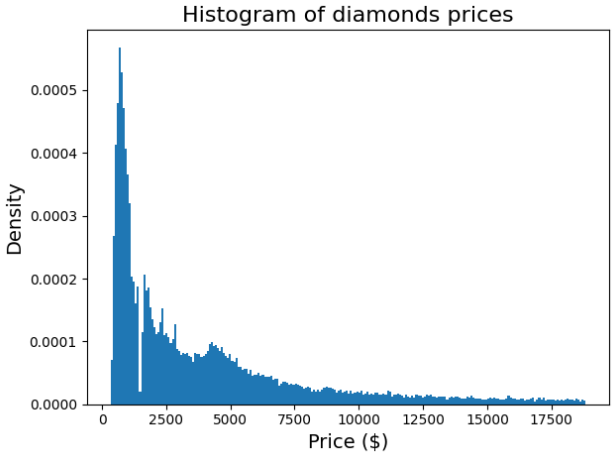

Figure 14: Heavy Right Skew in Diamond Prices

We use $d = 7$ features: six purely continuous predictors — `carat` (weight), `depth percentage`, `table width`, and `physical dimensions` x, y, z (in mm) — together with an intercept term. Notably, carat is approximately proportional to the product $x \cdot y \cdot z$(the volume scaled by diamond density), inducing a low-dimensional manifold structure with dense pockets for common small diamonds and sparse regions for rare large ones. This geometry is precisely the regime in which CA-ASI's coherence penalty is beneficial, as it discourages redundant sampling from over-represented pockets on the manifold.

The per-observation loss is $\ell_i(\theta) = y_i e^{-x_i^\top \theta} + x_i^\top \theta$ (where $y_i > 0$). The per-point Hessian is

$$\nabla^2 \ell_i(\theta) \;=\; y_i\, e^{-x_i^\top \theta}\, x_i x_i^\top \;=\; c_i\, x_i x_i^\top, \tag{11}$$

where $c_i \;=\; \hat{y}_i\, e^{-x_i^\top \hat{\theta}}$ and $\hat{\theta} := \arg\min_\theta \frac{1}{n} \sum_{i=1}^n \ell_\theta(x_i, f(x_i))$. Unlike exponential loss (where $y^2 = 1$ cancels the label), the response $\hat{y}_i$ remains in the curvature weight, encoding the scale of the predicted outcome. The square root factor is therefore

$$\hat{H}_i^{1/2} \;=\; \sqrt{c_i} \cdot x_i \;=\; \sqrt{\hat{y}_i} \cdot e^{-\frac{1}{2} x_i^\top \hat{\theta}} \cdot x_i. \tag{12}$$

Coherence Matrix Entry:

$$\begin{aligned} S_{ij} &= \|\hat{H}_i^{1/2} \hat{H}_j^{1/2}\|_F = \sqrt{c_i \cdot c_j} \cdot |x_i^\top x_j| \\ &= \sqrt{\hat{y}_i \hat{y}_j} \cdot \exp\Big(-\tfrac{1}{2}\big(x_i^\top \hat{\theta} + x_j^\top \hat{\theta}\big)\Big) \cdot |x_i^\top x_j|. \end{aligned} \tag{13}$$

Interpretation: Redundancy is the dot product $|x_i^\top x_j|$ weighted by $\sqrt{c_i c_j}$, where $c_i = \hat{y}_i / e^{\eta_i}$ is the ratio of the black-box prediction $\hat{y}_i$ to the GLM's predicted mean e^{η_i}. This weight is large when the black-box over-predicts relative to the GLM(i.e. the two models disagree strongly), and approaches unity when they agree ($\hat{y}_i \approx e^{\eta_i} \Rightarrow c_i \approx 1$).

Training and budget allocations are identical to those discussed in Appendix D.2.1, except we employ a 3-layered MLP as our black-box model in this case.

Figure 15: Diamonds Dataset in Batch Setting

Figure 16: Diamonds Dataset in Sequential Setting

From Figures 15, 16, CA-ASI uses the least amount of human labels, more diverse labels, and competitive width compared to ASI, while maintaining $\approx 90\%$ coverage. Methods like Uniform and ASI, while may provide tighter CIs, their coverage is extremely poor (close to 70%), while exhausting the entire budget on similar datapoints.

D.5 Runtime Analysis

D.5.1 Batch

A key challenge in our method lies in evaluating the redundancy score $\tilde{\rho}(x)$(see (2)) which captures the curvature overlap of each point with the rest of the dataset. This setting requires computing a full data coherence matrix $\mathbf{S} \in \mathbb{R}^{n \times n}$. We strictly adhere to this approach for our results in Section 6.

Anchor-based approximation. To reduce this overhead, we implement an anchor-based approximation that significantly lowers the computational burden while retaining geometric and curvature-aware inductive bias. Specifically, we sample a small subset of $B \ll n$ anchor points from the pool and compute coherence only with respect to these anchors. The approximation proceeds as follows:

1. **Anchor Selection**: Uniformly sample B representative points from the dataset.
2. **Partial Coherence Evaluation**: For each point x_i, compute curvature-based coherence scores S_{ij} with each anchor $x_j^{(\text{anchor})}$.
3. **Geometric Weighting**: Define a geometric affinity matrix $\mathbf{W} \in \mathbb{R}^{n \times B}$ using any lightweight distance metric of your choice(here we use RBF kernel for some $\sigma > 0$), such that:
$$W_{ij} = \frac{\exp\left(-\|x_i - x_j^{(\text{anchor})}\|/\sigma\right)}{\sum_{k=1}^{B} \exp\left(-\|x_i - x_k^{(\text{anchor})}\|/\sigma\right)}.$$
4. **Redundancy Aggregation**: Form the final score $\tilde{\rho}_i$ as a weighted average:
$$\tilde{\rho}_i = \sum_{j=1}^{B} W_{ij} \cdot S_{ij}.$$

Instead of evaluating an $n \times n$ matrix, we reduce it to $n \times B$ matrix, reducing runtime complexity from $\mathcal{O}(n^2 d)$ to $\mathcal{O}(nBd)$, enabling efficient scalability to larger pools. Since coherence is treated as a local property, this anchor-based approach respects the geometric structure of the dataset and is especially effective when coherence depends solely on covariates (e.g., in least-squares settings). Empirically, we test our anchor-approximation mechanism on the HTRU2 dataset with B=200: Approximation is roughly 20.1x faster (0.6765 ± 0.0812 s vs 0.0337 ± 0.0002s), indicating impressive scalability.

It is also worth mentioning that the coherence matrix $\mathbf{S}$ is symmetric in nature i.e $S_{ij} = S_{ji} \quad \forall i, j$. In practice, to compute $\tilde{\rho}_i$, we only need $\frac{N(N-1)}{2}$ entries(we skip the diagonal entries). All in all, our approach(Eq. (2)) presents a straightforward, full-pairwise coherence computation. However, several well-established approximations exist that can drastically reduce computational overhead. When the loss function is smooth and twice-differentiable, diagonal or low-rank Hessian approximations can be used [68, 45]. These approximations have been extensively studied in second-order optimization methods [44, 46] and are known to yield substantial speedups while effectively preserving local curvature information [45]. On a different note, instead of exhaustive comparisons, stochastic estimation methods like subsampling [7], bootstrapping, or random projections [53] can be employed to estimate coherence values over a representative subset of data [18].

Labeling vs. Computation: Economics of Active Learning

In many real-world deployments, the cost of acquiring a single high-quality label can dramatically exceed the cost of additional computation [9]. This disparity arises from the demand for skilled annotators, the time-intensive nature of manual review, and the sheer volume of data in modern applications. An equally critical but often overlooked constraint is safety. In many physical sciences and engineering domains, the physical act of querying an 'oracle' for a label corresponds to conducting real-world experiments that can be inherently hazardous, life-risking, or destructive. For example, synthesizing highly reactive or toxic chemicals [21], exploring extreme parameter spaces in high-power automated physical systems [66], automating quantum experiments [48], or collecting edge-case data in autonomous driving [26] all involve substantial physical risk. In such regimes, even marginal reductions in labeling requirements, as enabled by effective active learning strategies, translate into not only significant economic and resource savings, but also ensure safety(to some degree).

Our method introduces computational overhead due to the incorporation of redundancy-aware selection. Recognizing the dominant role of labeling expenses in active learning, any slight increase in runtime is often negligible compared to the potential savings in human labeling effort [57, 79, 28]. This motivates a critical re-framing of the core question in designing active learning systems: *Do we prioritize a few CPU cycles, or the cost of setting up an expensive real-life experiment, to label an instance?* In most practical and cost-sensitive scenarios, the latter unequivocally dominates. The former can be realistically mitigated to some extent, using parallelization, GPU acceleration [19],

and various approximation methods discussed previously, but the annotation charges are generally out of the practitioners control [65, 36].

To quantify this crucial trade-off empirically, we define the metric:

$$\Delta = \frac{n_{\text{ASI}} - n_{\text{CA-ASI}}}{T_{\text{CA-ASI}} - T_{\text{ASI}}}$$

where n_{ASI} and $n_{\text{CA-ASI}}$ denote the total number of labeled points required by ASI and CA-ASI, respectively, and T_{ASI} and $T_{\text{CA-ASI}}$ denote their corresponding runtimes. By defintion for $\beta > 0$, Δ is always non-negative in practice. This Δ score can be interpreted as the *labeling efficiency gain per unit of additional runtime*. A higher Δ indicates that CA-ASI achieves substantial label savings with minimal computational overhead, precisely the regime where redundancy-aware selection methods are most impactful for real-world applications.

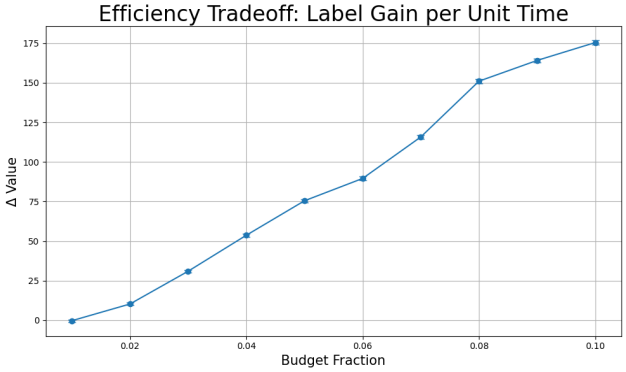

Figure 17: CA-ASI's impressive Label Gain per unit Time

Consequently, we compute Δ for the HTRU2 dataset under the same settings as those employed in Section 6. We then plot the Δ against the increasing budget. From Figure 17, we observe that the Δ improves as the budget increases. This suggests that CA-ASI is beneficial in the overall context, where label reduction has a more significant impact than runtime increase.

D.5.2 Sequential

Runtime Trade-offs in the Sequential Setting

While our Δ-metric (Eq. (D.5.1)) quantifies label-efficiency versus runtime in the batch setting, its intuition extends naturally to the sequential regime. However, β executes a nuanced dual role here, from the point of view of performance *and* runtime.

At the beginning of round t, $\mathbf{S}_t \in \mathbb{R}^{n_{(\text{lab},t)} \times n_{(\text{lab},t)}}$. At round t, we emphasize, we *do not* rebuild $\mathbf{S}_t$ from scratch. Instead, we compute only the new last row/column, shaped like an *inverted-outer* $L(\mathcal{O}(n_{(\text{lab},t)})$ operation). We then apply our sampling rule. If candidate is accepted, we append the row/column without touching the existing *inner* $n^2_{(\text{lab},t)}$ entries.

When data is introduced as a stream, β is introduced as a regularization parameter that trades off between model uncertainty and redundancy—favoring the selection of informative yet diverse examples. However, β also carries a compelling computational interpretation. At each time step t, our method maintains a coherence matrix $\mathbf{S}_t$ whose size depends on the number of labeled points seen thus far. A higher β encourages stricter pruning, leading to fewer accepted points, and therefore a *smaller and more sparsely updated* matrix $\mathbf{S}_t$ across rounds. This reduces both memory and computational overhead associated with updating and querying $\mathbf{S}_t$. Moreover, since fewer points are labeled under higher β, the frequency of model retraining decreases. This contrasts sharply with uniform and uncertainty-based strategies, where new labels are accepted more frequently and thus demand more frequent model updates. In scenarios where model fine-tuning is expensive (e.g., large language models or deep neural networks [34, 40]), CA-ASI may offer a compelling advantage by naturally amortizing the retraining cost over a sparser set of informative examples.

To highlight this, under the same setting as Section 6 for the HTRU2 Dataset, we plot the finetuning frequency of different sampling methods. We also plot the execution time of CA-ASI as a function of β for a budget of 2%. Figure 18 confirms our conjecture. Taken together, these observations motivate

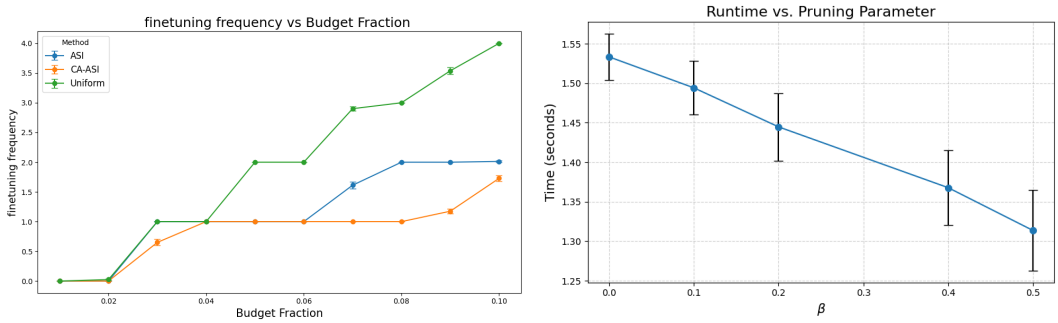

Figure 18: β's impact on finetuning frequency and execution speed on HTRU2 Dataset

viewing β not merely as a performance-control hyperparameter, but also as a runtime overhead controller in the sequential setting.