
Online Active Testing for Budget-Efficient Evaluation of Molecular Potentials

Hugo Schmutz
LIS UMR 7020 CNRS
Aix-Marseille University
13288 Marseille, France
hugo.schmutz@lis-lab.fr

Hachem Kadri
LIS UMR 7020 CNRS
Aix-Marseille University
13288 Marseille, France
hachem.kadri@lis-lab.fr

Thierry Artières
LIS UMR 7020 CNRS
École Centrale Méditerranée
13013 Marseille, France
thierry.artieres@lis-lab.fr

Abstract

Reliable evaluation of machine learning models in quantum chemistry can require a large number of expensive reference calculations. This challenge is particularly acute when molecular configurations are generated sequentially: each new configuration requires an immediate decision on whether to perform the quantum-mechanical evaluation, which can be orders of magnitude more expensive than generating the configuration itself. While active learning methods address this computational bottleneck during model training, existing active testing methods largely assume access to a pool of unlabelled samples and are therefore not directly suited to sequentially generated configurations. We introduce *Online Active Testing* (OAT) for estimating model risk when data arrive sequentially, and reference calculations are available only under a limited budget. OAT combines an unbiased importance-weighted estimator with a lightweight surrogate that guides evaluation decisions online. We derive the variance-optimal sampling rule and evaluate OAT on a molecular simulation task involving expensive quantum-chemistry calculations. On this task, OAT achieves the same estimation accuracy as uniform sampling using only $\sim 30\%$ of the reference calculations, compared with $\sim 23\%$ for an oracle strategy with access to the true loss. These results demonstrate that OAT can substantially reduce the computational cost of reliable model evaluation.

1 Introduction

In many scientific machine learning applications, obtaining training and evaluation data is expensive because targets must be computed using costly physical simulations. In molecular modelling, for example, obtaining accurate energies, forces, or other quantum-mechanical properties may require expensive electronic-structure calculations such as density functional theory [3]. When only a limited number of such simulations can be afforded, a natural question is which samples should be selected for reliable evaluation under a limited budget. This question is closely related to active learning, which selectively acquires labelled samples to improve a model. Active learning has been extensively studied and is commonly used to explore relevant regions of a physical or configurational space during the training phase. However, a related problem arises when the model is already trained, and

the goal is instead to estimate its performance. To this end, active testing selects a subset of samples to label in order to estimate a model’s risk accurately under a limited budget [21, 13, 14].

Prior active testing methods mainly consider a pool-based setting, where a collection of unlabelled samples is available and can be inspected before selecting which ones to evaluate. In many scientific applications, however, the expensive target must be acquired while the data are being generated. This situation arises naturally in high-throughput experiments and simulations, where observations arrive sequentially and acquiring their physical labels can be orders of magnitude more expensive than generating or storing the observations themselves. Molecular dynamics (MD) provides a natural example: configurations are generated sequentially along trajectories, and one may decide at predefined intervals whether to perform an expensive electronic-structure calculation to obtain a reference energy or force, or to trigger a correction of an ML-generated trajectory. Although configurations can be stored, performing such calculations for every configuration may be prohibitively costly. The relevant question is therefore how to allocate a limited budget of expensive evaluations online, using the information gathered from previous evaluations without requiring access to the labels of future configurations. In other high-volume, high-throughput scientific applications, even storing all observations may be impractical. For example, particle-collider experiments produce enormous numbers of collision events, of which only a small fraction can be retained for detailed analysis [12]. In such settings, the decision to acquire an expensive target must be made immediately for each incoming sample, without access to future observations. These applications motivate active testing methods that operate directly on a stream rather than relying on a fixed pool of unlabelled samples.

Main contributions We introduce *Online Active Testing (OAT)*, a framework for unbiased and sample-efficient model evaluation when data arrive sequentially, and expensive labels must be acquired online under a limited budget. OAT uses a lightweight surrogate to predict the loss of each incoming sample and adaptively determines whether to perform its expensive evaluation. Importance weighting [17] corrects for the resulting sampling bias, while the surrogate can be updated online from previously acquired evaluations. The surrogate is deliberately lightweight, making OAT suitable for high-throughput scientific simulation. We derive the corresponding anticipated variance-optimal sampling rule, which allocates evaluations according to the square root of the conditional second moment of the loss, and establish theoretical guarantees for the resulting estimator. We evaluate OAT on several benchmarks and focus in particular on a molecular simulation task involving expensive quantum-chemistry reference calculations. In this setting, OAT achieves the estimation accuracy of uniform sampling using only $\sim 30\%$ of the reference calculations, compared with $\sim 23\%$ for an oracle strategy with access to the true loss. These results demonstrate that OAT with lightweight surrogate-guided sampling can substantially reduce the computational cost of reliable model evaluation in scientific machine learning applications. We further show that OAT has substantially lower computational complexity than classical active testing strategies (see Appendix E), making it particularly suitable for high-volume, high-throughput scientific applications. These results demonstrate that lightweight surrogate-guided online sampling can substantially reduce the computational cost of reliable model evaluation.

2 Online active testing setting via adaptive importance sampling

We consider the problem of evaluating a fixed machine learning model when test data are generated sequentially and obtaining ground-truth targets is expensive. In scientific applications, x_t can represent a newly generated molecular configuration and y_t its expensive label or simulation output.

Risk estimation Let $(x, y) \sim p(x, y)$ denote the target data-generating distribution and let f_θ be a fixed model. We aim to estimate its risk $\mathcal{R}(\theta) = \mathbb{E}_{(x,y) \sim p}[L(\theta; x, y)]$ using as few expensive evaluations of y as possible. The model is fixed during testing and can be arbitrary: no parametric assumptions are made about f_θ . The framework applies to standard evaluation metrics such as mean squared error, mean absolute error, classification error, or log-loss.

Sequential data generation and sampling We observe an unlabelled sequence $(x_t)_{t \geq 1}$ drawn i.i.d. from the marginal $p(x)$. For each x_t , a target $y_t \sim p(y | x_t)$ exists but is initially unobserved. At time t , the learner must immediately decide whether to query y_t . We denote this decision by $r_t \sim \mathcal{B}(\beta_t)$, where $\beta_t := \beta(x_t) \in [0, 1]$ is the probability of querying y_t . If $r_t = 1$, the target and consequently the loss $\ell_t = L(\theta; x_t, y_t)$ are observed. Otherwise, the candidate is discarded and cannot be queried later. This captures the practical constraint of scientific simulation pipelines, where an expensive

computation must be launched when a configuration is encountered rather than after collecting a pool of candidates.

Adaptive sampling The sampling probability β_t may depend on the current input x_t and all information observed before time t . In particular, it can use the model prediction and a lightweight surrogate trained from previously acquired labels, $\beta_t = \beta(x_t, \mathcal{F}_{t-1})$, where $\mathcal{F}_{t-1}$ contains the observations and queried labels collected before t . Consequently, the queried samples are not distributed according to $p(x)$, and naive averaging of their losses is biased.

Unbiased risk estimate via adaptive importance sampling

The role of OAT is to exploit this adaptive sampling while preserving unbiasedness through importance weighting. Given a stream of T candidates, we use the estimator

$$\mathcal{R}_T(\theta) = \frac{1}{T} \sum_{t=1}^T \frac{r_t}{\beta_t} \ell_t. \quad (1)$$

This estimator is unbiased, $\mathbb{E}[\mathcal{R}_T(\theta)] = \mathcal{R}(\theta)$ (see Appendix A), provided $\beta_t > 0$ whenever ℓ_t can be non-zero [17, 7]. Using similar arguments as Beygelzimer et al. [2], the consistency of the estimator can be established in the stream setting (see Appendix B). The challenge is therefore to choose β_t so as to minimise the estimator variance while respecting a limited computational budget.

Budget and stopping criterion. We consider a stream of T generated candidates and constrain the expected number of expensive evaluations, $\mathbb{E}[n_T] = \sum_{t=1}^T \mathbb{E}[r_t] = \sum_{t=1}^T \beta_t \leq n_B$. This stopping criterion is preferable to stopping after exactly n_B . Indeed, in the latter, 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.

3 Variance-optimal adaptive sampling

The central question is how to allocate a limited budget across a stream of generated configurations. We therefore seek sampling probabilities β_t that minimise the uncertainty of the risk estimator for a fixed expected number of expensive evaluations. Conditioning on the generated inputs $\mathcal{D}_T = \{x_1, \dots, x_T\}$, we consider minimising the *anticipated variance* [11], i.e., the expected variance of the estimator with respect to the unknown targets. This isolates the uncertainty arising from the expensive evaluations and leads to a tractable allocation problem. Under the budget constraint, we prove in Appendix C that this derivation leads to

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

This 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 C. Simply clipping Equation 2 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. The optimal sampling strategy of Equation 2 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 pool-based work [5, 13]. The second moment decomposes as $\mathbb{E}_{y|x}[\ell_t^2] = V_{y|x}[\ell_t] + \mathbb{E}_{y|x}[\ell_t]^2$ revealing that the optimal strategy favours configurations that are either expected to be difficult for the model or for which the model’s error is particularly uncertain. This is especially relevant for scientific simulations, where different regions of configuration space can exhibit substantially different model errors and intrinsic uncertainty.

Baseline comparison 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 $\beta_t = n_B/T \cdot l_t(\theta)/\mathbb{E}[l_t(\theta)]$, known to be optimal in classical importance sampling [6, 5, 17] (but requires knowing y_t). The benefit of adaptive sampling can be quantified by comparing it with uniform evaluation. Let n_u be the number of labels required by uniform sampling to attain the same anticipated variance as the optimal strategy with budget n_B . Then, $n_B = n_u \cdot \rho_T$, with $\rho_T \leq 1$ (by

Algorithm 1 Online Active Testing

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

```
1: for  $t = 1$  to  $T$  do
2:   Observe  $x_t$ 
3:   Compute  $\beta_t = \beta(x_t)$  using  $f_\theta$  and  $\pi$ 
4:   Sample  $r_t \sim \text{Bernoulli}(\beta_t)$ 
5:   if  $r_t = 1$  then
6:     Query  $y_t$  and compute  $L(\theta; x_t, y_t)$ 
7:     Update  $\pi$  using  $(x_t, y_t)$ 
8:   end if
9: end for
10: return  $\mathcal{R}_T$ 
```

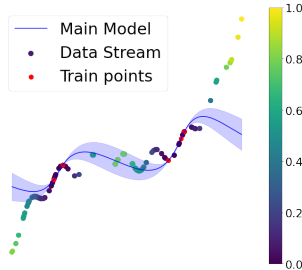


Figure 1: Second-moment active selection. Red points are used to train the Gaussian-process surrogate; colours indicate the resulting selection probability β_t^* .

Cauchy-Schwarz), with equality when the conditional second moments are constant across the stream. Thus, whenever the difficulty of the generated configurations is heterogeneous, adaptive evaluation can achieve the same estimation accuracy with fewer expensive simulations than uniform sampling.

4 Practical implementation: surrogate and normalisation

The constrained online settings and this optimal sampling strategy lead to practical considerations: a surrogate is needed to estimate the second moment of the loss, which we designed to be lightweight to support high-throughput data streams, and a normalisation constant must be updated.

Surrogate estimation for online adaptive sampling The optimal strategy in Equation 2 cannot be applied directly because the conditional second moment $\mathbb{E}_{y|x}[\ell_t(\theta)^2]$ is unknown. We approximate it with a surrogate model, π . Given a new configuration x_t , the surrogate provides a cheap estimate $\beta_t \propto n_B/T \cdot \sqrt{\mathbb{E}_{y_t|x_t \sim \pi}[\ell_t(\theta)^2]}$. This surrogate is adaptive: first initialised with the training and validation data and then retrained using all the labels acquired before time t .

Lightweight surrogate for high-throughput streams Kossen et al. [13] show that diversity (using a surrogate different from the evaluated model) and fidelity are key factors for sample efficiency. Building upon these observations, our goal is to design a lightweight surrogate suitable for high-throughput streams. We achieve diversity by using a different probabilistic model family (e.g., Random Forest or Gaussian Process) and fidelity by leveraging the latent representations of the evaluated model (e.g. Atomic Environment Vector of ANI models, Smith et al. 23). Only a single forward pass through the model is required and avoids the cost of additional large models or expensive retraining.

Normalisation constant estimation Compared to the pool-based case, the normalisation constant cannot be computed exactly; we approximate it online. In practice, to alleviate the cold-start issue, the validation set can be leveraged to initialise this estimate. 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. This also reduces computation cost.

We propose in Algorithm 1 the pseudocode of the method and in Figure 1 an illustration of the selection method, which favours samples that are either expected to be difficult for the model or for which the model’s error is particularly uncertain.

5 Experiments: application to molecular dynamics

We evaluate OAT on a realistic quantum-chemistry application arising in MD simulations. In MD, molecular configurations evolve sequentially along a trajectory, and machine-learning potentials can be used to predict molecular energies and forces at a much lower computational cost than electronic-structure calculations. However, obtaining accurate reference energies and forces, for example from density functional theory, remains expensive. A practical strategy is therefore to perform these quantum-chemistry calculations only at selected configurations along the trajectory, making the

allocation of the limited number of expensive evaluations a natural active testing problem. We consider the MELTS Fulvene dataset [3], which contains molecular geometries and their associated quantum-chemistry energies. The dataset is publicly available here. We train an ANI model [23] on 20% of the dataset and evaluate its performance on sequential streams of molecular configurations using a labelling budget of 200. Our surrogate is a Gaussian Process trained on the atomic environment vector (AEV) representations extracted from the ANI model. Results over 100 independent evaluation streams are shown in Figure 2. We report the median squared error of the risk estimate as a function of the number of observed configurations, which characterises typical performance while reducing the influence of occasional streams with unusually large errors (comparison with the mean is shown in Appendix D). We additionally report the relative labelling cost, defined as the fraction of labels required by OAT to reach the performance of uniform sampling [13].

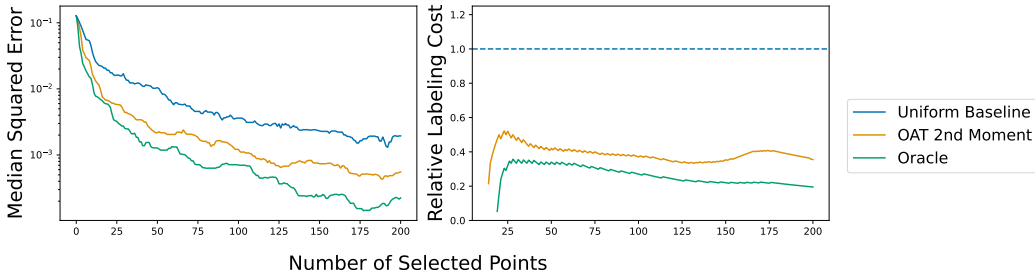


Figure 2: Online active testing for MD on the MELTS dataset using a GP surrogate to evaluate an ANI model. Left: Median squared error of the risk estimate (log scale) versus number of labelled samples. Right: relative labelling cost versus uniform baseline (dashed).

Additional experiments proving the statistical efficiency and variance-reduction properties of OAT in controlled settings, as well as its scalability to deep-learning models on image datasets, are provided in Appendix D, together with complete experimental details, computational resources, and runtime.

6 Limitations and Conclusion

Our experiments show that adaptive evaluation can substantially reduce the number of expensive simulations required to estimate model performance compared with uniform sampling. The remaining gap to the oracle strategy, which has access to the true loss, illustrates the importance of the surrogate used to guide sampling. Nevertheless, the lightweight surrogates used by OAT consistently recover a substantial fraction of the oracle benefit. When candidate configurations can be generated cheaply but high-fidelity targets require costly computations, the available computational budget need not be allocated accordingly. OAT can identify configurations that are likely to contribute most to the uncertainty of model evaluation, while importance weighting preserves an unbiased estimate of the risk. In this sense, OAT provides a building block for scientific workflows in which simulation and machine learning interact continuously. However, our sampling strategy targets a single evaluation metric for a single model and therefore does not directly optimise multiple objectives or allow model selection. An important extension is to couple OAT with adaptive data generation, where the simulation or configuration-generation process itself is influenced by previously acquired evaluations rather than producing a fixed stream. Handling non-stationary streams or distribution shifts is a natural future direction [22, 15].

Acknowledgments and Disclosure of Funding

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

We additionally thank Mario Barbatti and Matheus de Oliveira Bispo 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 LIS cluster operations team for maintaining the computing resources used in this study.

References

- [1] Kazuoki Azuma. Weighted sums of certain dependent random variables. *Tohoku Mathematical Journal, Second Series*, 19(3):357–367, 1967.
- [2] 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.
- [3] 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.
- [4] Leo Breiman. Random forests. *Machine Learning*, 45(1):5–32, 2001.
- [5] 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.
- [6] 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.
- [7] Ravi Ganti and Alexander Gray. Upal: Unbiased pool based active learning. In *Artificial Intelligence and Statistics*, pages 422–431. PMLR, 2012.
- [8] 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.
- [9] 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.
- [10] J. D. Hunter. Matplotlib: A 2D graphics environment. *Computing in Science & Engineering*, 9(3):90–95, 2007. doi: 10.1109/MCSE.2007.55.
- [11] 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.
- [12] Bruce Knuteson and Paul Padley. Statistical challenges with massive datasets in particle physics. *Journal of Computational and Graphical Statistics*, 12(4):808–828, 2003.
- [13] Jannik Kossen, Sebastian Farquhar, Yarin Gal, and Tom Rainforth. Active testing: Sample-efficient model evaluation. In *International Conference on Machine Learning*, pages 5753–5763. PMLR, 2021.
- [14] 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.
- [15] 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.
- [16] Wes McKinney et al. Data structures for statistical computing in python. *Scipy*, 445(1):51–56, 2010.
- [17] Art B. Owen. *Monte Carlo theory, methods and examples*. <https://artowen.su.domains/mc/>, 2013.
- [18] 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.

- [19] 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.
- [20] Carl Edward Rasmussen and Zoubin Ghahramani. Bayesian monte carlo. *Advances in Neural Information Processing Systems*, pages 505–512, 2003.
- [21] 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.
- [22] 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.
- [23] 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.
- [24] Guido Van Rossum and Fred L Drake Jr. *Python tutorial*, volume 620. Centrum voor Wiskunde en Informatica Amsterdam, The Netherlands, 1995.
- [25] 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>.

A Unbiasedness of the estimator

We begin by establishing the unbiasedness of the proposed estimator, which is a classical property of adaptive importance sampling estimators [17, 7], which we recall here for completeness. 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 $t \in \llbracket 1, T \rrbracket$, 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. 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 [17].

Proof. The proof is similar to the [7]. 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 $t \in \llbracket 1, T \rrbracket$. First, we remark that $\forall t \in \llbracket 1, T \rrbracket$, 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 $\mathcal{F}_{t-1}$ -measurable, thus, $\forall t \in \llbracket 1, T \rrbracket$, we can compute the expectation:

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

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

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

□

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.*

B Deviation and consistency

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 $t \in \llbracket 1, T \rrbracket$. Then, for any $\delta \in (0, 1)$,*

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

This result provides a probability deviation bound, implying that $\mathcal{R}_T(\theta)$ converges to $\mathcal{R}(\theta)$ as $T \rightarrow \infty$. 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.

Proof. Let’s consider the sequence of random variable $M_1, \dots, M_T$ defined for all $t \in \llbracket 1, T \rrbracket$ 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 $Z_t = \sum_{i=1}^t M_i$ (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

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

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 [1]:

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

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

C Optimal sampling strategy

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* [11], 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 $\mathcal{D}_T = \{x_1, \dots, x_T\}$. This quantity measures the expected uncertainty of the estimator due to label randomness, given a fixed stream of inputs. In particular, we show 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 3 (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*

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

Proof. We recall that we denote $\mathcal{D}_T = \{x_1, \dots, x_T\}$ as the set of all samples in the stream up to time T . Using the variance definition, we can show that

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

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

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

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:

$$\begin{aligned} \min_{\beta} \quad & \mathbb{E}_{y|x}[V_{\mathbf{r}}[\mathcal{R}_T(\theta)|\mathcal{D}_T]] \\ \text{s.t.} \quad & \sum_{t=1}^T \beta_t = n_B. \end{aligned} \quad (4)$$

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

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

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:

$$\Lambda(\beta, \lambda) = 0 \quad \Leftrightarrow \quad \begin{cases} \sum_{t=1}^T \beta_t = n_B \\ \frac{\mathbb{E}_{y_t|x_t}[\ell_t(\theta)^2]}{\beta_t^2} = \lambda \end{cases}. \quad (5)$$

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 second moment of the loss in $y \mid x$. $\square$

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:

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

On the other hand:

$$\mathbb{E}_{y|x}[V_{\mathbf{r}}[\mathcal{R}_T(\theta; \beta_u)|\mathcal{D}_T]] = \frac{1}{n_B T} \sum_{t=1}^T \mathbb{E}_{y|x}[\ell_t(\theta)^2] - \frac{1}{T^2} \sum_{t=1}^T \mathbb{E}_{y|x}[\ell_t(\theta)^2]. \quad (6)$$

Using the Cauchy-Schwarz inequality, we get that:

$$\mathbb{E}_{y|x}[V[\mathcal{R}_T(\theta; \beta^*)|\mathcal{D}_T]] \leq \mathbb{E}_{y|x}[V[\mathcal{R}_T(\theta; \beta_u)|\mathcal{D}_T]]. \quad (7)$$

The equality case of Cauchy-Schwarz yields to $\beta_T = C \beta_u$, 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. $\square$

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 $\sqrt{E_{y|x}[\ell_t(\theta)^2]} \propto E_{y|x}[\ell_t(\theta)]$. $\square$

C.1 Cleaner version of the Theorem

For simplicity, Theorem 3 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 4. *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 $\beta_t^*(x_t) = \min(1, c\sqrt{\mathbb{E}_{y|x}[\ell_t(\theta)^2]})$, with $c > 0$ the unique constant satisfying $\sum_t \min(1, c\sqrt{\mathbb{E}_{y|x}[\ell_t(\theta)^2]}) = n_B$.*

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 3:

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

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:

$$\Lambda(\beta, \lambda) = 0 \Leftrightarrow \begin{cases} \sum_{t=1}^T \beta_t = n_B \\ \frac{\mathbb{E}_{y_t|x_t}[\ell_t(\theta)^2]}{\beta_t^2} = \lambda + \mu_t \\ \mu_t (\beta_t - 1) = 0 \end{cases} . \quad (8)$$

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 $\sum_t \min(1, c\sqrt{\mathbb{E}_{y|x}[\ell_t(\theta)^2]}) = n_B$. □

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:

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

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 3 may lead to underspending the labels budget n_B as 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.

D Details on the experiments

We evaluate Online Active Testing (OAT) on synthetic and real-world datasets for both regression and classification, using both the 1st-moment (defined as $\beta_t \propto n_B/T \cdot \mathbb{E}_{y_t|x_t \sim \pi}[\ell_t(\theta)]$ [13]) and 2nd-moment (Equation 2) strategies, across a range of settings and examine the comparative advantages of different strategies. Performance is measured using the mean squared error or variance of the risk estimate, averaged over multiple independent runs.

D.1 Synthetic data:

We prove the unbiasedness and the variance reduction on toy examples (as in previous work [13]). We begin by demonstrating that online active testing on synthetic datasets enables sample-efficient evaluation of models. As illustrative examples, we actively evaluate a Gaussian process [20] and a linear regression model, as well as a random forest [4] for classification. In the regression setting, we estimate the squared error, while for classification, we estimate the cross-entropy loss and accuracy. The Gaussian process and random forest surrogates are retrained adaptively on all observed data. Figure 3 shows the performance of OAT with the 1st and 2nd moment sampling strategy compared to the uniform baseline and the oracle. OAT produces tighter confidence bands compared to uniform sampling in all settings (reflecting reduced variance) and approaches the oracle performance.

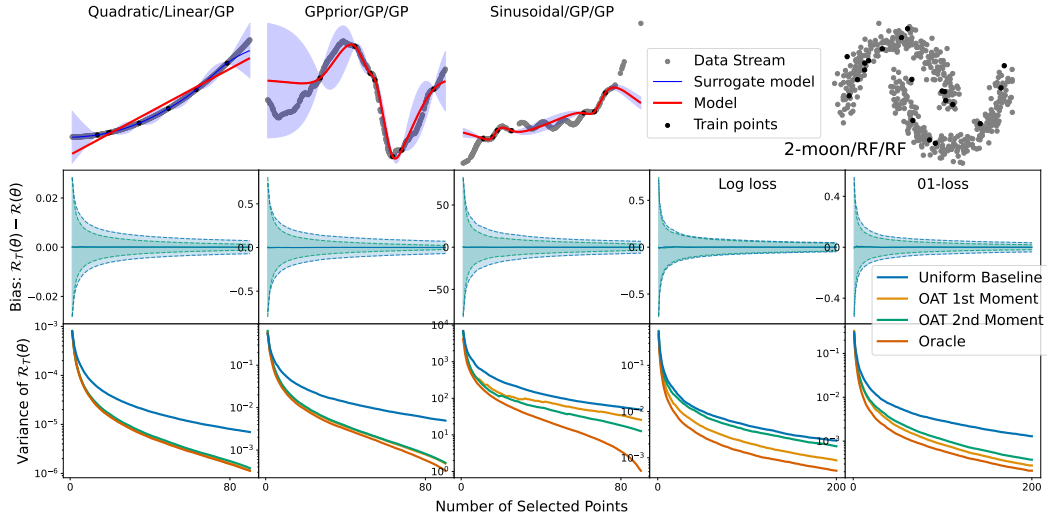


Figure 3: Online Active Testing improves risk estimation across diverse synthetic settings. Each column shows a different combination of data structure/model/surrogate. Top row: regression and classification scenarios. Bias (middle) and variance (bottom) of the risk estimation with respect to the number of seen points in the data stream under different loss functions (MSE for regression and log-loss and 0–1 loss for classification) and different sampling strategies.

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 ~ 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 cross-entropy 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 ~ 10 hours each.

D.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 1,000. 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 a 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 $\{5, 10, 20, 50, 80, 110, 140, 170\}$ datapoints.

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

The median squared error of the risk estimates for the different strategies (smoothed using a Gaussian kernel with a length scale of 10) and the relative labelling cost (the proportion of actively selected samples needed to get the same performance as a baseline [13]) of OAT compared to the uniform baseline are shown in Figure 4.

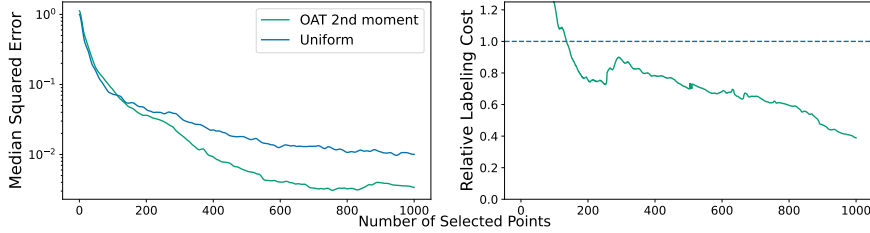


Figure 4: Online active testing on FashionMNIST. Top: MSE of the risk estimate (log-scale) versus number of labelled samples. Bottom: relative labelling cost versus uniform baseline (dashed). Left: log-loss; right: 0–1 loss.

D.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 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 ~ 2 hours.

We show sample of the Cell Count dataset in Figure 5 and the median squared error of the risk estimates for the different strategies (smoothed using a Gaussian kernel with a length scale of 10)

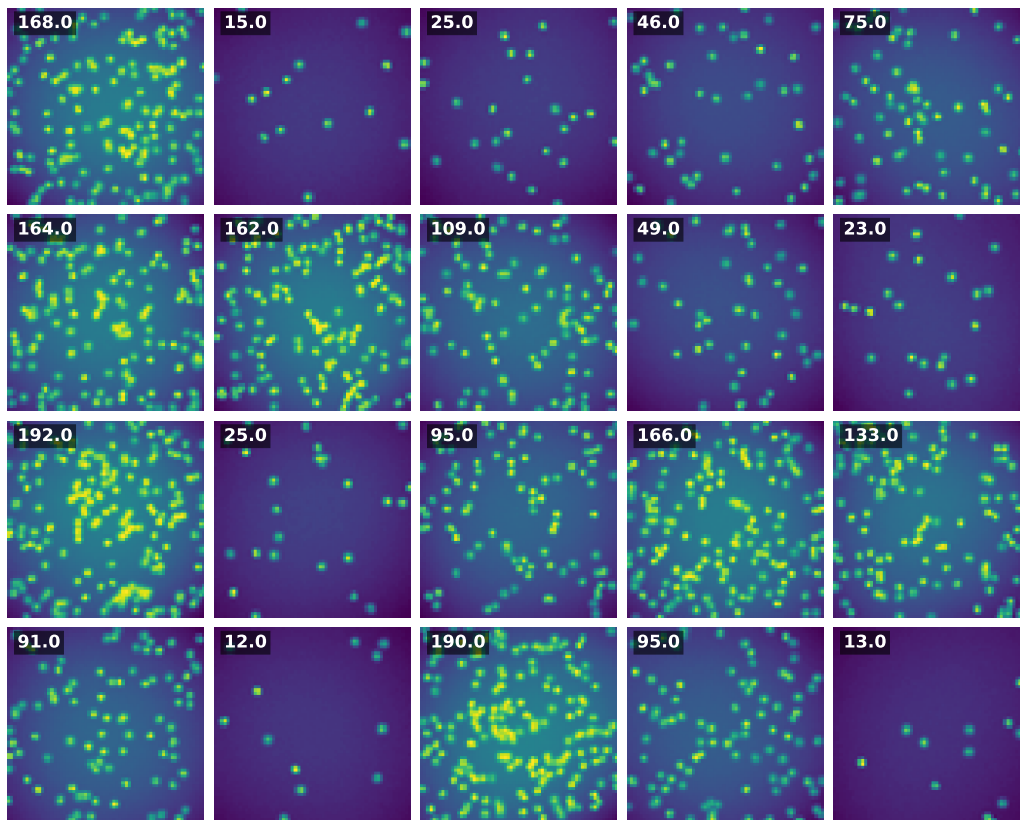


Figure 5: Example of the CellCount dataset.

and the relative labelling cost (the proportion of actively selected samples needed to get the same performance as a baseline [13]) of OAT compared to the uniform baseline are shown in Figure 6.

D.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 configurations 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 collecting $\{5, 10, 20, 50, 80, 110, 140, 170\}$ datapoints.

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

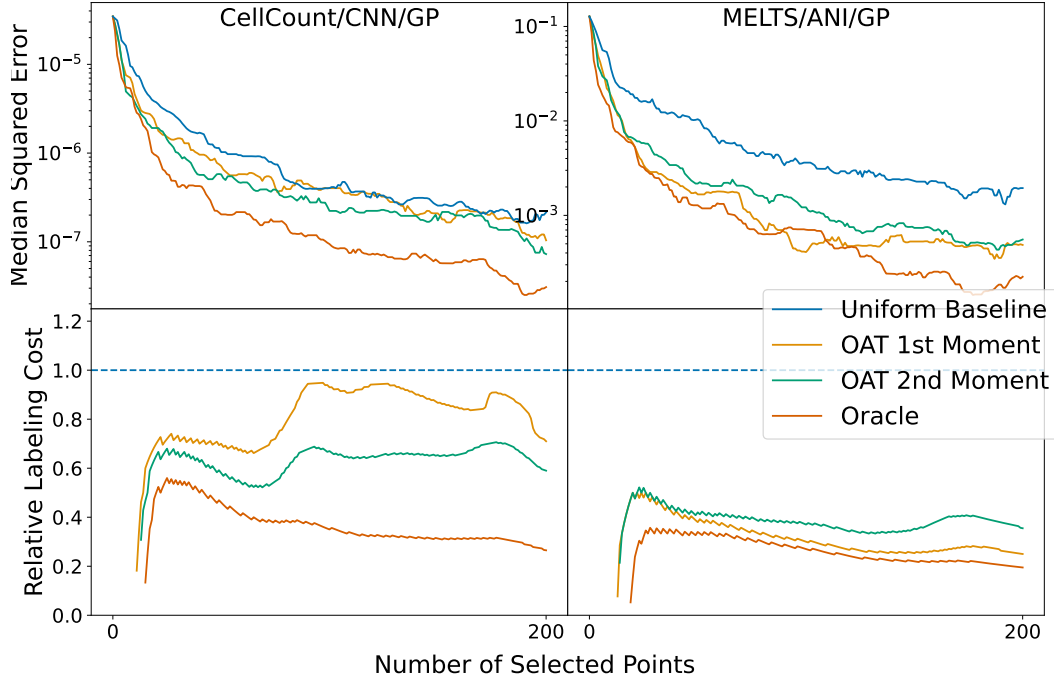


Figure 6: Online active testing for regression. Top: MSE of the risk estimate (log-scale) versus number of labelled samples. Bottom: relative labelling cost versus uniform baseline (dashed). Left: CellCount (CNN and GP surrogate). Right: MELTS energy prediction (ANI and GP surrogate).

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

The median squared error of the risk estimates for the different strategies (smoothed using a Gaussian kernel with a length scale of 10) and the relative labelling cost (the proportion of actively selected samples needed to get the same performance as a baseline [13]) of OAT compared to the uniform baseline are shown in Figure 6.

Comparison between median and mean squared error for risk estimation We chose to report the median squared error of the risk estimate as a function of the number of observed configurations, which characterises typical performance while reducing the influence of occasional streams with unusually large errors. To assess the robustness of this choice, we compare the results obtained using the median and mean squared error on the MELTS dataset. Figure 7 shows that the overall trends are consistent across the two metrics. Moreover, when using the mean squared error, the OAT methods appear to be even closer to the oracle, further highlighting their effectiveness in approaching the performance of the optimal sampling strategy.

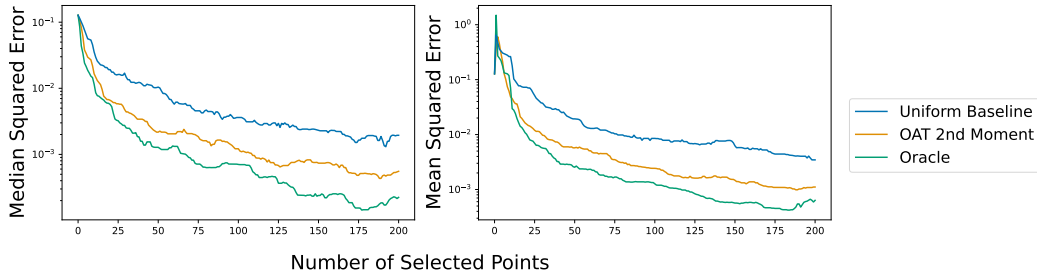


Figure 7: Comparison between median (Left) and mean (Right) squared error for the risk estimator with respect to the number of selected samples.

D.5 Computational library and tools

- Python [24]
- Scikit-learn [19]
- Numpy [9]
- Pytorch [18]
- Pandas [16]
- TorchANI [8]
- Seaborn [25]
- Matplotlib [10]

E Computational complexity

Computational efficiency is critical in online testing. In fact, the acquisition probability is lightweight: approximating the second-moment probability requires only evaluating the surrogate on the current sample and updating the running estimate of the normalising constant. No optimisation problem is solved online. 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 [13], 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