Online Active Testing: Adaptive Importance Sampling for Unbiased Risk Estimation in Data Streams

Hugo Schmutz
Aix-Marseille University,
CNRS, LIS, Marseille, France
hugo.schmutz@lis-lab.fr

Hachem Kadri
Aix-Marseille University,
CNRS, LIS, Marseille, France
hachem.kadri@lis-lab.fr

Thierry Artières
École Centrale Méditerranée
Aix-Marseille University,
CNRS, LIS, Marseille, France
thierry.artieres@lis-lab.fr

Abstract

We address the problem of evaluating machine learning models under a limited labelling budget in streaming environments, where data arrive sequentially. This setting is particularly critical for embedded and continuously deployed systems, such as autonomous driving or computer-aided diagnosis, where models must reliably assess their own performance online. We focus on a constrained online setting, where labelling decisions must be made immediately, and data cannot be stored. We introduce an *online active testing* (OAT) framework based on an adaptive importance sampling estimator that provides unbiased risk estimation for both regression and classification tasks. While importance sampling has been extensively studied in pool-based settings, we establish novel theoretical ground for the fully online setting, including deviation bounds and consistency guarantees. We derive an optimal sampling strategy in order to minimise the estimator's variance, which is driven by the expected second moment of the loss. This strategy is expressed for common loss functions, including mean squared error, mean absolute error, 0–1 loss, and log-loss, and analysed from a statistical perspective. We empirically demonstrate the effectiveness of OAT on synthetic benchmarks, deep learning tasks, and a real-world application involving the evaluation of machine learning models for molecular simulations, where data are generated sequentially and labels require expensive quantum chemistry computations, showing improved efficiency and reliable performance estimation under limited labelling budgets. In particular, on the real-world task, OAT requires only $\sim$35% of labels to match the accuracy of a naïve uniform baseline, compared to $\sim$20% for an oracle method that has access to the labels.

1 Introduction

While unlabelled data are typically abundant, obtaining labels can be costly. In many scientific fields, assigning a single label may require specialised expertise or weeks of laboratory work or computation [6]. Moreover, some labels are more informative than others, suggesting that examples should be selected strategically under limited budgets. For instance, in autonomous driving, much of the data is redundant, so labelling most of it yields little performance gain. Depending on data availability, two scenarios arise: (1) in the *pool-based* setting, the learner can select samples from a large unlabelled pool; (2) in the *online* (or *stream-based*) setting, data arrive sequentially, and

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

the learner must decide whether to query or discard it. This setup reflects real-world constraints, where decisions must be made without access to future data, and is relevant in applications such as manufacturing, fraud detection, clinical monitoring, and streaming perception systems [8]. This need for sequential decision-making is especially critical in embedded and continuously deployed systems, such as autonomous vehicles [59] and computer-aided diagnosis models [51]. These systems operate in dynamic environments where data distributions evolve. As a result, models must learn efficiently while continuously evaluating their performance as new data arrive. Without reliable and up-to-date estimates, systems cannot detect degradation, adapt to shifts, or ensure safety and reliability.

While much work has focused on selecting training data [1, 56], efficiently evaluating models under limited labelling budgets has received comparatively little attention [41]. In controlled research environments, evaluation often assumes access to large labelled test sets, which is unrealistic when test labels are as costly as training labels. Practitioners must therefore evaluate and compare models without exhaustive ground truth. In fact, evaluation is a key bottleneck for deploying active learning in real-world applications [42]. Poor evaluation can obscure issues such as the *cold start problem*, where active learning initially degrades performance [3, 18, 45]. Furthermore, selecting test points based on model or training data information introduces sample-selection bias [43, 11]. For example, querying only uncertain points [23] inflates estimated test loss, and this bias can impair model comparison and hyperparameter tuning. While such bias is often tolerable in training, it is far more harmful in evaluation [16]. To address this, *active testing* has been proposed to select which test examples to label for model selection and performance estimation. While pool-based active testing has been widely studied [53, 33, 34, 46], the online setting remains largely unexplored. Existing stream-based works mainly focus on model selection [29, 40, 30].

Main contributions In this work, we introduce an *online active testing (OAT)* framework that selects test points on the fly to produce accurate and reliable estimates of model performance under a limited labelling budget. OAT is designed for a practically motivated yet largely unexplored setting where test examples must be selected and labelled sequentially, without storage, to estimate model performance under labelling constraints (Section 3). We propose an adaptive importance sampling estimator that provides unbiased risk estimation, and establish a novel theoretical ground for the fully online setting. We derive non-asymptotic deviation bounds for the estimator, characterise its variance (Section 4), and minimise the anticipated variance under the labelling budget, yielding an optimal active sampling strategy based on the expected second moment of the loss. We characterise this strategy for several standard losses and analyse their statistical properties (Section 5). In practice, these sampling strategies rely on surrogate models. Building on [33] observations, we propose lightweight surrogates suitable for online deployment while achieving high fidelity and diversity. Finally, we empirically validate our approach for both regression and classification on synthetic data, deep learning benchmarks, and a real-world application in molecular modelling, demonstrating efficient and reliable performance estimation under limited labelling budgets (Section 6). In particular, for the real-world application, OAT achieves the same estimation accuracy as uniform sampling using only 35% of the labels, compared with 20% for a label-aware oracle, which demonstrates that OAT can significantly reduce evaluation cost.

2 Related work

Online active testing aims to estimate model performance on a data stream while minimising labelling costs. This task lies at the intersection of active evaluation, unbiased and online active learning. More marginally, it relates to stream sampling methods such as reservoir and sliding-window sampling, which control variance under memory constraints [10, 15, 27]

Active testing and model evaluation Active testing aims to estimate model performance using as few labelled examples as possible, rather than improving the model itself. Early work relied on self-normalised importance sampling [48], leading to foundational methods for active risk estimation [53–55, 52], later extended to joint training and testing [65]. These methods are typically consistent but not unbiased. Later work introduced stratified and budget-aware sampling schemes for large-scale evaluation [4, 14, 31, 35]. More recent methods propose unbiased pool-based estimators using adaptive importance sampling [33], as well as Bayesian approaches for evaluation and model selection [28, 47, 32]. Kossen et al. [34] proposed to directly learn a model to predict the risk over an unlabelled dataset, while Zrnic and Candès [66] proposed an unbiased estimator leveraging both labelled and unlabelled samples. However, these methods assume access to a static data pool.

Connection to active learning: unbiased active learning and data streams Active learning aims to improve models but suffers from selection bias when queried samples differ from the data distribution. Importance-weighted approaches mitigate this issue [5], including UPAL and related optimal sampling strategies [17, 25]. These ideas are closely related to evaluation under biased sampling, which requires unbiased estimators or principled corrections. In streaming settings, only a few works address active learning with self-normalised importance sampling [9], which is consistent but not unbiased. Stream-based active learning considers sequential data arrival and immediate querying decisions, making it challenging under distribution shift or non-stationarity [8]. Existing approaches remain limited, with few addressing online or drifting streams [57, 39].

3 Online active testing setting

We consider the problem of estimating the risk of a fixed model from a stream of unlabelled data, where labels are queried sequentially under a limited budget. We first present the setting of active testing, then detail the online setting.

3.1 Active testing objective: estimation of the risk

Let $p(x, y)$ be an unknown data-generating distribution over $\mathcal{X} \times \mathcal{Y}$. Our goal is to estimate an evaluation statistic of a pre-trained model $f_\theta : \mathcal{X} \to \mathcal{Y}$ using as few labels as possible. Evaluation statistics can be expressed as a risk $\mathcal{R}(\theta) = \mathbb{E}_{(x,y)\sim p}[L(\theta; x, y)]$ for an arbitrary loss function L. The model, f_θ, is fixed during the testing procedure, and no assumptions about the form of the model are made (parametric or non-parametric, stochastic or deterministic), but its predictions may be used for sampling test points. This framework applies to any supervised task, with L covering metrics such as accuracy, mean squared error, or log-likelihood.

3.2 Online active testing (OAT) setting

In the OAT setting, we observe an unlabelled sequence $(x_t)_{t\geq 1}$ drawn i.i.d. from the marginal $p(x)$. At each time step t, a label $y_t \sim p(y \mid x_t)$ exists, but remains unobserved unless explicitly queried. Formally, we consider a stream of pairs $\{(x_t, y_t) \sim p(x, y)\}_{t\geq 1}$ where only x_t is always observed. The learner decides whether to query y_t. In this article, we deliberately consider a restrictive but practically motivated setting, where the unselected data are never reconsidered.

For each sample x_t, we define a selection probability $\beta(x_t) \in [0, 1]$ and draw $r_t \sim \mathcal{B}(\beta_t)$ to decide whether to query y_t. We denote $\tilde{y}_t = r_t \odot y_t$ (following common notation [37, 60]), where $\tilde{y}_t = y_t$ if $r_t = 1$ and $\tilde{y}_t = NA$ (Not Available) otherwise. The loss is only observed when $r_t = 1$, and by slight abuse of notation, we write $L(\theta; x_t, \tilde{y}_t) = L(\theta; x_t, y_t)$.

Importantly, the selection probability β is allowed to be *adaptive*, meaning it may depend on past observations, therefore $\beta(x_t)$ really denotes $\beta(x_t \mid \{(x_\tau, r_\tau, \tilde{y}_\tau)\}_{\tau=1}^{t-1})$. For simplicity, we note $\beta_t = \beta(x_t)$ and assume β_t is $\mathcal{F}_{t-1} \cup \sigma(x_t)$ measurable, where $\mathcal{F}_{t-1}$ is the σ-algebra generated by all random variables up to time $t-1$ ($\mathcal{F}_t = \sigma(x_1, r_1, \tilde{y}_1, \ldots, x_t, r_t, \tilde{y}_t)$). Finally, the joint distribution of the active sampling process up to time T can be factorised as

$$p(\{x_t, y_t, r_t\}_{t=1}^T) = \Pi_{i=1}^T p(x_t)p(y_t|x_t)p(r_t|x_t, \mathcal{F}_{t-1}) = \Pi_{i=1}^T p(x_t)p(y_t|x_t),$$

with $p(r_t|x_t, \mathcal{F}_{t-1}) = \beta_t^{r_t}(1-\beta_t)^{1-r_t}$. As a result of this adaptive sampling mechanism, the distribution of the selected data differs from the marginal distribution $p(x)$, and the sequence (x_t, y_t, r_t) is no longer independent across time.

Stopping criteria In online active testing, labels are acquired sequentially as data arrive, raising the question of when to stop the acquisition process. We consider two standard stopping criteria: (1) Fixed stream horizon: stop after observing a total of T unlabelled samples $\{x_t\}_{t=1}^T$ from the stream. The number of queried labels, n_T, is then a random variable. To control the labelling cost, we design the selection strategy to satisfy the budget constraint $\mathbb{E}[n_T] \leq n_B$, where n_B is the target labelling budget. In practice, this budget can be enforced tightly or even met exactly. (2) Fixed labelling budget: stop once n_B labels have been collected. The number of observed samples, T, is then a random variable. While this guarantees that the labelling budget is exactly met, it introduces a key limitation: because the stopping time depends on the selection process, it is generally not possible to construct an unbiased estimator of the risk. Chu et al. [9] (in the context of online active learning) and

Sawade et al. [53] (in the context of pool-based active testing) proposed a self-normalised importance sampling estimator that is consistent but not unbiased.

The objective of this work is to build a "good" estimator of the risk $\mathcal{R}(\theta)$. By "good", we mean estimators that are unbiased and that achieve reduced variance through the adaptive importance sampling process β. Therefore, we focus on the first stopping criterion (after T points seen) to design an unbiased estimator. For variance reduction, we consider as a reference point the naïve uniform baseline that selects points uniformly at random from the stream.

4 Stream-based active testing via adaptive importance sampling

We now focus on the setting where the stopping criterion is the number of observed samples T. Let n_T denote the (random) number of queried labels up to time T, and assume a budget constraint $\mathbb{E}[n_T] \leq n_B$. At each time step t, upon observing x_t, we decide whether to query its label according to a sampling probability $\beta_t = p(r_t = 1 \mid x_t, \mathcal{F}_{t-1})$, where $r_t \in \{0, 1\}$ indicates whether the label is acquired. We define the adaptive importance sampling estimator of the risk as

$$\mathcal{R}_T(\theta) = \frac{1}{T} \sum_{t=1}^{T} \frac{p(x_t, y_t)}{p(x_t, y_t, r_t = 1)} r_t \, L(\theta; x_t, \tilde{y}_t) = \frac{1}{T} \sum_{t=1}^{T} \frac{r_t}{\beta_t} \ell_t(\theta), \tag{1}$$

where we used $p(x_t, y_t, r_t = 1) = \beta_t \, p(x_t, y_t)$ and denoted $\ell_t(\theta) = L(\theta; x_t, \tilde{y}_t)$.

This estimator has been widely studied in pool-based active learning [5, 17, 25], where the labelling budget can be enforced exactly by selecting one sample at each iteration. In contrast, in the online setting, time steps correspond to incoming unlabelled samples, and the number of queried labels n_T is random. Despite this additional source of randomness, we will show that the estimator retains desirable statistical properties in the streaming setting, including unbiasedness and controlled variance and allows us to design adaptive sampling strategies that significantly reduce the variance of the risk estimate compared to the naïve uniform baseline ($\beta_t = n_B/T$) while respecting the labelling budget.

4.1 An unbiased and consistent estimator for data stream

We begin by establishing the unbiasedness of the proposed estimator, which is a classical property of adaptive importance sampling estimators [48, 17], which we recall here for completeness. Then, we establish new deviation bounds that prove the consistency of the estimator. For generality, we consider an arbitrary sequence of selection probabilities $\{\beta_t\}_{t=1}^{T}$.

Theorem 1. ***Unbiasedness*** *Let $T \geq 1$. If $\beta_t > 0$ for all $t \in [\![1, T]\!]$, then $\mathcal{R}_T(\theta)$ is unbiased:*

$$\mathbb{E}_{\mathcal{F}_T}[\mathcal{R}_T(\theta)] = \mathcal{R}(\theta).$$

This result shows that $\mathcal{R}_T(\theta)$ remains unbiased for *any* adaptive selection strategy, provided that all selection probabilities are strictly positive (proof in Appendix A). Intuitively, this condition ensures that the induced sampling distribution $\beta_t \cdot p$ has the same support as the data-generating distribution p, which is necessary for valid importance weighting [48].

Remark 1. *The positivity condition on β_t can be relaxed. It is sufficient that $\beta_t > 0$ when $\ell_t(\theta) \neq 0$. For example, when using the 0–1 loss or mean squared error, only incorrectly predicted samples contribute to the risk, and therefore only those need to be assigned non-zero selection probabilities.*

Beyond unbiasedness, a desirable property for a "good" estimator is *consistency*, ensuring convergence to the true risk, we prove it under mild conditions on the loss and the sampling strategy.

Theorem 2. ***Deviation and consistency:*** *Assume that the loss function is non-negative and bounded, i.e., for all $(x, y) \sim p$, $0 \leq L(\theta; x, y) \leq M$ for some constant $M > 0$. Suppose further that there exists $\beta_{\min} > 0$ such that $\beta_t \geq \beta_{\min}$ for all $t \in [\![1, T]\!]$. Then, for any $\delta \in (0, 1)$,*

$$P\left(|\mathcal{R}_T(\theta) - \mathcal{R}(\theta)| > \frac{M}{\beta_{min}} \sqrt{\frac{2 \ln (2/\delta)}{T}} \right) \leq \delta.$$

This result provides a probability deviation bound, implying that $\mathcal{R}_T(\theta)$ converges to $\mathcal{R}(\theta)$ as $T \to \infty$. The proof follows by adapting the argument of Beygelzimer et al. [5], originally developed

for pool-based active learning with finite hypothesis classes and binary losses (see Appendix B). The boundedness assumption on L is standard for deriving concentration inequalities of Hoeffding type. It can be relaxed to sub-Gaussian or finite-variance assumptions, yielding slightly weaker guarantees via Bernstein-type inequalities. The condition $\beta_t \geq \beta_{\min} > 0$ ensures that importance weights remain bounded, preventing variance explosion. In practice, this corresponds to enforcing a minimal level of exploration in the sampling strategy.

4.2 Is $\mathcal{R}_{\mathcal{T}}$ an accurate risk estimate? Analysis of the variance

Since our estimator relies on adaptive importance sampling, a central objective is to control, and ideally minimise, its variance. A lower-variance estimator yields more accurate risk estimates, which directly improves model selection and evaluation reliability. Following ideas similar to Kossen et al. [33], we can derive an explicit expression for the variance.

Theorem 3. ***Variance expression:*** *Let $T > 0$, and assume $\beta_t > 0$ for all $t \in [\![1, T]\!]$. The variance of $\mathcal{R}_T(\theta)$ is given by*

$$V_{\mathcal{F}_T}[\mathcal{R}_T(\theta)] = \frac{V_{x,y}[L(\theta; x, y)]}{T} + \frac{1}{T^2} \sum_{t=1}^{T} \mathbb{E}_{x_t, y_t, \mathcal{F}_{t-1}} \left[\frac{1 - \beta_t}{\beta_t} L(\theta; x_t, y_t)^2 \right]. \tag{2}$$

The proof of this theorem is in Appendix C. This decomposition highlights two distinct sources of variance. The first term corresponds to the variance of the standard Monte Carlo estimator obtained when all labels are observed. This is the minimal variance achievable with full information. The second term captures the additional variance introduced by the active sampling process. It reflects the cost of observing only a subset of labels and depends explicitly on the sampling probabilities β_t.

Uniform baseline and oracle For comparison, we consider two reference strategies: (1) the baseline uniform strategy $\beta_t = n_B/T$. (2) the oracle strategy, which samples proportionally to the loss $\beta_t = n_B/T \cdot \ell_t(\theta)/\mathbb{E}[\ell_t(\theta)]$, known to be optimal in classical importance sampling [10, 9, 48] (but requires knowing y_t). The objective of an active sampling strategy is to improve over the uniform baseline and approach the oracle variance. These strategies respectively define upper and lower bounds for any valid active sampling strategy (achieving a variance lower than the uniform baseline). For further details, see Appendix C.

5 Actively sampling data points in a stream

In the previous sections, we established how to construct an unbiased estimator of the test risk from actively sampled data in a stream. We now turn to the design of *sampling strategies* that make this estimator as efficient as possible. For a fixed labelling budget n_B, the key objective is to minimise the variance of the estimator. Our approach is to first derive an idealised optimal strategy, and then discuss how it can be approximated in practice.

5.1 Optimal sampling strategy by minimising the anticipated variance

Directly minimising the variance expression is challenging due to its dependence on the unknown sample distribution $p(x)$. To obtain tractable solutions, we instead consider the *anticipated variance* [26], defined as the expected conditional variance of the estimator given the observed inputs, $\mathbb{E}_{y|x}[V[\mathcal{R}_T(\theta)|\mathcal{D}_T]]$, where $\mathcal{D}_T = \{x_1, \cdots, x_T\}$. This quantity measures the expected uncertainty of the estimator due to label randomness, given a fixed stream of inputs. In particular, we show in Appendix D that the optimisation of the anticipated variance under the budget constraint $\sum_{t=1}^{T} \beta_t = n_B$ yields the following optimal strategy.

Theorem 4 (Optimal sampling strategy). *Under the budget constraint $\sum_{t=1}^{T} \beta_t = n_B$, the optimal sampling probabilities with respect to the anticipated variance satisfy*

$$\beta_t^*(x_t) \propto n_B/T \cdot \sqrt{\mathbb{E}_{y|x}[\ell_t(\theta)^2]}. \tag{3}$$

Algorithm 1 Online Active Testing

Require: Model f_θ, budget n_B, horizon T and sampling strategy β, surrogate architecture π

Train surrogate model $\pi(y|x)$
for $t = 1$ to T **do**
  Observe unlabelled sample x_t
  Compute sampling probability $\beta_t = \beta(x_t)$ using f_θ and π (e.g. with Eq. 3)
  Sample $r_t \sim$ Bernoulli(β_t)
  if $r_t = 1$ **then**
    Query label y_t, compute loss $L(\theta; x_t, y_t)$
    Update surrogate model π using (x_t, y_t)
  end if
end for
return $\mathcal{R}_T$ (Equation 1)

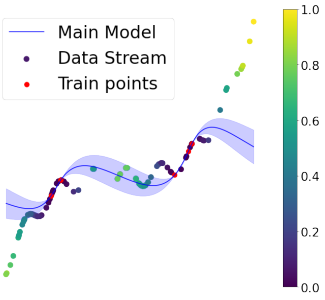


Figure 1: Illustration of second-moment active selection. A Gaussian Process is trained on the red points, while the remaining points correspond to the data stream. Colours indicate the selection probability under β_t^* with a Gaussian Process used as a surrogate model.

This result shows that optimal sampling is proportional to the *square root of the second moment of the loss*, rather than the expected loss (first moment) itself, as suggested in prior work [9, 33]. The second moment decomposes as $\mathbb{E}_{y|x}[\ell_t(\theta)^2] = V_{y|x}[\ell_t(\theta)] + \mathbb{E}_{y|x}[\ell_t(\theta)]^2$, revealing that the optimal strategy prioritises points with both large expected loss and high uncertainty in the loss. This Theorem is a simplification, as it ignores the constraint $\beta_t \in [0, 1]$. The details of the exact solution, which clips the probabilities at 1 and redistributes the remaining budget, are derived in Appendix D. Simply clipping Equation 3 without redistributing is still valid, since the budget is respected in expectation, $\mathbb{E}[n_T] \leq n_B$, but it may sample fewer points than allowed and thus yield a higher variance. In practice, the two coincide unless a few configurations have an exceptionally large loss relative to the budget.

Efficiency over naïve baseline and first-moment strategy The optimal strategy β_t^* strictly improves over the uniform baseline and the first-moment strategy in terms of anticipated variance and label complexity for a fixed estimation accuracy, unless: (i) the second moments $\mathbb{E}_{y|x}[\ell_t(\theta)^2]$ are constant across all inputs for the uniform sampling; (ii) the square roots of the second moments are proportional to the first moments across all inputs. See Appendix D for a proof and more details. Improvement arises because β_t^* adapts to the heterogeneity of the data through (i) the second moment of the loss compared to uniform sampling; (ii) high uncertainty in the loss compared to the first-moment strategy.

Label complexity We can quantify the efficiency gain of the proposed strategy relative to the naïve uniform baseline in terms of anticipated variance. Let n_u denote the number of labels required by uniform sampling to achieve the same estimation accuracy (i.e. the same variance) as our method with budget n_B. Then, $n_B = n_u \cdot \frac{\left(\sum_{t=1}^T \sqrt{\mathbb{E}_{y|x}[\ell_t(\theta)^2]}\right)^2}{\sum_{t=1}^T \mathbb{E}_{y|x}[\ell_t(\theta)^2]} = n_u \cdot \rho_T$. By Cauchy-Schwarz, $\rho_T \leq 1$, with equality only when all second moments are identical. In general, $\rho_T < 1$, implying that fewer labels are required compared to uniform sampling.

5.2 In practice: challenges of the online setting

In our online setting, data arrive sequentially, one example at a time x_t, and the learner must immediately decide whether to query its label y_t or discard it. The data are never reconsidered. This constraint fundamentally distinguishes the problem from the pool-based setting [33]. In particular, the optimal sampling strategy requires a normalisation constant that depends on all data, which cannot be computed exactly online and must therefore be approximated on the fly. Importantly, even if this approximation is imperfect, and thus the resulting sampling policy deviates from the optimal β_t^*, the risk estimator $\mathcal{R}_T$ remains unbiased, as unbiasedness only requires using the actual sampling probabilities in the importance weights. A second challenge is that the optimal policy β_t^* is intractable, as it depends on the unknown conditional distribution $y \mid x_t$. Following previous works [33, 66], we introduce a surrogate model $\pi(y \mid x)$ to approximate the second moment of the loss, yielding to

$$\beta_t \propto n_B/T \cdot \sqrt{\mathbb{E}_{y_t|x_t\sim\pi}[\ell_t(\theta)^2]}. \tag{4}$$

The surrogate π is updated adaptively during the stream using the queried samples, progressively improving its approximation of $y \mid x$. The normalisation constant is estimated online by evaluating

$\sqrt{\mathbb{E}_{y|x}[\ell_t(\theta)^2]}$ on incoming samples. Early in the stream, the strategy behaves close to uniform sampling, and gradually converges toward the optimal policy by refinement of the normalising constant. Importantly, because the strategy is adaptive, the normalisation constant evolves over time. We propose in Algorithm 1 the pseudo-code of the method and in Figure 1 an illustration of the selection method. We analyse the complexity of the proposed algorithm in Appendix E.1. We now give principled new derivations of the optimal strategy leading to acquisition functions for a variety of widely-used loss functions.

Derivation for the regression Substituting the squared error loss $\ell_t(\theta) = L(\theta; x_t, y_t) = (f_\theta(x_t) - y_t)^2$ into Equation 3 yields to:

$$\beta_t \propto n_B/T \cdot \sqrt{\underbrace{b(x_t)^4}_{(i)} + 6\underbrace{b(x_t)^2\sigma(x_t)^2}_{(ii)} - 4\underbrace{b(x_t)m_3(x_t)}_{(iii)} + \underbrace{m_4(x_t)}_{(iv)}}, \tag{5}$$

where $b(x_t) = f_\theta(x_t) - \mathbb{E}_{\pi(y|x_t)}[y]$ is the bias (prediction discrepancy), (3) $\sigma(x_t)^2 = V_{\pi(y|x_t)}[y]$ the predictive variance, and m_3, m_4[1] the third and fourth central moments under π.

As in Kossen et al. [33], both the bias and variance of the surrogate predictions appear, but here they also interact, and additional higher-order terms arise through the skewness (third moment) and kurtosis (fourth moment). (i) The first term grows rapidly with the bias, making the selection highly sensitive to systematic errors. (ii) The second term couples bias and variance, meaning that uncertainty has a stronger impact in regions where the model is already inaccurate. In contrast to Kossen et al. [33], variance is no longer independent. If the bias is zero (i.e., f_θ and π agree), this term vanishes. (iii) The third moment captures skewness: when bias and skewness have the same sign, the selection probability decreases, and when they differ, it increases (see Appendix E, Figure 5 for an illustrative example). (iv) The fourth moment reflects kurtosis (tail heaviness), promoting the selection of points with heavy-tailed or highly variable predictions, independently of f_θ.

Overall, the strategy prioritises regions with large bias, high uncertainty, or heavy-tailed behaviour. These quantities can be estimated using probabilistic models (e.g., Gaussian Processes [63] or Bayesian neural networks) or approximated via methods such as deep ensembles [36].

Remark 2. *If π is symmetric (e.g., Gaussian), then $m_3 = 0$, and for Gaussian π, $m_4 = 3\sigma^4$, simplifying the expression.*

The mean absolute error strategy simplifies to $\beta_t \propto n_B/T \cdot \sqrt{b(x_t)^2 + \sigma(x_t)^2}$, which is the same objective as [33] up to the square root scaling that amplifies small values, and mitigating overconfidence.

The **derivation for the classification** is presented and analysed in Appendix E for the log-loss and 0–1 loss. Other loss functions (e.g., Brier score) can be handled similarly, though Equation 3 remains the most general formulation.

Practical considerations: surrogate design and cold-start problem Kossen et al. [33] show that both diversity, obtained by using a surrogate different from the evaluated model, and fidelity are key to sample efficiency. Building on these insights, we design lightweight surrogates suited to online deployment: diversity is achieved through a different model family (e.g., Random Forest [7] or Gaussian Process), while fidelity for deep models is obtained by leveraging the evaluated model's latent representations. This requires only a single forward pass through the evaluated model and avoids additional large models or costly updates, which is essential for high-throughput streams.

Importantly, the surrogate is already trained at the start of the stream using the available labelled training data and, when available, the validation set. The main cold-start issue in OAT therefore concerns the normalisation constant, which depends on future stream observations and must be estimated online. OAT consequently behaves similarly to uniform sampling at the beginning of the stream, progressively approaching the variance-optimal policy as more observations become available. A small validation set can further initialise the normalisation estimate.

Both the surrogate design and normalisation strategy are evaluated in the ablations of Appendices F.2 and F.3. Alternative strategies, such as mini-batch processing or maintaining a reservoir of past samples, could further improve its accuracy, although they fall outside the scope of this work. Here, we deliberately focus on the most constrained setting, where discarded data cannot be reconsidered.

[1]The third central moment is given by $m_3(x_t) = \mathbb{E}_{\pi(y|x_t)}[(y - \mu(x_t))^3] = \kappa_3\sigma(x_t)^3$ with κ_3 the skewness of the surrogate's predictions and the fourth central moment $m_4(x_t) = \mathbb{E}_{\pi(y|x_t)}[(y - \mu(x_t))^4] = \kappa_4\sigma(x_t)^4$ with κ_4 the kurtosis of the surrogate's predictions.

6 Experiments

We evaluate OAT along three axes: first, statistical efficiency and variance reduction in controlled settings; second, scalability to deep models; and third, practical label savings in a real-world scientific application. Throughout, we compare OAT using the second moment strategy (Equation 3) with uniform sampling, the first-moment OAT strategy (defined as $\beta_t \propto n_B/T \cdot \mathbb{E}_{y_t|x_t\sim\pi}[\ell_t(\theta)]$ [33]), and with a label-aware oracle. Apart from the controlled experiments, we report the median squared error of the risk estimate as a function of the number of observed points across independent evaluation streams. We use the median to characterise the typical performance of the estimator while limiting the influence of occasional evaluation streams with unusually large errors. We additionally report the relative labelling cost, defined as the fraction of labels required by OAT to reach the performance of uniform sampling [33]. Appendix F provides complete experimental details, additional results, compute resources, and runtimes. Code is available at github.com/HugoSchmutz/OAT.

6.1 Controlled experiments: variance reduction and unbiasedness

We first empirically validate the statistical properties established theoretically and illustrate the sample-efficiency gains of OAT in controlled settings (as in previous work [33]). In particular, we assess whether the proposed estimators remain unbiased while reducing the variance of risk estimates compared with uniform sampling. As illustrative examples, we actively evaluate a Gaussian Process and a linear regression model, as well as a random forest for classification. In the regression setting, we estimate the squared error, while for classification, we estimate the log-loss and 0-1 loss. The Gaussian Process and random forest surrogates are retrained adaptively on all selected samples. Figure 2 shows the performance of OAT with the first- and second-moment sampling strategy compared to the uniform baseline and the oracle. Across all settings, OAT remains unbiased, while substantially producing tighter confidence bands compared to uniform sampling in all settings (reflecting reduced variance) and approaches the oracle performance. Going further, we compared OAT with another risk estimator by proposing an online adaptation of Active Risk Estimation (ARE) of Sawade et al. [53] (see Appendix G).

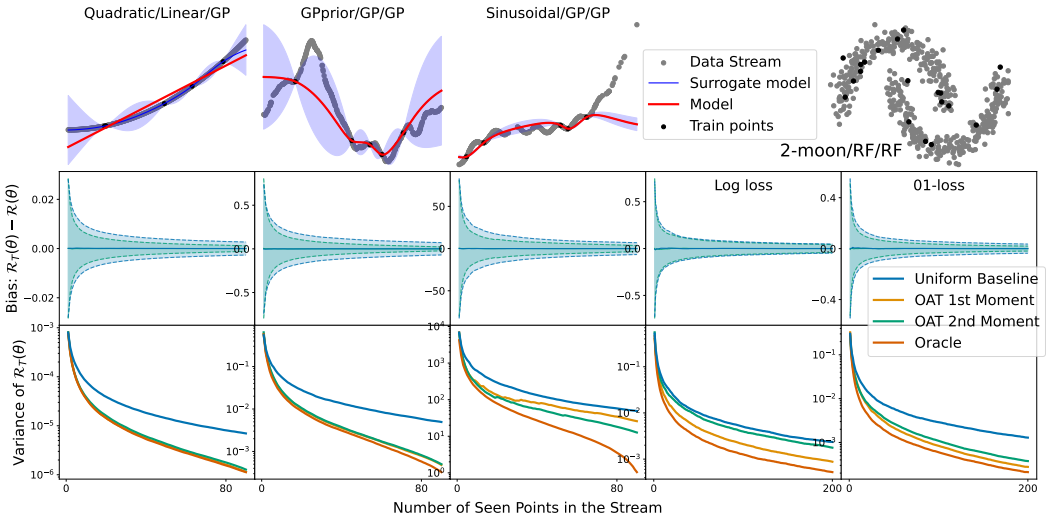


Figure 2: Online Active Testing improves risk estimation across diverse synthetic settings. Each column corresponds to a different combination of data structure, prediction model, and surrogate. Top row: regression and classification scenarios at the onset of the data stream (grey dots), showing the training data (black), the model under evaluation (red), and the surrogate model (blue). Middle and bottom rows: bias and variance of the risk estimator, respectively, as a function of the number of observations seen in the stream, for different loss functions (MSE for regression; log-loss and 0–1 loss for classification) and sampling strategies.

6.2 Scaling to deep models for images

We next ask whether the gains observed in controlled settings persist when the evaluated model is a deep neural network and the input space is high-dimensional, considering both classification and regression tasks on image data. For both tasks, we report the median squared error of the risk estimates for the different strategies (smoothed using a Gaussian kernel) and the relative labelling cost of OAT compared to the uniform baseline, highlighting the improved sample efficiency of the proposed approach.

Regression: CellCount prediction This task consists of predicting the number of cells (between 1 and 200) from synthetic greyscale fluorescence microscopy [12, 13] (64×64, see Appendix F.3). We trained a simple convolutional neural network (CNN) and evaluated the trained model on data streams with a labelling budget of 200 samples. The surrogate model used is a Gaussian Process trained using the hidden representation of the CNN. Figure 3 reports the median squared error of the risk estimates across 100 independent test streams, along with the corresponding relative labelling cost.

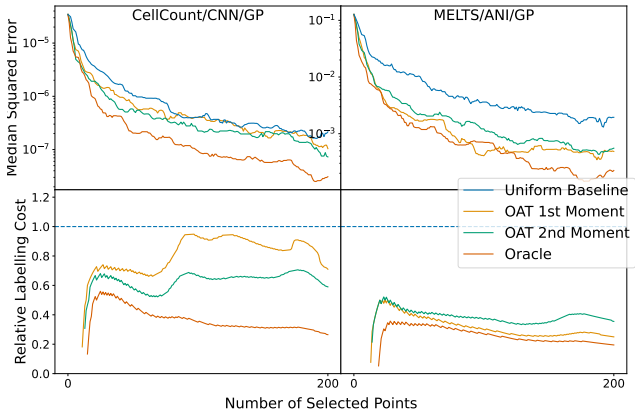


Figure 3: Online active testing for regression. Top: Median squared error of the risk estimate (log-scale) versus number of labelled samples. Bottom: relative labelling cost versus uniform baseline (dashed). Left: CellCount (CNN and GP surrogate). Right: MELTS energy prediction (ANI and GP).

Classification: FashionMNIST We consider the classification task of FashionMNIST [64] with a restricted training set of 250 samples. A ResNet-18 [22] is trained on this limited dataset, and we then assess its performance using OAT on a test data stream. Evaluation is performed under the log-loss with a labelling budget of $1,000$ samples, and the surrogate model is a Random Forest pre-trained on the training set and incrementally refined using the adaptively collected evaluation points. More details and results are given in Appendix F.2.

When does the second moment matter?

The second-moment strategy is theoretically preferable to the first-moment strategy, as established in Section 5. In practice, however, the two moments are highly correlated, and their estimation relies on a surrogate model whose errors can reduce the expected benefit of second-moment sampling. We therefore investigate whether the second moment can provide a practical advantage in challenging surrogate regimes. In particular, we consider an overconfident surrogate and find that the first-moment strategy can, in some cases, perform worse than uniform sampling, whereas the second-moment strategy still provides a substantial improvement. This experiment, detailed in Appendix F.2, illustrates the practical robustness of the second-moment criterion when surrogate estimates of the loss are unreliable.

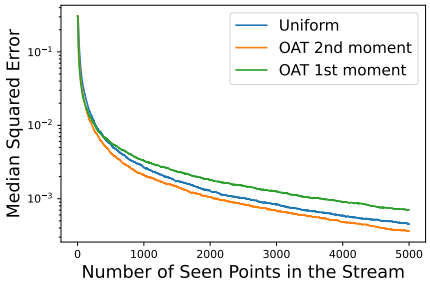


Figure 4: While the first-moment strategy may underperform the uniform baseline due to surrogate overconfidence, the second-moment strategy achieves better performance on FashionMNIST.

6.3 Real-world application to quantum chemistry data for molecular simulation (MELTS)

Finally, we evaluate OAT in a real scientific setting where acquiring a label requires an expensive quantum-chemistry computation. Molecular dynamics (MD) provides a particularly relevant example: molecular configurations are generated sequentially along trajectories, while obtaining accurate reference energies and forces requires electronic-structure calculations that can become a major

computational bottleneck. Although configurations can be stored, performing these calculations for every configuration may be prohibitively costly. This motivates methods that allocate a limited budget of expensive evaluations online, using information gathered from previous observations without requiring access to future configurations. Beyond estimating model performance, such evaluations can also inform decisions such as whether to correct a machine-learning-generated trajectory.

Machine-learning interatomic potentials aim to alleviate this computational burden by approximating the molecular potential-energy surface from atomic configurations. However, generating the reference energies and forces needed to train and evaluate these models remains expensive. This has motivated active learning approaches such as MELTS [6], which select configurations to label for model training. OAT addresses the complementary problem of efficiently evaluating an already-trained model on a stream of configurations. In this context, configurations arise dynamically along simulation trajectories and may not be available in advance, naturally connecting to membership-query synthesis settings [8]. We consider the MELTS Fulvene dataset [6], consisting of molecular geometries along with their associated reference energies. The dataset is publicly available[2]. We train an ANI model [58] on 20% of the dataset and evaluate the trained model on data streams using a labelling budget of 200. The surrogate model is a Gaussian Process trained on representations extracted from the ANI model. Results over 100 repetitions are shown in Figure 3. OAT matches the accuracy of uniform sampling using only approximately 35% of its labels (compared with 20% for the label-aware oracle). In a setting where each acquired label requires an expensive quantum-chemistry computation, this represents a substantial reduction in evaluation cost.

7 Limitations and Conclusion

Adaptive strategies substantially reduce the estimation error compared to uniform sampling, confirming the advantage of prioritising informative samples under a fixed labelling budget. Interestingly, while the first-moment ([33]) and second-moment (ours, Equation 3) strategies perform similarly in some regimes, the second-moment approach tends to be more robust across experiments, especially when the surrogate is overconfident, suggesting that incorporating uncertainty helps mitigate surrogate approximation errors. Finally, in some settings, the gap between OAT and the oracle highlights the impact of an imperfect surrogate. However, OAT strategies consistently capture a significant fraction of the oracle benefit, demonstrating their effectiveness.

We introduced OAT, a framework for unbiased and sample-efficient evaluation of machine learning models in streaming settings, where sampling decisions are irrevocable, and future or discarded samples cannot be revisited. Our findings highlight a key principle: effective sampling should jointly account for uncertainty to explore poorly understood regions of the input space, and loss magnitude to focus on regions where the model is likely to wrongly predict. This perspective opens the door to designing robust, adaptive evaluation strategies that remain effective under real-world constraints. We discuss the societal impact of this work in Appendix H.

Despite its theoretical guarantees, our approach has several limitations. First, the sampling strategy is derived for a single loss function, which prevents simultaneous optimisation for multiple evaluation metrics (for example, accuracy and calibration). Second, the effectiveness of the method depends on the quality of the surrogate model used to approximate $\mathbb{E}_{y|x}[\ell^2]$. In low-data regimes, poorly calibrated models can trigger the cold-start issue and lead to suboptimal sampling decisions. Empirically, the benefit of second-moment sampling over first-moment is not consistent. This suggests that, in practice, the benefits of variance-aware sampling can be limited by approximation errors and model misspecification or calibration.

Several directions could extend this work. A natural extension is to consider non-stationary or drifting data streams [57, 39]. Another promising direction is to directly learn the risk (like Kossen et al. 34 in the pool-based setting) or its moments in a deep learning setting using dedicated models (e.g., evidential or energy-based approaches), bypassing the need for explicit surrogate distributions. Addressing the normalising constant cold-start problem is also critical; hybrid strategies that interpolate between uniform sampling and adaptive importance sampling, similar to the method of Li et al. [38], could provide more stable performance in early stages. Finally, OAT could also be extended to manage both multi-loss and multi-model evaluation or model ranking.

[2] `https://zenodo.org/records/16900694`

Acknowledgments and Disclosure of Funding

This work received support from the French government under the France 2030 investment plan, as part of the Initiative d'Excellence d'Aix-Marseille Université - A*MIDEX (AMX-22-IN1-48).

We additionally thank Mario Barbatti (Aix-Marseille University, CNRS, ICR) and Matheus de Oliveira Bispo (Aix-Marseille University, CNRS, ICR) for their time and insightful discussions on the challenges of machine-learning potentials and molecular simulations, which inspired us to consider the problem of online active testing.

We gratefully acknowledge the Laboratoire d'Informatique et des Systèmes (Aix-Marseille Université, CNRS UMR 7020) cluster operations team for maintaining the computing resources used in this study.

References

[1] Les Atlas, David Cohn, and Richard Ladner. Training connectionist networks with queries and selective sampling. *Advances in Neural Information Processing Systems*, 2, 1989.

[2] Kazuoki Azuma. Weighted sums of certain dependent random variables. *Tohoku Mathematical Journal, Second Series*, 19(3):357–367, 1967.

[3] Javad Zolfaghari Bengar, Joost van de Weijer, Bartlomiej Twardowski, and Bogdan Raducanu. Reducing label effort: Self-supervised meets active learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1631–1639, 2021.

[4] Paul N Bennett and Vitor R Carvalho. Online stratified sampling: evaluating classifiers at web-scale. In *Proceedings of the 19th ACM International Conference on Information and Knowledge Management*, pages 1581–1584, 2010.

[5] Alina Beygelzimer, Sanjoy Dasgupta, and John Langford. Importance weighted active learning. In *Proceedings of the 26th annual International Conference on Machine Learning*, pages 49–56, 2009.

[6] Matheus de Oliveira Bispo, Rafael Souza Mattos, Max Pinheiro Jr, Bidhan Chandra Garain, Pavlo O Dral, and Mario Barbatti. MELTS: Fully automated active learning for fewest-switches surface hopping dynamics. *Journal of Chemical Theory and Computation*, 2025.

[7] Leo Breiman. Random forests. *Machine Learning*, 45(1):5–32, 2001.

[8] Davide Cacciarelli and Murat Kulahci. Active learning for data streams: a survey. *Machine Learning*, 113(1):185–239, 2024.

[9] Wei Chu, Martin Zinkevich, Lihong Li, Achint Thomas, and Belle Tseng. Unbiased online active learning in data streams. In *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 195–203, 2011.

[10] Edith Cohen, Nick Duffield, Haim Kaplan, Carsten Lund, and Mikkel Thorup. Stream sampling for variance-optimal estimation of subset sums. In *Proceedings of the 20th annual ACM-SIAM Symposium on Discrete Algorithms*, pages 1255–1264. SIAM, 2009.

[11] Sanjoy Dasgupta and Daniel Hsu. Hierarchical sampling for active learning. In *Proceedings of the 25th International Conference on Machine learning*, pages 208–215, 2008.

[12] Xin Ding, Yongwei Wang, Zuheng Xu, William J Welch, and Z Jane Wang. CCGAN: Continuous conditional generative adversarial networks for image generation. In *International Conference on Learning Representations*, 2021.

[13] Xin Ding, Yongwei Wang, Zuheng Xu, William J Welch, and Z Jane Wang. Continuous conditional generative adversarial networks: Novel empirical losses and label input mechanisms. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(7):8143–8158, 2022.

[14] Gregory Druck and Andrew McCallum. Toward interactive training and evaluation. In *Proceedings of the 20th ACM International Conference on Information and Knowledge Management*, pages 947–956, 2011.

[15] Nick Duffield, Carsten Lund, and Mikkel Thorup. Learn more, sample less: control of volume and variance in network measurement. *IEEE Transactions on Information Theory*, 51(5):1756–1775, 2005.

[16] Sebastian Farquhar, Yarin Gal, and Tom Rainforth. On statistical bias in active learning: How and when to fix it. In *International Conference on Learning Representations*, 2021.

[17] Ravi Ganti and Alexander Gray. UPAL: Unbiased pool based active learning. In *Artificial Intelligence and Statistics*, pages 422–431. PMLR, 2012.

[18] Mingfei Gao, Zizhao Zhang, Guo Yu, Sercan Ö Arık, Larry S Davis, and Tomas Pfister. Consistency-based semi-supervised active learning: Towards minimizing labeling cost. In *European Conference on Computer Vision*, pages 510–526. Springer, 2020.

[19] Xiang Gao, Farhad Ramezanghorbani, Olexandr Isayev, Justin S Smith, and Adrian E Roitberg. TorchANI: a free and open source PyTorch-based deep learning implementation of the ANI neural network potentials. *Journal of Chemical Information and Modeling*, 60(7):3408–3415, 2020.

[20] Tilmann Gneiting and Adrian E Raftery. Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, 102(477):359–378, 2007.

[21] Charles R Harris, K Jarrod Millman, Stéfan J Van Der Walt, Ralf Gommers, Pauli Virtanen, David Cournapeau, Eric Wieser, Julian Taylor, Sebastian Berg, Nathaniel J Smith, et al. Array programming with NumPy. *Nature*, 585(7825):357–362, 2020.

[22] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016.

[23] Neil Houlsby, Ferenc Huszár, Zoubin Ghahramani, and Máté Lengyel. Bayesian active learning for classification and preference learning. *arXiv preprint arXiv:1112.5745*, 2011.

[24] J. D. Hunter. Matplotlib: A 2D graphics environment. *Computing in Science & Engineering*, 9 (3):90–95, 2007. doi: 10.1109/MCSE.2007.55.

[25] Henrik Imberg, Johan Jonasson, and Marina Axelson-Fisk. Optimal sampling in unbiased active learning. In *Proceedings of the 23rd International Conference on Artificial Intelligence and Statistics*, volume 108 of *PMLR*, pages 559–569. PMLR, 26–28 Aug 2020.

[26] Cary T Isaki and Wayne A Fuller. Survey design under the regression superpopulation model. *Journal of the American Statistical Association*, 77(377):89–96, 1982.

[27] Rajesh Jayaram, David P Woodruff, and Samson Zhou. Truly perfect samplers for data streams and sliding windows. In *Proceedings of the 41st ACM SIGMOD-SIGACT-SIGAI Symposium on Principles of Database Systems*, pages 29–40, 2022.

[28] Disi Ji, Robert L Logan, Padhraic Smyth, and Mark Steyvers. Active bayesian assessment of black-box classifiers. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 7935–7944, 2021.

[29] Mohammad Reza Karimi, Nezihe Merve Gürel, Bojan Karlaš, Johannes Rausch, Ce Zhang, and Andreas Krause. Online active model selection for pre-trained classifiers. In *International Conference on Artificial Intelligence and Statistics*, pages 307–315. PMLR, 2021.

[30] Parnian Kassraie, Nicolas Emmenegger, Andreas Krause, and Aldo Pacchiano. Anytime model selection in linear bandits. *Advances in Neural Information Processing Systems*, 36:29478–29514, 2023.

[31] Namit Katariya, Arun Iyer, and Sunita Sarawagi. Active evaluation of classifiers on large datasets. In *2012 IEEE 12th International Conference on Data Mining*, pages 329–338. IEEE, 2012.

[32] Justin Kay, Grant Van Horn, Subhransu Maji, Daniel Sheldon, and Sara Beery. Consensus-driven active model selection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4594–4604, 2025.

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

[34] Jannik Kossen, Sebastian Farquhar, Yarin Gal, and Thomas Rainforth. Active surrogate estimators: An active learning approach to label-efficient model evaluation. *Advances in Neural Information Processing Systems*, 35:24557–24570, 2022.

[35] Anurag Kumar and Bhiksha Raj. Classifier risk estimation under limited labeling resources. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, pages 3–15. Springer, 2018.

[36] Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. *Advances in Neural Information Processing Systems*, 30, 2017.

[37] Marine Le Morvan, Nicolas Prost, Julie Josse, Erwan Scornet, and Gaël Varoquaux. Linear predictor on linearly-generated data with missing values: non consistency and solutions. In *International Conference on Artificial Intelligence and Statistics*, pages 3165–3174. PMLR, 2020.

[38] Puheng Li, Tijana Zrnic, and Emmanuel Candes. Robust sampling for active statistical inference. In D. Belgrave, C. Zhang, H. Lin, R. Pascanu, P. Koniusz, M. Ghassemi, and N. Chen, editors, *Advances in Neural Information Processing Systems*, volume 38, pages 68727–68756. Curran Associates, Inc., 2025.

[39] Sanmin Liu, Shan Xue, Jia Wu, Chuan Zhou, Jian Yang, Zhao Li, and Jie Cao. Online active learning for drifting data streams. *IEEE Transactions on Neural Networks and Learning Systems*, 34(1):186–200, 2021.

[40] Xuefeng Liu, Fangfang Xia, Rick L Stevens, and Yuxin Chen. Contextual active online model selection with expert advice. In *ICML2022 Workshop on Adaptive Experimental Design and Active Learning in the Real World*. ICML, 2022.

[41] David Lowell, Zachary C Lipton, and Byron C Wallace. Practical obstacles to deploying active learning. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 21–30, 2019.

[42] Carsten Lüth, Till Bungert, Lukas Klein, and Paul Jaeger. Navigating the pitfalls of active learning evaluation: A systematic framework for meaningful performance assessment. *Advances in Neural Information Processing Systems*, 36:9789–9836, 2023.

[43] David JC MacKay. Information-based objective functions for active data selection. *Neural Computation*, 4(4):590–604, 1992.

[44] Wes McKinney et al. Data structures for statistical computing in Python. *9th Python in Science Conference (Scipy)*, 445(1):51–56, 2010.

[45] Sudhanshu Mittal, Maxim Tatarchenko, Özgün Çiçek, and Thomas Brox. Parting with illusions about deep active learning. *arXiv preprint arXiv:1912.05361*, 2019.

[46] Subhojyoti Mukherjee, Ardhendu S. Tripathy, and Robert Nowak. Chernoff sampling for active testing and extension to active regression. In Gustau Camps-Valls, Francisco J. R. Ruiz, and Isabel Valera, editors, *Proceedings of the 25th International Conference on Artificial Intelligence and Statistics*, volume 151, pages 7384–7432. PMLR, 28–30 Mar 2022.

[47] Patrik Okanovic, Andreas Kirsch, Jannes Kasper, Torsten Hoefler, Andreas Krause, and Nezihe Merve Gürel. All models are wrong, some are useful: Model selection with limited labels. In *The 28th International Conference on Artificial Intelligence and Statistics*, 2025.

[48] Art B. Owen. *Monte Carlo theory, methods and examples*. `https://artowen.su.domains/mc/`, 2013.

[49] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in Neural Information Processing Systems*, 32, 2019.

[50] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. Scikit-learn: Machine learning in Python. *The Journal of Machine Learning Research*, 12:2825–2830, 2011.

[51] Nicholas Petrick, Berkman Sahiner, Samuel G Armato III, Alberto Bert, Loredana Correale, Silvia Delsanto, Matthew T Freedman, David Fryd, David Gur, Lubomir Hadjiiski, et al. Evaluation of computer-aided detection and diagnosis systems. *Medical Physics*, 40(8):087001, 2013.

[52] Christoph Sawade. *Active evaluation of predictive models*. Universitätsverlag Potsdam, 2013.

[53] Christoph Sawade, Niels Landwehr, Steffen Bickel, and Tobias Scheffer. Active risk estimation. In *Proceedings of the 27th International Conference on Machine Learning*, pages 951–958, 2010.

[54] Christoph Sawade, Niels Landwehr, and Tobias Scheffer. Active estimation of F-measures. *Advances in Neural Information Processing Systems*, 23, 2010.

[55] Christoph Sawade, Niels Landwehr, and Tobias Scheffer. Active comparison of prediction models. *Advances in Neural Information Processing Systems*, 25, 2012.

[56] Burr Settles. Active learning literature survey. Technical report, University of Wisconsin-Madison Department of Computer Sciences, 2009.

[57] Jicheng Shan, Hang Zhang, Weike Liu, and Qingbao Liu. Online active learning ensemble framework for drifted data streams. *IEEE Transactions on Neural Networks and Learning Systems*, 30(2):486–498, 2019.

[58] Justin S Smith, Olexandr Isayev, and Adrian E Roitberg. ANI-1: an extensible neural network potential with DFT accuracy at force field computational cost. *Chemical Science*, 8(4):3192–3203, 2017.

[59] Sedat Sonko, Emmanuel Augustine Etukudoh, Kenneth Ifeanyi Ibekwe, Valentine Ikenna Ilojianya, and Cosmas Dominic Daudu. A comprehensive review of embedded systems in autonomous vehicles: Trends, challenges, and future directions. *World Journal of Advanced Research and Reviews*, 21(1):2009–2020, 2024.

[60] Aude Sportisse, Hugo Schmutz, Olivier Humbert, Charles Bouveyron, and Pierre-Alexandre Mattei. Are labels informative in semi-supervised learning? Estimating and leveraging the missing-data mechanism. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett, editors, *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 32521–32539. PMLR, 23–29 Jul 2023.

[61] Guido Van Rossum and Fred L Drake Jr. *Python tutorial*, volume 620. Centrum voor Wiskunde en Informatica Amsterdam, The Netherlands, 1995.

[62] Michael L. Waskom. Seaborn: statistical data visualization. *Journal of Open Source Software*, 6(60):3021, 2021. doi: 10.21105/joss.03021. URL `https://doi.org/10.21105/joss.03021`.

[63] Christopher KI Williams and Carl Edward Rasmussen. *Gaussian processes for machine learning*, volume 2. MIT press, Cambridge, MA, 2006.

[64] Han Xiao, Kashif Rasul, and Roland Vollgraf. Fashion-MNIST: a novel image dataset for benchmarking machine learning algorithms, 2017.

[65] Dayou Yu, Weishi Shi, and Qi Yu. Actively testing your model while it learns: Realizing label-efficient learning in practice. In *Advances in Neural Information Processing Systems*, 2023.

[66] Tijana Zrnic and Emmanuel J. Candès. Active statistical inference. In *Proceedings of the 41st International Conference on Machine Learning*. PMLR, 2024.

A Proof of Theorem 1, unbiasedness of the estimator

Theorem 1 ***Unbiasedness*** *Let $T \geq 1$. If $\beta_t > 0$ for all $t \in [\![1, T]\!]$, then $\mathcal{R}_T(\theta)$ is unbiased:*

$$\mathbb{E}_{\mathcal{F}_T}[\mathcal{R}_T(\theta)] = \mathcal{R}(\theta).$$

Proof. The proof is similar to [17]. We have to consider the expectation with respect to all the random variables involved, i.e. all couples (x_t, y_t, r_t) for $t \in [\![1, T]\!]$. First, we remark that $\forall t \in [\![1, T]\!]$, we have that $L(\theta; x_t, y_t) = \ell_t(\theta)$ is $\mathcal{F}_t$-measurable,

$$\mathbb{E}[\ell_t(\theta)|\mathcal{F}_t] = \ell_t(\theta).$$

Additionally, β_t is $\mathcal{F}_{t-1} \cup \sigma(x_t)$-measurable, thus, $\forall t \in [\![1, T]\!]$, we can compute the expectation:

$$\begin{aligned}
\mathbb{E}_{\mathcal{F}_T}\left[\frac{r_t}{\beta(x_t)}\ell_t(\theta)\right] &= \mathbb{E}_{\mathcal{F}_{t-1},x_t,y_t,r_t}\left[\frac{r_t}{\beta(x_t)}\ell_t(\theta)\right] \\
&= \mathbb{E}_{x_t,y_t,\mathcal{F}_{t-1}}\left[\mathbb{E}_{r_t}\left[\frac{r_t}{\beta(x_t)}\ell_t(\theta)|\mathcal{F}_{t-1}, x_t, y_t\right]\right] \\
&= \mathbb{E}_{x_t,y_t,\mathcal{F}_{t-1}}\left[\frac{1}{\beta_t}\mathbb{E}_{r_t}\left[r_t|\mathcal{F}_{t-1}, x_t, y_t\right]\ell_t(\theta)\right] \\
&= \mathbb{E}_{x_t,y_t,\mathcal{F}_{t-1}}\left[\frac{1}{\beta_t}\mathbb{E}_{r_t}\left[r_t|\mathcal{F}_{t-1}, x_t, y_t\right]\ell_t(\theta)\right] \\
&= \mathbb{E}_{x_t,y_t,\mathcal{F}_{t-1}}\left[\ell_t(\theta)\right] = \mathbb{E}_{\mathcal{F}_{t-1}}\left[\mathcal{R}(\theta)\right] = \mathcal{R}(\theta).
\end{aligned}$$

Finally, by linearity of the expectation, we have to compute the expectation of the estimator:

$$\begin{aligned}
\mathbb{E}_{\mathcal{F}_T}[\mathcal{R}_T(\theta)] &= \frac{1}{T}\sum_{t=1}^{T}\mathbb{E}_{\mathcal{F}_T}\left[\ell_t(\theta)\right] \\
&= \frac{1}{T}\sum_{t=1}^{T}\mathbb{E}_{x_t,y_t,r_t,\mathcal{F}_{t-1}}\left[\frac{r_t}{\beta(x_t)}\ell_t(\theta)\right] \\
&= \frac{1}{T}\sum_{t=1}^{T}\mathcal{R}(\theta) = \mathcal{R}(\theta).
\end{aligned}$$

□

B Proof of Theorem 2, deviation and consistency

Theorem 2 ***Deviation and consistency:*** *Assume that the loss function is non-negative and bounded, i.e., for all* $(x, y) \sim p$, $0 \leq L(\theta; x, y) \leq M$ *for some constant* $M > 0$. *Suppose further that there exists* $\beta_{\min} > 0$ *such that* $\beta_t \geq \beta_{\min}$ *for all* $t \in [\![1, T]\!]$. *Then, for any* $\delta \in (0, 1)$,

$$P\left(|\mathcal{R}_T(\theta) - \mathcal{R}(\theta)| > \frac{M}{\beta_{min}}\sqrt{\frac{2\ln(2/\delta)}{T}}\right) < \delta.$$

Proof. Let's consider the sequence of random variable $M1, \cdots, M_T$ defined for all $t \in [\![1, T]\!]$ as:

$$M_t = \frac{r_t}{\beta_t}\ell_t(\theta) - \mathcal{R}(\theta).$$

Since $\beta_t \geq \beta_{min}$ and L is bounded and, then $\mathcal{R}(\theta)$ is too, we have that $|M_t| \leq M/\beta_{min}$. In what follow we prove that the sequence $Z_t = \sum_{i=1}^{t} M_t$ (with $Z_0 = 0$) is a martingale. For all t, we have:

$$\mathbb{E}_{\mathcal{F}_t}[|Z_t|] \leq tM/\beta_{min} < +\infty,$$

and

$$\begin{aligned}
\mathbb{E}_{\mathcal{F}_t}[Z_t|Z_0, \cdots, Z_{t-1}] &= \mathbb{E}_{x_t, y_t, r_t}[M_t + Z_{t-1}|Z_0, \cdots, Z_{t-1}] \\
&= Z_{t-1} + \mathbb{E}_{x_t, y_t, r_t}[M_t|Z_0, \cdots, Z_{t-1}] \\
&= Z_{t-1} + \mathbb{E}_{x_t, y_t, r_t}\left[\frac{r_t}{\beta_t}\ell_t(\theta)|Z_0, \cdots, Z_{t-1}\right] - \mathcal{R}(\theta) \\
&= Z_{t-1} + \mathcal{R}(\theta) - \mathcal{R}(\theta) \\
&= Z_{t-1}.
\end{aligned}$$

Therefore, the sequence $(Z_t)_{t\geq 0}$ is a martingale. Additionally, we have $|Z_{t+1} - Z_t| = |M_{t+1}| \leq M/\beta_{min}$.

Using $Z_T = T(\mathcal{R}_T(\theta) - \mathcal{R}(\theta))$ and applying Azuma's inequality [2]:

$$P\left(|\mathcal{R}_T(\theta) - \mathcal{R}(\theta)| > \frac{\lambda M}{\beta_{min}\sqrt{T}}\right) = P\left(|Z_T| > \frac{\lambda M\sqrt{T}}{\beta_{min}}\right) \leq 2e^{-\lambda^2/2}.$$

Setting $\lambda = \sqrt{2\ln(2/\delta)}$ leads to the theorem's formulation. □

C Proof of Theorem 3, computation of the variance

Theorem 3 ***Variance expression:*** *Let $T > 0$, and assume $\beta_t > 0$ for all $t \in [\![1, T]\!]$. The variance of $\mathcal{R}_T(\theta)$ is given by*

$$V_{\mathcal{F}_T}[\mathcal{R}_T(\theta)] = \frac{V_{x,y}[L(\theta; x, y)]}{T} + \frac{1}{T^2}\sum_{t=1}^{T} \mathbb{E}_{x_t, y_t, \mathcal{F}_{t-1}}\left[\frac{1-\beta_t}{\beta_t} L(\theta; x_t, y_t)^2\right].$$

Proof. We denote $\mathcal{S}_T$ the stream of samples (always observed) and labels (observed or unobserved) of the stream at time T, $\mathcal{S}_T = \{(x_t, y_t)\}_{t \leq T}$.

By the definition of the variance and unbiasedness of the estimator, we have:

$$\begin{aligned} V_{\mathcal{F}_T}[\mathcal{R}_T(\theta)] &= \mathbb{E}_{\mathcal{F}_T}[\mathcal{R}_T(\theta)^2] - \mathbb{E}_{\mathcal{F}_T}[\mathcal{R}_T(\theta)]^2 \\ &= \mathbb{E}_{\mathcal{F}_T}[\mathcal{R}_T(\theta)^2] - \mathcal{R}(\theta)^2, \end{aligned}$$

which, by the tower property of expectations, by denoting $\boldsymbol{r} = \{r_1, \cdots, r_T\}$ and the definition of the variance, yields:

$$\begin{aligned} V_{\mathcal{F}_T}[\mathcal{R}_T(\theta)] &= \mathbb{E}_{\mathcal{S}_T}[\mathbb{E}_{\boldsymbol{r}}[\mathcal{R}_T(\theta)^2|\mathcal{S}_T]] - \mathcal{R}(\theta)^2 \\ &= \mathbb{E}_{\mathcal{S}_T}\left[V_{\boldsymbol{r}}[\mathcal{R}_T(\theta)|\mathcal{S}_T] + \mathbb{E}_{\boldsymbol{r}}[\mathcal{R}_T(\theta)|\mathcal{S}_T]^2\right] - \mathcal{R}(\theta)^2 \\ &= \mathbb{E}_{\mathcal{S}_T}[V_{\boldsymbol{r}}[\mathcal{R}_T(\theta)|\mathcal{S}_T]] + \mathbb{E}_{\mathcal{S}_T}[\mathbb{E}_{\boldsymbol{r}}[\mathcal{R}_T(\theta)|\mathcal{S}_T]^2] - \mathcal{R}(\theta)^2. \end{aligned}$$

The expectation of $\mathcal{R}_T(\theta)$ knowing $\mathcal{S}_T$ is equal to the Monte Carlo estimator of the risk, observing all labels:

$$\mathbb{E}_{\boldsymbol{r}}[\mathcal{R}_T(\theta)|\mathcal{S}_T] = \frac{1}{T}\sum_{t=1}^{T} \ell_t(\theta) := \mathcal{R}_T^{MC}(\theta).$$

The stream being independent and identically distributed, the Monte Carlo estimator is also an unbiased estimator of $\mathcal{R}(\theta)$ and has a variance $V_{\mathcal{S}_T}[\mathcal{R}_T^{MC}(\theta)] = V_{x,y}[L(\theta; x, y)]/T$. Thus, by definition of the variance:

$$\begin{aligned} V[\mathcal{R}_T(\theta)] &= \mathbb{E}_{\mathcal{S}_T}[V_{\boldsymbol{r}}[\mathcal{R}_T(\theta)|\mathcal{S}_T]] + \mathbb{E}_{\mathcal{S}_T}[\mathcal{R}_T^{MC}(\theta)^2] - \mathcal{R}(\theta)^2 \\ &= \mathbb{E}_{\mathcal{S}_T}[V_{\boldsymbol{r}}[\mathcal{R}_T(\theta)|\mathcal{S}_T]] + \mathbb{E}_{\mathcal{S}_T}[\mathcal{R}_T^{MC}(\theta)^2] - \mathbb{E}_{\mathcal{S}_T}[\mathcal{R}_T^{MC}(\theta)]^2 \\ &= \mathbb{E}_{\mathcal{S}_T}[V_{\boldsymbol{r}}[\mathcal{R}_T(\theta)|\mathcal{S}_T]] + V_{\mathcal{S}_T}[\mathcal{R}_T^{MC}(\theta)] \\ &= \mathbb{E}_{\mathcal{S}_T}[V_{\boldsymbol{r}}[\mathcal{R}_T(\theta)|\mathcal{S}_T]] + \frac{V_{x,y}[L(\theta; x, y)]}{T}. \end{aligned}$$

From this point, we need to consider the first term $\mathbb{E}_{\mathcal{S}_T}[V_{\boldsymbol{r}}[\mathcal{R}_T(\theta)|\mathcal{S}_T]]$ and we will start by attacking the variance $V_{\boldsymbol{r}}[\mathcal{R}_T(\theta)|\mathcal{S}_T]$. By definition:

$$\begin{aligned} V_{\boldsymbol{r}}[\mathcal{R}_T(\theta)|\mathcal{S}_T] &= \mathbb{E}_{\boldsymbol{r}}[(\mathcal{R}_T(\theta) - \mathbb{E}_{\boldsymbol{r}}[\mathcal{R}_T(\theta)|\mathcal{S}_T])^2|\mathcal{S}_T] \\ &= \mathbb{E}_{\boldsymbol{r}}[(\mathcal{R}_T(\theta) - \mathcal{R}_T^{MC}(\theta))^2|\mathcal{S}_T] \\ &= \mathbb{E}_{\boldsymbol{r}}\left[\left(\frac{1}{T}\sum_{t=1}^{T}\left(\frac{r_t}{\beta_t} - 1\right)\ell_t(\theta)\right)^2 \mid \mathcal{S}_T\right] \\ &= \frac{1}{T^2}\sum_{t,t'=1}^{T} \mathbb{E}_{r_t, r_{t'}}\left[\left(\frac{r_t}{\beta_t} - 1\right)\left(\frac{r_{t'}}{\beta_{t'}} - 1\right)\ell_t(\theta)\ell_{t'}(\theta)|\mathcal{S}_T\right]. \end{aligned}$$

We attack this term by first considering the terms for which $t \neq t'$. We will assume, without loss of generality, that $t' < t$, noting that by symmetry the same set of arguments can be similarly applied when $t < t'$.

Using the tower property on the first term and the fact that $t' < t$, thus $(r_{t'}/\beta(x_{t'}) - 1)\ell_{t'}(\theta)$ is deterministic given $\mathcal{F}_{t-1}$ and β_t is deterministic given $\mathcal{F}_{t-1}$ and $\mathcal{S}_T$, we have:

$$\begin{aligned}
&\mathbb{E}_{r_t, r_{t'}}\left[\left(\frac{r_t}{\beta_t} - 1\right)\left(\frac{r_{t'}}{\beta_{t'}} - 1\right)\ell_t(\theta)\ell_{t'}(\theta)|\mathcal{S}_T\right] \\
&\quad = \mathbb{E}_{\mathcal{F}_{t-1}}\left[\mathbb{E}_{r_t, r_{t'}}\left[\left(\frac{r_t}{\beta_t} - 1\right)\left(\frac{r_{t'}}{\beta_{t'}} - 1\right)\ell_t(\theta)\ell_{t'}(\theta)|\mathcal{F}_{t-1}, \mathcal{S}_T\right]|\mathcal{S}_T\right] \\
&\quad = \mathbb{E}_{\mathcal{F}_{t-1}}\left[\left(\frac{r_{t'}}{\beta_{t'}} - 1\right)\ell_{t'}(\theta)\ell_t(\theta)\mathbb{E}_{r_t}\left[\frac{r_t}{\beta_t} - 1|\mathcal{F}_{t-1}, \mathcal{S}_T\right]|\mathcal{S}_T\right] \\
&\quad = 0.
\end{aligned}$$

Finally, we consider the terms with $t = t'$ noting β_t is deterministic given $\mathcal{F}_{t-1}$ and $\mathcal{S}_T$:

$$\begin{aligned}
&\mathbb{E}_{r_t}\left[\left(\frac{r_t}{\beta_t} - 1\right)^2\ell_t(\theta)^2|\mathcal{S}_T\right] \\
&\mathbb{E}_{r_t}\left[\left(\frac{r_t}{\beta_t} - 1\right)^2|\mathcal{S}_T\right]\ell_t(\theta)^2 \\
&\quad = \mathbb{E}_{\mathcal{F}_{t-1}}\left[V_{r_t}\left[\left(\frac{r_t}{\beta_t} - 1\right)|\mathcal{F}_{t-1}, \mathcal{S}_T\right]|\mathcal{S}_T\right]\ell_t(\theta)^2 \\
&\quad = \mathbb{E}_{\mathcal{F}_{t-1}}\left[\frac{1}{\beta_t^2}V_{r_t}\left[r_t|\mathcal{F}_{t-1}, \mathcal{S}_T\right]|\mathcal{S}_T\right]\ell_t(\theta)^2 \\
&\quad = \mathbb{E}_{\mathcal{F}_{t-1}}\left[\frac{1-\beta_t}{\beta_t}\ell_t(\theta)^2|\mathcal{D}_T\right] \\
&\quad = \frac{1-\beta_t}{\beta_t}\ell_t(\theta)^2.
\end{aligned}$$

Finally, putting everything together, we get:

$$V_{\mathcal{F}_T}[\mathcal{R}_T(\theta)] = \frac{V_{x,y}[L(\theta; x, y)]}{T} + \frac{1}{T^2}\sum_{t=1}^{T}\mathbb{E}_{x_t, y_t, \mathcal{F}_{t-1}}\left[\frac{1-\beta_t}{\beta_t}L(\theta; x_t, y_t)^2\right].$$

□

Uniform baseline and oracle variance analysis We consider the baseline uniform strategy $\beta_t = n_B/T$. In this case, the variance simplifies to

$$V_{\mathcal{F}_T}[\mathcal{R}_T(\theta)] = \frac{V_{x,y}[L(\theta; x, y)]}{n_B} + \left(\frac{1}{n_B} - \frac{1}{T}\right)\mathcal{R}(\theta)^2.$$

The first term corresponds to the effective sample variance based on n_B labels, while the second term is an importance-weighting penalty. This penalty shows that inverse-probability weighting inflates variance when the expected loss is large. In the limit $n_B = T$, we recover the standard Monte Carlo variance.

On the other hand, a well-known oracle with access to the true labels would sample proportionally to the loss $\beta_t = n_B/T \cdot \ell_t(\theta)/\mathbb{E}[\ell_t(\theta)]$. This strategy is known to be optimal in classical importance sampling [10, 9, 48]. However, it is not implementable in practice since the label y_t is unknown at the time of selection. This strategy yields to a variance

$$V_{\mathcal{F}_T}[\mathcal{R}_T(\theta)] = \left(\frac{1}{n_B} - \frac{1}{T}\right)\mathcal{R}(\theta)^2.$$

The possible improvement for an active sampling strategy lies between these two variances in a range of $V_{x,y}[L(\theta; x, y)]/n_B$.

D Proof of Theorem 4, optimal sampling strategy

Theorem 4 ***Optimal sampling strategy*** *Under the budget constraint $\sum_{t=1}^{T} \beta_t = n_B$, the optimal sampling probabilities satisfy*

$$\beta_t^*(x_t) \propto n_B/T \cdot \sqrt{\mathbb{E}_{y|x}[\ell_t(\theta)^2]}.$$

Proof. We recall that we denote $\mathcal{D}_T = \{x_1, \cdots, x_T\}$ as the set of all samples in the stream up to time T. From the proof of Theorem 3, we have that

$$V_{\boldsymbol{r}}[\mathcal{R}_T(\theta)|\mathcal{D}_T] = \frac{1}{T^2}\sum_{t=1}^{T}\frac{1-\beta_t}{\beta_t}\ell_t(\theta)^2 = \frac{1}{T^2}\sum_{t=1}^{T}\frac{\ell_t(\theta)^2}{\beta_t} - \frac{1}{T^2}\sum_{t=1}^{T}\ell_t(\theta)^2.$$

Therefore, we can express the anticipated variance of our estimator as

$$\mathbb{E}_{y|x}[V_{\boldsymbol{r}}[\mathcal{R}_T(\theta)|\mathcal{D}_T]] = \frac{1}{T^2}\sum_{t=1}^{T}\frac{\mathbb{E}_{y_t|x_t}[\ell_t(\theta)^2]}{\beta_t} - \frac{1}{T^2}\sum_{t=1}^{T}\mathbb{E}_{y_t|x_t}[\ell_t(\theta)^2]].$$

Our objective is to minimise this anticipated variance under the budget constraint $\sum_{t=1}^{T} \beta_t = n_B$. We consider the following optimisation problem:

$$\begin{aligned} \min_{\beta} \quad & \mathbb{E}_{y|x}[V_{\boldsymbol{r}}[\mathcal{R}_T(\theta)|\mathcal{D}_T]] \\ \text{s.t.} \quad & \sum_{t=1}^{T}\beta_t = n_B. \end{aligned} \tag{6}$$

Since the second term is constant with respect to β, optimisation reduces to minimising the first term:

$$\arg\min_{\beta} \mathbb{E}_{y|x}[V_{\boldsymbol{r}}[\mathcal{R}_T(\theta)|\mathcal{D}_T]] = \arg\min_{\beta}\sum_{t=1}^{T}\frac{\mathbb{E}_{y_t|x_t}[\ell_t(\theta)^2]}{\beta_t}.$$

Using the method of Lagrange multipliers, we introduce the auxiliary function

$$\Lambda(\beta,\lambda) = \sum_{t=1}^{T}\frac{\mathbb{E}_{y_t|x_t}[\ell_t(\theta)^2]}{\beta_t} + \lambda\left(\sum_{t=1}^{T}\beta_t - n_B\right).$$

Then, critical points of the Lagrangian are found by solving the system of equations:

$$\Lambda(\beta,\lambda) = 0 \quad \Leftrightarrow \quad \begin{cases} \sum_{t=1}^{T}\beta_t = n_B \\ \frac{\mathbb{E}_{y_t|x_t}[\ell_t(\theta)^2]}{\beta_t^2} = \lambda \end{cases}. \tag{7}$$

Then, we get, $\forall t$, $\beta_t = \sqrt{\frac{\mathbb{E}_{y_t|x_t}[\ell_t(\theta)^2]}{\lambda}}$. And by the first condition:

$$\sum_{t=1}^{T}\beta_t = n_B \quad \Leftrightarrow \quad \beta_t = n_B\,\frac{\sqrt{\mathbb{E}_{y_t|x_t}[\ell_t(\theta)^2]}}{\sum_{t'=1}^{T}\sqrt{\mathbb{E}_{y_{t'}|x_{t'}}[\ell_{t'}(\theta)^2]}}.$$

This concludes the proof. The optimal strategy is to select points proportionally to the square root of the second moment of loss in $y \mid x$. □

Remark (Efficiency over uniform baseline) Unless the second moments $\mathbb{E}_{y|x}[\ell_t(\theta)^2]$ are constant across all inputs, the optimal strategy β_t^* strictly improves over the uniform baseline. In particular, it achieves both a lower anticipated variance and a lower label complexity for a fixed estimation accuracy. The improvement arises because β_t^* adapts to the heterogeneity of the data through the second moment of the loss. When all second moments are equal, no such adaptation is possible, and the uniform sampling strategy becomes optimal.

Proof. Let compare the anticipated variance of $\mathcal{R}_T(\theta)$ at β^* (denoted by $\mathcal{R}_T(\theta;\beta^*)$) with the estimator obtained with uniform distribution $\beta_{u,t} = n_B/T$ for any $t \leq T$ (denoted by $\mathcal{R}_T(\theta;\beta_u)$). Both these strategies lead to unbiased estimators; we can compare their anticipated variance.

For the optimal strategy, we get:

$$\mathbb{E}_{y|x}[V_{\boldsymbol{r}}[\mathcal{R}_T(\theta;\beta^*)|\mathcal{D}_T]] = \frac{1}{n_B T^2}\left(\sum_{t=1}^{T}\sqrt{\mathbb{E}_{y|x}[\ell_t(\theta)^2]}\right)^2 - \frac{1}{T^2}\sum_{t=1}^{T}\mathbb{E}_{y|x}[\ell_t(\theta)^2]].$$

On the other hand:

$$\mathbb{E}_{y|x}[V_{\boldsymbol{r}}[\mathcal{R}_T(\theta;\beta_u)|\mathcal{D}_T]] = \frac{1}{n_B T}\sum_{t=1}^{T}\mathbb{E}_{y|x}[\ell_t(\theta)^2] - \frac{1}{T^2}\sum_{t=1}^{T}\mathbb{E}_{y|x}[\ell_t(\theta)^2]]. \tag{8}$$

Using the Cauchy-Schwarz inequality, we get that:

$$\mathbb{E}_{y|x}[V[\mathcal{R}_T(\theta;\beta^*)|\mathcal{D}_T]] \leq \mathbb{E}_{y|x}[V[\mathcal{R}_T(\theta;\beta_u)|\mathcal{D}_T]]. \tag{9}$$

The equality case of Cauchy-Schwarz yields $\beta_T = C\,\beta_u$, with $C \geq 0$ (actually, with the budget constraint, we have $C = 1$). We conclude that the optimal strategy β^* has an anticipated variance strictly smaller than the uniform baseline as long as all second moments are not equal. □

Remark (Efficiency over first-moment strategy) Unless the second moments $\mathbb{E}_{y|x}[\ell_t(\theta)^2]$ are proportional to the square of the first moments across all inputs, the optimal strategy β_t^* strictly improves over the first moment strategy in terms of anticipated variance.

Proof. Since the second term of this anticipated variance characterisation is constant with respect to the sampling policy, we will compare only the first term between the first- and second-moment strategies. The second-moment strategy leads to $\beta_t = n_B \frac{\sqrt{E_{y_t|x_t}[\ell_t(\theta)^2]}}{\sum_{t'=1}^{T}\sqrt{E_{y_{t'}|x_{t'}}[\ell_{t'}(\theta)^2]}}$.

The first term of the anticipated variance for the second moment strategy is $\frac{1}{n_B T^2}\left(\sum_{t=1}^{T}\sqrt{E_{y_t|x_t}[\ell_t(\theta)^2]}\right)^2$.

On the other hand, the first-moment strategy leads to $\beta_t = n_B \frac{E_{y_t|x_t}[\ell_t(\theta)]}{\sum_{t'=1}^{T} E_{y_{t'}|x_{t'}}[\ell_{t'}(\theta)]}$ and therefore the first term of the anticipated variance is $\frac{1}{n_B T^2}\left(\sum_{t=1}^{T}\frac{E_{y_t|x_t}[\ell_t(\theta)^2]}{E_{y_t|x_t}[\ell_t(\theta)]}\right)\left(\sum_{t=1}^{T} E_{y_t|x_t}[\ell_t(\theta)]\right)$.

By applying Cauchy-Schwarz, we prove the second-moment strategy is never worse in terms of anticipated variance than the first-moment strategy, and the equality condition is given by $\sqrt{E_{y|x}[\ell_t(\theta)^2]} \propto E_{y|x}[\ell_t(\theta)]$. □

D.1 Cleaner version of the Theorem

For simplicity, Theorem 4 omits the condition that the β_t are probabilities and thus, $\beta_t \in]0, 1=$. In particular, the previous optimal sampling strategy can break this condition when the square root of the second moment of a sample x_t is large relative to the others:

$$n_B\ \sqrt{\mathbb{E}_{y_t|x_t}[\ell_t(\theta)^2]} > \sum_{t'=1}^{T}\sqrt{\mathbb{E}_{y_{t'}|x_{t'}}[\ell_{t'}(\theta)^2]}.$$

We propose here a corrected version of the previous Theorem and an interpretation of the changes:

Theorem 5. *Cleaner version of the optimal sampling Theorem Under the budget constraint* $\sum_{t=1}^{T}\beta_t = n_B$ *and* $\beta_t \in]0, 1]$ *the optimal sampling probabilities satisfy* $\beta_t^*(x_t) = \min\left(1, c\sqrt{\mathbb{E}_{y|x}[\ell_t(\theta)^2]}\right)$*, with* $c > 0$ *the unique constant satisfying* $\sum_t \min\left(1, c\sqrt{\mathbb{E}_{y|x}[\ell_t(\theta)^2]}\right) = n_B$.

In particular, if we have for all t, $n_B \sqrt{\mathbb{E}_{y_t|x_t}[\ell_t(\theta)^2]} \leq \sum_{t'=1}^{T} \sqrt{\mathbb{E}_{y_{t'}|x_{t'}}[\ell_{t'}(\theta)^2]}$, then we fall back in the conclusion of Theorem 4:

$$\beta_t^* = n_B \frac{\sqrt{\mathbb{E}_{y_t|x_t}[\ell_t(\theta)^2]}}{\sum_{t'=1}^{T} \sqrt{\mathbb{E}_{y_{t'}|x_{t'}}[\ell_{t'}(\theta)^2]}}.$$

We propose below a sketch of this proof

Proof. The lower bound constraint of β_t is never active as the objective diverges in 0(as long as $\mathbb{E}_{y_t|x_t}[\ell_t(\theta)^2] > 0$, otherwise sample x_t does not contribute and can be given $\beta_t = 0$). Therefore, only the upper bound matters.

The Lagrangian becomes:

$$\Lambda(\beta, \lambda, \{\mu_t\}_{t=1}^T) = \sum_{t=1}^{T} \frac{\mathbb{E}_{y_t|x_t}[\ell_t(\theta)^2]}{\beta_t} + \lambda \left(\sum_{t=1}^{T} \beta_t - n_B \right) + \sum_{t=1}^{T} \mu_t(\beta_t - 1).$$

We find that critical points should satisfy:

$$\Lambda(\beta, \lambda) = 0 \quad \Leftrightarrow \quad \begin{cases} \sum_{t=1}^{T} \beta_t = n_B \\ \frac{\mathbb{E}_{y_t|x_t}[\ell_t(\theta)^2]}{\beta_t^2} = \lambda + \mu_t \\ \mu_t(\beta_t - 1) = 0 \end{cases} . \tag{10}$$

Analysing, $\beta_t < 1$ and $\beta_t = 1$ cases, we find that:

$$\beta_t^* = \min\left(1, c\sqrt{\mathbb{E}_{y|x}[\ell_t(\theta)^2]}\right),$$

where c is the unique value such that $\sum_t \min\left(1, c\sqrt{\mathbb{E}_{y|x}[\ell_t(\theta)^2]}\right) = n_B$.

□

Interpretation, risk and practical challenger Samples x_t with large second moment are systematically sampled. The remaining budget is then allocated to the other points according to the following formula:

$$\beta_t^* = \left(n_B - |S|\right) \frac{s_t}{\sum_{t' \in U} s_{t'}},$$

where $S = \left\{ x_t | n_B \sqrt{\mathbb{E}_{y_t|x_t}[\ell_t(\theta)^2]} > \sum_{t'=1}^{T} \sqrt{\mathbb{E}_{y_{t'}|x_{t'}}[\ell_{t'}(\theta)^2]} \right\}$, the saturated set (points systematically sampled) and U the unsaturated samples, the complement of S.

Computing β_t^* according to Theorem 4 may lead to underspending the label budget n_B because we do not reallocate the budget of the saturated samples, and ultimately a higher variance. Note that the unbiasedness of the estimator is preserved. This risk is particularly important in large sampling streams ($n_B \sim T$) or heavy-tailed losses.

In practice, c is not computable before observing all the second moments of the stream and therefore is not computable in the restrictive setting of this paper. A practical approach can be to compute $\beta_t = \min(1, c_t s_t)$ with c_t that is updated according to the saturated points encountered. The study of the empirical impact of this cleaner version is left as future work.

E Optimal strategy derivations and analysis

In this section, we recall the analysis of the optimal strategy for regression presented in Section 5 with an illustrative figure to better understand the bias-3rd moment interaction. Additionally, we present the derivation of the optimal strategy for two classification losses, the log-loss and 0–1 loss.

Derivation for the regression Substituting the squared error loss $\ell_t(\theta) = L(\theta; x_t, y_t) = (f_\theta(x_t) - y_t)^2$ into Equation 3 yields to:

$$\beta_t \propto n_B/T \cdot \sqrt{\underbrace{b(x_t)^4}_{(i)} + 6\underbrace{b(x_t)^2\sigma(x_t)^2}_{(ii)} - 4\underbrace{b(x_t)m_3(x_t)}_{(iii)} + \underbrace{m_4(x_t)}_{(iv)}}, \tag{11}$$

where $b(x_t) = f_\theta(x_t) - \mathbb{E}_{\pi(y|x_t)}[y]$ is the bias, (3) $\sigma(x_t)^2 = V_{\pi(y|x_t)}[y]$ the predictive variance, and m_3, m_4[3] the third and fourth central moments under π.

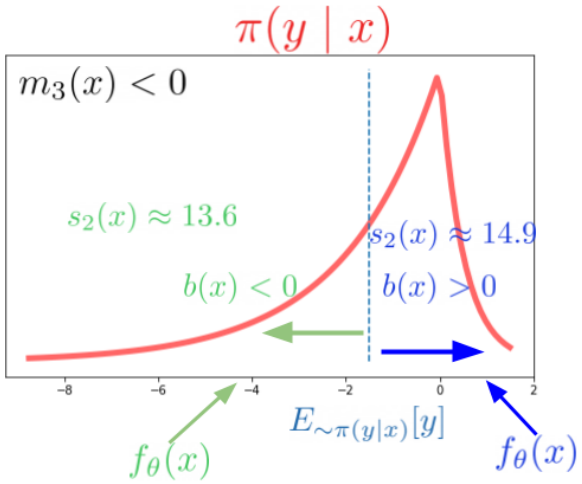


Figure 5: Illustrative example of the effect of the bias and skewness interaction on the selection probability with an asymmetric Laplace distribution. s_2 is the second moment of the loss.

We analyse this expression through its statistical components and highlight key effects. As in Kossen et al. [33], both the bias and variance of the surrogate predictions appear, but here they also interact, and additional higher-order terms arise through the skewness (third moment) and kurtosis (fourth moment). (i) The first term grows rapidly with the bias, making the selection highly sensitive to systematic errors. (ii) The second term couples bias and variance, meaning that uncertainty has a stronger impact in regions where the model is already inaccurate. In contrast to Kossen et al. [33], variance is no longer independent. If the bias is zero (i.e., f_θ and π agree), this term vanishes. (iii) The third moment captures skewness: when bias and skewness have the same sign, the selection probability decreases, and when they differ, it increases (see Figure 5). (iv) The fourth moment reflects kurtosis (tail heaviness), promoting the selection of points with heavy-tailed or highly variable predictions, independently of f_θ.

Overall, the strategy prioritises regions with large bias, high uncertainty, or heavy-tailed behaviour. These quantities can be estimated using probabilistic models (e.g., Gaussian Processes or Bayesian neural networks) or approximated via methods such as deep ensembles [36].

Remark 3. *If π is symmetric (e.g., Gaussian), then $m_3 = 0$, and for Gaussian π, $m_4 = 3\sigma^4$, simplifying the expression.*

For the mean absolute error, the strategy simplifies to $\beta_t \propto n_B/T \cdot \sqrt{b(x_t)^2 + \sigma(x_t)^2}$, which is the same objective as [33] up to the square root scaling that amplifies small values, so overconfidence in π.

[3]The third central moment is given by $m_3(x_t) = \mathbb{E}_{\pi(y|x_t)}[(y - \mu(x_t))^3] = \kappa_3\sigma(x_t)^3$ with κ_3 the skewness of the surrogate's predictions and the fourth central moment $m_4(x_t) = \mathbb{E}_{\pi(y|x_t)}[(y - \mu(x_t))^4] = \kappa_4\sigma(x_t)^4$ with κ_4 the kurtosis of the surrogate's predictions.

Derivation for the classification For probabilistic classifiers $f_\theta(x_t) = \{p_\theta(y|x_t)\}_{y=1}^C$, C the number of classes, evaluated with log-loss, we obtain from Equation 3

$$\beta_t \propto n_B/T \cdot \sqrt{\underbrace{CE(\pi, p_\theta; x_t)^2}_{model\ disagreement} + \underbrace{V_{\pi(y|x_t)}(\log(p_\theta(y|x_t)))}_{uncertainty}}, \tag{12}$$

where $CE(\pi, p_\theta; x_t) = -\sum_{y=1}^C \pi(y|x_t) \log(p_\theta(y|x_t))$. The first term captures disagreement between the model and the surrogate, while the second term emphasises uneven or sharp probability assignments. Compared to standard approaches [33], this formulation places stronger emphasis on "catastrophic" mismatches, where the model assigns very low probability to likely labels.

Classification performance is often evaluated through the 0-1 loss, which leads to

$$\beta_t \propto n_B/T \cdot \sqrt{1 - \pi(\arg\max_y p_\theta(y|x_t) \mid x_t)}. \tag{13}$$

This formulation is similar to Kossen et al. [33], up to the square root scaling, which prioritises points where the predicted class is uncertain under π.

Remark 4. *Other scoring rules (e.g., Brier score [20]) can be handled similarly, though Equation 3 remains the most general formulation.*

E.1 Computational complexity

Computational efficiency is critical in online testing. In fact, the acquisition probability is lightweight: Equation 4 only requires evaluating the surrogate on the current sample and updating the running estimate of the normalising constant. No optimisation is required at the acquisition step. In our implementation, this corresponds to one forward pass through the predictive model for latent feature extraction (already required for prediction), one evaluation of the lightweight surrogate, and negligible additional computation.

The main computational cost comes from surrogate updates after acquiring labels. With retraining after each acquisition, the total complexity is bounded by $\mathcal{O}(n_B|\mathcal{D}_{train}| + n_B^2 + T)$, where n_B is the labelling budget and T the stream length. Thus, OAT is lighter compared with existing pool-based active testing frameworks 33, Appendix B.2.

In its non-adaptive form, the computation cost comes only from the second-moment estimation, yielding a total complexity bounded by $\mathcal{O}(T)$, i.e. similar to the uniform baseline complexity.

F Details on the experiments

We evaluate Online Active Testing (OAT) on synthetic and real-world datasets for both regression and classification. Performance is measured using the mean squared error or variance of the risk estimate, averaged over multiple independent runs.

In particular, our experiments have been conducted to validate our theoretical findings: unbiasedness, variance reduction, and approaching the label-aware oracle. The evaluation covers regression, classification, multiple loss functions, synthetic and image datasets, deep learning models, and a real-world molecular-simulation task. The experiments span the following scales:

Table 1: Dataset Configuration and Model Settings

Dataset	Dimension	Stream	Budget	Repetition	Model	Surrogate
Synthetic regression	1	90	25	20,000	Linear&GP	GP
Two-moons	2	200	50	2,500	RF	RF
FashionMNIST	28x28	5,000	1,000	500	ResNet-18	RF
CellCount	64x64	2,000	200	100	CNN	GP
MELTS	36	2,000	200	100	ANI	GP

F.1 Synthetic data:

Regression tasks Several datasets have been generated according to the following principles:

- Quadratic data from a non-linear function $y = x^2$ with x uniformly sampled.
- GP prior data is sampled from a Gaussian Process (GP) with a Matérn kernel, with x uniformly sampled with smoothness of 1.5.
- Sinusoidal data from a non-linear function $y = \sin(10x) + x^3$ with x sampled from a mixture of two Gaussians.

A total of $20,000$ repetitions of sampling a dataset of 100 points, training the main model with 10 samples and adaptively selecting data points to label in the stream of 90 points have been performed using a budget constraint of 25. For all tasks, the surrogate model is a Gaussian Process, updated online with selected points using a Matérn with smoothness parameter $\nu = 1.5$. All experiments are run on a CPU (Intel i9-10885H) in $\sim$ 30 minutes each.

Classification tasks We consider the standard two-moons dataset. A total of $2,500$ repetitions of sampling a dataset of 500 points, training the main Random Forest model with 50 samples and adaptively selecting data points to label in the stream of 200 points have been performed using a budget constraint of 50. For both log-loss and accuracy estimation, the surrogate model is a Random Forest, updated online with selected points. Both experiments are run on a CPU (Intel i9-10885H) in $\sim$ 10 hours each.

F.2 FashionMNIST

We evaluate OAT on image classification using FashionMNIST. A total of 500 repetitions of sampling a training dataset of 250 points, training the main ResNet model, and adaptively selecting data points to labels in the stream of $5,000$ points have been performed using a budget constraint of $1,0000$. A small validation set of 50 points is used during training to perform early stopping. The main model is a 1D ResNet trained with cross-entropy loss using SGD (learning rate 0.1, momentum 0.9, weight decay 5×10^{-4}) for a total of 160 epochs and batch size of 64. Evaluation is performed under the log-loss using a Random Forest as a surrogate model, pre-trained on the training set and incrementally updated using adaptively selected evaluation samples. Due to computational limitations, surrogate models are updated at predefined milestones during the stream: after collection of $\{5, 10, 20, 50, 80, 110, 140, 170\}$ data points.

Both experiments are run on a GPU (Nvidia A40 48GB) in $\sim$ 12 hours each.

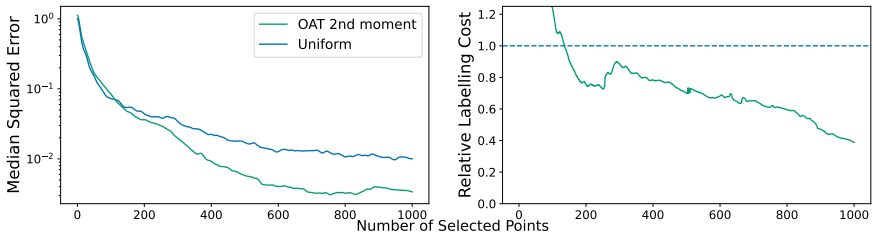


Figure 6: Online active testing on FashionMNIST. Left: Median squared error of the risk estimate (log-scale) versus number of labelled samples. Right: relative labelling cost versus uniform baseline (dashed).

Second-moment efficiency To isolate the benefit of the second moment, we sharpen the Random Forest surrogate predictions in the FashionMNIST experiment, which may correspond to an overconfident surrogate. Sharpening does not significantly reduce the estimation of tail probabilities relevant to the squared loss, while degrading the first-moment approximation. We report the mean squared estimation error in Figure 7 (Left). The mean squared estimation errors below show that second-moment sampling still significantly improves over uniform sampling, while first-moment sampling becomes significantly worse. At the end of the stream, second-moment sampling significantly outperforms uniform sampling (p=1.4e-24 according to a Student t-test), whereas first-moment sampling significantly underperforms it (p=1.1e-11).

Normalising constant ablation: We propose an additional experiment in which we intentionally "cheat" by computing the normalisation constant using the entire stream on the FashionMNIST dataset, allowing us to isolate its impact. We report the mean squared estimation error in Figure 7 (Right). For all selected time steps in the Figure, a Student's t-test shows no significant difference between the two estimators.

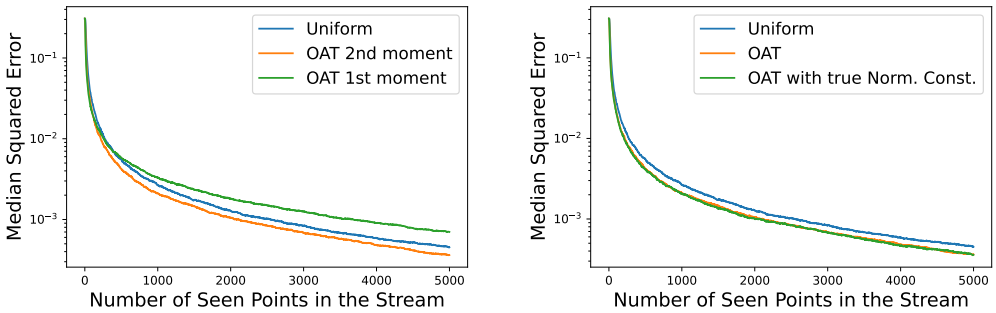


Figure 7: (Left) Demonstration of the efficiency of the second moment sampling strategy over the first-moment strategy in a scenario with an overconfident surrogate. (Right) Ablation study on the normalisation constant online estimation.

F.3 CellCount

We consider a regression task on microscopy images. This dataset is synthetically generated and consists of 64×64 greyscale images on which between one and 200 cells are counted. The training set of $200,000$ images is restricted to $20,000$ images, on which we train a simple CNN and evaluate this CNN with a stream of $2,000$ images with a labelling budget of 200. We repeat the evaluation 100 times. The main model is a small CNN trained with MSE loss for 30 epochs using Adam (learning rate 10^{-3}). The surrogate model is a Gaussian Process trained on a low-dimensional representation obtained from the extraction of the CNN features and the reduction to 15 dimensions using Principal Component Analysis. We use a composite kernel corresponding to a constant kernel multiplied by an RBF kernel, plus a white kernel. This kernel assumes the target function is smooth and

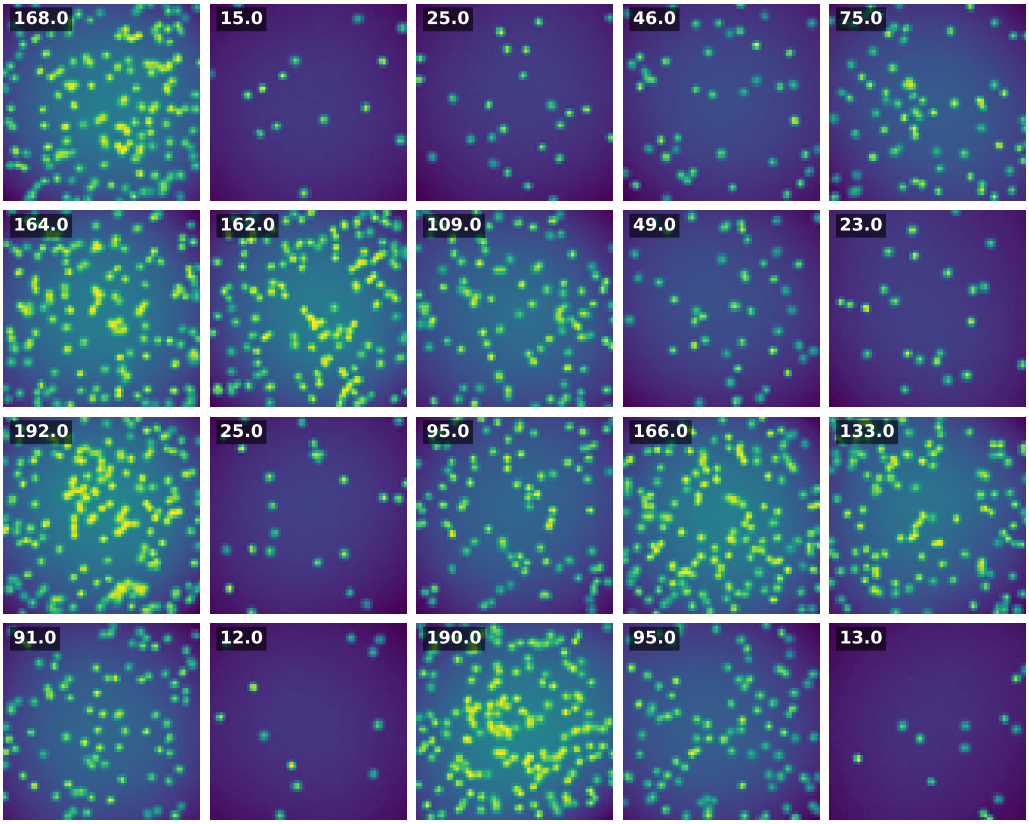


Figure 8: Example of the CellCount dataset.

varies continuously with the input, while allowing for independent noise. Due to computational limitations, surrogate models are updated at predefined milestones during the stream: after collection of $\{5, 10, 20, 30, 40, 50, 75, 100, 130, 160, 180\}$ data points.

This experiment is run on a GPU (NVIDIA Quadro T2000 4GB) in $\sim$ 2 hours.

Surrogate model ablation We compare our lightweight surrogate using latent representations with a surrogate trained directly on raw images on the CellCount dataset. The latent surrogate achieves lower estimation error while requiring approximately six times less training time. We report the mean squared estimation error in Table 2. Note that the surrogate trained on raw data is also 6 times longer to train. At the end of the stream, the raw-input surrogate is significantly worse than the one using latent representations (p=1.2e-3 according to a Student t-test).

Table 2: Performance Comparison across Stream Steps between a lightweight surrogate using latent representations of the model to evaluate and a surrogate using the raw input data points. Median squared error values $\times 10^5$.

Stream steps	500	1000	1500	2000
Latent representation	1.90	0.95	0.62	0.47
Raw input	2.38	1.13	0.82	0.58

F.4 Melts: molecular simulation

We evaluate OAT on a quantum chemistry task using molecular energy prediction. The training set is restricted to 20% of the $3,408$ geometries of the fulvene molecule (consisting of 12 atoms: 6 carbon, 6 hydrogen) dataset, on which we train a TorchANI (ANI-1x) model to predict molecular energies.

We evaluate this ANI model with a stream of $2,000$ data points with a labelling budget of 200. The main model is trained for a total of 500 epochs using MSE loss and Adam (learning rate 10^{-3}).

A key challenge in this setting is that molecular geometries are represented as 3D coordinates of atoms (of dimension $3 \times N_{\text{atoms}}$, with N_{atoms} the number of atoms in the molecule), which are not invariant to rotations, translations, or permutations of atoms. To address this, invariant representations are required. In our case, we leverage the atomic environment vectors (AEVs) provided by the ANI framework, which encode local atomic environments in a way that is invariant to these transformations.

The surrogate model is a Gaussian Process trained on atomic environment vectors extracted from the ANI model, averaged per molecule, resulting in a 384-dimensional representation. We also use a composite kernel corresponding to a constant kernel multiplied by an RBF kernel, plus a white kernel. This kernel assumes the target function is smooth and varies continuously with the input, while allowing for independent noise. Due to computational limitations, surrogate models are updated at predefined milestones during the stream: after collection of $\{5, 10, 20, 50, 80, 110, 140, 170\}$ data points.

This experiment was run on a GPU (NVIDIA Quadro T2000 4GB) in ~ 2 hours.

The dataset for this experiment is publicly available at the following link: `https://zenodo.org/records/16900694`.

F.5 Computational library and tools

- Python [61]
- Scikit-learn [50]
- Numpy [21]
- Pytorch [49]
- Pandas [44]
- TorchANI [19]
- Seaborn [62]
- Matplotlib [24]

G Comparison to Active Risk Estimation

We propose an adaptation of Active Risk Estimation (ARE, Sawade et al. 53) for online active testing, as ARE is originally designed for the pool-based setting. We first introduce ARE for the pool-based setting, then detail our online adaptation and empirically compare this version to our OAT framework. An important note is that ARE relies on a self-normalised importance sampling estimator; therefore, ARE is biased for finite samples, but it benefits from consistency guarantees and represents one of the strongest existing alternatives. However, an advantage of ARE is that it relies only on the original model to approximate $y|x$.

G.1 Pool-based ARE and online adaptation

In this subsection, we consider the pool-based setting. After selecting M data points according to the sampling policy β, ARE is based on a self-normalised importance sampling estimator given by:

$$\mathcal{R}_M^{ARE}(\theta) = \frac{1}{\sum_{i=1}^{M} \frac{1}{\beta(x_i)}} \sum_{i=1}^{M} \frac{1}{\beta(x_i)} L(\theta; x_i, y_i). \tag{14}$$

This estimator is not unbiased but benefits from consistency when sampling with replacement is allowed. Sawade et al. [53] derive an optimal sampling strategy in terms of deviation from the true $\mathcal{R}(\theta)$:

$$\beta(x) \propto \sqrt{\int \left[L(\theta; x, y) - \mathcal{R}(\theta)\right]^2 p(y|x) dy}.$$

The true risk is the unknown quantity to estimate during active testing; it is not accessible and should be approximated. They do so by computing the expected risk according to the model output distribution p_θ over the entire pool $\mathcal{D}_{pool}$:$\hat{\mathcal{R}}(\theta) = \frac{1}{M} \sum_{\mathcal{D}_{pool}} \sum_{y \in \mathcal{Y}} L(\theta; x, y) p_\theta(y|x)$.

Online adaptation of ARE We now consider the setting described in Section 3 and propose a self-normalised importance sampling estimator at a fixed horizon T:

$$\mathcal{R}_T^{ARE}(\theta) = \frac{1}{\sum_{t=1}^{T} \frac{r_t}{\beta(x_t)}} \sum_{t=1}^{T} \frac{r_t}{\beta(x_t)} L(\theta; x_t, y_t). \tag{15}$$

Note that ARE is a consistent estimator in the pool-based setting when sampling with replacement, which is not possible in our online setting.

Finally, the expected risk according to the model output distribution p_θ cannot be computed over the entire stream; therefore, it is updated online when new data points arrive in the same fashion as the normalising constant of OAT's sampling strategy.

G.2 Online setting: comparison between OAT and ARE

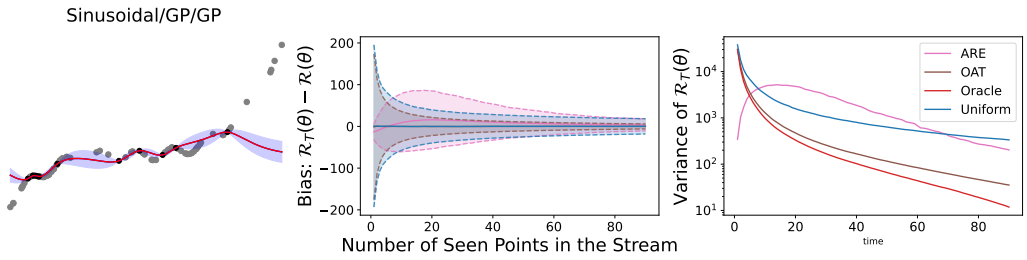


Figure 9: Online active testing on the Sinusoidal dataset compared to Active Risk Estimation [53]. Left: data distribution and model to evaluate. Bias (middle) and variance (right) of the risk estimation with respect to the number of seen points in the data stream for the MSE under different sampling strategies.

We propose to compare the online version of ARE and OAT on the sinusoidal dataset of Section 6 in a stream of 100 points and a budget of 10 samples, with 10,000 repetitions. ARE does not require surrogate models, so we decided to compare it with the non-adaptive OAT with second moment, using the model to evaluate as a surrogate. We report in Figure 9 the data distribution, the bias and the variance of the risk estimation for each method. We observe a bias for the ARE method at the beginning of the stream, while OAT is unbiased. ARE struggles to improve over uniform sampling in terms of variance, while OAT is always a lower-variance estimator. In terms of mean squared estimation error after seeing the entire stream, OAT significantly outperforms ARE at the end of the stream according to a Student's t-test (p=6.3e-63).

H Societal impact of the present research

As stated in the introduction, online active testing finds application in the continuous evaluation of embedded and continuously deployed systems such as autonomous vehicles and computer-aided diagnosis models. These systems operate in dynamic environments where data distributions evolve over time. Reliable and sample-efficient performance estimation can help detect model degradation or distribution shifts, thereby improving safety, reliability, and trustworthiness. In high-stakes applications, earlier detection of failures may reduce risks and potentially save lives.

In scientific domains, active learning is increasingly used, yet evaluation under limited labelling budgets remains challenging [6]. By enabling more reliable model evaluation, OAT can help these practitioners to better evaluate their models and accelerate scientific discovery with societal impact (negative or positive).

NeurIPS Paper Checklist

1. **Claims**

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

   Answer: [Yes]

   Justification: The abstract and introduction clearly state the paper's contributions and scope, and these claims are directly supported by the theoretical analysis in Sections 4 and 5 and empirical results presented in Section 6.

   Guidelines:

   - The answer [N/A] 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 [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

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

   Answer: [Yes]

   Justification: The paper includes a dedicated discussion of limitations in Section 7, covering assumptions and results of the experiments compared to theoretical theorems, as well as sensitivity to the choice of the surrogate model and data accessibility.

   Guidelines:

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

3. **Theory assumptions and proofs**

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

Answer: [Yes]

Justification: All theoretical results are stated with explicit assumptions, and complete proofs are provided in the appendix.

Guidelines:

- The answer [N/A] 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: The paper provides a detailed description of datasets, preprocessing, model architectures, training procedures, and evaluation protocols sufficient to reproduce the main results in Section F. The code is also available in the supplementary materials.

Guidelines:

- The answer [N/A] 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: All data used in the experimental part is open access. We provide the code in the supplementary materials.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.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://neurips.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) necessary to understand the results?

   Answer: [Yes]

   Justification: All relevant experimental details, including data splits, hyperparameters, optimiser settings, and model configurations, are specified in Appendix F.

   Guidelines:

   - The answer [N/A] 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: Results are reported over multiple runs with different random seeds, and we include mean and standard deviation (error bars) to reflect variability due to initialisation and data sampling (see Section 6).

   Guidelines:

   - The answer [N/A] 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: The paper specifies the hardware used (GPU/CPU type) and training time per experiment in Appendix F.

   Guidelines:

   - The answer [N/A] 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: The research adheres to the NeurIPS Code of Ethics.

   Guidelines:

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

10. **Broader impacts**

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

    Answer: [Yes]

Justification: Societal impact of the present work has been discussed in Section H of the Appendix. Being suited for the evaluation of deployed systems such as autonomous vehicles, OAT has a potentially positive societal impact in terms of improving the safety, reliability, and trustworthiness of these systems.

Guidelines:

- The answer [N/A] means that there is no societal impact of the work performed.
- If the authors answer [N/A] 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., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: The work does not involve releasing high-risk models or sensitive datasets, and therefore, no specific safeguards are required.

    Guidelines:

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

12. **Licenses for existing assets**

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

    Answer: [Yes]

    Justification: All datasets, codebases, and models used are properly cited, and their licenses and terms of use are respected.

    Guidelines:

- The answer [N/A] 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: [N/A]

    Justification: The paper does not introduce new datasets or publicly released models requiring additional documentation.

    Guidelines:

    - The answer [N/A] 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: [N/A]

    Justification: The study does not involve human subjects or crowdsourced data.

    Guidelines:

    - The answer [N/A] 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: [N/A]

Justification: The research does not involve human subjects.

Guidelines:

- The answer [N/A] 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 rigor, or originality of the research, declaration is not required.

Answer: [N/A]

Justification: LLMs are not used as part of the core methodology of this work.

Guidelines:

- The answer [N/A] 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 in the NeurIPS handbook for what should or should not be described.