Active Measurement: Efficient Estimation at Scale

Max Hamilton* **Jinlin Lai*** **Wenlong Zhao** **Subhransu Maji†** **Daniel Sheldon†**
Manning College of Information & Computer Sciences
University of Massachusetts, Amherst
{jmhamilton,jinlinlai,wenlongzhao,smaji,sheldon}@cs.umass.edu

Abstract

AI has the potential to transform scientific discovery by analyzing vast datasets with little human effort. However, current workflows often do not provide the accuracy or statistical guarantees that are needed. We introduce *active measurement*, a human-in-the-loop AI framework for scientific measurement. An AI model is used to predict measurements for individual units, which are then sampled for human labeling using importance sampling. With each new set of human labels, the AI model is improved and an unbiased Monte Carlo estimate of the total measurement is refined. Active measurement can provide precise estimates even with an imperfect AI model, and requires little human effort when the AI model is very accurate. We derive novel estimators, weighting schemes, and confidence intervals, and show that active measurement reduces estimation error compared to alternatives in several measurement tasks.

1 Introduction

AI offers a transformative approach to scientific discovery, empowering scientists to analyze vast datasets in ways that traditional methods cannot achieve [22, 38, 41]. Applications include species identification from images and audio for biodiversity monitoring [1, 5, 19, 39], disease diagnosis in medical imaging [2, 37], classifying galaxies in astronomy [23], assessing crop health in agriculture [13, 14], and myriad other applications in fields such as remote sensing, microscopy, and neuroscience. In these applications, the typical goal is to make measurements to answer science or policy questions, which therefore must be precise. But AI models are far from perfect: they may introduce bias or have unacceptable error rates, and do not offer the statistical guarantees that scientists desire.

As a concrete example, consider counting birds in the high-resolution photograph in Fig. 1 of a large flock of Tree Swallows in Old Lyme, CT in September, 2018. Scientists would like to estimate the population size and track it over time from photographs like this [e.g., 4, 7, 18]. Computer vision is an excellent tool for this task, but current workflows are poorly suited to scientific measurement. A scientist might be able to estimate the count with relatively low effort using a pre-trained object detector adapted with a few labeled examples. However, substantial additional effort is needed to obtain a model that is accurate enough for scientific measurement and gain confidence in its performance. A typical scenario would be to first tune and validate the detector's performance using labeled image tiles. To obtain bounds on count performance, more validation is needed. After substantial effort, the scientist may find that the count is $20{,}358 \pm 8000$ birds, which is not accurate enough to track populations over time. What can be done? Under current practice, the scientist returns to model development. For challenging tasks, it may be impossible to achieve precise enough measurements.

*† Denotes authors with equal contribution

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

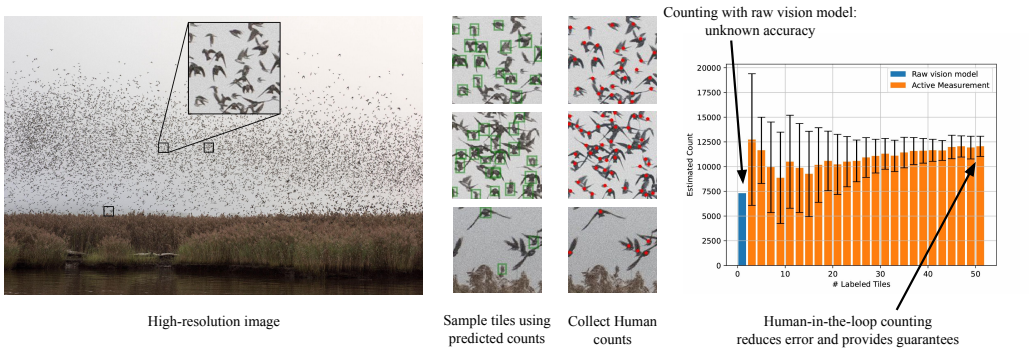


Figure 1: **Active measurement.** An AI model predicts counts for each image tile, which are then used to form a proposal distribution to selectively sample tiles for human labeling. Ground-truth counts are used both to improve the AI model and produce a Monte Carlo estimate of the final measurement.

We introduce a new paradigm of *active measurement* where the scientist uses AI to *interactively* solve the measurement task (Fig. 1). Active measurement is based on Monte Carlo estimation. In each step, the scientist receives an unbiased estimate of the true count *and* an estimate of its error and is asked to provide new labels. These labels are used *both* to reduce Monte Carlo error of the estimated measurement and to improve the AI model for future rounds of estimation. The process continues until the error is sufficiently low — in this example the estimates are 11,977 $\pm$ 1,076 after labeling 50 tiles in the image, closely aligning with the true count of 12,486 birds. Unlike existing workflows, arbitrarily low error is possible with enough labels even with an imperfect model. On the other hand, if the model approaches perfect performance, very accurate estimates are possible with a small amount of labeled data.

Active measurement builds on adaptive importance sampling [28], active testing [20, 26], and prior work on covariate and detector-based counting [25, 32], but extends each of these works in different directions. We introduce the active measurement framework, derive new estimators, and show that they are unbiased and consistent. We contribute novel approaches for confidence interval construction that are tailored to the active measurement setting, and derive novel weighting schemes to combine estimates obtained in each step to minimize the overall estimation variance. We show empirically that our techniques provide accurate confidence intervals and reduce estimation error compared to prior methods on several scientific measurement tasks.

Code for this paper is available at: `https://github.com/cvl-umass/active-measurement`.

2 Active measurement

We consider the following scientific measurement task. There is a set Ω of $N = |\Omega|$ individual units (e.g., image tiles), with unknown ground-truth measurement $f(s) \geq 0$ for each $s \in \Omega$. For $S \subseteq \Omega$, define $F(S) = \sum_{s \in S} f(s)$. We seek to estimate the total measurement $F(\Omega)$ across all units.

Active measurement combines human labeling with AI predictions. Suppose that, at time t, the scientist has labeled units $\mathcal{D}_t \subseteq \Omega$ and therefore knows the ground-truth value $F(\mathcal{D}_t) = \sum_{s \in \mathcal{D}_t} f(s)$. Further, suppose an AI model is available (e.g., trained on $\mathcal{D}_t$) to predict $g(s) \approx f(s)$ for other units. Let q_t be the probability distribution proportional to g on $\Omega \setminus \mathcal{D}_t$. The base estimator for active measurement is:

$$\hat{F}_t = F(\mathcal{D}_t) + \frac{f(s_t)}{q_t(s_t)}, \quad s_t \sim q_t. \tag{1}$$

The second term is an importance sampling estimate satisfying $\mathbb{E}_{q_t}[f(s_t)/q_t(s_t)] = F(\Omega \setminus \mathcal{D}_t)$, from which it follows that $\mathbb{E}[\hat{F}_t] = F(\Omega)$. This estimator is inspired by active testing [20], which we discuss more below. It is unbiased for any proposal distribution q_t, but has minimum variance (of zero) when $q_t \propto f$, which motivates the choice $q_t \propto g \approx f$.

The active measurement framework, shown in Alg. 1, uses this estimator in each step of a human-in-the-loop process. The set $\mathcal{D}_1$ contains initially labeled units, which may be empty. At the start of step t, units in $\mathcal{D}_t$ have been labeled, and an *acquisition distribution* q_t is used to sample a new unit for labeling. Our typical acquisition strategy is to train an AI model on $\mathcal{D}_t$ to generate predictions

Algorithm 1 Active measurement

Require: Initially labeled units $\mathcal{D}_1 \subseteq \Omega$, acquisition distribution q_1, weight sequences α_τ, β_τ

1: **for** $t = 1, 2, ..., T$ **do**
2: Sample $s_t \sim q_t(\cdot)$ and obtain $f(s_t)$
3: Form IS estimate

$$\hat{F}_t = F(\mathcal{D}_t) + \frac{f(s_t)}{q_t(s_t)} \tag{2}$$

4: Combine estimates as $\hat{F}_{1:t} = \sum_{\tau=1}^{t} \bar{\alpha}_\tau \hat{F}_\tau$
5: Get variance estimates $\{\widehat{\mathrm{Var}}_\tau\}_{\tau=1}^{t}$ using Alg. 2
6: Combine variances as $\widehat{\mathrm{Var}}_{1:t} = \sum_{\tau=1}^{t} \bar{\alpha}_\tau^2 \widehat{\mathrm{Var}}_\tau$
7: Update $\mathcal{D}_{t+1} = \mathcal{D}_t \cup \{s_t\}$
8: Update acquisition distribution q_{t+1} over $\Omega \setminus \mathcal{D}_{t+1}$, e.g., by updating an AI model using $\mathcal{D}_{t+1}$
9: **end for**

Algorithm 2 Variance estimation

Require: All variables from Alg. 1

1: **for** $\tau = 1, 2, ..., t$ **do**
2: Get mean estimate $\hat{G}_{\tau,t} = \hat{F}_{1:t-1} - F(\mathcal{D}_\tau)$
3: **for** $r = \tau, \ldots, t$ **do**
4: Form single variance estimator

$$\widehat{\mathrm{Var}}_{\tau,r} = \sum_{s \in \mathcal{D}_r \setminus \mathcal{D}_\tau} q_\tau(s) \left(\frac{f(s)}{q_\tau(s)} - \hat{G}_{\tau,t} \right)^2 + \frac{q_\tau(s_r)}{q_r(s_r)} \left(\frac{f(s_r)}{q_\tau(s_r)} - \hat{G}_{\tau,t} \right)^2 \tag{3}$$

5: **end for**
6: $\widehat{\mathrm{Var}}_\tau = \sum_{r=\tau}^{t} \bar{\beta}_r \widehat{\mathrm{Var}}_{\tau,r}$
7: **end for**

$g(s)$ for all units $s \in \Omega \setminus \mathcal{D}_t$ and then choose $q_t \propto g$. In Line 3, the newly labeled unit is used to form an importance sampling estimate with $\mathbb{E}[\hat{F}_t] = F(\Omega)$ using Eq. (1). Line 4 combines the estimators from each step into the estimator $\hat{F}_{1:t}$ with normalized weights that satisfy $\sum_{\tau=1}^{t} \bar{\alpha}_\tau = 1$, which incorporates all of the labeled samples and represents the model's overall estimate at time t. Hereafter, for any weighting scheme, we use α_τ to denote a sequence of unnormalized weights for $1 \leq \tau \leq N$ and $\bar{\alpha}_\tau$ for their normalized counterparts $\bar{\alpha}_\tau = \alpha_\tau / \sum_{r=1}^{t} \alpha_r$, with t implicit from context.

Proposition 1. *The combined estimator $\hat{F}_{1:t} = \sum_{\tau=1}^{t} \bar{\alpha}_\tau \hat{F}_\tau$ is unbiased: $\mathbb{E}[\hat{F}_{1:t}] = F(\Omega)$.*

At step t, we also generate a variance estimate from Alg. 2, which is derived in § 5. Alg. 2 takes $\mathcal{O}(t^2)$ time naively, but can be improved to $\mathcal{O}(t)$ using the streaming algorithm in § B.2, so that step t of active measurement takes $\mathcal{O}(t)$ time overall. In practical settings we expect the time to be dominated by labeling and updating an AI model. We next discuss related work before returning to weighting schemes and variance estimation.

3 Related work

Active measurement is closely related to three existing lines of work. Active testing [12, 20] interactively estimates the test loss of an AI model by sampling unlabeled points according to predicted losses from a surrogate model and then using the newly acquired labels to form an importance-sampling based estimator of the loss *and* update the surrogate model. The sampling strategy and form of the estimator in Eq. (1) is identical to that used by active testing, with the significant difference that the estimand is different: active measurement seeks to directly estimate a scientific measurement while iteratively refining the AI model itself, while active testing estimates a test loss and refines a surrogate model. Compared to active testing, we also contribute novel weighting schemes (§ 4) and show that these reduce error, and novel approaches for variance estimation and confidence interval construction (§ 5), which was not considered in active testing.

Active testing and active measurement are both versions of adaptive importance sampling (AIS) [6, 27, 28]. AIS improves an importance-sampling estimator $f(s)/q(s)$ by iteratively refining the proposal q and forming a weighted combination of the estimators made with different proposal distributions. Active measurement modifies AIS to sample *without* replacement from a finite sample space, which leads to the estimator in Eq. (1), compared to the usual importance-sampling estimator supported on all of Ω. Sampling without replacement leads to novel considerations in the selection of weights (§ 4) and in variance estimation (§ 5). The DISCount (detector-based importance sampling) method of Perez et al. [32], which builds on [25], estimates the total counts in scientific collections using AI model predictions for the importance sampling proposal distribution. Active measurement shares the same goal, and advances on DISCount by interactively refining the AI model with acquired labels and by sampling without replacement, which both can significantly reduce estimation error.

The concept of combining an AI model with limited human effort is related to semi-supervised learning [8], where the model is trained with using a combination of labeled and unlabeled data.

Most recently, prediction-powered inference (PPI) [3] was proposed to give valid statistical inference using a small amount of labeled data. Unlike PPI, which assumes data is iid and seeks to estimate a population parameter, active measurement samples data interactively and non-uniformly as aided by AI and seeks to estimate a scientific measurement on a finite dataset. We show in the experiments that active measurement outperforms a PPI-motivated baseline. In a similar spirit, active machine learning has been used to construct large datasets with crowd-sourcing [21, 30, 40]. We have a different objective of achieving accurate estimation with a small amount of labeled data.

4 Weighting schemes

What is an appropriate weight sequence α_τ to combine the estimators in active measurement? An important observation is that the estimators in each step, though not independent, are uncorrelated:

Proposition 2. *For any* $1 \leq \tau < r \leq t$, $\mathrm{Cov}(\hat{F}_\tau, \hat{F}_r) = 0$.

Thus, the combined variance is $\mathrm{Var}[\hat{F}_{1:t}] = \sum_{\tau=1}^{t} \bar{\alpha}_\tau^2 \mathrm{Var}[\hat{F}_\tau]$, and the weighting that achieves minimum variance uses inverse-variance weighting, i.e., $\alpha_\tau = 1/\mathrm{Var}[\hat{F}_\tau]$. However, in practice we don't know $\mathrm{Var}[\hat{F}_\tau]$ and it is difficult to estimate (see § 5), so we first consider fixed weighting approaches that approximate this principle under assumptions about how $\mathrm{Var}[\hat{F}_\tau]$ changes due to: (1) adapting the model, and (2) the shrinking sample space.

Square root law. In AIS, a simple and near-optimal alternative is the square root law $\alpha_\tau^{\mathrm{SQRT}} = \sqrt{\tau}$ [29]. To derive this law, it is assumed that variance reduces at the rate $\mathrm{Var}[\hat{F}_\tau] \propto \tau^{-y}$ due to adaption of the proposal distribution, but the rate $y \in [0, 1]$ is unknown. The square root law works as a conservative strategy that is within a factor of $9/8$ of the optimal variance even without knowing the rate y: if the variance with an optimal weighting strategy for rate y at step t is $\mathrm{Var}_{\mathrm{opt}}(y, t)$ and $\hat{F}_{1:t}^{\mathrm{SQRT}}$ is the estimator using weights $\alpha_\tau^{\mathrm{SQRT}}$ (we will use similar notation for other weighting schemes below), then $\sup_{t\geq 1} \sup_{y\in[0,1]} \frac{\mathrm{Var}[\hat{F}_{1:t}^{\mathrm{SQRT}}]}{\mathrm{Var}_{\mathrm{opt}}(y,t)} \leq \frac{9}{8}$. However, in our context, the model $\mathrm{Var}[\hat{F}_\tau] \propto \tau^{-y}$ does not account for shrinking variance due to sampling without replacement, which is not optimal.

LURE weights. In active testing, a weighting scheme known as LURE was introduced for the pool-based setting with sampling without replacement [12]. If the number of units is $|\Omega| = N$, the LURE weights are $\alpha_\tau^{\mathrm{LURE}} = \frac{1}{(N-\tau)(N-\tau+1)}$. The original motivation was to treat samples equally in the combined estimate, but we can reinterpret these weights as accounting for variance reduction due to the shrinking sample space:

Proposition 3. *If there exist constants* $0 < A \leq B$ *such that for any* $s \in \Omega$, $A \leq f(s) \leq B$ *and* $A \leq g(s) \leq B$, *then there exists a constant* $C > 0$ *such that* $\mathrm{Var}[\hat{F}_\tau] \leq C(N-\tau)(N-\tau+1)$.

For brevity, define $w_\tau = \frac{1}{(N-\tau)(N-\tau+1)}$. This proposition indicates that the variance is order $1/w_\tau$ when only considering the sample space reduction, so the LURE weights of w_τ are well justified by inverse variance weighting. However, in our setting, LURE weights neglect variance reduction due to adaptation of the AI model, which is also not optimal.

Combined weights. We propose to combine the above two weighting schemes and use $\alpha_\tau^{\mathrm{COMB}} = w_\tau\sqrt{\tau}$. Fig. 2 demonstrates the growth of different weighting schemes. When τ is small, the detector has not been well-tuned so weighting schemes that discourage early estimators, like the square root law, are preferred. As τ becomes closer to N, the sample space reduction becomes more significant, which is addressed by the LURE weights. Our combined weights unify the best of the two, accounting for the two sources of variance reduction. The following result bounds our worst-case estimation error under our variance reduction model, in a way similar to the square root law for AIS.

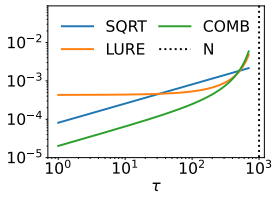


Figure 2: Normalized weights as functions of τ for $t = 700$ and $N = 1000$.

Proposition 4. *If* $\mathrm{Var}[\hat{F}_\tau] \propto \tau^{-y}/w_\tau$, *where* $y \in [0, 1]$ *and* w_τ *are non-decreasing, then the estimate* $\hat{F}_{1:t}^{\mathrm{COMB}} = \sum_{\tau=1}^{t} \bar{\alpha}_\tau^{\mathrm{COMB}} \hat{F}_\tau$, *where* $\alpha_\tau^{\mathrm{COMB}} = w_\tau\sqrt{\tau}$, *satisfies*

$$\sup_{t\geq 1}\sup_{y\in[0,1]}\frac{\mathrm{Var}[\hat{F}^{\mathrm{COMB}}_{1:t}]}{\mathrm{Var}_{\mathrm{opt}}(y,t)}\leq\frac{9}{8}, \tag{4}$$

where $\mathrm{Var}_{\mathrm{opt}}(y,t)$ *is the estimation variance when the estimators are weighted proportionally to the inverse of the ground-truth variances.*

In §7, we show empirically that the combined weighting scheme produces lower estimation error than either of the two components alone.

Inverse variance weighting. In AIS, it has been advised not to approximate inverse variance weighting using *estimated* variances [29], due to the unreliability of variance estimates with only a finite number of samples in each round of AIS and the potentially large bias this could introduce. In the next section, we introduce consistent estimates of (conditional) variances and show that these can make inverse variance weighting practical and achieve even better estimation error than combined weighting scheme in some settings.

5 Variance estimation and confidence intervals

Along with an unbiased and consistent estimator $\hat{F}_{1:t}$ we typically want a consistent estimate of $\mathrm{Var}[\hat{F}_{1:t}]$ to understand the error and construct confidence intervals. However, this is complicated by the fact that the estimators $\hat{F}_\tau$ are not iid. Rather, each depends on the samples drawn in the previous step, and we only have a *single* random draw for each τ. As a result, we will be able to form an unbiased but not a consistent estimate of $\mathrm{Var}[\hat{F}_{1:t}]$, unlike in the iid case (but similar to AIS [43]).

Martingale convergence. Fortunately, we can turn to martingale arguments to instead estimate a conditional variance that is appropriate for confidence intervals. For this argument, consider the centered sequence of estimators $\tilde{F}_{1:t}=\sum_{\tau=1}^{t}\alpha_\tau(\hat{F}_\tau-F(\Omega))$ where the weights are unnormalized. Because each $\hat{F}_t$ is unbiased, the displacement $\tilde{F}_{1:t}-\tilde{F}_{1:t-1}=\alpha_t(\hat{F}_t-F(\Omega))$ has zero mean, making the sequence $\tilde{F}_{1:t}$ a martingale. A martingale central limit theorem can be used to show that averaging over enough of these zero-mean displacements gives a limiting behavior similar to iid averaging. Stated informally, we have the following:

Proposition 5 (informal)**.** *Let* $V^2_{1:t}=\sum_{\tau=1}^{t}\bar{\alpha}^2_\tau\,\mathrm{Var}[F_\tau|\,\mathcal{D}_\tau]$ *be the conditional variance of* $\hat{F}_{1:t}$*. Under suitable regularity conditions,* $\frac{\hat{F}_{1:t}-F(\Omega)}{V_{1:t}}\xrightarrow{D}\mathcal{N}(0,1)$.

We provide a formal statement and proof in § A.5, which requires constructing a triangular array and letting the number of active measurement steps T and domain size N go to infinity jointly. See [43] and [33] for similar results for AIS, i.e., sampling *with* replacement. This result motivates forming estimators of the *conditional* variance for use in confidence intervals.

Novel conditional variance estimators. We first focus on the conditional variance of a single estimator $\hat{F}_\tau$, which is the variance of the importance weights $f(s)/q_\tau(s)$ where $s\sim q_\tau$:

$$\mathrm{Var}[\hat{F}_\tau|\,\mathcal{D}_\tau]=\sum_{s\in\Omega\setminus\mathcal{D}_\tau}q_\tau(s)\left(\frac{f(s)}{q_\tau(s)}-F(\Omega\setminus\mathcal{D}_\tau)\right)^2 \tag{5}$$

In importance sampling, we typically estimate this variance over iid samples from q_τ, but in active measurement we have only a single sample $s_\tau\sim q_\tau$. We therefore derive a novel variance estimator that uses samples $s_\tau, s_{\tau+1},\dots s_t$ to estimate $\mathrm{Var}[\hat{F}_\tau|\,\mathcal{D}_\tau]$ following an importance sampling approach similar to that of the base estimator in Eq. (1). For $r\geq\tau$, let

$$\widehat{\mathrm{Var}}_{\tau,r}=\sum_{s\in\mathcal{D}_r\setminus\mathcal{D}_\tau}q_\tau(s)\left(\frac{f(s)}{q_\tau(s)}-F(\Omega\setminus\mathcal{D}_\tau)\right)^2+\frac{q_\tau(s_r)}{q_r(s_r)}\left(\frac{f(s_r)}{q_\tau(s_r)}-F(\Omega\setminus\mathcal{D}_\tau)\right)^2. \tag{6}$$

The first term of Eq. 6 is the exact sum of the terms from Eq. 5 for $s\in\{s_\tau,\dots,s_{r-1}\}$, and the second term is an importance sampling estimate using $s_r\sim q_r$ for the sum of the remaining terms. We then mix the variance estimates $\widehat{\mathrm{Var}}_{\tau,r}$ for each r to get the combined estimator $\widehat{\mathrm{Var}}_\tau:=\sum_{r=\tau}^{t}\bar{\beta}_r\,\widehat{\mathrm{Var}}_{\tau,r}$, where β_r is another sequence of weights; for simplicity, we will always use the LURE weights $\beta^{\mathrm{LURE}}_r=w_r$ to model the shrinking sample space.

Proposition 6. *The estimators* $\widehat{\mathrm{Var}}_\tau$ *and* $\widehat{\mathrm{Var}}_{\tau,r}$ *for* $\tau \le r \le t$ *satisfy* $\mathbb{E}[\widehat{\mathrm{Var}}_\tau | \mathcal{D}_\tau] = \mathbb{E}[\widehat{\mathrm{Var}}_{\tau,r} | \mathcal{D}_\tau] = \mathrm{Var}[\hat{F}_\tau | \mathcal{D}_\tau]$ *and* $\mathbb{E}[\widehat{\mathrm{Var}}_\tau] = \mathbb{E}[\widehat{\mathrm{Var}}_{\tau,r}] = \mathrm{Var}[\hat{F}_\tau]$.

Further, beacuse it averages over many individual estimates, the error of $\widehat{\mathrm{Var}}_\tau$ for estimating the conditional variance converges as $t \to N$:

Proposition 7. *With the same settings as Prop. 3, the variance estimate for* $\hat{F}_\tau$ *weighted by the LURE scheme* $\widehat{\mathrm{Var}}_\tau^{\mathrm{LURE}} = \sum_{r=\tau}^t \bar{\beta}_r^{\mathrm{LURE}} \widehat{\mathrm{Var}}_{\tau,r}$ *satisfies that* $\mathrm{Var}\left[\widehat{\mathrm{Var}}_\tau^{\mathrm{LURE}} | \mathcal{D}_\tau\right] \lesssim \frac{1}{t-\tau+1} \cdot \min\left(1, \frac{(N-t)^2}{t-\tau+1}\right)$.

The notation $\lesssim$ hides multiplicative constants with respect to t. When $t \ll N$, we have the usual Monte Carlo rate of $(t-\tau+1)^{-1}$ (recall that $t-\tau+1$ is the number of samples). On the other hand, when $t \to N$, a faster rate is achieved thanks to sampling without replacement. This result contrasts with the usual practice of variance estimation for AIS, where only the samples from one stage are used to estimate the conditional variance, and the stagewise estimators do not converge.

Confidence intervals. Several practical steps remain to construct confidence intervals. The full variance estimation procedure is shown include Alg. 2. Previously, we assumed knowledge of $F(\Omega \setminus \mathcal{D}_\tau)$ for estimating the conditional variance, but in practice we use the plug-in estimator $\hat{G}_{\tau,t} = \hat{F}_{1:t-1} - F(\mathcal{D}_\tau) \approx F(\Omega \setminus \mathcal{D}_\tau)$ from step t of active measurement. The full conditional variance estimator is then $\widehat{\mathrm{Var}}_{1:t} = \sum_{\tau=1}^t \bar{\alpha}_\tau^2 \widehat{\mathrm{Var}}_\tau \approx \sum_{\tau=1}^t \bar{\alpha}_\tau^2 \mathrm{Var}[F_\tau | \mathcal{D}_\tau] = V_{1:t}^2$, which according to Prop. 5 is the appropriate quantity for the martingale central limit theorem. The naive implementation to compute $\widehat{\mathrm{Var}}_{1:t}$ using Alg. 2 takes $\mathcal{O}(t^2)$ time but we can reduce the complexity to $\mathcal{O}(t)$ with a streaming algorithm; see §B.2 for details. Finally, a $1-\alpha$ confidence interval is constructed as $\hat{F}_{1:t} \pm z_{\alpha/2} \widehat{\mathrm{Var}}_{1:t}^{1/2}$, where $z_{\alpha/2}$ is the $1-\alpha/2$ quantile of the standard normal distribution.

We will also consider confidence intervals formed using the "simple" conditional variance estimator $\widehat{\mathrm{Var}}_{1:t}^{\mathrm{simp}} = \sum_{\tau=1}^t \bar{\alpha}_\tau^2 (\hat{F}_\tau - \hat{F}_{1:t})^2$, which is the motivated by the fact the sum of squared deviations $\sum_{\tau=1}^t \bar{\alpha}_\tau^2 (\hat{F}_\tau - F(\Omega))^2$ of the martingale sequence converges to the conditional variance $V_{1:t}^2$ under the same regularity conditions as Prop. 5; see § A.5 for details. When needed, we will denote the estimator $\widehat{\mathrm{Var}}_{1:t}$ from Alg. 1 as $\widehat{\mathrm{Var}}_{1:t}^{\mathrm{cond}}$ to emphasize that it is based on estimating conditional variances for each τ.

Conditional inverse variance weighting. Prior AIS literature has advised against using inverse variance weighting with estimated variances [29]. However, our analysis suggests that weighting with estimated *conditional* variances may be appropriate. First, Prop. 5 can be read as $\hat{F}_{1:t} \approx \mathcal{N}(F(\Omega), V_{1:t}^2)$, which motivates choosing weight sequences α_τ to minimize the *conditional* variance $V_{1:t}^2 = \sum_{\tau=1}^t \bar{\alpha}_\tau^2 \mathrm{Var}[\hat{F}_\tau | \mathcal{D}_\tau]$, the optimal choice being $\alpha_\tau \propto 1/\mathrm{Var}[\hat{F}_\tau | D_\tau]$. Second, Prop. 7 controls the error of the conditional variance estimators $\widehat{\mathrm{Var}}_\tau \approx \mathrm{Var}[F_\tau | D_\tau]$. Because the error decays with the number of samples $t-\tau+1$, estimates for $\tau \ll t$ will be more reliable than those for $\tau \approx t$. This motivates our proposed scheme, which uses inverse estimated variances for $\tau \le \gamma t$ and continues the sequence using the $\alpha_\tau^{\mathrm{COMB}}$ weights for $\tau > \gamma t$, where $\gamma \in (0,1)$ is a hyperparameter:

$$\alpha_\tau^{\mathrm{INV}} = \begin{cases} 1/\widehat{\mathrm{Var}}_\tau & 1 \le \tau \le \gamma t \\ \alpha_\tau^{\mathrm{COMB}} \cdot \frac{1/\widehat{\mathrm{Var}}_{\gamma t}}{\alpha_{\gamma t}^{\mathrm{COMB}}} & \gamma t \le \tau \le t \end{cases}. \tag{7}$$

The constant factor $(1/\widehat{\mathrm{Var}}_{\gamma t})/(\alpha_{\gamma t}^{\mathrm{COMB}})$ for $\tau \ge \gamma t$ ensures that the weight sequence is continuous at $\tau = \gamma t$. We show in our experiments that these weights can produce even better estimation results than the α^{COMB} weights for suitable settings of γ. Disadvantages are that $\hat{F}_{1:t}$ is no longer unbiased when using estimated weights, and that confidence intervals may have poorer coverage.

6 Experiments

We mainly experiment with two tasks: 1) counting birds in high-resolution images of Tree Swallow flocks, and 2) counting roosting birds in weather radar images across multiple radar stations and years. The first task is analogous to common counting problems in microscopy or medical imaging, where an object detector trained on a few examples may perform reasonably well. In contrast, the second task is considerably more challenging, as it requires a custom detector—typically the result

of substantial community effort involving dataset annotation, model training, and validation. To demonstrate domain generality, we also performed experiments on malaria-infected cell counting and damaged-building counting from satellite images. For all IS-based methods, we use proposal distributions proportional to predicted counts using an object detector.

Counting birds in high-resolution images. We aim to estimate the total number of birds in two separate images—"sky" and "reeds"—each with different difficulty levels. The reeds image is more challenging due to its higher bird density and more complex background (see Fig. 1 and Fig. A1). We divide the sky and reeds images into tiles of size 200×200 and 160×160 pixels, respectively, and manually annotate the birds in each tile using the VGG annotator [11]. This results in 925 tiles for sky and 1,426 tiles for reeds, which serve as the annotation units in our experiments. In total, the ground truth bird count is 5,682 for the sky image and 12,486 for the reeds image.

To detect birds, we train a Faster R-CNN [35] detector with a ResNet-50 [17] backbone pre-trained on ImageNet [9], using the Detectron2 [42] library. The model is fine-tuned with a single A16 GPU for 400 iterations and a learning rate of 0.001 on the annotated tiles. The detector performs reasonably well, with average error rates of 9.5% and 33.1% when trained on 50 randomly selected tiles. We also experimented with few-shot detection models [34, 36] that do not require training, but found that a standard Faster R-CNN detector outperforms them when even a few labeled tiles are available. See § D.1 for a comparison.

Counting roosting birds in radar. Another scientific application involves estimating bird counts from weather radar. Birds often congregate overnight in large numbers. Their mass departure from roosting sites in the morning can be detected by weather radar and the roosts leave visible signatures in radar-collected data channels such as reflectivity (Fig. A2) [5, 10]. Scientists can use these signals to estimate total bird counts, providing valuable data to analyze long-term migration trends.

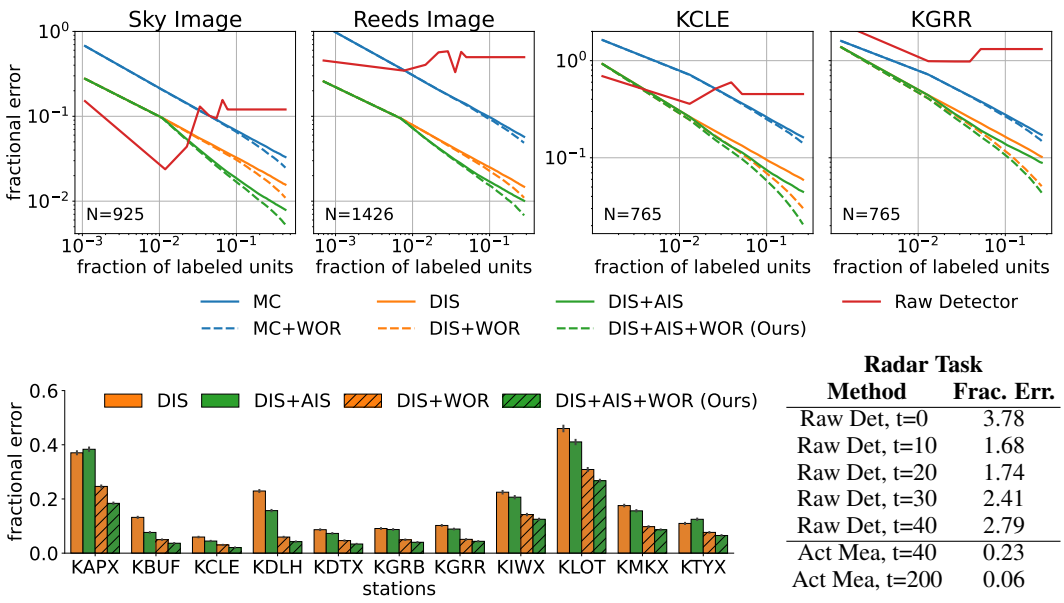
We adapt the experimental setup of DISCount [32], leveraging expert annotated roosts from the Great Lakes analysis in [5] and [10] as ground truth. Our objective is to automatically estimate the total of the daily bird counts for each of 11 radar stations in the Great Lakes region, for dates between June 1st and October 31st over the five-year period from 2015 to 2019.

We follow Perez et al. [31] to pretrain a roost detector on a manually labeled training set from multiple radar stations that do not include the Great Lakes stations. It is based on a Faster R-CNN architecture with a ResNet-101 backbone and an adapter layer to handle radar channels across elevations and time steps. Detections from consecutive timestamps are assembled into tracks and bird counts are estimated from the tracks based on the radar geometry and reflectivity of the birds. We provide details in § C.1. The detection task is challenging and the best model only achieves 56% mean average precision. Thus, [5, 10] invested substantial manual screening time for scientific analyses that require high precision. Our method substantially reduces this screening effort in estimating the bird counts.

We adapt the detector to each station with learning rate 10^{-4} for 3000 iterations on a single A16 GPU. To reduce overfitting, we use a mix of 80% pretraining data and 20% station-specific labeled data. The model is finetuned every 10 samples for the first 40 samples, after which performance saturates.

Counting malaria-infected cells. The Malaria Cell dataset (image set BBBC041v1, available from the Broad Bioimage Benchmark Collection [24]) comprises 1,364 images (about 80,000 cells). This dataset contains microscopy images of blood smears used for detecting and classifying malaria parasites, aimed at enabling the development and evaluation of automated methods for parasite detection, quantification, and stage classification. For this evaluation, we focus on counting the number of infected cells, which are around 5% of the dataset. We use the same settings as our sky and reeds image experiments. For our initial model we finetune the default Faster R-CNN network on three randomly selected cell images.

Counting damaged buildings. For damaged building detection we focus on the Palu Tsunami subset of xBD [15], which contains 113 satellite images of the shoreline before and after the Palu Tsunami. We count the number of damaged buildings, which corresponds to a label of "major-damage" or "destroyed". Only the post disaster images are given to the model. We use the same hyperparameters as before. The initial model is trained on 5 randomly selected images.



Radar Task

Method	Frac. Err.
Raw Det, t=0	3.78
Raw Det, t=10	1.68
Raw Det, t=20	1.74
Raw Det, t=30	2.41
Raw Det, t=40	2.79
Act Mea, t=40	0.23
Act Mea, t=200	0.06

Figure 3: **Estimation error on two measurement tasks.** Top: Fractional error of the estimated count as the percentage of labeled tiles increases, averaged over 10,000 runs for the counting birds in the "sky" and "reeds" images, and counting roosting birds in KCLE and KGRR radar stations. Bottom: Fractional error of the estimated count after 200 labeled days for the 11 radar stations, using different estimators. The bottom-right table shows that the geometric average fractional error across stations for the raw detector and active measurement across iterations. We see that both the adaptation and the sampling without replacement are beneficial, and quickly outperform the detector and baselines for both tasks.

Baselines. We compare our work against DISCount [32], which we will denote DIS. We also investigate the impact of the two major extensions individually: sampling without replacement (DIS+WOR) and fine-tuning the detector similar to adaptive importance sampling (DIS+AIS). Active measurement combines all of these, so we denote it as DIS+AIS+WOR. We also compare against simple Monte Carlo (MC), i.e., DIS with uniform q [32], and MC without replacement (MC+WOR). Lastly we show the performance of the raw detector predicted counts after each step of adaptation.

Since averaging over many trials is computationally expensive—particularly when detector fine-tuning (+AIS) is involved—we approximate the effect of interactive model adaptation using a "fixed checkpoint" approach that uses a fixed sequence of detectors pre-trained on an increasing number of labels sampled according to a fixed strategy. Specifically, we sample annotation units uniformly from those with non-zero counts for different sizes. These fixed model checkpoints are then reused across all trials, allowing us to run a significantly larger number of evaluations. In a real-world deployment, a practitioner would typically fine-tune the detector using the same samples on which annotations are collected during active measurement. However, we find that the estimates produced using this approximation match those obtained via a full end-to-end setup, as discussed in § D.3.

Evaluation metrics. We evaluate our methods using fractional error and CI coverage. Specifically, we estimate them using the *fractional error* $1/M \sum_{m=1}^{M} |\hat{F}_{1:t}^{(m)} - F(\Omega)|/F(\Omega)$ and *coverage* $1/M \sum_{m=1}^{M} \mathbb{I}\big[|\hat{F}_{1:t}^{(m)} - F(\Omega)| \leq z_{\alpha/2}\big(\widehat{\mathrm{Var}}_{1:t}^{(m)}\big)^{1/2}\big]$, where M is the number of trials, $F(\Omega)$ is the ground truth measurement, and $(\hat{F}_{1:t}^{(m)}, \widehat{\mathrm{Var}}_{1:t}^{(m)})$ are the estimates from the mth trial.

7 Results

Main results. Our main results for both tasks are presented in Fig. 3. The top row illustrates that simply using raw detector predictions (i.e., directly interpreting model outputs as estimates) can lead to suboptimal performance, particularly when the detector is biased or exhibits irreducible

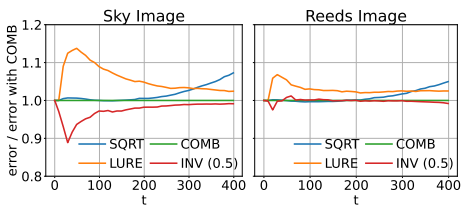


Figure 4: Relative errors compared to α^{COMB} weighting. Other fixed weighting strategies (α^{SQRT}, α^{LURE}) are worse, but inverse variance weighting (denoted by INV ($\gamma = 0.5$)) may achieve lower error.

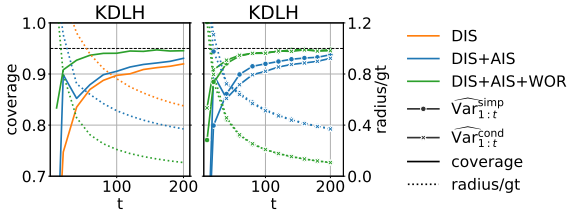


Figure 5: Coverage and radius (relative to the ground-truth) of CIs on the roost data for station KDLH as a function of t (from 5,000 replications), built with either variance estimators from § 5. The left panel uses the $\widehat{\mathrm{Var}}^{\mathrm{cond}}_{1:t}$ estimator. We achieve the desired coverage with narrower CIs.

errors. In the sky and reeds images, the detector performance improves with additional training but typically saturates. In contrast, our method demonstrates greater robustness to biased predictions and continues to improve as detector performance improves. Compared to other unbiased estimators, active measurement (AIS+DIS+WOR) outperforms DISCount (DIS) with as little as 1% of labeled tiles, and further gains are achieved by sampling without replacement once approximately 10% of tiles have been labeled. These improvements are particularly notable in the challenging reeds image, where the detector's baseline performance leaves considerable room for enhancement.

On the radar task, shown in the top-right panels and the bottom row, the detector performance (summarized in the table) is substantially worse. Fine-tuning on station-specific data reduces the fractional error from 3.78 to approximately 2, averaged across stations, but performance again saturates. However, active measurement (AIS+DIS+WOR) continues to yield improvements. Improvements are consistent across all radar stations and are substantially better than DISCount (DIS).

While the fractional error of the raw detector saturates quickly, we observe that additional fine-tuning continues to improve the proposal distribution, which in turn leads to better AIS performance (see § D.2 for details). In Fig. A4, we also compare true end-to-end training with the more computationally efficient fixed checkpoint scheme. We observe that fractional errors between the two methods are not significantly different, especially when fewer tiles are labeled. With a higher number of labeled tiles, we see greater variance arising from the choice of the specific checkpoint; however, the relative performance remains consistent.

Different weighting schemes. We now examine the performance of different weighting schemes in active measurement. According to our theory, we expect that our proposed weights α^{COMB} work consistently well for all t. In Fig. 4, we find that this is the case for the counting problem in the high resolution images. When t is small, the visual detector works poorly, so weighting schemes that assign equal weights early-on, like α^{LURE}, work worse than α^{COMB}. As t increases, the reduction of the sample space is more significant, so weighting schemes that do not consider this, like the α^{SQRT} weights, works increasingly worse. For large t, as the performance gains from fine-tuning saturate, the performance of LURE weights will eventually converge to COMB weights, but are never better. We also test the inverse variance weighting scheme with different hyperparameters γ. The full results can be found in Fig. A5. In general, a conservative $\gamma = 0.3$ brings little benefit, while an aggressive $\gamma = 0.9$ can be detrimental because each individual estimator may not be accurate enough. In the middle, we find that $\gamma = 0.5$ performs well and can achieve even smaller error than α^{COMB}.

Confidence intervals. We evaluate variance estimators and confidence intervals by looking at CI coverage and width over 5,000 trials on the radar counting problem. Results from one station are shown in Fig. 5, and the full results can be found in Fig. A7. Overall, we find that the DISCount baseline undercovers *and* has wider confidence intervals, due to its higher estimation error—we expect width to be proportional to error if the coverage is correct. With adaptation, both metrics improve, but are different with different estimators. We compare the two variance estimators $\widehat{\mathrm{Var}}^{\mathrm{simp}}_{1:t}$ and $\widehat{\mathrm{Var}}^{\mathrm{cond}}_{1:t}$ in the second panel. In most of our experiments, the coverage with them both improves with more samples and converges to the desired confidence level.

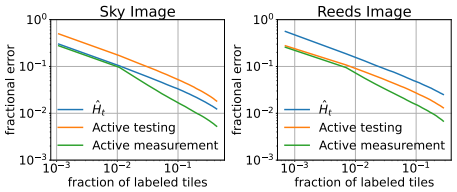


Figure 6: Fractional error compared with other baselines ($\hat{H}$ is motivated by PPI).

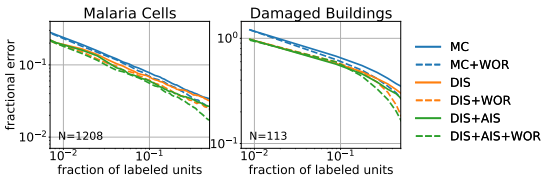


Figure 7: Estimation error on additional datasets, averaged over 1,000 end-to-end trials.

Comparison with other baselines. We use an experiment to illustrate the differences between active measurement and other baselines. Assume the loss function is $|f(s) - g(s)|$ for any $s \in \Omega$. Active testing will use an acquisition distribution proportional to the approximate loss to select units to label. Even if we know the ground-truth loss values, this acquisition distribution is worse than active measurement because of the mismatch between the proposal distribution and the estimand. Similarly, another estimator, motivated by PPI, is $\hat{H}_t = \sum_{s\in\Omega} g(s) - \frac{N}{t}\sum_{\tau=1}^{t}(g(s_\tau) - f(s_\tau))$ with uniformly sampled labeled data. This estimator is also worse than active measurement because it requires a fixed detector and does not select the best labels for estimation. This is shown in Fig. 6.

Additional results. Additional experimental results on counting malaria cells and counting damaged buildings are in Fig. 7. Here we run active measurement with end-to-end settings and average over 1,000 trials. From the first figure, we observe trends similar to those seen in the image and radar experiments for malaria cell counting. DIS consistently outperforms MC, while AIS shows an early reduction in error as the model improves. WOR also demonstrates steady improvement, with its impact becoming more pronounced after roughly 10% of the data is labeled. The results on the second dataset follow similar trends: WOR provides the most benefit when a larger fraction is labeled, while AIS is more helpful early on. We also note that our performance is lower than the numbers reported in DISCount [32], as they use a detector specifically trained for damaged building detection, whereas we use a simple ImageNet-pretrained backbone fine-tuned on just 5 satellite images. These results show that active measurement is capable of improving scientific estimation tasks across different domains.

8 Conclusions, Limitations, Future Work

We introduced *active measurement*, a framework that interactively leverages AI models to achieve precise scientific measurements. In contrast to traditional workflows, which are prone to errors from fully automated methods, active measurement integrates Monte Carlo estimation and model adaptation with iterative human labeling. This approach yields unbiased estimates with calibrated confidence intervals. We formally derive the unbiasedness and consistency of our estimators and propose novel techniques for sequential update weighting and uncertainty quantification. Empirical results on two measurement tasks show that active measurement not only reduces estimation error but also provides reliable uncertainty estimates, outperforming existing methods.

One limitation of our approach is that the estimates may still not be precise enough for all applications; users should carefully validate the safety of any AI method in high-stakes scenarios. Another is that AI measurement has potential to be misused, e.g., to target vulnerable populations; our work does not enable new measurements but can make them more accurate.

There are several promising directions for future work. Beyond simple fine-tuning, unsupervised, semi-supervised, and transductive approaches could improve predictions across the full dataset. The current importance sampling strategy prioritizes units with high counts, but early-stage measurements could adopt more balanced sampling to better support detector training—for example, through active learning. Although model adaptation improves performance, there is a substantial computational overhead; in our experiments, we fine-tuned deep networks for detection tasks on GPUs. To enable interactivity, future systems must support fast and lightweight model updates. Approaches such as few-shot counting models and in-context learning architectures may offer practical solutions for efficient adaptation. Another promising opportunity is to consider correlation among images. As an example, in the reeds image we have observed spatial auto-correlation in the residuals of the AI model. Fitting a Gaussian process can exploit this spatial structure to improve predictions for unlabeled tiles.

Acknowledgments and Disclosure of Funding

We thank Maria C. T. D. Belotti for helpful discussions and contributing the algorithm to estimate bird counts from radar data. We also thank Jody Dole for kindly granting permission to use the sky and reeds photographs featured in our main experiment. This work is supported in part by NSF grants #2504073, #2406687, #2329927, and #2210979.

References

[1] Merlin Sound ID. `https://merlin.allaboutbirds.org/sound-id/`, accessed 13 October 2023.

[2] Ravi Aggarwal, Viknesh Sounderajah, Guy Martin, Daniel SW Ting, Alan Karthikesalingam, Dominic King, Hutan Ashrafian, and Ara Darzi. Diagnostic accuracy of deep learning in medical imaging: a systematic review and meta-analysis. *NPJ digital medicine*, 4(1):65, 2021.

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

[4] Carlos Arteta, Victor Lempitsky, and Andrew Zisserman. Counting in the wild. In *European conference on computer vision*, pages 483–498. Springer, 2016.

[5] Maria Carolina TD Belotti, Yuting Deng, Wenlong Zhao, Victoria F Simons, Zezhou Cheng, Gustavo Perez, Elske Tielens, Subhransu Maji, Daniel Sheldon, Jeffrey F Kelly, et al. Long-term analysis of persistence and size of swallow and martin roosts in the US Great Lakes. *Remote Sensing in Ecology and Conservation*, 2023.

[6] Monica F Bugallo, Victor Elvira, Luca Martino, David Luengo, Joaquin Miguez, and Petar M Djuric. Adaptive importance sampling: The past, the present, and the future. *IEEE Signal Processing Magazine*, 34(4):60–79, 2017.

[7] Jeffrey J Buler, Lori A Randall, Joseph P Fleskes, Wylie C Barrow Jr, Tianna Bogart, and Daria Kluver. Mapping wintering waterfowl distributions using weather surveillance radar. *PloS one*, 7(7):e41571, 2012.

[8] O. Chapelle, B. Scholkopf, and A. Zien, Eds. Semi-supervised learning [book reviews]. *IEEE Transactions on Neural Networks*, 20(3):542–542, 2009. doi: 10.1109/TNN.2009.2015974.

[9] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255, 2009. doi: 10.1109/CVPR.2009.5206848.

[10] Yuting Deng, Maria Carolina TD Belotti, Wenlong Zhao, Zezhou Cheng, Gustavo Perez, Elske Tielens, Victoria F Simons, Daniel R Sheldon, Subhransu Maji, Jeffrey F Kelly, et al. Quantifying long-term phenological patterns of aerial insectivores roosting in the Great Lakes region using weather surveillance radar. *Global Change Biology*, 29(5):1407–1419, 2023.

[11] A. Dutta, A. Gupta, and A. Zissermann. VGG image annotator (VIA), 2016. Version: 2.0.12, Accessed: Mar. 2025.

[12] Sebastian Farquhar, Yarin Gal, and Tom Rainforth. On statistical bias in active learning: How and when to fix it. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*, 2021.

[13] Fernando Fuentes-Peñailillo, Karen Gutter, Ricardo Vega, and Gilda Carrasco Silva. Transformative technologies in digital agriculture: Leveraging internet of things, remote sensing, and artificial intelligence for smart crop management. *Journal of Sensor and Actuator Networks*, 13 (4):39, 2024.

[14] Demin Gao, Quan Sun, Bin Hu, and Shuo Zhang. A framework for agricultural pest and disease monitoring based on internet-of-things and unmanned aerial vehicles. *Sensors*, 20(5):1487, 2020.

[15] Ritwik Gupta, Bryce Goodman, Nirav Patel, Ricky Hosfelt, Sandra Sajeev, Eric Heim, Jigar Doshi, Keane Lucas, Howie Choset, and Matthew Gaston. Creating xBD: A dataset for assessing building damage from satellite imagery. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, pages 10–17, 2019.

[16] Peter Hall and Christopher C Heyde. *Martingale limit theory and its application*. Academic press, 2014.

[17] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition, 2015. URL `https://arxiv.org/abs/1512.03385`.

[18] Jason W Horn and Thomas H Kunz. Analyzing NEXRAD doppler radar images to assess nightly dispersal patterns and population trends in Brazilian free-tailed bats (tadarida brasiliensis). *Integrative and Comparative Biology*, 48(1):24–39, 2008.

[19] Justin Kay, Peter Kulits, Suzanne Stathatos, Siqi Deng, Erik Young, Sara Beery, Grant Van Horn, and Pietro Perona. The Caltech fish counting dataset: A benchmark for multiple-object tracking and counting. In *European Conference on Computer Vision*, pages 290–311. Springer, 2022.

[20] Jannik Kossen, Sebastian Farquhar, Yarin Gal, and Tom Rainforth. Active testing: Sample-efficient model evaluation. In *International Conference on Machine Learning*, pages 5753–5763. PMLR, 2021.

[21] Adriana Kovashka, Olga Russakovsky, Li Fei-Fei, Kristen Grauman, et al. Crowdsourcing in computer vision. *Foundations and Trends® in computer graphics and Vision*, 10(3):177–243, 2016.

[22] Nico Lang, Walter Jetz, Konrad Schindler, and Jan Dirk Wegner. A high-resolution canopy height model of the earth. *Nature Ecology & Evolution*, pages 1–12, 2023.

[23] Chris J Lintott, Kevin Schawinski, Anže Slosar, Kate Land, Steven Bamford, Daniel Thomas, M Jordan Raddick, Robert C Nichol, Alex Szalay, Dan Andreescu, et al. Galaxy Zoo: morphologies derived from visual inspection of galaxies from the Sloan Digital Sky Survey. *Monthly Notices of the Royal Astronomical Society*, 389(3):1179–1189, 2008.

[24] Vebjorn Ljosa, Katherine L Sokolnicki, and Anne E Carpenter. Annotated high-throughput microscopy image sets for validation. *Nature methods*, 9(7):637, 2012.

[25] Chenlin Meng, Enci Liu, Willie Neiswanger, Jiaming Song, Marshall Burke, David Lobell, and Stefano Ermon. IS-Count: Large-scale object counting from satellite images with covariate-based importance sampling. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 12034–12042, 2022.

[26] Phuc Nguyen, Deva Ramanan, and Charless Fowlkes. Active testing: An efficient and robust framework for estimating accuracy. In *International Conference on Machine Learning*, pages 3759–3768. PMLR, 2018.

[27] Man-Suk Oh and James O Berger. Adaptive importance sampling in Monte Carlo integration. *Journal of statistical computation and simulation*, 41(3-4):143–168, 1992.

[28] Art B Owen. Monte Carlo theory, methods and examples, 2013.

[29] Art B Owen and Yi Zhou. The square root rule for adaptive importance sampling. *ACM Transactions on Modeling and Computer Simulation (TOMACS)*, 30(2):1–12, 2020.

[30] Genevieve Patterson, Grant Van Horn, Serge Belongie, Pietro Perona, and James Hays. Tropel: Crowdsourcing detectors with minimal training. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, volume 3, pages 150–159, 2015.

[31] Gustavo Perez, Wenlong Zhao, Zezhou Cheng, Maria Carolina T. D. Belotti, Yuting Deng, Victoria F. Simons, Elske Tielens, Jeffrey F. Kelly, Kyle G. Horton, Subhransu Maji, and Daniel Sheldon. Using spatio-temporal information in weather radar data to detect and track communal bird roosts. *bioRxiv*, 2022. doi: 10.1101/2022.10.28.513761. URL `https://www.biorxiv.org/content/early/2022/10/31/2022.10.28.513761`.

[32] Gustavo Perez, Subhransu Maji, and Daniel Sheldon. DISCount: counting in large image collections with detector-based importance sampling. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 22294–22302, 2024.

[33] François Portier and Bernard Delyon. Asymptotic optimality of adaptive importance sampling. *Advances in Neural Information Processing Systems*, 31, 2018.

[34] Viresh Ranjan, Udbhav Sharma, Thu Nguyen, and Minh Hoai. Learning to count everything. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3394–3403, 2021.

[35] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks, 2016. URL `https://arxiv.org/abs/1506.01497`.

[36] Khoi Nguyen Thanh Nguyen, Chau Pham and Minh Hoai. Few-shot Object Counting and Detection. In *Proceedings of the European Conference on Computer Vision 2022*, 2022.

[37] Arun James Thirunavukarasu, Darren Shu Jeng Ting, Kabilan Elangovan, Laura Gutierrez, Ting Fang Tan, and Daniel Shu Wei Ting. Large language models in medicine. *Nature medicine*, 29(8):1930–1940, 2023.

[38] Devis Tuia, Benjamin Kellenberger, Sara Beery, Blair R Costelloe, Silvia Zuffi, Benjamin Risse, Alexander Mathis, Mackenzie W Mathis, Frank van Langevelde, Tilo Burghardt, et al. Perspectives in machine learning for wildlife conservation. *Nature communications*, 13(1):792, 2022.

[39] Grant Van Horn, Oisin Mac Aodha, Yang Song, Yin Cui, Chen Sun, Alex Shepard, Hartwig Adam, Pietro Perona, and Serge Belongie. The iNaturalist species classification and detection dataset. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8769–8778, 2018.

[40] Mike Walmsley, Chris Lintott, Tobias Géron, Sandor Kruk, Coleman Krawczyk, Kyle W Willett, Steven Bamford, Lee S Kelvin, Lucy Fortson, Yarin Gal, William Keel, Karen L Masters, Vihang Mehta, Brooke D Simmons, Rebecca Smethurst, Lewis Smith, Elisabeth M Baeten, and Christine Macmillan. Galaxy Zoo DECaLS: Detailed visual morphology measurements from volunteers and deep learning for 314,000 galaxies. *Monthly Notices of the Royal Astronomical Society*, 509(3):3966–3988, September 2021. ISSN 1365-2966. doi: 10.1093/mnras/stab2093.

[41] Hanchen Wang, Tianfan Fu, Yuanqi Du, Wenhao Gao, Kexin Huang, Ziming Liu, Payal Chandak, Shengchao Liu, Peter Van Katwyk, Andreea Deac, et al. Scientific discovery in the age of artificial intelligence. *Nature*, 620(7972):47–60, 2023.

[42] Yuxin Wu, Alexander Kirillov, Francisco Massa, Wan-Yen Lo, and Ross Girshick. Detectron2. `https://github.com/facebookresearch/detectron2`, 2019.

[43] Yi Zhou. *Adaptive importance sampling for integration*. Phd thesis, Stanford University, 1998.

NeurIPS Paper Checklist

1. **Claims**

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

   Answer: [Yes]

   Justification: Our derivation of estimators, weighting schemes, and confidence intervals is included in Section 2. We show that our method results in lower estimation error in Figure 3. The coverages of our confidence intervals are verified in Figure 5.

   Guidelines:

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

2. **Limitations**

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

   Answer: [Yes]

   Justification: Section 6 discusses limitations of our work, such as fine-tuning overhead and suboptimal detector improvement in the early stages of measurement.

   Guidelines:

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

3. **Theory assumptions and proofs**

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

Answer: [Yes]

Justification: We provide a proof for all of our propositions and state the assumptions clearly in the appendix.

Guidelines:

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

4. **Experimental result reproducibility**

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

Answer: [Yes]

Justification: We provide pseudo code in section 2 for the Active measurement and variance estimation algorithms. In section 4 we provide the details of counting birds in high-resolution images, which uses an off-the-shelf detector, and counting birds in radar, which closely follows Perez et al. [32]. Finally we link to our code on github at the end of the introduction.

Guidelines:

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

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

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

   Answer: [Yes]

   Justification: We provide access to our code and documentation on github with a link at the end of the introduction.

   Guidelines:

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

6. **Experimental setting/details**

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

   Answer: [Yes]

   Justification: We provide the important details in the results section, and more specific information in the appendix.

   Guidelines:

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

7. **Experiment statistical significance**

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

   Answer: [Yes]

   Justification: Yes, an important contribution of our work is that it provides error bars for the measurement task. We include these error bars in our results and evaluate them.

   Guidelines:

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

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

8. **Experiments compute resources**

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

   Answer: [Yes]

   Justification: We mention the compute hardware used in the experiments section, and provide more detail in the appendix.

   Guidelines:

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

9. **Code of ethics**

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

   Answer: [Yes]

   Justification: Justification: We conform with the NeurIPS Code of Ethics to the best of our ability.

   Guidelines:

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

10. **Broader impacts**

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

    Answer: [Yes]

    We discuss both positive and negative societal impacts in Section 8.

    Guidelines:

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

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

11. **Safeguards**

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

    Answer: [NA]

    Justification: The paper poses no such risks. We do not plan to release pre-trained models, as running our method is focused on the fine-tuning procedure itself, not in the final weights of a trained model.

    Guidelines:

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

12. **Licenses for existing assets**

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

    Answer: [Yes]

    Justification: The NEXRAD radar data that we use is open to the public and can be used as desired, and the roost and sky images were used with permission from the photographer. We adapt code from Perez et al. [31] and will give credit in the publicly released repository.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.

- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

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

    Answer: [Yes]

    Justification: Documentation for the data and code is provided with the code submission.

    Guidelines:

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

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

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

    Answer: [NA]

    Justification: The paper does not involve crowdsourcing or research with human subjects.

    Guidelines:

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

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

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

    Answer: [NA]

    Justification: The paper does not involve crowdsourcing or research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.

- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

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

    Answer: [NA]

    Justification: The paper's method does not involve LLMs.

    Guidelines:

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

A Proof of the theoretical results

In this section, we restate each theoretical result and provide proofs.

A.1 Proof of Proposition 1

Proposition 1. *The combined estimator* $\hat{F}_{1:t} = \sum_{\tau=1}^{t} \bar{\alpha}_\tau \, \hat{F}_\tau$ *is unbiased:* $\mathbb{E}[\hat{F}_{1:t}] = F(\Omega)$.

Proof. We first show that each individual estimator $\hat{F}_\tau$ is unbiased.

$$\begin{aligned}
\mathbb{E}_{\mathcal{D}_\tau, s_\tau}[\hat{F}_\tau] &= \mathbb{E}_{\mathcal{D}_\tau, s_\tau}\left[F(D_\tau) + \frac{f(s_\tau)}{q_\tau(s_\tau)}\right] \\
&= \mathbb{E}_{\mathcal{D}_\tau}\left[F(D_\tau) + \mathbb{E}_{s_\tau}\left[\frac{f(s_\tau)}{q_\tau(s_\tau)}\right]\right] \\
&= \mathbb{E}_{\mathcal{D}_\tau}\left[F(D_\tau) + \sum_{s_\tau \in \Omega \setminus \mathcal{D}_\tau} q_\tau(s_\tau)\frac{f(s_\tau)}{q_\tau(s_\tau)}\right] \\
&= \mathbb{E}_{\mathcal{D}_\tau}\left[F(D_\tau) + F(\Omega \setminus \mathcal{D}_\tau)\right] \\
&= F(\Omega).
\end{aligned}$$

Therefore, for the combined estimator

$$\begin{aligned}
\mathbb{E}[\hat{F}_{1:t}] &= \mathbb{E}\left[\sum_{\tau=1}^{t} \bar{\alpha}_\tau \, \hat{F}_\tau\right] \\
&= \sum_{\tau=1}^{t} \bar{\alpha}_\tau \, \mathbb{E}[\hat{F}_\tau] \\
&= \sum_{\tau=1}^{t} \bar{\alpha}_\tau \, F(\Omega) \\
&= F(\Omega).
\end{aligned}$$

□

A.2 Proof of proposition 2

Proposition 2. *For any* $1 \leq \tau < r \leq t$, $\mathrm{Cov}(\hat{F}_\tau, \hat{F}_r) = 0$.

Proof.

$$\begin{aligned}
\mathrm{Cov}[\hat{F}_\tau, \hat{F}_r] &= \mathbb{E}[(\hat{F}_\tau - F(\Omega))(\hat{F}_r - F(\Omega))] \\
&= \mathbb{E}_{\mathcal{D}_r}[(\hat{F}_\tau - F(\Omega)) \, \mathbb{E}_{s_r}[\hat{F}_r - F(\Omega)]] \\
&= \mathbb{E}_{\mathcal{D}_r}\left[(\hat{F}_\tau - F(\Omega)) \sum_{s_r \in \Omega \setminus \mathcal{D}_r} q_r(s_r)\left(F(\mathcal{D}_r) + \frac{f(s_r)}{q_r(s_r)} - F(\Omega)\right)\right] \\
&= \mathbb{E}_{\mathcal{D}_r}\left[(\hat{F}_\tau - F(\Omega))\left(F(\mathcal{D}_r) + \sum_{s_r \in \Omega \setminus \mathcal{D}_r} q_r(s_r)\frac{f(s_r)}{q_r(s_r)} - F(\Omega)\right)\right] \\
&= \mathbb{E}_{\mathcal{D}_r}\Big[(\hat{F}_\tau - F(\Omega)) \cdot 0\Big] \\
&= 0.
\end{aligned}$$

□

A.3 Proof of proposition 3

Proposition 3. *If there exists constants $0 < A \le B$ such that for any $s \in \Omega$, $A \le f(s) \le B$ and $A \le g(s) \le B$, then there exists a constant $C > 0$ such that $\mathrm{Var}[\hat{F}_\tau] \le C(N-\tau)(N-\tau+1)$.*

Proof. To start, we first bound the sampling distribution implied by the detector measurements $g(s)$. For unit $s \in \Omega \backslash \mathcal{D}_\tau$, the sampling distribution is $q_\tau(s) = \frac{g(s)}{\sum_{s' \in \Omega \backslash \mathcal{D}_\tau} g(s')}$. By the lower and upper bounds, we can see that $\frac{A}{(N-\tau+1)B} \le q_\tau(s) \le \frac{B}{(N-\tau+1)A}$. Also, we can bound $F(\Omega) - F(\mathcal{D}_\tau) = \sum_{s \in \Omega \backslash \mathcal{D}_\tau} f(s)$ with the same technique, which is $(N-\tau+1)A \le F(\Omega) - F(\mathcal{D}_\tau) \le (N-\tau+1)B$.

Next, we write the variance of $\hat{F}_\tau$ in terms of expectations.

$$\begin{aligned}
\mathrm{Var}[\hat{F}_\tau] &= \mathbb{E}_{\mathcal{D}_\tau, s_\tau}\left[\left(F(\mathcal{D}_\tau) + \frac{f(s_\tau)}{q_\tau(s_\tau)} - F(\Omega)\right)^2\right] \\
&= \mathbb{E}_{\mathcal{D}_\tau, s_\tau}\left[\frac{f(s_\tau)^2}{q_\tau(s_\tau)^2} + (F(\mathcal{D}_\tau) - F(\Omega))^2 + 2(F(\mathcal{D}_\tau) - F(\Omega))\frac{f(s_\tau)}{q_\tau(s_\tau)}\right] \\
&\le \mathbb{E}_{\mathcal{D}_\tau, s_\tau}\left[\frac{(N-\tau+1)^2 B^4}{A^2} + (N-\tau+1)^2 B^2 - 2(N-\tau+1)^2 \frac{A^3}{B}\right] \\
&= (N-\tau+1)^2 \frac{B^3(B^2+A^2) - 2A^5}{A^2 B}.
\end{aligned}$$

When $\tau < N$, we further have

$$\begin{aligned}
\mathrm{Var}[\hat{F}_\tau] &= (N-\tau+1)^2 \frac{B^3(B^2+A^2) - 2A^5}{A^2 B} \\
&= (N-\tau)(N-\tau+1)\frac{N-\tau+1}{N-\tau}\frac{B^3(B^2+A^2) - 2A^5}{A^2 B} \\
&\le (N-\tau)(N-\tau+1)\frac{2B^3(B^2+A^2) - 4A^5}{A^2 B}.
\end{aligned}$$

Also note that

$$\begin{aligned}
\mathrm{Var}[\hat{F}_N] &= \mathbb{E}_{\mathcal{D}_N, s_N}\left[\left(F(\mathcal{D}_N) + \frac{f(s_N)}{q_N(s_N)} - F(\Omega)\right)^2\right] \\
&= \mathbb{E}_{\mathcal{D}_N, s_N}\left[\left(F(\mathcal{D}_N) + \frac{f(s_N)}{1} - F(\Omega)\right)^2\right] \\
&= \mathbb{E}_{\mathcal{D}_N, s_N}\left[\left(F(\Omega) - F(\Omega)\right)^2\right] = 0.
\end{aligned}$$

This means that for all $\tau \le N$, $\mathrm{Var}[\hat{F}_\tau] \le C(N-\tau)(N-\tau+1)$, where $C = \frac{2B^3(B^2+A^2)-4A^5}{A^2 B}$. □

A.4 Proof of proposition 4

Proposition 4. *If $\mathrm{Var}[\hat{F}_\tau] \propto \tau^{-y}/w_\tau$, where $y \in [0,1]$, and w_τ are non-decreasing, then the estimate $\hat{F}^{\mathrm{comb}}_{1:t} = \sum_{\tau=1}^t \alpha^{\mathrm{comb}}_\tau \hat{F}_\tau$, where $\alpha^{\mathrm{comb}}_\tau \propto w_\tau \sqrt{\tau}$ and $\sum_{\tau=1}^t \alpha^{\mathrm{comb}}_\tau = 1$, satisfy*

$$\sup_{t \ge 1} \sup_{y \in [0,1]} \frac{\mathrm{Var}[\hat{F}^{\mathrm{comb}}_{1:t}]}{\mathrm{Var}_{\mathrm{opt}}(y,t)} \le \frac{9}{8},$$

where $\mathrm{Var}_{\mathrm{opt}}(y,t)$ is the estimation variance when the estimators are weighted proportional to inverse of ground-truth variances.

Proof. Denote the variance of combined estimator weighted by $\alpha^{(x)}_\tau \propto w_\tau \tau^x$ by $\mathrm{Var}_x(y,t)$, where $x \in [0,1]$. An observation is that $\mathrm{Var}[\hat{F}^{\mathrm{comb}}_{1:t}] = \mathrm{Var}_{0.5}(y,t)$, and $\mathrm{Var}_{\mathrm{opt}}(y,t) = \mathrm{Var}_y(y,t)$. So we

shall bound $\frac{\text{Var}_{0.5}(y,t)}{\text{Var}_y(y,t)}$. Note that for any $x \in [0,1]$,

$$\begin{aligned}\text{Var}_x(y,t) &\propto \frac{\sum_{\tau=1}^t w_\tau^2 \tau^{2x} \cdot \tau^{-y}/w_\tau}{\left(\sum_{\tau=1}^t w_\tau \tau^x\right)^2} \\ &= \frac{\sum_{\tau=1}^t w_\tau \tau^{2x-y}}{\left(\sum_{\tau=1}^t w_\tau \tau^x\right)^2}.\end{aligned}$$

Therefore, let

$$l_x(y,t) = \frac{\text{Var}_x(y,t)}{\text{Var}_y(y,t)} = \frac{\left(\sum_{\tau=1}^t w_\tau \tau^{2x-y}\right)\left(\sum_{\tau=1}^t w_\tau \tau^y\right)}{\left(\sum_{\tau=1}^t w_\tau \tau^x\right)^2},$$

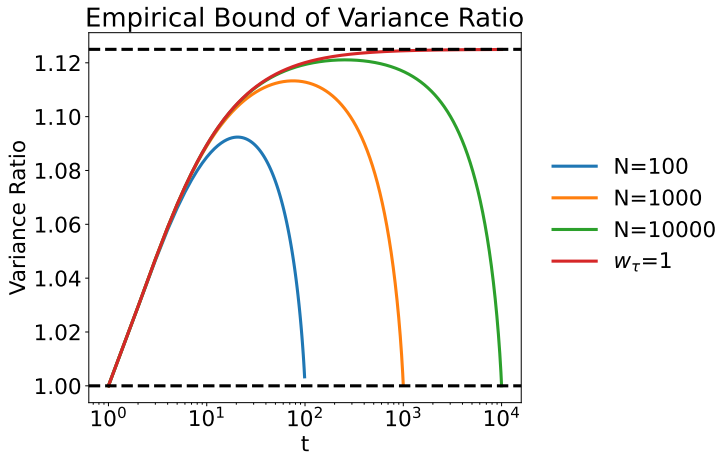
we have

$$\frac{\partial^2 l_x(y,t)}{\partial y^2} = \frac{\sum_{\tau=1}^t \sum_{\tau'=1}^t w_\tau w'_\tau \tau^{2x-y}\tau'^y (\log\tau - \log\tau')^2}{\left(\sum_{\tau=1}^t w_\tau \tau^x\right)^2} > 0.$$

So $l_x(y,t)$ is strictly convex over y. Also note that $l_x(y,t)$ is symmetric around $y = x$. So $\sup_{y\in[0,1]} l_x(y,t) = \begin{cases} l_x(1,t) & x \le 1/2 \\ l_x(0,t) & x \ge 1/2 \end{cases}$. In the combined estimator we chose $x = 0.5$, which leads to the supremum $l_{0.5}(1,t)$ for any t. Furthermore, $x = 0.5$ is the best assumption for the mixing weights. The derivation is the same as Owen and Zhou [29] and we omit it here. In such case

$$\sup_{y\in[0,1]} l_{0.5}(y,t) = l_{0.5}(1,t) = \frac{\left(\sum_{\tau=1}^t w_\tau\right)\left(\sum_{\tau=1}^t w_\tau \tau\right)}{\left(\sum_{\tau=1}^t w_\tau \tau^{0.5}\right)^2} = \frac{\sum_{\tau=1}^t \frac{w_\tau}{\left(\sum_{\tau'=1}^t w_{\tau'}\right)}\tau}{\left(\sum_{\tau=1}^t \frac{w_\tau}{\left(\sum_{\tau'=1}^t w_{\tau'}\right)}\tau^{0.5}\right)^2}.$$

We can plot this specific function for a few different choices of w_τ and t,



The lines labeled $N = x$ correspond to the combined weighting scheme proposed in our main method (see Section 4), and $w_\tau = 1$ is a uniform weighting. We can see empirically that these are all bounded above by $\frac{9}{8}$.

To prove this bound, we define the following weights,

$$\hat{w}_\tau = \frac{w_\tau}{\left(\sum_{\tau'=1}^t w_{\tau'}\right)}, \qquad \beta_\tau = (\hat{w}_\tau - \hat{w}_{\tau-1})(t - \tau + 1).$$

Since w_τ are non-decreasing, $\hat{w}_\tau - \hat{w}_{\tau-1} \geq 0$ and thus $\beta_\tau \geq 0$. We can also cancel out terms of a telescoping series to see that $\sum_{\tau=1}^t \beta_\tau = \sum_{\tau=1}^t \hat{w}_\tau = 1$. Therefore both $\hat{w}_\tau$ and β_τ are non-negative weights that sum to 1. Note we define $\hat{w}_0 = 0$.

Plugging these in, and borrowing a trick from Gabriel's Staircase, we can rewrite as,

$$\frac{\sum_{\tau=1}^t \frac{w_\tau}{\left(\sum_{\tau'=1}^t w_{\tau'}\right)}\tau}{\left(\sum_{\tau=1}^t \frac{w_\tau}{\left(\sum_{\tau'=1}^t w_{\tau'}\right)}\tau^{0.5}\right)^2} = \frac{\sum_{\tau=1}^t \hat{w}_\tau \tau}{\left(\sum_{\tau=1}^t \hat{w}_\tau \tau^{0.5}\right)^2} = \frac{\sum_{\tau=1}^t \beta_\tau \frac{1}{t-\tau+1}\sum_{\tau'=\tau}^t \tau'}{\left(\sum_{\tau=1}^t \beta_\tau \frac{1}{t-\tau+1}\sum_{\tau'=\tau}^t \tau'^{0.5}\right)^2}.$$

We can bound this expression by considering the supremum over all possible positive weights λ_τ,

$$\begin{aligned}\frac{\sum_{\tau=1}^t \beta_\tau \frac{1}{t-\tau+1}\sum_{\tau'=\tau}^t \tau'}{\left(\sum_{\tau=1}^t \beta_\tau \frac{1}{t-\tau+1}\sum_{\tau'=\tau}^t \tau'^{0.5}\right)^2} &\leq \sup_{\lambda\in\mathbb{R}^t_{\geq 0}} \frac{\sum_{\tau=1}^t \frac{\lambda_\tau}{\sum_{\tau'=1}^t \lambda_{\tau'}} \frac{1}{t-\tau+1}\sum_{\tau'=\tau}^t \tau'}{\left(\sum_{\tau=1}^t \frac{\lambda_\tau}{\sum_{\tau'=1}^t \lambda_{\tau'}} \frac{1}{t-\tau+1}\sum_{\tau'=\tau}^t \tau'^{0.5}\right)^2} \\ &= \sup_{\lambda\in\mathbb{R}^t_{\geq 0}} \frac{\left(\sum_{\tau=1}^t \lambda_\tau\right)\left(\sum_{\tau=1}^t \frac{\lambda_\tau}{t-\tau+1}\sum_{\tau'=\tau}^t \tau'\right)}{\left(\sum_{\tau=1}^t \frac{\lambda_\tau}{t-\tau+1}\sum_{\tau'=\tau}^t \tau'^{0.5}\right)^2}.\end{aligned}$$

We can simplify by using the closed form expression for sum of consecutive integers and a lower bound for the sum of consecutive square roots from Owen and Zhou [29],

$$\begin{aligned}&\leq \sup_{\lambda\in\mathbb{R}^t_{\geq 0}} \frac{\left(\sum_{\tau=1}^t \lambda_\tau\right)\left(\sum_{\tau=1}^t \frac{\lambda_\tau}{t-\tau+1}\frac{(t+\tau)(t-\tau+1)}{2}\right)}{\left(\sum_{\tau=1}^t \frac{\lambda_\tau}{t-\tau+1}\frac{(t+0.5)^{1.5}-(\tau-0.5)^{1.5}}{1.5}\right)^2} \\ &= \sup_{\lambda\in\mathbb{R}^t_{\geq 0}} \frac{9}{8}\frac{\left(\sum_{\tau=1}^t \lambda_\tau\right)\left(\sum_{\tau=1}^t \lambda_\tau(t+\tau)\right)}{\left(\sum_{\tau=1}^t \frac{\lambda_\tau}{t-\tau+1}\left((t+0.5)^{1.5}-(\tau-0.5)^{1.5}\right)\right)^2}. \qquad (8)\end{aligned}$$

Define this last expression inside the supremum as $f(\lambda, t)$. Taking the derivative with respect to each λ_n,

$$\begin{aligned}\frac{\partial f(\lambda,t)}{\partial \lambda_n} = &\frac{\left(\sum_{\tau=1}^t \lambda_\tau\,(t+\tau)\right) + (t+n)\left(\sum_{\tau=1}^t \lambda_\tau\right)}{\left(\sum_{\tau=1}^t \lambda_\tau \frac{1}{t-\tau+1}\left((t+0.5)^{1.5}-(\tau-0.5)^{1.5}\right)\right)^2} \\ &- \frac{2\left(\sum_{\tau=1}^t \lambda_\tau\right)\left(\sum_{\tau=1}^t \lambda_\tau\,(t+\tau)\right)\left(\frac{1}{t-n+1}\left((t+0.5)^{1.5}-(n-0.5)^{1.5}\right)\right)}{\left(\sum_{\tau=1}^t \lambda_\tau \frac{1}{t-\tau+1}\left((t+0.5)^{1.5}-(\tau-0.5)^{1.5}\right)\right)^3}.\end{aligned}$$

Setting this equal to 0 and rearranging,

$$\frac{\partial f(\lambda,t)}{\partial \lambda_n} = 0$$

$$\frac{\left(\sum_{\tau=1}^t \frac{\lambda_\tau}{\sum_{\tau'=1}^t \lambda_{\tau'}}\tau\right) + (2t+n)}{\left(\frac{1}{t-n+1}\left((t+0.5)^{1.5}-(n-0.5)^{1.5}\right)\right)} = \frac{2\left(\sum_{\tau=1}^t \lambda_\tau\,(t+\tau)\right)}{\left(\sum_{\tau=1}^t \frac{\lambda_\tau}{t-\tau+1}\left((t+0.5)^{1.5}-(\tau-0.5)^{1.5}\right)\right)}. \qquad (9)$$

Note the right hand side does not depend on n, the index of the partial derivative. Thus at a critical point, the left hand side will equal the same value for all n. We can define this left side as a function $g(n, t, \lambda)$.

To simplify, we make the following substitutions:

$$\begin{aligned}c_1(t,\lambda) &= \left(\sum_{\tau=1}^t \frac{\lambda_\tau}{\sum_{\tau'=1}^t \lambda_{\tau'}}\tau\right) - 0.5 \\ u(n) &= \sqrt{n-0.5} \\ w(t) &= \sqrt{t+0.5}\end{aligned}$$

Since $n \geq 1$, $t \geq 1$, and $n \leq t$, we can note that $u(n) > 0$, $w(t) > 0$, $u(n) < w(t)$, and $0.5 \leq c_1(t, \lambda) \leq t - 0.5$. This alleviates concerns over dividing by zero going forward.

Applying these to $g(n, t, \lambda)$,

$$g(n,t,\lambda) = \frac{c_1(t,\lambda) + 2w(t)^2 + u(n)^2}{\frac{w(t)^3 - u(n)^3}{w(t)^2 - u(n)^2}} = \frac{(w(t) + u(n))\left(c_1(t,\lambda) + 2w(t)^2 + u(n)^2\right)}{w(t)^2 + w(t)u(n) + u(n)^2}.$$

Next we can consider the partial derivative over n and set it equal to zero, noting that $u'(n) = \frac{1}{2u(n)}$,

$$\frac{\partial g(n,t,\lambda)}{\partial n} = \frac{u(n)(u(n)^3 + 2w(t)u(n)^2 + (2w(t)^2 - c_1(t,\lambda))u(n) - 2w(t)(w(t)^2 + c_1(t,\lambda))}{2u\left(u(n)^2 + w(t)u(n) + w(t)^2\right)^2}$$

$$u(n)^3 + 2w(t)u(n)^2 + (2w(t)^2 - c_1(t,\lambda))u(n) - 2w(t)(w(t)^2 + c_1(t,\lambda)) = 0.$$

This is a cubic equation over $u(n)$. Looking at the coefficients, 1 and $2w(t)$ are both positive. Making a substitution and using a bound on c_1, $2w(t)^2 - c_1(t,\lambda) \geq 2t + 1 - t + 0.5 = t + 1.5 > 0$. Clearly the last term is negative. By Descartes' rule of signs, the coefficients of this polynomial have 1 sign change and thus at most 1 positive real root over $u(n)$. Since $u(n)$ is monotonic over n and positive for all $n \geq 1$, we see that there exists at most one n such that $\frac{\partial g(n,t,\lambda)}{\partial n} = 0$, for any given t and λ. From this we can deduce that the equation $g(n, t, \lambda) = h(t, \lambda)$ for any function $h(t, \lambda)$ that does not depend on n has at most 2 solutions over n.

Applying this to Eq. 9 and noting that the left hand side is equal to $g(n, t, \lambda)$, we can conclude that this equation can have no more than 2 solutions over n. Thus the partial derivative of $f(\lambda, t)$ with respect to λ can have at most 2 components with value 0 at the same time. We can also observe $f(c\lambda, t) = f(\lambda, t)$ for all $c \in \mathbb{R}^+$, thus the supremum must occur at some finite point. Since we're optimizing over the positive orthant, that finite point must have all non-zero components with gradient zero. Thus there can be at most 2 non-zero components of λ at the supremum. Continuing from Eq. 8, we now have the upper bound,

$$\begin{aligned}
&\sup_{\lambda \in \mathbb{R}^t_{\geq 0}} \frac{9}{8} \frac{\left(\sum_{\tau=1}^t \lambda_\tau\right)\left(\sum_{\tau=1}^t \lambda_\tau (t+\tau)\right)}{\left(\sum_{\tau=1}^t \frac{\lambda_\tau}{t-\tau+1}\left((t+0.5)^{1.5} - (\tau - 0.5)^{1.5}\right)\right)^2} \\
&\quad = \sup_{\lambda \in \mathbb{R}^t_{\geq 0}, 1 \leq i < j \leq t} \frac{9}{8} \frac{(\lambda_i + \lambda_j)\left(\lambda_i(t+i) + \lambda_j(t+j)\right)}{\left(\frac{\lambda_i}{t-i+1}((t+0.5)^{1.5} - (i-0.5)^{1.5}) + \frac{\lambda_j}{t-j+1}((t+0.5)^{1.5} - (j-0.5)^{1.5})\right)^2}. \qquad (10)
\end{aligned}$$

We will denote this function inside the supremum as z. We can replace λ_i and λ_j with variables x, y and make similar substitutions as before,

$$u(i) = \sqrt{i - 0.5}$$
$$v(j) = \sqrt{j - 0.5}$$
$$w(t) = \sqrt{t + 0.5}$$

For more compact notation we will ignore arguments going forward. Making these substitutions yields,

$$z(x,y,i,j,t) = \frac{9}{8} \frac{(x+y)(x(w^2+u^2) + y(w^2+v^2))}{\left(\frac{x(w^3-u^3)}{w^2-u^2} + \frac{y(w^3-v^3)}{w^2-v^2}\right)^2}.$$

Using a computer algebra system, we can expand this as,

$$z(x,y,i,j,t) = \frac{9}{8}\left(1 - \frac{p}{q}\right)$$

$$\begin{aligned}
p = &-u^4v^2xy - 2u^4vwxy - u^4w^2xy + 2u^3v^3xy + 2u^3v^2wxy - u^2v^4xy + 2u^2v^3wxy + \\
&u^2v^2w^2x^2 + 4u^2v^2w^2xy + u^2v^2w^2y^2 + 2u^2vw^3x^2 + 2u^2vw^3xy + u^2w^4x^2 + \\
&u^2w^4xy - 2uv^4wxy + 2uv^2w^3xy + 2uv^2w^3y^2 - v^4w^2xy + v^2w^4xy + v^2w^4y^2 \\
q = &\left(x(w^2 + wu + u^2)(w+v) + y(w^2 + wv + v^2)(w+u)\right)^2.
\end{aligned}$$

By definition $x, y \geq 0$ and $1 \leq i < j \leq t$, so $u, w, v > 0$, $u < w$, and $v < w$. Thus we can see that $q > 0$. Rearranging terms,

$$\begin{aligned} p = u^3v^2\left(2w-u\right)xy + 2u^2vw\left(w^2-u^2\right)xy + u^2w^2\left(w^2-u^2\right)xy \\ +u^2v^2\left(4w^2-v^2\right)xy + 2uv^2w\left(w^2-v^2\right)xy + v^2w^2\left(w^2-v^2\right)xy \\ +2u^3v^3xy + 2u^2v^3wxy + u^2v^2w^2x^2 + u^2v^2w^2y^2 + 2u^2vw^3x^2 \\ +u^2w^4x^2 + 2uv^2w^3y^2 + v^2w^4y^2. \end{aligned}$$

Since $w > u$ and $w > v$ we can see that p is a sum of positive terms so $p > 0$. Putting this together we have $\frac{p}{q} > 0$. Therefore continuing from Eq. 10,

$$= \sup_{t\geq 1, x, y \geq 0, 1 \leq i < j \leq t} z(x, y, i, j, t) = \sup_{t\geq 1, x, y \geq 0, 1 \leq i < j \leq t} \frac{9}{8}(1 - \frac{p}{q}) \leq \frac{9}{8}.$$

This proves our bound. In conclusion, we first show that the variance ratio is bounded by another ratio involving 2 weighted averages. We then use the fact that the weights are non-decreasing to transform the expressions into a weighted average of unweighted averages. Using closed forms and approximations we find a simpler upper bound. We consider the supremum of this expression over all possible weights, and observe that the maximum can only occur with at most 2 non-zero weights. Finally we analyze this 2 weight case and conclude it is bounded by $\frac{9}{8}$. Thus we have,

$$\sup_{t\geq 1} \sup_{y \in [0,1]} \frac{\mathrm{Var}[\hat{F}^{\mathrm{comb}}_{1:t}]}{\mathrm{Var}_{\mathrm{opt}}(y, t)} \leq \frac{9}{8},$$

□

A.5 Proof of Proposition 5 (martingale CLT)

To construct the formal martingale CLT, we need to let the number of active measurement steps T and the domain size N go to infinity jointly. Suppose that for each N, we set the number of active measurement steps to be T_N ($T_N < N$) where T_N is non-decreasing with N. By letting $N \to \infty$ and $T_N \to \infty$, we have a martingale array $S_{N,t} = \sum_{\tau=1}^{t} X_{N,\tau} = \sum_{\tau=1}^{t} \tilde{\alpha}_\tau(\hat{F}_\tau - F(\Omega))$, for $t \leq T_N$. Here $\tilde{\alpha}_\tau = \alpha_\tau / \sqrt{\sum_{\tau'=1}^{T_N} \alpha_{\tau'}^2 \,\mathrm{Var}[\hat{F}_{\tau'}]}$ is normalized using total variances. It is direct to check that $\mathbb{E}[S_{N,t+1} - S_{N,t} | S_{N,t}] = \mathbb{E}[X_{N,t+1} | S_{N,t}] = 0$. With addition assumptions, we are able to construct a limiting theorem for active measurement.

Proposition 5 (formal)**.** *Assume that (1)* $\sum_{\tau=1}^{T_N} \tilde{\alpha}_\tau^2 \,\mathrm{Var}[\hat{F}_\tau | \,\mathcal{D}_\tau] \to \eta^2$ *as* $N \to \infty$*; (2)* $P(\eta^2 < \infty) = 1$ *and* $P(\eta^2 > 0) = 1$*; (3)* $\sum_{\tau=1}^{T_N} \mathbb{E}[X_{N,\tau}^2 \mathbb{I}(|X_{N,\tau}| > \epsilon) | \,\mathcal{D}_\tau] \xrightarrow{p} 0$ *for any* $\epsilon > 0$*. We have two central limit theorems*

$$\frac{S_{N,T_N}}{\sqrt{\sum_{\tau=1}^{T_N} X_{N,\tau}^2}} \xrightarrow{D} \mathcal{N}(0,1), \tag{11}$$

$$\frac{S_{N,T_N}}{\sqrt{\sum_{\tau=1}^{T_N} \tilde{\alpha}_\tau^2 \,\mathrm{Var}[\hat{F}_\tau | \,\mathcal{D}_\tau]}} \xrightarrow{D} \mathcal{N}(0,1). \tag{12}$$

Proof. We have met all the conditions for Theorem 3.3 and Corollary 3.2 in Hall and Heyde [16]. Then we directly have

$$\frac{S_{N,T_N}}{U_{N,T_N}} \xrightarrow{D} \mathcal{N}(0,1),$$
$$\frac{S_{N,T_N}}{V_{N,T_N}} \xrightarrow{D} \mathcal{N}(0,1),$$

where

$$U_{N,T_N} = \sum_{\tau=1}^{T_N} X_{N,\tau}^2$$
$$V_{N,T_N} = \sum_{\tau=1}^{T_N} \tilde{\alpha}_\tau^2 \,\mathrm{Var}[\hat{F}_\tau | \,\mathcal{D}_\tau].$$

□

The formal proposition employs a different normalizer from active measurement. It is useful for bounding the discrepancy between conditional and total variances. For active measurement, we have the following corollary.

Corollary 1. *With the same conditions as Prop. 5, we also have that*

$$\frac{\hat{F}_{1:T_N} - F(\Omega)}{\sqrt{\sum_{\tau=1}^{T_N} \bar{\alpha}_\tau^2 (\hat{F}_\tau - F(\Omega))^2}} \xrightarrow{D} \mathcal{N}(0,1), \tag{13}$$

$$\frac{\hat{F}_{1:T_N} - F(\Omega)}{\sqrt{\sum_{\tau=1}^{T_N} \bar{\alpha}_\tau^2 \operatorname{Var}[\hat{F}_\tau | \mathcal{D}_\tau]}} \xrightarrow{D} \mathcal{N}(0,1). \tag{14}$$

Proof. Basically the equations are the same as Eq. 11 and 12, but with different normalizers. After multiplying $\frac{\sqrt{\sum_{\tau'=1}^{T_N} \alpha_{\tau'}^2 \operatorname{Var}[\hat{F}_{\tau'}]}}{\sum_{\tau'=1}^{T_N} \alpha_{\tau'}}$ to both the numerator and denominator, we get Eq. 13 and 14. □

This leads to our informal version of the martingale CLT. Both $\operatorname{Var}_{1:t}^{\text{cond}}$ and $\operatorname{Var}_{1:t}^{\text{simp}}$ can be constructed using the corollary.

We also demonstrate our assumptions in Prop. 5 with a simple example. If $\operatorname{Var}[\hat{F}_\tau] = \sigma_\tau^2 = \lambda_1 \tau^{-y} / w_\tau$ and $\operatorname{Var}[\hat{F}_\tau | \mathcal{D}_\tau] = \lambda_2 \tau^{-y} / w_\tau$ where $y \in [0,1]$ and $0 < \lambda_2 \leq \lambda_1$, then with COMB weights

$$\begin{aligned}\sum_{\tau=1}^{T_N} \tilde{\alpha}_\tau^2 \operatorname{Var}[\hat{F}_\tau | \mathcal{D}_\tau] &= \frac{\sum_{\tau=1}^{T_N} \lambda_2 w_\tau \tau^{1-y}}{\sum_{\tau=1}^{T_N} \lambda_1 w_\tau \tau^{1-y}} \\ &= \frac{\lambda_2}{\lambda_1}.\end{aligned}$$

It is clear that $0 < \lambda_2/\lambda_1 < \infty$, so the first two assumptions are met. For the third assumption, let $T_N = N/2$ and further assume for some $\delta > 0$, $\sup_{\tau,N} \frac{\mathbb{E}[(\hat{F}_\tau - F(\Omega))^{2+\delta} | \mathcal{D}_\tau]}{\sigma_\tau^{2+\delta}} < \infty$. The additional assumption implies the uniform boundedness of $\frac{(\hat{F}_\tau - F(\Omega))^2}{\sigma_\tau^2}$, which is common in the literature [43]. Then note that

$$\begin{aligned}\frac{\max_\tau \alpha_\tau^2 \sigma_\tau^2}{\sum_{\tau=1}^{T_N} \alpha_\tau^2 \sigma_\tau^2} &= \frac{\lambda_1 w_{T_N} T_N^{1-y}}{\sum_{\tau=1}^{T_N} \lambda_1 w_\tau \tau^{1-y}} \\ &\leq \frac{\lambda_1 w_{T_N} T_N^{1-y}}{\frac{\lambda_1}{T_N} \left(\sum_{\tau=1}^{T_N} w_\tau\right) \left(\sum_{\tau=1}^{T_N} \tau^{1-y}\right)} \\ &= \frac{N T_N^{1-y}}{(N - T_N + 1) \left(\sum_{\tau=1}^{T_N} \tau^{1-y}\right)} \\ &\leq \frac{N T_N^{1-y}}{(N - T_N + 1) T_N^{2-y}/(2-y)} \to 0.\end{aligned} \tag{15}$$

Therefore, for any τ,

$$\begin{aligned}\mathbb{E}[X_{N,\tau}^2 \mathbb{I}(|X_{N,\tau}| > \epsilon) | \mathcal{D}_\tau] &= \mathbb{E}\left[\frac{\alpha_\tau^2 (\hat{F}_\tau - F(\Omega))^2}{\sum_{\tau'=1}^{T_N} \alpha_{\tau'}^2 \sigma_{\tau'}^2} \mathbb{I}(|X_{N,\tau'}| > \epsilon) | \mathcal{D}_\tau\right] \\ &= \frac{\alpha_\tau^2 \sigma_\tau^2}{\sum_{\tau'=1}^{T_N} \alpha_{\tau'}^2 \sigma_{\tau'}^2} \mathbb{E}\left[\frac{(\hat{F}_\tau - F(\Omega))^2}{\sigma_\tau^2} \mathbb{I}(|X_{N,\tau'}| > \epsilon) | \mathcal{D}_\tau\right].\end{aligned}$$

Note that

$$\mathbb{E}\left[\frac{(\hat{F}_\tau - F(\Omega))^2}{\sigma_\tau^2}\mathbb{I}(|X_{N,\tau'}| > \epsilon)|\,\mathcal{D}_\tau\right] = \mathbb{E}\left[\frac{(\hat{F}_\tau - F(\Omega))^2}{\sigma_\tau^2}\mathbb{I}\left(\frac{|\hat{F}_\tau - F(\Omega)|}{\sigma_\tau} > \epsilon\frac{\sqrt{\sum_{\tau'=1}^{T_N}\alpha_{\tau'}^2\sigma_{\tau'}^2}}{\alpha_\tau\sigma_\tau}\right)|\,\mathcal{D}_\tau\right].$$

Observe that $\frac{\mathbb{E}[(\hat{F}_\tau - F(\Omega))^2|\,\mathcal{D}_\tau]}{\sigma_\tau^2} = \frac{\lambda_2}{\lambda_1}$, but $\frac{\sqrt{\sum_{\tau'=1}^{T_N}\alpha_{\tau'}^2\sigma_{\tau'}^2}}{\alpha_\tau\sigma_\tau} \to \infty$ for any τ according to Eq. 15. By the uniform boundedness, then for any $\epsilon_0 > 0$, there exists N_0 such that for any $N > N_0$,

$$\mathbb{E}\left[\frac{(\hat{F}_\tau - F(\Omega))^2}{\sigma_\tau^2}\mathbb{I}(|X_{N,\tau'}| > \epsilon)|\,\mathcal{D}_\tau\right] < \epsilon_0.$$

Therefore,

$$\begin{aligned}\sum_{\tau=1}^{T_N}\mathbb{E}[X_{N,\tau}^2\mathbb{I}(|X_{N,\tau}| > \epsilon)|\,\mathcal{D}_\tau] &< \sum_{\tau=1}^{T_N}\frac{\alpha_\tau^2\sigma_\tau^2}{\sum_{\tau'=1}^{T_N}\alpha_{\tau'}^2\sigma_{\tau'}^2}\epsilon_0\\ &= \epsilon_0.\end{aligned}$$

This shows that the third assumption is also met.

A.6 Proof of proposition 6

Proposition 6. *The estimators* $\widehat{\mathrm{Var}}_\tau$ *and* $\widehat{\mathrm{Var}}_{\tau,r}$ *for* $\tau \le r \le t$ *satisfy* $\mathbb{E}[\widehat{\mathrm{Var}}_\tau|\,\mathcal{D}_\tau] = \mathbb{E}[\widehat{\mathrm{Var}}_{\tau,r}|\,\mathcal{D}_\tau] = \mathrm{Var}[\hat{F}_\tau|\,\mathcal{D}_\tau]$ *and* $\mathbb{E}[\widehat{\mathrm{Var}}_\tau] = \mathbb{E}[\widehat{\mathrm{Var}}_{\tau,r}] = \mathrm{Var}[\hat{F}_\tau]$.

Proof. We first show the unbiasedness of individual variance estimators $\widehat{\mathrm{Var}}_{\tau,r}$.

$$\begin{aligned}&\mathbb{E}[\widehat{\mathrm{Var}}_{\tau,r}|\,\mathcal{D}_\tau]\\ &= \mathbb{E}_{\mathcal{D}_r\setminus\mathcal{D}_\tau,s_r}\left[\sum_{s\in\mathcal{D}_r\setminus\mathcal{D}_\tau} q_\tau(s)\left(\frac{f(s)}{q_\tau(s)} - F(\Omega\setminus\mathcal{D}_\tau)\right)^2 + \frac{q_\tau(s_r)}{q_r(s_r)}\left(\frac{f(s_r)}{q_\tau(s_r)} - F(\Omega\setminus\mathcal{D}_\tau)\right)^2|\,\mathcal{D}_\tau\right]\\ &= \mathbb{E}_{\mathcal{D}_r\setminus\mathcal{D}_\tau}\left[\sum_{s\in\mathcal{D}_r\setminus\mathcal{D}_\tau} q_\tau(s)\left(\frac{f(s)}{q_\tau(s)} - F(\Omega\setminus\mathcal{D}_\tau)\right)^2 + \mathbb{E}_{s_r}\left[\frac{q_\tau(s_r)}{q_r(s_r)}\left(\frac{f(s_r)}{q_\tau(s_r)} - F(\Omega\setminus\mathcal{D}_\tau)\right)^2\right]|\,\mathcal{D}_\tau\right]\\ &= \mathbb{E}_{\mathcal{D}_r\setminus\mathcal{D}_\tau}\left[\sum_{s\in\mathcal{D}_r\setminus\mathcal{D}_\tau} q_\tau(s)\left(\frac{f(s)}{q_\tau(s)} - F(\Omega\setminus\mathcal{D}_\tau)\right)^2 + \sum_{s_r\in\Omega\setminus\mathcal{D}_r} q_\tau(s_r)\left(\frac{f(s_r)}{q_\tau(s_r)} - F(\Omega\setminus\mathcal{D}_\tau)\right)^2|\,\mathcal{D}_\tau\right]\\ &= \mathbb{E}_{\mathcal{D}_r\setminus\mathcal{D}_\tau}\left[\sum_{s\in\Omega\setminus\mathcal{D}_\tau} q_\tau(s)\left(\frac{f(s)}{q_\tau(s)} - F(\Omega\setminus\mathcal{D}_\tau)\right)^2|\,\mathcal{D}_\tau\right]\\ &= \mathrm{Var}[\hat{F}_\tau|\,\mathcal{D}_\tau].\end{aligned}$$

Also, note that

$$\begin{aligned}\mathrm{Var}_{\mathcal{D}_\tau}[\mathbb{E}[\hat{F}_\tau|\,\mathcal{D}_\tau]] &= \mathrm{Var}_{\mathcal{D}_\tau}\left[\mathbb{E}\left[F(\mathcal{D}_\tau) + \frac{f(s_\tau)}{q(s_\tau)}|\,\mathcal{D}_\tau\right]\right]\\ &= \mathrm{Var}_{\mathcal{D}_\tau}\left[\sum_{s_\tau\in\Omega\setminus\mathcal{D}_\tau} q(s_\tau)\left(F(\mathcal{D}_\tau) + \frac{f(s_\tau)}{q(s_\tau)}\right)\right]\\ &= \mathrm{Var}_{\mathcal{D}_\tau}\left[F(\mathcal{D}_\tau) + \sum_{s_\tau\in\Omega\setminus\mathcal{D}_\tau} q(s_\tau)\frac{f(s_\tau)}{q(s_\tau)}\right]\\ &= \mathrm{Var}_{\mathcal{D}_\tau}\left[F(\Omega)\right]\\ &= 0.\end{aligned}$$

Therefore, we further have

$$\begin{aligned}\mathbb{E}[\mathbb{E}[\widehat{\mathrm{Var}}_{\tau,r}|\,\mathcal{D}_\tau]] &= \mathbb{E}_{\mathcal{D}_\tau}[\mathrm{Var}[\hat{F}_\tau|\,\mathcal{D}_\tau]] \\ &= \mathrm{Var}[\hat{F}_\tau] - \mathrm{Var}_{\mathcal{D}_\tau}[\mathbb{E}[\hat{F}_\tau|\,\mathcal{D}_\tau]] \\ &= \mathrm{Var}[\hat{F}_\tau].\end{aligned}$$

Since $\widehat{\mathrm{Var}}_\tau$ is convex on $\widehat{\mathrm{Var}}_{\tau,r}$ with the weights, we also have $\mathbb{E}[\widehat{\mathrm{Var}}_\tau|\,\mathcal{D}_\tau] = \mathrm{Var}[\hat{F}_\tau|\,\mathcal{D}_\tau]$ and $\mathbb{E}[\widehat{\mathrm{Var}}_\tau] = \mathrm{Var}[\hat{F}_\tau]$. □

A.7 Proof of proposition 7

Proposition 7. *With the same settings as Prop. 3, the variance estimate for $\hat{F}_\tau$ weighted by the LURE scheme $\widehat{\mathrm{Var}}_\tau^{\mathrm{LURE}} = \sum_{r=\tau}^{t} \bar{\beta}_r^{\mathrm{LURE}}\,\widehat{\mathrm{Var}}_{\tau,r}$ satisfies that* $\mathrm{Var}\left[\widehat{\mathrm{Var}}_\tau^{\mathrm{LURE}}|\,\mathcal{D}_\tau\right] \lesssim \frac{1}{t-\tau+1} \cdot \min\left(1, \frac{(N-t)^2}{t-\tau+1}\right)$.

Proof. We first derive the normalized LURE weights.

$$\begin{aligned}\sum_{r=\tau}^{t} \beta_r^{\mathrm{LURE}} &= \sum_{r=\tau}^{t} \frac{1}{(N-r)(N-r+1)} \\ &= \sum_{r=\tau}^{t} \frac{1}{N-r} - \frac{1}{N-r+1} \\ &= \frac{1}{N-t} - \frac{1}{N-\tau+1} \\ &= \frac{t-\tau+1}{(N-t)(N-\tau+1)}.\end{aligned}$$

So the normalized LURE weights satisfy

$$\bar{\beta}_r^{\mathrm{LURE}} = \frac{(N-t)(N-\tau+1)}{(t-\tau+1)(N-r)(N-r+1)}.$$

We then bound the variance of individual variance estimators. Observe that

$$\begin{aligned}\mathrm{Var}[\widehat{\mathrm{Var}}_{\tau,r}|\,\mathcal{D}_\tau] &= \mathrm{Var}[\mathbb{E}[\widehat{\mathrm{Var}}_{\tau,r}|D_r]|\,\mathcal{D}_\tau] + \mathbb{E}[\mathrm{Var}[\widehat{\mathrm{Var}}_{\tau,r}|D_r]|\,\mathcal{D}_\tau] \\ &= \mathrm{Var}[\mathrm{Var}[\hat{F}_\tau]|\,\mathcal{D}_\tau] + \mathbb{E}[\mathrm{Var}[\widehat{\mathrm{Var}}_{\tau,r}|D_r]|\,\mathcal{D}_\tau] \\ &= 0 + \mathbb{E}\left[\mathrm{Var}\left[\frac{q_\tau(s_r)}{q_r(s_r)}\left(\frac{f(s_r)}{q_\tau(s_r)} - F(\Omega\backslash\,\mathcal{D}_\tau)\right)^2 |D_r\right]|\,\mathcal{D}_\tau\right] \\ &\le \mathbb{E}\left[\mathbb{E}\left[\frac{q_\tau(s_r)^2}{q_r(s_r)^2}\left(\frac{f(s_r)}{q_\tau(s_r)} - F(\Omega\backslash\,\mathcal{D}_\tau)\right)^4 |D_r\right]|\,\mathcal{D}_\tau\right] \\ &\le \mathbb{E}\left[\mathbb{E}\left[\frac{q_\tau(s_r)^2}{q_r(s_r)^2}\cdot 8\cdot\left(\frac{f(s_r)^4}{q_\tau(s_r)^4} + F(\Omega\backslash\,\mathcal{D}_\tau)^4\right)|D_r\right]|\,\mathcal{D}_\tau\right] \\ &\le \mathbb{E}\left[\mathbb{E}\left[\frac{B^4(N-r+1)^2}{A^4(N-\tau+1)^2}\cdot 8\cdot\left(\frac{B^4(N-\tau+1)^4B^4}{A^4} + B^4(N-\tau+1)^4\right)|D_r\right]|\,\mathcal{D}_\tau\right] \\ &= \frac{8B^8(B^4+A^4)}{A^8}(N-r+1)^2(N-\tau+1)^2.\end{aligned}$$

Therefore, let $D = \frac{8B^8(B^4+A^4)}{A^8}$, we can say that $\mathrm{Var}[\widehat{\mathrm{Var}}_{\tau,r}|\,\mathcal{D}_\tau] \le D(N-\tau+1)^2(N-r+1)^2$. Next, we show that for every two variance estimators $\mathrm{Cov}(\widehat{\mathrm{Var}}_{\tau,r_1}, \widehat{\mathrm{Var}}_{\tau,r_2}|\,\mathcal{D}_\tau) = 0$ if $\tau \le r_1 < r_2 \le N$.

$$\begin{aligned}\mathrm{Cov}(\widehat{\mathrm{Var}}_{\tau,r_1}, \widehat{\mathrm{Var}}_{\tau,r_2}|\,\mathcal{D}_\tau) &= \mathbb{E}[(\widehat{\mathrm{Var}}_{\tau,r_1} - \mathrm{Var}_{s_\tau}[\hat{F}_\tau|\,\mathcal{D}_\tau])(\widehat{\mathrm{Var}}_{\tau,r_2} - \mathrm{Var}_{s_\tau}[\hat{F}_\tau|\,\mathcal{D}_\tau])|\,\mathcal{D}_\tau] \\ &= \mathbb{E}_{\mathcal{D}_{r_2}\backslash\,\mathcal{D}_\tau}[(\widehat{\mathrm{Var}}_{\tau,r_1} - \mathrm{Var}_{s_\tau}[\hat{F}_\tau|\,\mathcal{D}_\tau])\,\mathbb{E}_{s_{r_2}}[(\widehat{\mathrm{Var}}_{\tau,r_2} - \mathrm{Var}_{s_\tau}[\hat{F}_\tau|\,\mathcal{D}_\tau])]] \\ &= \mathbb{E}_{\mathcal{D}_{r_2}\backslash\,\mathcal{D}_\tau}[(\widehat{\mathrm{Var}}_{\tau,r_1} - \mathrm{Var}_{s_\tau}[\hat{F}_\tau|\,\mathcal{D}_\tau])\cdot 0] \\ &= 0.\end{aligned}$$

Therefore (when $t < N$),

$$\begin{aligned}\mathrm{Var}\left[\widehat{\mathrm{Var}}_\tau^{\mathrm{LURE}}|\,\mathcal{D}_\tau\right] &= \sum_{r=\tau}^{t} \frac{(N-t)^2(N-\tau+1)^2}{(t-\tau+1)^2(N-r)^2(N-r+1)^2}\,\mathrm{Var}[\widehat{\mathrm{Var}}_{\tau,r}] \\ &\le \sum_{r=\tau}^{t} \frac{(N-t)^2(N-\tau+1)^2}{(t-\tau+1)^2(N-r)^2(N-r+1)^2} D(N-\tau+1)^2(N-r+1)^2 \\ &= \frac{D(N-t)^2(N-\tau+1)^4}{(t-\tau+1)^2}\sum_{r=\tau}^{t}\frac{1}{(N-r)^2}.\end{aligned}$$

There are two ways to bound the right hand side. When t is far from N, we can bound with

$$\begin{aligned}\mathrm{Var}\left[\widehat{\mathrm{Var}}_\tau^{\mathrm{LURE}}|\,\mathcal{D}_\tau\right] &\le \frac{D(N-t)^2(N-\tau+1)^4}{(t-\tau+1)^2}\sum_{r=\tau}^{t}\frac{1}{(N-r)^2} \\ &\le \frac{D(N-t)^2(N-\tau+1)^4}{(t-\tau+1)^2}\sum_{r=\tau}^{t}\frac{1}{(N-t)^2} \\ &= \frac{D(N-\tau+1)^4}{t-\tau+1}.\end{aligned}$$

When t is close to N, by the fact that $\sum_{x=1}^{\infty}\frac{1}{x^2} = \frac{\pi^2}{6}$, we have

$$\begin{aligned}\mathrm{Var}\left[\widehat{\mathrm{Var}}_\tau^{\mathrm{LURE}}|\,\mathcal{D}_\tau\right] &\le \frac{D(N-t)^2(N-\tau+1)^4}{(t-\tau+1)^2}\sum_{r=\tau}^{t}\frac{1}{(N-r)^2} \\ &\le \frac{D(N-\tau+1)^4}{t-\tau+1}\cdot\frac{\pi^2(N-t)^2}{6(t-\tau+1)}.\end{aligned}$$

The above two inequalities still hold when $t = N$, where LURE will assign infinite weights on the exact variance $\widehat{\mathrm{Var}}_{\tau,N} = \mathrm{Var}[\hat{F}_\tau|D_\tau]$, so $\mathrm{Var}[\widehat{\mathrm{Var}}_\tau^{\mathrm{LURE}}|\,\mathcal{D}_\tau] = 0$. In summary, we see that $\mathrm{Var}\left[\widehat{\mathrm{Var}}_\tau^{\mathrm{LURE}}|\,\mathcal{D}_\tau\right] \le \frac{E}{t-\tau+1}\cdot\min\left(1, \frac{(N-t)^2}{(t-\tau+1)}\right)$, for $E = \frac{\pi^2 D(N-\tau+1)^2}{6}$. Thus $\mathrm{Var}\left[\widehat{\mathrm{Var}}_\tau^{\mathrm{LURE}}|\,\mathcal{D}_\tau\right] \lesssim \frac{1}{t-\tau+1}\cdot\min\left(1, \frac{(N-t)^2}{(t-\tau+1)}\right)$.

□

B Details of variance estimation

B.1 The plug-in mean in variance estimation

One practical gap is the incorporation of a plug-in mean instead of the ground-truth mean in Eq. 6. We show that the error introduced by this trick is controlled. First of all, the variance of the plug-in mean could be controlled with the following lemma.

Lemma 1. *For $\hat{F}_{1:t}^{\mathrm{LURE}} = \sum_{\tau=1}^{t}\bar{\alpha}_\tau^{\mathrm{LURE}}\,\hat{F}_\tau$, its variance satisfies that there exists a constant $E > 0$, $\mathrm{Var}[\hat{F}_{1:t}^{\mathrm{LURE}}] \le \frac{E(N-t)}{t}$.*

Proof. By Prop. 3,

$$\mathrm{Var}[\hat{F}_\tau] \le C(N-\tau)(N-\tau+1).$$

Then

$$\begin{aligned}
\mathrm{Var}[\hat{F}_{1:t}^{\mathrm{LURE}}] &= \sum_{\tau=1}^{t} \left(\bar{\alpha}_\tau^{\mathrm{LURE}}\right)^2 \mathrm{Var}[\hat{F}_\tau] \\
&\leq \sum_{\tau=1}^{t} \frac{N^2(N-t)^2}{t^2(N-\tau)^2(N-\tau+1)^2} \cdot C(N-\tau)(N-\tau+1) \\
&= \frac{CN^2(N-t)^2}{t^2} \sum_{\tau=1}^{t} \frac{1}{(N-\tau)(N-\tau+1)} \\
&= \frac{CN^2(N-t)^2}{t^2} \frac{t}{N(N-t)} \\
&= \frac{CN(N-t)}{t}.
\end{aligned}$$

Let $E = CN$, then $\mathrm{Var}[\hat{F}_{1:t}^{\mathrm{LURE}}] \leq \frac{E(N-t)}{t}$. □

This directly implies that the plug-in mean constructed from the LURE estimate will not be far from $F(\Omega)$. Denote the variance estimate with the plug-in mean by

$$\widetilde{\mathrm{Var}}_{\tau,t} = \sum_{s \in \mathcal{D}_t \setminus \mathcal{D}_\tau} q_\tau(s) \left(\frac{f(s)}{q_\tau(s)} - \hat{G}_{\tau,t}\right)^2 + \frac{q_\tau(s_t)}{q_t(s_t)} \left(\frac{f(s_t)}{q_\tau(s_t)} - \hat{G}_{\tau,t}\right)^2.$$

We control the error with the following proposition.

Proposition 8. *Given the same assumptions as Prop. 3, with at least probability $1-p$, there exists a constant H dependent on p such that $|\widetilde{\mathrm{Var}}_{\tau,t} - \widehat{\mathrm{Var}}_{\tau,t}| \leq H(N-\tau+1)\sqrt{\frac{N-t+1}{t-1}}$.*

Proof. First, note that

$$\hat{G}_{\tau,t} - F(\Omega \setminus \mathcal{D}_\tau) = \hat{F}_{1:t}^{\mathrm{LURE}} - F(\Omega).$$

By Chebyshev's inequality, with at least probability $1-p$,

$$|\hat{G}_{\tau,t} - F(\Omega \setminus \mathcal{D}_\tau)| = |\hat{F}_{1:t-1}^{\mathrm{LURE}} - F(\Omega)| \leq \sqrt{\frac{E(N-t+1)}{p(t-1)}}.$$

Also,

$$\begin{aligned}
|\hat{G}_{\tau,t} + F(\Omega \setminus \mathcal{D}_\tau)| &\leq |\hat{G}_{\tau,t} - F(\Omega \setminus \mathcal{D}_\tau)| + 2F(\Omega \setminus \mathcal{D}_\tau) \\
&\leq \sqrt{\frac{E(N-t+1)}{p(t-1)}} + 2B(N-\tau+1) \\
&\leq I(N-\tau+1),
\end{aligned}$$

where I is a constant dependent on p. We expand the difference directly.

$$
\begin{aligned}
|\widetilde{\mathrm{Var}}_{\tau,t} - \widehat{\mathrm{Var}}_{\tau,t}| =& \Bigg| \sum_{s\in\mathcal{D}_t\setminus\mathcal{D}_\tau} q_\tau(s)(\hat{G}^2_{\tau,t} - F(\Omega\setminus\mathcal{D}_\tau)^2) - 2f(s)(\hat{G}_{\tau,t} - F(\Omega\setminus\mathcal{D}_\tau)) \\
&+ \frac{q_\tau(s_t)}{q_t(s_t)}\left(\hat{G}^2_{\tau,t} - F(\Omega\setminus\mathcal{D}_\tau)^2\right) - 2\frac{f(s_t)}{q_t(s_t)}(\hat{G}_{\tau,t} - F(\Omega\setminus\mathcal{D}_\tau))\Bigg| \\
\le& \sum_{s\in\mathcal{D}_t\setminus\mathcal{D}_\tau} q_\tau(s)|\hat{G}^2_{\tau,t} - F(\Omega\setminus\mathcal{D}_\tau)^2| + 2f(s)|\hat{G}_{\tau,t} + F(\Omega\setminus\mathcal{D}_\tau)| \\
&+ \frac{q_\tau(s_t)}{q_t(s_t)}\left|\hat{G}^2_{\tau,t} - F(\Omega\setminus\mathcal{D}_\tau)^2\right| + 2\frac{f(s_t)}{q_t(s_t)}|\hat{G}_{\tau,t} - F(\Omega\setminus\mathcal{D}_\tau)| \\
\le& \sqrt{\frac{E(N-t+1)}{p(t-1)}}\Bigg(\sum_{s\in\mathcal{D}_t\setminus\mathcal{D}_\tau}\frac{B}{(N-\tau+1)A}I(N-\tau+1) + 2B \\
&+ \frac{B^2(N-t+1)}{A^2(N-\tau+1)}I(N-\tau+1) + 2\frac{B^2(N-t+1)}{A}\Bigg) \\
=& \left(\frac{IB}{A} + 2B\right)\left(t-\tau+\frac{B}{A}(N-t+1)\right)\sqrt{\frac{E(N-t+1)}{p(t-1)}} \\
\le& \frac{(I+2A)B^2}{A^2}\sqrt{\frac{E}{p}}(N-\tau+1)\sqrt{\frac{N-t+1}{t-1}}.
\end{aligned}
$$

Let $H = \frac{(I+2A)B^2}{A^2}\sqrt{\frac{E}{p}}$, then $|\widetilde{\mathrm{Var}}_{\tau,t} - \widehat{\mathrm{Var}}_{\tau,t}| \le H(N-\tau+1)\sqrt{\frac{N-t+1}{t-1}}$. □

Comparing the derivation of Prop. 7 and Prop. 8, the L1 error introduced by the plug-in mean goes to 0 at a faster rate than the standard deviation of $\widehat{\mathrm{Var}}_{\tau,t}$, indicating that the error is negligible in our evaluation when t is large enough.

B.2 The streaming algorithm of variance estimation

Alg. 2 has a summation for each $1 \le \tau \le t$, amounting to a complexity of $\mathcal{O}(t^2)$. In practice, we would reuse part of previous computation at step $t-1$ to accelerate the computation at step t. Observe that

$$
\begin{aligned}
\widetilde{\mathrm{Var}}_{\tau,t} =& \sum_{s\in\mathcal{D}_t\setminus\mathcal{D}_\tau} q_\tau(s)\left(\frac{f(s)}{q_\tau(s)} - \hat{G}_{\tau,t}\right)^2 + \frac{q_\tau(s_t)}{q_t(s_t)}\left(\frac{f(s_t)}{q_\tau(s_t)} - \hat{G}_{\tau,t}\right)^2 \\
=& \sum_{s\in\mathcal{D}_t\setminus\mathcal{D}_\tau}\left(\frac{f(s)^2}{q_\tau(s)} + q_\tau(s)\hat{G}^2_{\tau,t} - 2f(s)\hat{G}_{\tau,t}\right) + \frac{q_\tau(s_t)}{q_t(s_t)}\left(\frac{f(s_t)}{q_\tau(s_t)} - \hat{G}_{\tau,t}\right)^2 \\
=& \sum_{s\in\mathcal{D}_t\setminus\mathcal{D}_\tau}\frac{f(s)^2}{q_\tau(s)} + \hat{G}^2_{\tau,t}\sum_{s\in\mathcal{D}_t\setminus\mathcal{D}_\tau} q_\tau(s) - 2\hat{G}_{\tau,t}\sum_{s\in\mathcal{D}_t\setminus\mathcal{D}_\tau} f(s) + \frac{q_\tau(s_t)}{q_t(s_t)}\left(\frac{f(s_t)}{q_\tau(s_t)} - \hat{G}_{\tau,t}\right)^2.
\end{aligned}
$$

This is a polynomial with respect to our plug-in mean $\hat{G}_{\tau,t}$. Therefore, we could maintain the three summations in the coefficients $\sum_{s\in\mathcal{D}_t\setminus\mathcal{D}_\tau}\frac{f(s)^2}{q_\tau(s)}$, $\sum_{s\in\mathcal{D}_t\setminus\mathcal{D}_\tau} q_\tau(s)$ and $\sum_{s\in\mathcal{D}_t\setminus\mathcal{D}_\tau} f(s)$ for each τ during the algorithm to speed-up the computation of each individual $\widetilde{\mathrm{Var}}_{\tau,t}$. In addition, our final estimate is

$$
\widetilde{\mathrm{Var}}_\tau = \sum_{r=\tau}^{t}\bar{\beta}_r\,\widetilde{\mathrm{Var}}_{\tau,r} = \frac{\sum_{r=\tau}^{t}\beta_r\widetilde{\mathrm{Var}}_{\tau,r}}{\sum_{r=\tau}^{t}\beta_r}.
$$

The numerator is still a polynomial with respect to the plug-in mean whose coefficients can be written as summations. We would further maintain the three coefficients and $\sum_{r=\tau}^{t}\beta_r$ to avoid the loop to compute $\widehat{\mathrm{Var}}_\tau$. With the two techniques, our approach is summarized as Alg. 3.

Algorithm 3 Streaming variance estimation update

Require: Sample s_t, label $f(s_t)$ and estimate $\hat{F}_{1:t-1}$
Set $x_t = y_t = z_t = a_t = b_t = c_t = u_t = 0$
for $\tau = 1, 2, ..., t$ **do**
  Get estimated mean $\hat{G}_{\tau,t} = \hat{F}_{1:t-1} - F(\mathcal{D}_\tau)$
  $a_\tau \leftarrow a_\tau + \beta_t \left(x_\tau + \frac{f(s_t)^2}{q_t(s_t) q_\tau(s_t)} \right)$
  $b_\tau \leftarrow b_\tau + \beta_t \left(y_\tau + \frac{f(s_t)}{q_t(s_t)} \right)$
  $c_\tau \leftarrow c_\tau + \beta_t \left(z_\tau + \frac{q_\tau(s_t)}{q_t(s_t)} \right)$
  $u_\tau \leftarrow u_\tau + \beta_t$
  $\widetilde{\mathrm{Var}}_\tau = (a_\tau - 2 b_\tau \hat{G}_{\tau,t} + c_\tau \hat{G}_{\tau,t}^2)/u_\tau$
  $x_\tau \leftarrow x_\tau + \frac{f(s_t)^2}{q_\tau(s_t)}$
  $y_\tau \leftarrow y_\tau + f(s_t)$
  $z_\tau \leftarrow z_\tau + q_\tau(s_t)$
end for

C Experimental settings

C.1 Estimating birds in radar data

Weather radar preliminaries. The NEXRAD radar network contains 159 high-resolution radars and covers most of the U.S. territories. Each radar station typically scans the surrounding atmosphere every 4-10 minutes to collect weather data. This is achieved by rotating the radar antenna around the vertical axis at multiple elevation angles. At each elevation, the radar performs "sweeps" to collect several radar products, such as, reflectivity and radial velocity. These signals also record objects like bird roosts. We render radar data for 300km $\times$ 300km regions into 600 $\times$ 600 pixel arrays, on which we train a detector model to automatically predict bounding boxes for roosts.

Visualization. Fig. A2 (left) illustrates weather radar scans capturing bird roosts. Fig. A2 (right) visualizes the reflectivity data at $0.5\circ$ elevation of a radar scan that happened at the KCLE weather radar station on August 19 2015 at 10:31:22 UTC, which is a randomly sampled radar scan that captures bird roosts. The circular and semicircular patterns are massive amounts of birds departing from their overnight roosting locations.

Detector, tracking, and filtering. We follow Perez et al. [31] to pretrain a station-agnostic spatiotemporal roost detector. We also follow their configuration for deploying the detector on the Great Lakes radar stations, which are excluded from the pretraining data, assembling model predicted boundings boxes into roost tracks, and filtering tracks with too few detected time frames or too low of a max or mean detection confidence score. Since each weather radar collects data every few minutes, the same roost is often captured by multiple consecutive scans of a radar station. To avoid double counting, it is necessary to assemble detections of the same roost into a track.

Perez et al. [31] deploys the roost predition system on radar data from 12 radar stations in the Great Lakes region and let human experts screen the system predictions. They focus on the time window from 30 minutes to 90 minutes after the local sunrise for every station-day.

Following Deng et al. [10], we focus on 11 stations, including KAPX, KBUF, KCLE, KDLH, KDTX, KGRB, KGRR, KLOT, KMKX, KTYX, KIWX, since KMQT does not clearly observe roosts in the human screened data. Our goal is to estimate the number of birds in June to October in the 2015-2019 five-year period at these 11 stations. We use screened data from Perez et al. [31] as the ground truth bounding boxes for estimating bird counts. We let our station-specific finetuned checkpoints predict bounding boxes for the same time periods on the same station-day for our estimations.

Station-day bird count estimation.

- Per-sweep counts. For each predicted bounding box by the detector model in a radar scan in a station-day, we enumerate over the radar sweeps at all elevation within 5000km height and

follow Belotti et al. [5] to estimate the bird count of this bounding box geographical region in this sweep.

- Per-bounding-box counts. We summarize, for each detection, radar sweeps taken across multiple elevations. In order to prevent double counting of birds in regions sampled twice by two consecutive beams, we bin the sweep elevations into $1\circ$ bins. We then take the average of sweeps that fall in the same bin. Lastly, we sum counts across all bins.
- Per-track counts. We obtain summaries across detections by finding the median count within each track and selecting two scans before and after the median. We then calculate the average count within this set of 5 scans.
- Per-day counts. We sum over per-track counts to obtain a per-day bird count at a station-day.

C.2 Additional Finetuning And Sampling Details

To fine-tune our models, we use the Detectron2 library with modified configurations. Specifically, we disable learning rate warmup, set the batch size to 8, and set FILTER_EMPTY_ANNOTATIONS to False. For the high-resolution image experiments, we use the faster_rcnn_R_50_FPN_3x.yaml configuration with the default ImageNet-pretrained weights.

Fine-tuning the image detector takes approximately 20 minutes on a single NVIDIA A16 GPU. We perform fine-tuning 8 times, resulting in a total end-to-end runtime of just under 3 hours per image. For the roost detector, each fine-tuning run takes about 2 hours on a single A16 GPU. We fine-tune 4 times, totaling roughly 8 hours of compute per station for roost counting. In both cases, the time required to compute count and variance estimates is negligible.

In practice, the measurements from the detectors could be as low as 0, making the labeling of some units impossible and introducing huge variances to the estimation. In our image counting tasks, we set $g(s) \leftarrow \max(g(s), 1)$ to make sure that each tile can be sampled. In our radar counting tasks, we set $g(s) \rightarrow g(s) + 1000$ to cover every day in every station.

C.3 Labeling Birds in Images

We visualize the Reeds and Sky images in figure A1. To collect ground truth, we manually label both images entirely. We use a tile size of 200 pixels for sky and 160 pixels for reeds, resulting in a total of 925 and 1426 tiles respectively. To label we randomly group the tiles into batches of 50, and have labelers annotate each batch using VGG annotator. To speed up the process, we label each bird with a single point at its center of mass. This results in about 1 second of human effort per labeled point. We only label a bird on the edge if a majority of it is visible in the tile. In total, our ground truth count for sky is 5682 birds and 12486 birds for reeds.

To convert point annotations into approximate bounding boxes, we use a two-stage heuristic algorithm. First, we apply Otsu thresholding to each tile and compute the average size of a square bounding box by dividing the number of foreground pixels by the number of labeled points. In the second stage, we center these boxes at each labeled point, mask out foreground pixels outside the boxes, and recompute the average box size using the remaining foreground pixels.

D Additional experimental results

D.1 Comparison to few-shot models

Figure A3 presents results in terms of the Hellinger distance between the predicted proposal distribution P and the optimal sampling distribution Q (proportional to ground truth counts). We use Hellinger distance because it accommodates zero-probability entries. It is defined as:

$$H(P, Q) = \frac{1}{\sqrt{2}} \sqrt{\sum_{i=1}^{k} (\sqrt{p_i} - \sqrt{q_i})^2}.$$

Where p_i, q_i are the probabilities assigned to unit i for P and Q, respectively. Unlike raw counts, the Hellinger distance is unaffected by constant-factor over or under prediction and thus gives a better idea of how good the model is for the estimation process.

Figure A1: The Sky image (left) and Reeds image (right) used for our detection experiments. Note the resolution has been reduced for this visualization.

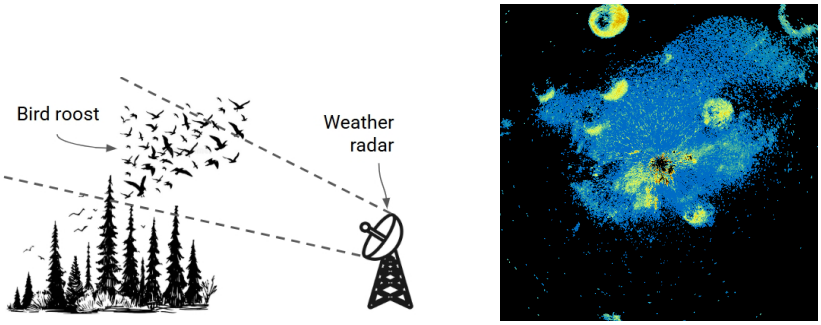


Figure A2: (Left) Weather radars can detect bird roosts departing from their nighttime roosting locations. (Right) Roosts appear as circular patterns in images rendered from radar products, such as the reflectivity at 0.5° channel as visualized.

On the left, we compare the proposal distributions of the sky image raw predictions from our fine-tuned detector and the few-shot FamNet model [34], which estimates object counts from a few examples without additional training. FamNet struggles to adapt with more examples, while the fine-tuned model achieves lower Hellinger distance even with a single labeled tile.

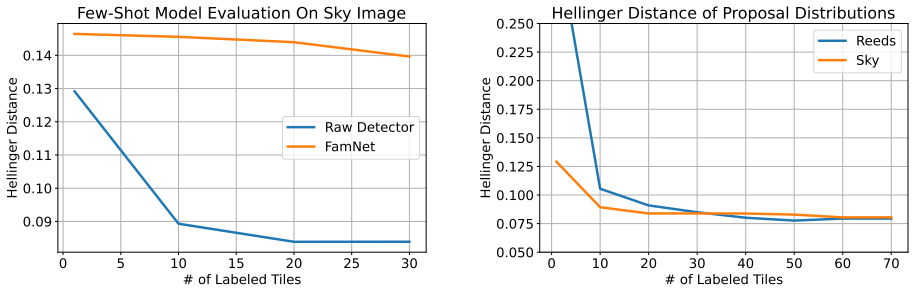


Figure A3: Left: Comparing the performance of raw predictions from fine-tuned detectors vs. few-shot FamNet[34] model, based on Hellinger distance to ground truth. The FamNet model performs worse and does not improve as much with more labeled tiles. Right: Hellinger distance between the raw predictions of the fine-tuned detector and the optimal sampling distribution for both images, averaged over 100 checkpoints. Unlike the raw counts which saturate quickly, additional training continues to improve the proposal distribution, especially on the harder Reeds image.

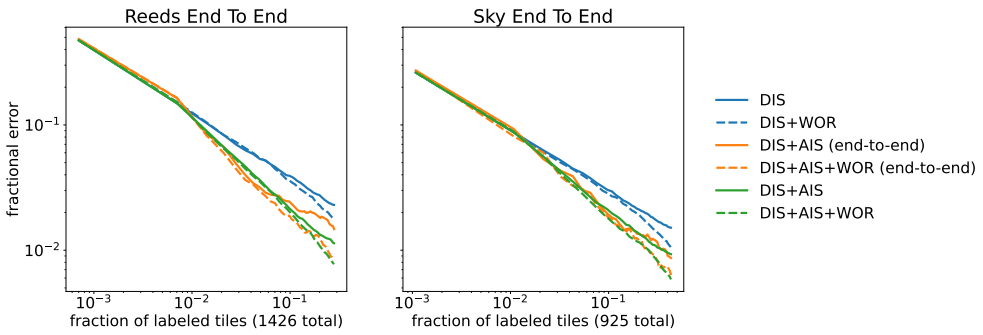


Figure A4: Comparing our fixed checkpoint scheme against true end-to-end training, averaged over 1000 trials. While the fixed checkpoint approach introduces slightly higher variance, particularly at higher label counts, it preserves relative trends between methods. This supports its use for our large-scale evaluation.

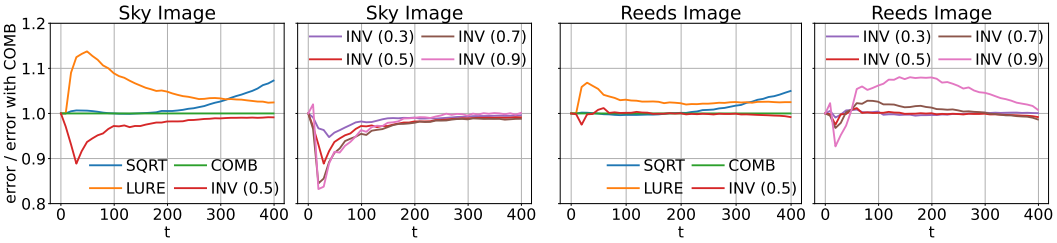


Figure A5: Relative errors compared to α^{COMB} weighting. Other fixed weighting strategies (α^{SQRT}, α^{LURE}) are worse, but inverse variance weighting (denoted by INV (γ)) may achieve lower error.

D.2 Fine-Tuning Improves Proposal Distribution

On the right of Figure A3, we track the Hellinger distance over the course of finetuning, averaged over 100 checkpoints. While raw detector counts tend to saturate early, the Hellinger distance of the proposal distribution continues to improve, especially for the harder Reeds image, before eventually saturating.

D.3 True End-to-End Finetuning

To approximate the effect of interactive model adaptation in our main results, we use a "fixed checkpoint" approach: a predefined sequence of detectors trained on a progressively larger number of labels. This allows us to conduct a significantly larger number of trials than would be feasible with full end-to-end training.

In practice, however, a practitioner would retrain the detector as new samples are labeled during active measurement. To assess how well our "fixed checkpoint" approach approximates this more realistic scenario, we compare it to end-to-end training averaged over 1,000 trials (Figure A4).

The results indicate that performance under both approaches is similar, particularly when using fewer tiles, and that the relative trends between methods are preserved. We do observe slightly higher variance in the fixed checkpoint setting, likely due to the specific checkpoint chosen.

D.4 Different weighting schemes for all experiments

We compare different weighting schemes in both image counting and radar counting experiments. The results are in Fig. A5 and A6. From the first and the third columns, we conclude that the α_{COMB} weights work consistently better than α_{SQRT} and α_{LURE}. In the second and the fourth columns, the performances of inverse variance weighting with different γ are compared. In general, a conservative $\gamma = 0.3$ brings little benefit, while an aggressive $\gamma = 0.9$ can be detrimental because each individual estimator may not be accurate enough.

D.5 Confidence intervals for all experiments

In Fig. A7, we compare the CI coverages for all experiments. In most of the experiments, the coverage for active measurement improves as the number of labeled samples increases (except on KIWX and KTYX). This behavior is typical in importance sampling-based methods. Out of the 11 radar stations, we achieve near perfect coverage (around 0.95) in 6, good coverage (around 0.9) in 2 and fair coverage (around 0.8) in 3. Comparing the two variance estimators $\mathrm{Var}_{1:t}^{\mathrm{simp}}$ and $\mathrm{Var}_{1:t}^{\mathrm{cond}}$, we find that the $\mathrm{Var}_{1:t}^{\mathrm{simp}}$ works better, especially on stations with bad coverage. We conclude that the CIs in active measurement usually have the desired properties for measuring the uncertainty.

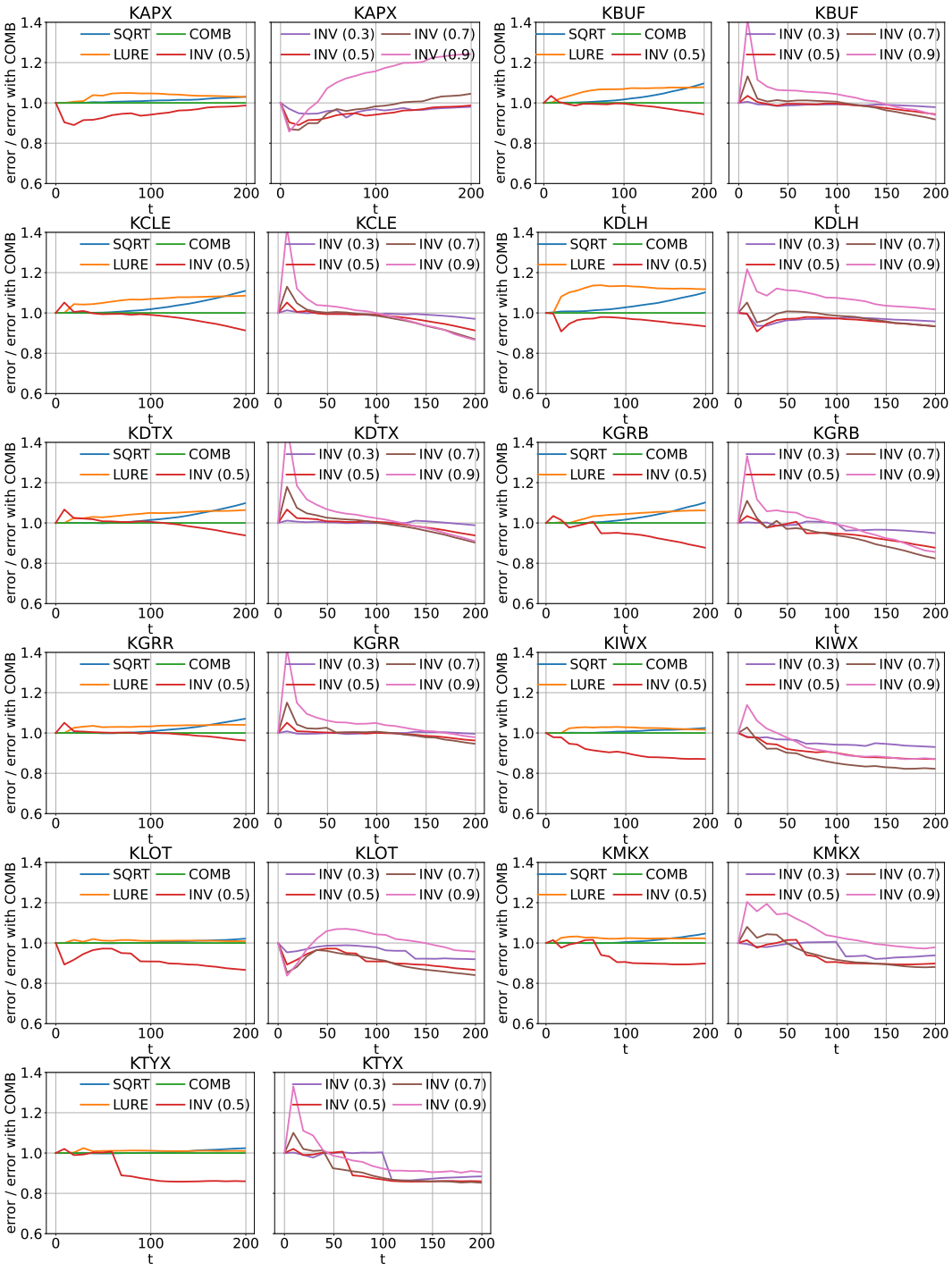


Figure A6: Relative errors compared to α^{COMB} weighting, for the roost counting problem on all 11 radar stations.

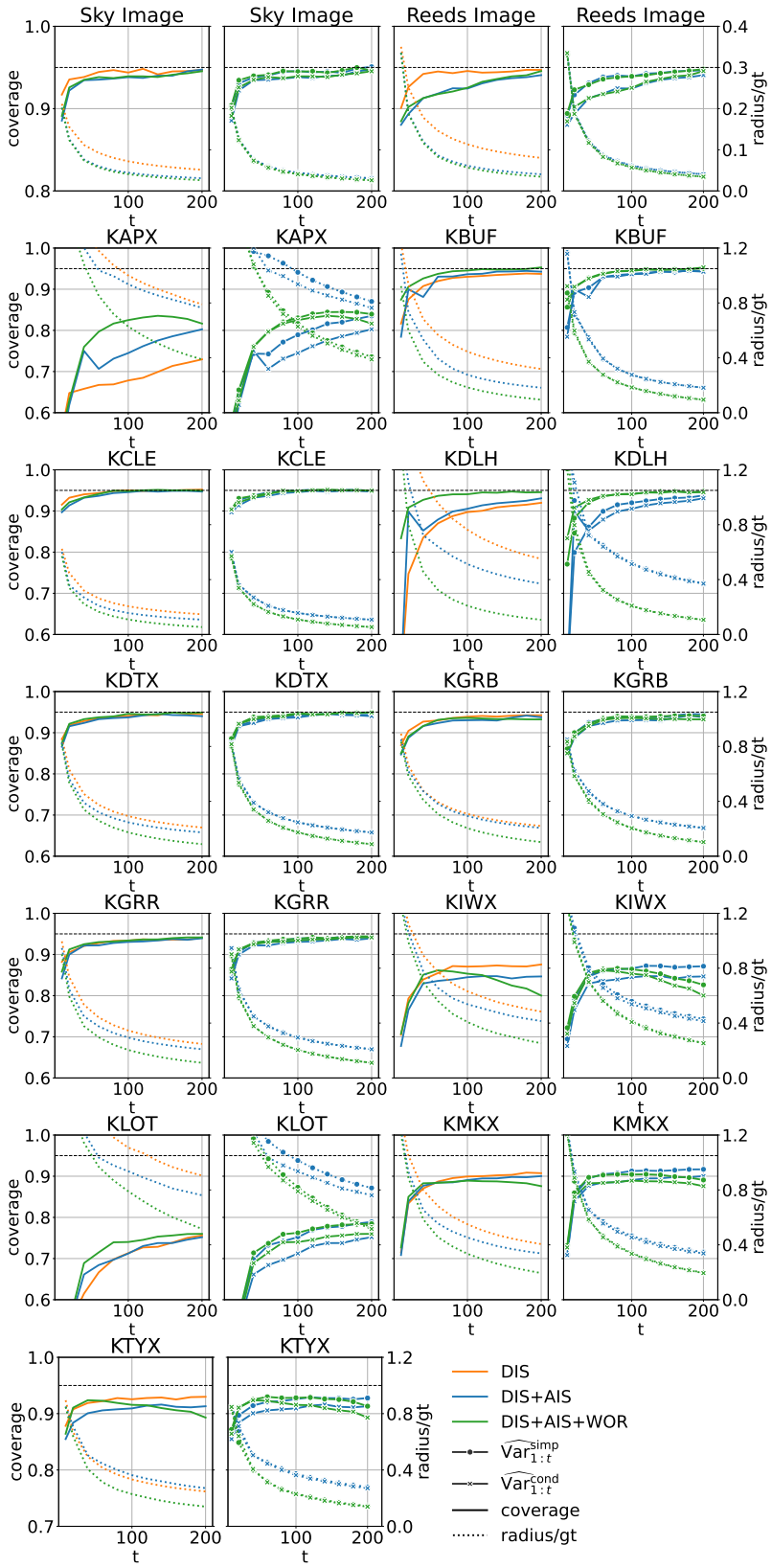


Figure A7: Full CI coverage results for counting on the two images and on all 11 radar stations.