

# Prediction–Powered Active Testing

Kianoosh Ashouritaklimi<sup>1</sup>, Valentin Kilian<sup>1</sup>, Daolang Huang<sup>2</sup>, Tom Rainforth<sup>1</sup>, François Caron<sup>1</sup>

<sup>1</sup>Department of Statistics, University of Oxford, <sup>2</sup>Department of Computer Science, Aalto University.  
Correspondence to: ashouritaklimi@stats.ox.ac.uk

**Abstract** Active testing provides a label–efficient approach to risk estimation by adaptively selecting which test points should be labelled. However, existing estimators fail to exploit the informative predictions of powerful black–box models, even though such predictions are increasingly available in settings where labels remain expensive. To address this, we propose **Prediction–Powered Active Testing (PPAT)**, a novel label–efficient risk estimation framework that combines the unbiased LURE estimator (Farquhar et al., 2021) with a prediction–powered control variate. Rather than using proxy predictions as biased pseudo–labels, PPAT uses them to residualise the loss, preserving unbiasedness while reducing variance. Beyond the estimator itself, PPAT also changes which points should be acquired: we derive oracle and practical surrogate–based acquisition rules tailored to reducing the variance of our estimator. Moreover, we establish asymptotic normality for PPAT, yielding asymptotically valid confidence intervals and thus a principled estimate of the uncertainty around our estimates. Across tabular regression and image–classification tasks, PPAT outperforms existing methods in risk estimation, while its confidence intervals attain the target coverage with substantially fewer labels and smaller widths.

## 1 Introduction

Active learning (Settles, 2009; Liu et al., 2022; Atlas et al., 1989; Angluin, 1988) reduces the cost of training, but in many applications the cost of *evaluation* is just as important. In many domains, large unlabelled test pools are often easy to obtain, while ground–truth labels remain expensive and require expert annotation. Active testing (Sawade et al., 2010a,b; Yilmaz et al., 2021; Katariya et al., 2012; Yu et al., 2023; Kossen et al., 2021, 2022; Berrada et al., 2025) addresses this bottleneck by adaptively sampling which test points to label while correcting for the bias introduced by adaptive sampling. Crucially, adaptive sampling can achieve lower variance than random sampling under suitable acquisition strategies (Farquhar et al., 2021), while remaining unbiased.

In many settings, such as image classification (Radford et al., 2021; Li et al., 2023; Jia et al., 2021; Zhai et al., 2023), language modelling (Brown et al., 2020; Wang et al., 2021; Touvron et al., 2023; Team et al., 2023), and biomolecular structure prediction (Jumper et al., 2021; Abramson et al., 2024; Corso et al., 2022; Prat et al., 2025), we also have access to a powerful black–box predictor that can produce cheap predictions on the entire test pool. A natural question here is: *can we leverage these predictions to produce more accurate, label–efficient estimates of our risk?* One approach would be to replace the true, unknown labels with the predictions of the black–box model, similar to the work of Kossen et al. (2022). This, however, generally results in biased estimates of the risk. While it can yield accurate estimates of the risk when the labelling budget is very small and the black–box predictor is already highly accurate, the resulting estimator is generally biased, with the bias becoming increasingly problematic as the labelling budget grows.

To address this, we propose **Prediction–Powered Active Testing (PPAT)**, which combines the unbiased importance–weighted LURE estimator (Farquhar et al., 2021) with a prediction–powered control variate built from cheap black–box predictions (Angelopoulos et al., 2023a). Rather than using these predictions as biased pseudo–labels, PPAT uses them to *residualise* the loss: it subtracts, from each true loss, the loss induced by the black–box prediction after centring it by its average over the test pool, thereby preserving the unbiasedness of

active testing. Our approach introduces a tuning parameter  $\lambda$  that controls the strength of this correction and recovers LURE as the special case  $\lambda = 0$ , making PPAT a principled extension of the LURE estimator. We show that this control-variate formulation can reduce variance relative to LURE and discuss practical choices of  $\lambda$ .

Beyond the estimator itself, PPAT also changes what should be sampled. Rather than reusing active-testing proposals designed for the raw loss, we derive proposals tailored to the residualised objective induced by our estimator. This leads to an oracle proposal and a practical surrogate acquisition rule that samples points expected to be most informative for reducing the variance of the resulting risk estimate. We also establish asymptotic normality for the PPAT estimator, which allows us to construct asymptotically valid confidence intervals. Empirically, we evaluate PPAT extensively on tabular regression and image-classification tasks. We find that PPAT outperforms existing methods in risk estimation, while also attaining the target coverage level with substantially fewer labels and narrower confidence intervals.

## 2 Background

Throughout this paper, we reserve capital letters for random variables and lowercase letters for observations. For readability, we state the results in the main body using the simplified notation introduced below; their formal versions are given in §E.

Let  $\mathcal{D} = \{\mathbf{x}_1, \dots, \mathbf{x}_N\}$  be an unlabelled test pool of size  $N$ , where each input  $\mathbf{x}_i \in \mathcal{X}$  has a true, unknown label  $y_i \in \mathcal{Y}$ . We treat both the inputs and labels as fixed quantities. Furthermore, let  $f : \mathcal{X} \rightarrow \mathcal{Y}$  be a fixed function we wish to evaluate for a given loss function  $\mathcal{L} : \mathcal{Y} \times \mathcal{Y} \rightarrow \mathbb{R}$ . We define the loss for each test data point as  $\ell_i := \mathcal{L}(f(\mathbf{x}_i), y_i)$  for  $i = 1, \dots, N$ , and the sequence of these losses as  $\ell_{1:N} := (\ell_1, \dots, \ell_N)$ . For a fixed labelling budget  $M \ll N$ , our objective is to estimate the test pool risk

$$R = \frac{1}{N} \sum_{i=1}^N \ell_i \quad (1)$$

based on the inputs  $\mathcal{D}$  and on  $M$  queried labels  $y_{I_1}, \dots, y_{I_M}$ , with indices  $I_1, \dots, I_M$ . A naive approach is to sample the indices randomly without replacement from  $[N] := \{1, \dots, N\}$  and use the unweighted estimator

$$\hat{R}_{\text{unw}} = \frac{1}{M} \sum_{m=1}^M \ell_{I_m}. \quad (2)$$

This estimator is unbiased but typically suffers from high variance (Farquhar et al., 2021; Kossen et al., 2021).

### 2.1 Active Testing and the LURE Estimator

Active testing (AT) (Kossen et al., 2021) aims to estimate the test pool risk  $R$  more efficiently by sequentially selecting the most useful points to label. More formally, for each acquisition round  $m$ , let  $S_m \subseteq \{1, \dots, N\}$  denote the remaining indices to be sampled and  $\mathcal{F}_{m-1}$  the  $\sigma$ -algebra generated by all the information available before round  $m$  (e.g.  $\mathcal{D}$ , past queried indices and labels). AT selects the  $m$ th index to query by sampling from a proposal  $Q_m$ :

$$Q_m(\cdot) := \mathbb{P}(I_m = \cdot \mid \mathcal{F}_{m-1}) \quad \text{supported on } S_m. \quad (3)$$

To estimate the risk, AT uses the LURE (*Levelled Unbiased Risk Estimator*) estimator of Farquhar et al. (2021):

$$\hat{R}_{\text{LURE}} = \frac{1}{M} \sum_{m=1}^M V_m \ell_{I_m}, \quad (4)$$

where  $I_m \mid I_{1:m-1} \sim Q_m$  and

$$V_m = 1 + \frac{N - M}{N - m} \left( \frac{1}{(N - m + 1) Q_m(I_m)} - 1 \right). \quad (5)$$

This is an *unbiased* estimate of  $R$  and, when  $\mathcal{L}$  is nonnegative, has strictly *lower variance* than random sampling without replacement under the oracle proposal  $Q_m^*(i) \propto \ell_i$  (Farquhar et al., 2021, Thm. 5). For general real-valued losses, the variance is instead minimised greedily at each step using the oracle *myopic* proposal  $Q_m^*(i) \propto |\ell_i|$  (see §E.4 for further details). In practise, the true labels are not known, and so  $Q_m^*$  is approximated<sup>1</sup> with

$$Q_m^{\text{AT}}(i) \propto \sqrt{\mathbb{E}_{\pi_m(Y|\mathbf{x}_i)}[\mathcal{L}(f(\mathbf{x}_i), Y)^2]}, \quad (6)$$

where  $\pi_m(Y \mid \mathbf{x})$  is a *surrogate model* used to approximate  $p(Y \mid \mathbf{x})$  at round  $m$  (Kossen et al., 2021).

## 2.2 Active Surrogate Estimators

*Active Surrogate Estimators* (ASEs, Kossen et al. (2022)) provide an alternative approach to AT that replaces importance-weighted Monte Carlo estimation with an interpolation-based surrogate estimate of the loss. Rather than estimating the risk directly from the queried losses, ASEs learn a surrogate model for the unknown test-time label distribution and use it to impute losses across the full test pool. Concretely, after  $M$  acquisitions, the estimator is

$$\widehat{R}_{\text{ASE}} = \frac{1}{N} \sum_{i=1}^N \mathbb{E}_{Y \sim \pi_M(\cdot|\mathbf{x}_i)}[\mathcal{L}(f(\mathbf{x}_i), Y)].$$

In contrast to LURE-based AT, the acquisition in ASE is designed to improve the surrogate itself rather than to produce an unbiased importance-weighted estimator. To this end, Kossen et al. (2022) propose the *Expected Weighted Disagreement (XWED)* acquisition function, which prefers points with high epistemic uncertainty under  $\pi_m$  that are also expected to contribute strongly to the final risk estimate (see §D.3.1 for additional details).

A key advantage of ASEs is that they can make more effective use of surrogate generalisation and allow for more flexible acquisition strategies, including deterministic ones. However, they do not retain the finite-sample unbiasedness of LURE: surrogate misspecification and finite-sample estimation error enter the final estimator directly. Consequently, their performance depends strongly on the quality of the learned surrogate and they suffer from weaker theoretical guarantees than LURE-based AT.

## 2.3 Control-Variate Estimator

We conclude this background section with a short review of control variate estimators. Let  $\widehat{R}$  be an unbiased estimator of  $R$ . Let  $\widehat{C}$  be a zero-mean random variable, called the *control variable* (Glasserman, 2004). For  $\lambda \in \mathbb{R}$ , the control variate estimator,  $\widehat{R}_{\text{cv}}$ , of  $R$  is defined as

$$\widehat{R}_{\text{cv}} = \widehat{R} - \lambda \widehat{C}. \quad (7)$$

The centred variable  $\widehat{C}$  acts as a control-variate correction to the estimator  $\widehat{R}$ . By design,  $\widehat{R}_{\text{cv}}$  is unbiased if  $\widehat{R}$  is, with variance  $\text{Var}(\widehat{R}_{\text{cv}}) = \text{Var}(\widehat{R}) - 2\lambda \text{Cov}(\widehat{R}, \widehat{C}) + \lambda^2 \text{Var}(\widehat{C})$ . Moreover,  $\widehat{R}_{\text{cv}}$  has lower variance than  $\widehat{R}$  whenever  $\lambda \in (\min(0, 2\lambda^*), \max(0, 2\lambda^*))$ , where  $\lambda^* = \text{Cov}(\widehat{R}, \widehat{C})/\text{Var}(\widehat{C})$  results in the largest variance reduction. At  $\lambda^*$ ,  $\text{Var}(\widehat{R}_{\text{cv}}) = (1 - \rho^2)\text{Var}(\widehat{R})$ , where  $\rho$  is the correlation between  $\widehat{R}$  and  $\widehat{C}$ . Thus, a larger absolute correlation implies a larger variance reduction.

<sup>1</sup>Note that when the true loss is unknown, the oracle  $Q_m^* \propto |\ell_i|$  is better approximated by the root-mean-square score (6) rather than by an expected absolute value; see §D.1 for more details.

### 3 Prediction–Powered Active Testing

We now describe our approach, **Prediction–Powered Active Testing (PPAT)**, which leverages a cheap proxy labeller  $g$  to reduce the variance of the LURE estimator in a similar spirit to Angelopoulos et al. (2023a,b).

Assume we can evaluate a cheap proxy labeller  $g : \mathcal{X} \rightarrow \mathcal{Y}$  on all test inputs, yielding proxy labels  $\tilde{y}_i = g(\mathbf{x}_i)$ . Define the *proxy loss*  $\tilde{\ell}_i := \mathcal{L}(f(\mathbf{x}_i), \tilde{y}_i)$  and the associated proxy size- $N$  test risk  $\tilde{R} := \frac{1}{N} \sum_{i=1}^N \tilde{\ell}_i$ . We first note that, for any  $\lambda \in \mathbb{R}$ , the test risk can be written alternatively as

$$R = \frac{1}{N} \sum_{i=1}^N (\ell_i - \lambda(\tilde{\ell}_i - \tilde{R})),$$

that is, an average of the *residualised losses*  $(\ell_i - \lambda(\tilde{\ell}_i - \tilde{R}))_{i=1}^N$ . We propose to apply the LURE weighting scheme to this residualised representation of  $R$ , yielding our PPAT estimator:

$$\hat{R}_{\text{PPAT}}(\lambda) = \frac{1}{M} \sum_{m=1}^M V_m(\ell_{I_m} - \lambda(\tilde{\ell}_{I_m} - \tilde{R})). \quad (8)$$

Our estimator recovers the standard LURE estimator (4) as a special case when  $\lambda = 0$ . Moreover, noting that

$$\hat{R}_{\text{PPAT}}(\lambda) = \hat{R}_{\text{LURE}} - \lambda \hat{C}_{\text{LURE}},$$

the PPAT estimator can be interpreted as a control–variate correction of  $\hat{R}_{\text{LURE}}$ , with the control variate being

$$\hat{C}_{\text{LURE}} = \frac{1}{M} \sum_{m=1}^M V_m(\tilde{\ell}_{I_m} - \tilde{R}),$$

where  $\hat{C}_{\text{LURE}}$  is itself a zero–mean LURE-style estimator<sup>2</sup>. This control–variate view suggests that suitable choices of  $\lambda$  should reduce variance relative to LURE. We make this more precise in the next section, where we study the finite–sample properties of the PPAT estimator.

#### 3.1 Finite–Sample Properties of PPAT

The results below hold for any  $1 \leq M < N$ , and any choice of proposals  $Q_m$ . We first note that our estimator naturally inherits the unbiasedness of LURE.

**Proposition 3.1** (Unbiasedness). *For any fixed  $\lambda \in \mathbb{R}$ , we have  $\mathbb{E}[\hat{R}_{\text{PPAT}}(\lambda)] = R$ .*

*Proof.* See §E.3.

Secondly, as  $\hat{R}_{\text{PPAT}}(\lambda)$  is a control–variate estimator, it follows that it achieves lower variance than  $\hat{R}_{\text{LURE}}$  for a range of  $\lambda$  values.

**Proposition 3.2** (Variance of  $\hat{R}_{\text{PPAT}}$ ). *Let  $\lambda \in \mathbb{R}$ . We have*

$$\text{Var}(\hat{R}_{\text{PPAT}}(\lambda)) = \text{Var}(\hat{R}_{\text{LURE}}) - 2\lambda \text{Cov}(\hat{R}_{\text{LURE}}, \hat{C}_{\text{LURE}}) + \lambda^2 \text{Var}(\hat{C}_{\text{LURE}}). \quad (9)$$

<sup>2</sup>Indeed, we have that  $\mathbb{E}[\hat{C}_{\text{LURE}}] = \frac{1}{N} \sum_{i=1}^N (\tilde{\ell}_i - \tilde{R}) = 0$ .

Thus,  $\text{Var}(\widehat{R}_{\text{PPAT}}(\lambda)) < \text{Var}(\widehat{R}_{\text{LURE}})$  whenever  $\lambda \in (\min\{0, 2\lambda^*\}, \max\{0, 2\lambda^*\})$  and  $\text{Var}(\widehat{C}_{\text{LURE}}) > 0$ , where

$$\lambda^* = \frac{\text{Cov}(\widehat{R}_{\text{LURE}}, \widehat{C}_{\text{LURE}})}{\text{Var}(\widehat{C}_{\text{LURE}})}. \quad (10)$$

In particular,  $\lambda^*$  minimises the variance of  $\widehat{R}_{\text{PPAT}}(\lambda)$ .

*Proof.* See §E.3.

Note that under  $\lambda^*$ , the variance of the PPAT estimator can also be written as  $\text{Var}(\widehat{R}_{\text{LURE}})(1 - \rho^2)$ , where  $\rho$  is the correlation between  $\widehat{R}_{\text{LURE}}$  and  $\widehat{C}_{\text{LURE}}$ . Here, a larger correlation between these two terms results in a larger reduction in variance relative to LURE.

While Prop. 3.2 shows that PPAT can achieve lower variance than LURE, this depends critically on our choice of  $\lambda$  and holds only when both LURE and PPAT use the same proposals. Moreover, although it is possible to adopt the same LURE proposal,  $Q_m^{\text{AT}}$ , for our approach and tune only  $\lambda$ , this will typically be suboptimal as it selects points by their raw loss  $\ell_i$ , whereas the variance of PPAT is driven by the residual  $\ell_i - \lambda(\ell_i - \widetilde{R})$ . In the following sections, we derive a proposal tailored to our PPAT estimator and discuss practical choices of  $\lambda$ .

## 3.2 PPAT Proposal

While the previous results hold for any valid proposal  $Q_m$ , we would ideally like to choose the proposal that minimises the variance of the PPAT estimator. As the exact finite-horizon variance-minimising proposal is intractable, we instead consider proposals that minimise the alternative, myopic proxy described below.

Firstly, note that since  $\widehat{R}_{\text{PPAT}}$  is itself a LURE estimator applied to the residualised losses  $\ell_i - \lambda(\ell_i - \widetilde{R})$ , its sampling variance decomposes across rounds in the same way as that of LURE (see §E.4 for further details). In particular, conditionally on  $I_{1:m-1}$ , the  $m$ th contribution to this variance is proportional to

$$\text{Var}\left(\frac{\ell_{I_m} - \lambda(\ell_{I_m} - \widetilde{R})}{N Q_m(I_m)} \mid I_{1:m-1}\right). \quad (11)$$

Thus, at round  $m$ , a natural myopic oracle is the proposal which minimises this conditional variance over the remaining pool  $S_m$ . This oracle admits a simple closed form.

**Proposition 3.3** (Myopic oracle proposal). *Fix  $\lambda \in \mathbb{R}$ . For each round  $m$ , among all proposals supported on  $S_m$ , the myopic proxy (11) is minimised by*

$$Q_{m,\lambda}^*(i) = \frac{|\ell_i - \lambda(\ell_i - \widetilde{R})|}{\sum_{j \in S_m} |\ell_j - \lambda(\ell_j - \widetilde{R})|}.$$

*Proof.* See §E.4.

In practice, this oracle proposal cannot be used, as it depends on the unknown labels  $y_i$  through the true losses  $\ell_i$ . As in §2.1, we approximate the oracle using the surrogate predictive distribution  $\pi_m(\cdot \mid \mathbf{x}_i)$ , giving

the surrogate score<sup>3</sup>

$$a_{m,\lambda}(i) := \sqrt{\mathbb{E}_{Y \sim \pi_m(\cdot | \mathbf{x}_i)} \left[ \left( \mathcal{L}(f(\mathbf{x}_i), Y) - \lambda(\tilde{\ell}_i - \tilde{R}) \right)^2 \right]}, \quad (12)$$

and the PPAT proposal

$$Q_{m,\lambda}^{\text{PPAT}}(i) = \frac{a_{m,\lambda}(i)}{\sum_{j \in S_m} a_{m,\lambda}(j)}, \quad i \in S_m. \quad (13)$$

To ensure strict positivity, this proposal is mixed with a uniform distribution over  $S_m$  like in (Kossen et al., 2021), replacing  $Q_{m,\lambda}^{\text{PPAT}}(i)$  by  $(1 - \varepsilon) Q_{m,\lambda}^{\text{PPAT}}(i) + \varepsilon/|S_m|$  for a small  $\varepsilon \in (0, 1)$ .

### 3.3 Choosing $\lambda$

As mentioned previously, the variance reduction achieved by PPAT depends critically on the choice of  $\lambda$ . More specifically, Prop. 3.2 shows that PPAT achieves lower variance than LURE only for  $\lambda \in (\min\{0, 2\lambda^*\}, \max\{0, 2\lambda^*\})$ ; a  $\lambda$  with the wrong sign, or with too large a magnitude, lies outside this range and instead inflates the variance relative to LURE. Selecting  $\lambda$  appropriately is therefore crucial, but is also further complicated by the fact that it affects both the PPAT estimator *and* the PPAT proposal. Indeed, for given proposals  $Q_m$ , the control-variate coefficient  $\lambda^*$  in (10) depends on the sampling law they induce, which cannot be estimated from a single active-testing run. Moreover,  $\lambda$  also parameterises our proposal,  $Q_{m,\lambda}^{\text{PPAT}}$ , and so shapes the sampling law on which  $\lambda^*$  itself depends.

These considerations motivate two practical strategies for choosing  $\lambda$ : a *heuristic* choice of fixed values, and an *online* estimate updated from the labels acquired during testing.

**Heuristic, fixed choices of  $\lambda$ .** In a similar spirit to Angelopoulos et al. (2023a), we propose the fixed values  $\lambda = 1$  and  $\lambda = \frac{1}{2}$ , where we expect  $\lambda = 1$  to be most effective when the proxy is strongly predictive of the true loss, and use  $\lambda = \frac{1}{2}$  as a more conservative choice. To see why these are sensible, first note that  $\lambda^*$  can be written as  $\lambda^* = \rho \sigma_R / \sigma_C$ , where  $\rho$  is the correlation between  $\hat{R}_{\text{LURE}}$  and  $\hat{C}_{\text{LURE}}$  and  $\sigma_R, \sigma_C$  are their standard deviations. When the proxy losses vary on a scale comparable to the true losses, we have  $\sigma_R \approx \sigma_C$ , so that  $\lambda^* \approx \rho$ . Moreover, for an informative proxy whose loss tracks the true loss, we expect  $\rho > 0$ . Under these conditions, PPAT reduces variance over LURE when  $\lambda \in (0, 2\rho)$  and, consequently,  $\lambda = 1$  yields a reduction when  $\rho > \frac{1}{2}$ , while  $\lambda = \frac{1}{2}$  does so over the wider range  $\rho > \frac{1}{4}$ . For a genuinely useful proxy, we expect it to be strongly correlated with the true loss, giving  $\rho > \frac{1}{2}$ , in which case both choices reduce variance; a less useful proxy, with only moderate correlation  $\frac{1}{4} < \rho \leq \frac{1}{2}$ , falls outside the range for  $\lambda = 1$  but is still handled by the more conservative  $\lambda = \frac{1}{2}$ .

**Online estimation of  $\lambda$ .** As  $\lambda^*$  depends on the sampling law induced by the proposals  $Q_m$ , it cannot be estimated from a single active-testing run. We therefore target an alternative coefficient  $\lambda^\dagger$  that (i) does not depend on the proposal, and so can be estimated online from the acquired labels, and (ii) flattens the residualised losses  $\ell_i - \lambda c_i$  across the pool, which minimises an upper bound on the estimator’s variance. Concretely, writing  $c_i = \tilde{\ell}_i - \tilde{R}$ , we define

$$\lambda^\dagger := \arg \min_{\lambda \in \mathbb{R}} \frac{1}{N} \sum_{i=1}^N (\ell_i - \lambda c_i)^2 = \frac{N^{-1} \sum_{i=1}^N \ell_i c_i}{N^{-1} \sum_{i=1}^N c_i^2},$$

which is well-defined whenever  $N^{-1} \sum_{i=1}^N c_i^2 > 0$ . Here,  $\lambda^\dagger$  can be seen as choosing the amount of proxy correction to subtract so that the residualised losses are as flat as possible across the pool. Moreover, under mild assumptions about our proposal, it can be shown that this minimises an upper bound on the variance of our estimator:

<sup>3</sup>As in §2.1, the root-mean-square score (12) – rather than an expected absolute value – is the more appropriate surrogate of the oracle here.

**Proposition 3.4** (Variance bound for PPAT). *Suppose that our proposals satisfy a uniform overlap condition: there exists  $\beta > 0$  such that*

$$Q_m(i) \geq \frac{\beta}{N}, \quad m = 1, \dots, M, \quad i \in S_m.$$

Then, for every  $\lambda \in \mathbb{R}$ ,

$$\text{Var}(\widehat{R}_M^{\text{PPAT}}(\lambda)) \leq \frac{1}{\beta M} \frac{N}{N - M + 1} \left( \frac{1}{N} \sum_{i=1}^N (\ell_i - \lambda c_i)^2 \right).$$

Moreover, whenever  $N^{-1} \sum_{i=1}^N c_i^2 > 0$ , this bound is minimised over  $\lambda$  by

$$\lambda^\dagger = \frac{N^{-1} \sum_{i=1}^N \ell_i c_i}{N^{-1} \sum_{i=1}^N c_i^2}.$$

*Proof.* See §E.3.1.

In particular, we expect this bound to become tight when the labelling budget is small relative to the pool size and the proposal is not overly concentrated on a small subset of points (see §E.3.1 for further discussion). Additionally, we empirically verify the suitability of  $\lambda^\dagger$  as an alternative to  $\lambda^*$  in §C.1.

Crucially,  $\lambda^\dagger$  can be estimated online during active testing. As the denominator is known on the full pool, it suffices to estimate the numerator, which we do via its LURE estimate, giving the plug-in estimate

$$\widehat{\lambda}_M = \frac{M^{-1} \sum_{m=1}^M V_m \ell_{I_m} c_{I_m}}{N^{-1} \sum_{i=1}^N c_i^2}. \quad (14)$$

While using this plug-in estimate of  $\lambda$  yields a biased estimator of the risk, we show in §E.5.4 that the resulting estimator remains consistent, so that its bias vanishes asymptotically. In practice, updating  $\widehat{\lambda}_M$  after every label can be unstable, since each update changes the residualised loss used for sampling; we therefore update only after every  $k$  newly collected labels, keeping  $\lambda$  fixed in between.

## 4 Asymptotic Properties and Approximate Confidence Intervals

While the previous sections have focused on point estimation of the risk, in practice it is typically beneficial to establish the estimator’s asymptotic behavior and provide a measure of uncertainty around this estimate. After confirming the consistency of our estimator, we therefore turn to the construction of asymptotically valid confidence intervals (CIs) for PPAT – an aspect that has been largely neglected in prior work on active testing.<sup>4</sup>

All expectations and probabilities for the following result are over the active sampling process, conditionally on the finite test pool. For each budget  $M$ , let  $N_M$  denote the pool size and  $R_M = N_M^{-1} \sum_{i=1}^{N_M} \ell_i$  the test pool risk. We write  $\widehat{R}_{M,\text{PPAT}}(\lambda)$  for the PPAT estimator at budget  $M$  and tuning parameter  $\lambda \in \mathbb{R}$ , with  $\widehat{R}_{M,\text{LURE}} = \widehat{R}_{M,\text{PPAT}}(0)$  the corresponding LURE estimator. We state the main result informally below; its full statement and proof are deferred to §E.

<sup>4</sup>A notable exception is Berrada et al. (2025), who bootstrap the reweighted losses to obtain confidence intervals for LURE. However, the reweighted losses are not i.i.d. – they arise from active sampling without replacement – and their bootstrap resampling ignores this dependence, so the resulting intervals lack theoretical guarantees.

**Theorem 4.1** (Consistency and asymptotic normality of PPAT, informal). *Fix  $\lambda \in \mathbb{R}$ . Suppose that our proposals satisfy a uniform overlap condition (i.e. for each  $1 \leq m \leq M$ , there exists  $\beta > 0$  s.t.  $Q_m(i) \geq \beta/N_M$  for all  $i \in S_m$ ) and that the pool ratio  $N_M/M$  converges to some  $\alpha > 1$ . Then:*

1. (Consistency.) *Assume the finite-pool averages  $N_M^{-1} \sum_{i=1}^{N_M} \ell_i^2$  and  $N_M^{-1} \sum_{i=1}^{N_M} \tilde{\ell}_i^2$  are both bounded. Then we have  $\widehat{R}_{M,\text{PPAT}}(\lambda) - R_M \rightarrow_p 0$ .*
2. (Asymptotic normality.) *Assume the finite-pool averages  $N_M^{-1} \sum_{i=1}^{N_M} \ell_i^4$  and  $N_M^{-1} \sum_{i=1}^{N_M} \tilde{\ell}_i^4$  are both bounded. Write  $\zeta_i(\lambda) = \ell_i - \lambda(\tilde{\ell}_i - \widetilde{R}_M)$  for the residualised loss, with  $\widetilde{R}_M = N_M^{-1} \sum_{i=1}^{N_M} \tilde{\ell}_i$ , and let  $\gamma_{M,m} = \frac{N_M(N_M-M)}{(N_M-m)(N_M-m+1)}$ . Assume the predictable quadratic variation*

$$\sigma_M^2(\lambda) = \frac{1}{M} \sum_{m=1}^M \gamma_{M,m}^2 s_m^2(\lambda), \quad s_m^2(\lambda) = \frac{1}{N_M^2} \left[ \sum_{i \in S_m} \frac{\zeta_i(\lambda)^2}{Q_m(i)} - \left( \sum_{i \in S_m} \zeta_i(\lambda) \right)^2 \right], \quad (15)$$

converges in probability to some  $\sigma_{\text{PPAT}}^2(\lambda) \in (0, \infty)$ . Then

$$\sqrt{M} \{ \widehat{R}_{M,\text{PPAT}}(\lambda) - R_M \} \Rightarrow \mathcal{N}(0, \sigma_{\text{PPAT}}^2(\lambda)),$$

and  $\sigma_{\text{PPAT}}^2(\lambda)$  admits the plug-in estimator

$$\widehat{\sigma}_M^2(\lambda) = \frac{1}{M} \sum_{m=1}^M \gamma_{M,m}^2 (A_m(\lambda) - \widehat{R}_{M,\text{PPAT}}(\lambda))^2, \quad A_m(\lambda) = \frac{1}{N_M} \left( \frac{\zeta_{I_m}(\lambda)}{Q_m(I_m)} + \sum_{t=1}^{m-1} \zeta_{I_t}(\lambda) \right),$$

which is computable from the queried labels and satisfies  $\widehat{\sigma}_M^2(\lambda) \rightarrow_p \sigma_{\text{PPAT}}^2(\lambda)$ .

Moreover, analogous consistency and asymptotic normality results hold when  $\lambda$  is replaced by the plug-in estimate  $\widehat{\lambda}_M$  from (14).

*Proof.* See §E.5.3-E.5.4.

Crucially, Thm. 4.1 immediately yields the asymptotic  $(1 - \delta)$ -CI

$$\widehat{R}_{M,\text{PPAT}}(\lambda) \pm z_{1-\delta/2} \widehat{\sigma}_M(\lambda) / \sqrt{M}, \quad (16)$$

where  $z_{1-\delta/2}$  denotes the  $(1 - \delta/2)$ -quantile of the standard normal distribution. As LURE is the special case  $\lambda = 0$ , this also yields, as a by-product, an asymptotic CI for LURE.

Note that the uniform overlap condition is the usual positivity requirement for importance-weighted active testing (Farquhar et al., 2021) and can be enforced by mixing with a uniform proposal (Kossen et al., 2021), while the moment conditions are bounded finite-pool averages requiring no distributional assumption, as we work conditionally on the test pool. Moreover, the condition  $N_M/M \rightarrow \alpha > 1$  is the usual finite-population asymptotic regime (Farquhar et al., 2021). The remaining assumption – that the predictable quadratic variation  $\sigma_M^2(\lambda)$  converges to a finite, strictly positive limit – is the standard condition for the martingale central limit theorem and is needed only to show asymptotic normality; we discuss this assumption in detail in §E.5.5. We provide empirical support for this assumption indirectly through the coverage experiments in §5, where we show that PPAT can attain nominal coverage substantially faster than competing strategies, while also producing substantially narrower intervals.

## 5 Experiments

We evaluate our approach on real-world regression and classification datasets. We provide a description of our experimental setup and results below with full details in §B.

## 5.1 Experimental Setup

**Datasets.** For regression, we use the following UCI datasets (Dheeru and Karra Taniskidou, 2017): `Keggundirected`, `Keggdirected`, `Sml`, `Bike`. For each dataset, we randomly sample a fixed training set of 250 points to train the model  $f$  whose risk we wish to estimate, and use the remaining points as the test pool. For classification, we use `CIFAR-10`, `CIFAR-100` (Krizhevsky, 2009), and `Tiny-ImageNet` (Le and Yang, 2015). We respect the original train/test splits, using the training set to train  $f$  and treating the test set as the test pool. Additional dataset details are given in §B.

**Models and proxies.** For regression,  $f$  is a Gaussian process regression model with an RBF kernel whose hyperparameters are optimised using the marginal likelihood, and the active-testing surrogate is a Bayesian linear regression model with the default `Scikit-Learn` hyperparameters (Pedregosa et al., 2011). For classification, we follow the standard pretrained-encoder pipeline: inputs are mapped to CLIP embeddings (Radford et al., 2021), and  $f$  is a linear classifier trained on these frozen embeddings. The surrogate is a Laplace-approximated Bayesian neural network (MacKay, 1992; Daxberger et al., 2021): an MLP on CLIP embeddings with three hidden layers of width 128 for `CIFAR-10` and `CIFAR-100`, and five hidden layers of width 128 for `Tiny-ImageNet`.

For our proxy, we use strong off-the-shelf predictors suited to each task. For regression, we use the `TabPFN-2.5` foundation model (Grinsztajn et al., 2025) and for classification we use the zero-shot predictions of the `ViT-L-14` CLIP model from the `sentence-transformers` library (Reimers and Gurevych, 2019). Further implementation details and ablations over surrogate and proxy choices are deferred to §B and C respectively.

**Baselines.** We compare `PPAT` with `LURE` (Kossen et al., 2021), `ASE` (Kossen et al., 2022), and `Random`, which samples randomly without replacement and uses the unweighted estimator in (2). Within each experimental setting, all active methods are run with the same surrogate model. For `PPAT`, we report the fixed choices  $\lambda = 1$  and  $\lambda = 0.5$ , as well as the plug-in estimate  $\hat{\lambda}_M$ , which we initialise at 0.5 and update after every 100 acquired points. For `ASE`, we use 100 Monte Carlo samples to compute the acquisition scores; further details about the baselines are given in §A and §B.

**Active-testing setup and metrics.** In all experiments, we use a labelling budget of  $M = 500$  and run 1000 independent trials. For regression, the loss is squared error and the surrogate is updated after every newly acquired label. For classification, the loss is cross-entropy and, following prior work (Kossen et al., 2021, 2022), we keep the surrogate fixed during acquisition. Following Kossen et al. (2021, 2022), we report the median squared error, where the error is the difference between the estimated risk and the true risk on the full test pool. Mean error, confidence-interval widths, and further ablations are provided in §C.

## 5.2 PPAT Improves Risk Estimation

**Regression.** Fig. 1 shows that, across all datasets, `PPAT` achieves lower median squared error than `Random`, `LURE`, and `ASE` for all three choices of  $\lambda$ . The only exception occurs at very small labelling budgets, where `ASE` can sometimes attain lower error, for example on `Keggundirected`. However, this comes at the cost of using a **biased** risk estimate.

For the plug-in choice  $\hat{\lambda}_M$ , we occasionally observe a small transient increase in error at early labelling budgets (e.g. on `Keggundirected`). This is expected, since  $\hat{\lambda}_M$  is updated from the actively collected labels during acquisition, which introduces finite-sample bias relative to fixed- $\lambda$  `PPAT` and also changes the proposal used to sample future points. As the number of acquired labels grows, however,  $\hat{\lambda}_M$  stabilises and this bias diminishes (see §C.3.1). Importantly, we also see that `PPAT` with  $\hat{\lambda}_M$  generally converges to `PPAT` with the best performing fixed  $\lambda$ , suggesting that targeting  $\lambda^\dagger$  is an effective choice (see also §C.1).

**Classification.** Fig. 2 shows the same pattern on the classification datasets: `PPAT` consistently achieves lower median squared error than `Random`, `LURE`, and `ASE` for all choices of  $\lambda$ , with the exception of

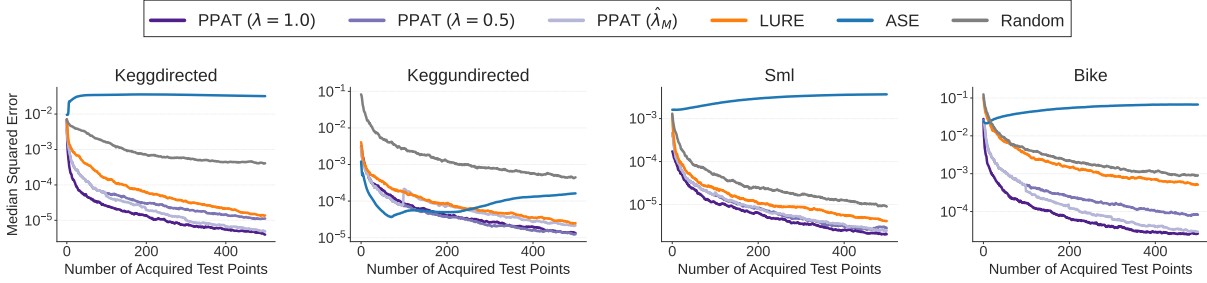


Figure 1: Regression experiments comparing **PPAT**, **LURE**, **ASE**, and **Random**. Plots show the median squared error across 1000 trials.

**Tiny-ImageNet**, where **PPAT** with  $\lambda = 1$  performs comparably to **Random**. We further note that **Random** outperforms **LURE** on these datasets, likely a result of the surrogate poorly tracking the losses. Indeed, since **LURE** builds its proposal, and hence its importance weights, from the surrogate’s expected losses, a surrogate that poorly tracks the true losses can result in importance weights that inflate the variance of **LURE**. In contrast, **PPAT** is less affected by this issue: it importance-weights the residualised loss rather than the raw loss, and since the proxy explains part of the loss variation, the residuals have smaller spread, making the variance less sensitive to errors in the surrogate.

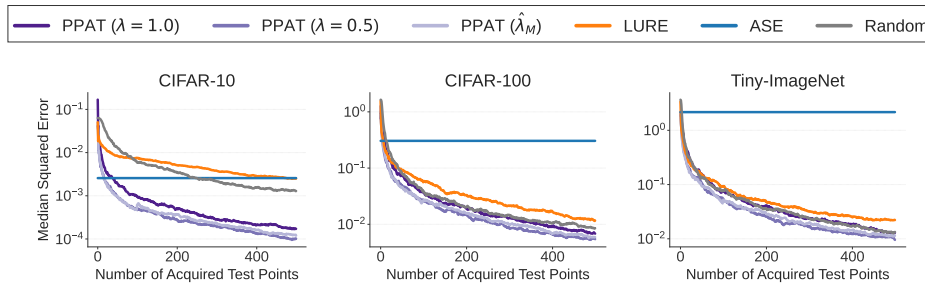


Figure 2: Classification experiments comparing **PPAT**, **LURE**, **ASE**, and **Random**. Plots show the median squared error across 1000 trials.

Overall, these results indicate that **PPAT** improves risk estimation not only for tabular regression, but also for more challenging, higher-dimensional image-classification tasks.

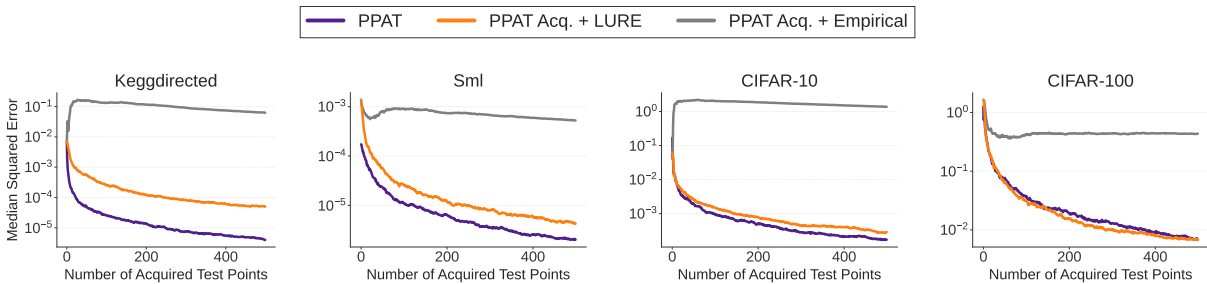


Figure 3: Experiments on UCI and classification datasets studying the influence of our estimator. We compare **PPAT** with  $\lambda = 1$  to using the **PPAT** acquisition strategy,  $Q^{\text{PPAT}}$ , but with the **LURE** estimate of the risk (**PPAT Acq. + LURE**) and the empirical estimate of the risk (**PPAT Acq. + Empirical**). Plots show the median squared error across 1000 trials.

### 5.3 Disentangling the Components of PPAT

Beyond the choice of  $\lambda$ , which we have discussed in §5.2 and §C.1, there are two main components that can affect the behaviour of **PPAT**: the choice of acquisition strategy and the choice of the PPI-like estimator. To disentangle their contributions, we study the effect of using different estimators in Fig. 3 and different acquisition strategies in Fig. 4 on four representative datasets.

The notable improvements observed over both different acquisition strategies (Fig. 4) and different estimators (Fig. 3) show that the gains of **PPAT** do not come merely from using a better, PPI-like estimator. Instead, they come from combining our PPI-like estimator with an acquisition strategy tailored to the corresponding residualised objective.

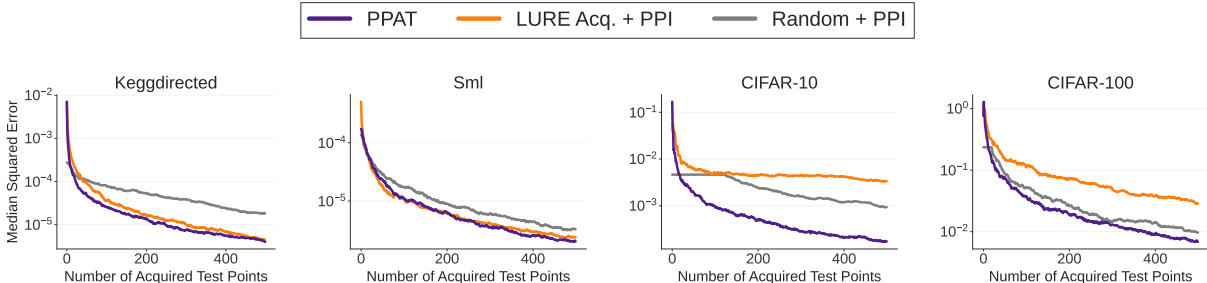


Figure 4: Experiments on UCI and classification datasets studying the influence of our acquisition strategy. We compare **PPAT** with  $\lambda = 1$  to using the PPI estimator with the LURE acquisition strategy (**LURE Acq. + PPI**) and random sampling (**Random + PPI**). Plots show the median squared error across 1000 trials.

### 5.4 Coverage

Beyond reducing estimation error, we also want our CIs to be well-calibrated: across repeated active-testing runs, a nominal  $(1 - \delta)$ -CI should contain the true test-pool risk with frequency close to  $(1 - \delta)$ . Fig. 5 shows the coverage for the CIs presented in §4 for **PPAT**, **LURE**, and **Random** on four representative datasets. **ASE** is excluded as it **cannot** provide asymptotically valid CIs without additional assumptions about  $f$  and  $\pi_m(\cdot|\mathbf{x})$ . We see that on all datasets, our approach achieves the desired coverage level for all choices of  $\lambda$ . More importantly, it achieves the desired coverage at either the same or a significantly faster rate than all other approaches. Additionally, our approach attains significantly smaller widths for the confidence intervals than the competing approaches (see §C.3.2).

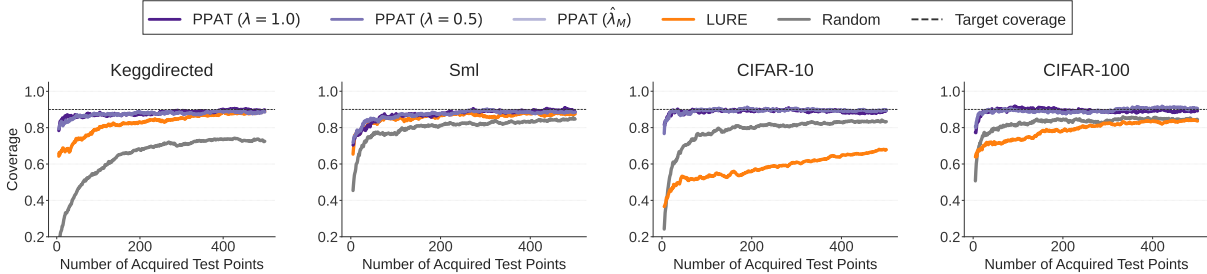


Figure 5: Representative experiments on UCI and classification datasets for the coverage of **PPAT**, **Random**, **LURE**. Plots show the coverage of asymptotic confidence intervals at a target error level of  $\delta = 0.1$  across 1000 trials. The **ASE** baseline is excluded as it does **not** provide asymptotically valid confidence intervals. The width of the intervals are shown in Appendix C.3.2.

## 6 Related Work

---

We review the most closely related work here, and defer a more extensive discussion to §A.

**Active testing.** PPAT builds directly on the active-testing line of work initiated by Farquhar et al. (2021), who introduced the *Levelled Unbiased Risk Estimator (LURE)*, and Kossen et al. (2021), who developed it into a sample-efficient model-evaluation framework. A closely related approach is that of *Active Surrogate Estimators (ASE)* (Kossen et al., 2022), which, like PPAT, exploit predictions over the entire pool. ASE, however, does so in a fundamentally different way: rather than correcting an importance-weighted estimator with observed labels, it learns a surrogate for the conditional loss and imputes losses across the pool via interpolation. While this can be successful at very low labelling budgets, ASE loses the theoretical guarantees of LURE, since surrogate misspecification or finite-sample estimation error enters the risk estimate directly. More recently, Berrada et al. (2025) scaled LURE-based active testing to large language model (LLM) evaluation using fixed, in-context-learning surrogates. Our approach is complementary: it can incorporate the same in-context surrogates while additionally leveraging cheap black-box predictions to improve the estimator itself.

**Earlier work on label-efficient evaluation.** Earlier work considered more restricted sampling schemes. Stratification-based methods (Bennett and Carvalho, 2010; Ji et al., 2021; Katariya et al., 2012; Kumar and Raj, 2018) partitioned the test pool into strata, for example using a simple non-adaptive measure of model confidence, and sampled uniformly within each. PPAT is complementary here: once a stratification is fixed, it can be applied within each stratum to further improve label efficiency. A different line of work considered importance and Poisson sampling: Sawade et al. (2010a) used importance sampling *with* replacement, which is suboptimal relative to estimators that explicitly account for sampling without replacement (see Appendix D of Kossen et al. (2021)); Yilmaz et al. (2021) proposed a Poisson-sampling approach, but it is restricted to particular evaluation metrics, uses a non-adaptive acquisition rule, and relies on ratio-type estimates that are generally biased. Nguyen et al. (2018) studied a more specialised active-testing setting focused on human vetting of noisy labels.

## 7 Discussion

---

In this work, we introduced **Prediction-Powered Active Testing (PPAT)**, a label-efficient framework for risk estimation that combines the unbiased LURE estimator with a prediction-powered control variate. Rather than imputing labels, PPAT uses cheap proxy predictions to residualise the loss, preserving unbiasedness while reducing variance. We also established asymptotic normality for PPAT, yielding asymptotic confidence intervals that provide a principled quantification of uncertainty around the risk estimate. Our results show that PPAT improves risk estimation across tabular regression and image-classification tasks, while attaining target coverage with substantially fewer labels and producing narrower confidence intervals than competing approaches.

PPAT also suggests several natural directions for future work. One is to tighten our variance bound by identifying additional, practically reasonable assumptions on the proposals under which it more closely reflects the exact variance. Another is to complement our asymptotic confidence-interval theory with non-asymptotic guarantees that more directly characterise finite-budget behaviour. A further extension is to consider the setting where the proxy model is also updated as labels are acquired. Since PPAT only requires proxy losses over the pool and uses them through a centred control variate, our approach and theoretical results transfer naturally to this setting.

## Acknowledgments

---

KA is supported by the EPSRC Centre for Doctoral Training in Modern Statistics and Statistical Machine Learning (EP/S023151/1). VK is supported by the Clarendon Funds Scholarship. DH is supported by the Research Council of Finland (Flagship programme: Finnish Center for Artificial Intelligence FCAI, 359207). TR is supported by the EPSRC grant EP/Y037200/1. The authors are grateful to Stefano Cortinovis for helpful discussions.

## References

---

- Josh Abramson, Jonas Adler, Jack Dunger, Richard Evans, Tim Green, Alexander Pritzel, Olaf Ronneberger, Lindsay Willmore, Andrew J Ballard, Joshua Bambrick, Sebastian Bodenstern, David A Evans, Chia-Chun Hung, Michael O’Neill, David Reiman, Kathryn Tunyasuvunakool, Zachary Wu, Akvile Zengulyte, Eirini Arvaniti, Charles Beattie, Ottavia Bertolli, Alex Bridgland, Alexey Cherepanov, Miles Congreve, Alexander Imani Cowen-Rivers, Andrew Cowie, Michael Figurnov, Fabian B Fuchs, Hannah Gladman, Rishub Jain, Yousuf A. Khan, Caroline M R Low, Kuba Perlin, Anna Potapenko, Pascal Savy, Sukhdeep Singh, Adrian Stecula, Ashok Thillaisundaram, Catherine Tong, Sergei Yakneen, Ellen D. Zhong, Michal Zielinski, Augustin Židek, Vic-613 tor Bapst, Pushmeet Kohli, Max Jaderberg, Demis Hassabis, and John M. Jumper. Accurate structure prediction of biomolecular interactions with alphafold 3. *Nature*, 630:493 – 500, 2024. URL <https://api.semanticscholar.org/CorpusID:269633210>. (page 1)
- Guillaume Alain and Yoshua Bengio. Understanding intermediate layers using linear classifier probes. *arXiv preprint arXiv:1610.01644*, 2016. (page 24)
- Anastasios N Angelopoulos, Stephen Bates, Clara Fanfjiang, Michael I Jordan, and Tijana Zrnica. Prediction-powered inference. *Science*, 382(6671):669–674, 2023a. (pages 1, 4, 6, 21, and 25)
- Anastasios N Angelopoulos, John C Duchi, and Tijana Zrnica. Ppi++: Efficient prediction-powered inference. *arXiv preprint arXiv:2311.01453*, 2023b. (pages 4 and 21)
- Dana Angluin. Queries and concept learning. *Mach. Learn.*, 2(4):319–342, April 1988. ISSN 0885-6125. doi: 10.1023/A:1022821128753. URL <https://doi.org/10.1023/A:1022821128753>. (page 1)
- Shir Ashury-Tahan, Ariel Gera, Benjamin Sznajder, Leshem Choshen, Liat Ein-Dor, and Eyal Shnarch. Label-efficient model selection for text generation. *arXiv preprint arXiv:2402.07891*, 2024. (page 21)
- Les Atlas, David Cohn, and Richard Ladner. Training connectionist networks with queries and selective sampling. In D. Touretzky, editor, *Advances in Neural Information Processing Systems*, volume 2. Morgan-Kaufmann, 1989. URL [https://proceedings.neurips.cc/paper\\_files/paper/1989/file/b1a59b315fc9a3002ce38bbe070ec3f5-Paper.pdf](https://proceedings.neurips.cc/paper_files/paper/1989/file/b1a59b315fc9a3002ce38bbe070ec3f5-Paper.pdf). (page 1)
- Heejung Bang and James M. Robins. Doubly robust estimation in missing data and causal inference models. *Biometrics*, 61(4):962–973, 12 2005. ISSN 0006-341X. doi: 10.1111/j.1541-0420.2005.00377.x. URL <https://doi.org/10.1111/j.1541-0420.2005.00377.x>. (page 22)
- 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, CIKM ’10*, page 1581–1584, New York, NY, USA, 2010. Association for Computing Machinery. ISBN 9781450300995. doi: 10.1145/1871437.1871677. URL <https://doi.org/10.1145/1871437.1871677>. (page 12)
- Gabrielle Berrada, Jannik Kossen, Freddie Bickford Smith, Muhammed Razzak, Yarin Gal, and Tom Rainforth. Scaling up active testing to large language models. *arXiv preprint arXiv:2508.09093*, 2025. (pages 1, 7, and 12)
- Lukas Biewald. Experiment tracking with weights and biases, 2020. URL <https://www.wandb.com/>. Software available from wandb.com. (page 26)
- Patrick Billingsley. *Probability and Measure*. John Wiley & Sons, New York, 3 edition, 1995. (page 49)
- Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020. (page 1)

- Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PmLR, 2020. (page 24)
- William G. Cochran. *Sampling Techniques*. John Wiley & Sons, 3 edition, 1977. (page 22)
- Gabriele Corso, Hannes Stärk, Bowen Jing, Regina Barzilay, and Tommi Jaakkola. Diffdock: Diffusion steps, twists, and turns for molecular docking. *arXiv preprint arXiv:2210.01776*, 2022. (page 1)
- Erik Daxberger, Agustinus Kristiadi, Alexander Immer, Runa Eschenhagen, Matthias Bauer, and Philipp Hennig. Laplace redux-effortless bayesian deep learning. *Advances in neural information processing systems*, 34:20089–20103, 2021. (pages 9, 24, and 26)
- Dua Dheeru and Efi Karra Taniskidou. UCI machine learning repository, 2017. URL <http://archive.ics.uci.edu/ml>. (page 9)
- Trefor Evans. uci\_datasets: Regression datasets from the uci repository with standardized test-train splits. [https://github.com/treforevans/uci\\_datasets](https://github.com/treforevans/uci_datasets), 2022. (page 26)
- Hadi Fanaee-T. Bike Sharing. UCI Machine Learning Repository, 2013. DOI: <https://doi.org/10.24432/C5W894>. (page 26)
- Sebastian Farquhar, Yarin Gal, and Tom Rainforth. On statistical bias in active learning: How and when to fix it. *arXiv preprint arXiv:2101.11665*, 2021. (pages 1, 2, 3, 8, 12, 37, and 60)
- Adam Fisch, Joshua Maynez, R Hofer, Bhuwan Dhingra, Amir Globerson, and William W Cohen. Stratified prediction-powered inference for effective hybrid evaluation of language models. *Advances in Neural Information Processing Systems*, 37:111489–111514, 2024. (page 21)
- Paul Glasserman. *Monte Carlo Methods in Financial Engineering*, volume 53. Springer, New York, 2004. (page 3)
- Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, et al. Bootstrap your own latent—a new approach to self-supervised learning. *Advances in neural information processing systems*, 33:21271–21284, 2020. (page 24)
- Léo Grinsztajn, Klemens Flöge, Oscar Key, Felix Birkel, Philipp Jund, Brendan Roof, Benjamin Jäger, Dominik Safaric, Simone Alessi, Adrian Hayler, et al. TabPFN-2.5: Advancing the state of the art in tabular foundation models. *arXiv preprint arXiv:2511.08667*, 2025. (pages 9 and 23)
- Peter Hall, Eugene Lukács, Zygmunt William Birnbaum, and Chris C. Heyde. Martingale limit theory and its application. 1980. URL <https://api.semanticscholar.org/CorpusID:120148048>. (page 61)
- Max Hamilton, Jinlin Lai, Wenlong Zhao, Subhansu Maji, and Daniel Sheldon. Active measurement: Efficient estimation at scale. *arXiv preprint arXiv:2507.01372*, 2025. (page 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. (page 26)
- Yuheng Huang, Jiayang Song, Qiang Hu, Felix Juefei-Xu, and Lei Ma. Actracer: Active testing of large language model via multi-stage sampling. *ACM Transactions on Software Engineering and Methodology*, 35(3):1–30, 2026. (page 21)
- J. D. Hunter. Matplotlib: A 2d graphics environment. *Computing in Science & Engineering*, 9(3):90–95, 2007. doi: 10.1109/MCSE.2007.55. (page 26)

- Alexander Immer, Maciej Korzepa, and Matthias Bauer. Improving predictions of bayesian neural nets via local linearization. In *International conference on artificial intelligence and statistics*, pages 703–711. PMLR, 2021. (page 24)
- 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. (page 12)
- Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *International conference on machine learning*, pages 4904–4916. PMLR, 2021. (page 1)
- John M. Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Žídek, Anna Potapenko, Alex Bridgland, Clemens Meyer, Simon A A Kohl, Andy Ballard, Andrew Cowie, Bernardino Romera-Paredes, Stanislav Nikolov, Rishub Jain, Jonas Adler, Trevor Back, Stig Petersen, David Reiman, Ellen Clancy, Michal Zielinski, Martin Steinegger, Michalina Pacholska, Tamas Berghammer, Sebastian Bodenstein, David Silver, Oriol Vinyals, Andrew W. Senior, Koray Kavukcuoglu, Pushmeet Kohli, and Demis Hassabis. Highly accurate protein structure prediction with alphafold. *Nature*, 596:583 – 589, 2021. URL <https://api.semanticscholar.org/CorpusID:235959867>. (page 1)
- 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, 2012. doi: 10.1109/ICDM.2012.161. (pages 1 and 12)
- Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. (page 24)
- Simon Kornblith, Jonathon Shlens, and Quoc V Le. Do better imagenet models transfer better? In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2661–2671, 2019. (page 24)
- 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. (pages 1, 2, 3, 6, 8, 9, 12, 23, 24, 25, 33, and 35)
- 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. (pages 1, 3, 9, 12, 23, 24, 25, 33, and 35)
- Alex Krizhevsky. Learning multiple layers of features from tiny images. pages 32–33, 2009. URL <https://www.cs.toronto.edu/~kriz/learning-features-2009-TR.pdf>. (pages 9, 24, and 26)
- 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. (page 12)
- Ya Le and Xuan S. Yang. Tiny imagenet visual recognition challenge. 2015. URL <https://api.semanticscholar.org/CorpusID:16664790>. (pages 9, 24, and 26)
- Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pages 19730–19742. PMLR, 2023. (page 1)
- Peng Liu, Lizhe Wang, Rajiv Ranjan, Guojin He, and Lei Zhao. A survey on active deep learning: From model driven to data driven. *ACM Comput. Surv.*, 54(10s), September 2022. ISSN 0360-0300. doi: 10.1145/3510414. URL <https://doi.org/10.1145/3510414>. (page 1)

- David J. C. MacKay. Bayesian interpolation. *Neural Computation*, 4(3):415–447, 05 1992. ISSN 0899-7667. doi: 10.1162/neco.1992.4.3.415. URL <https://doi.org/10.1162/neco.1992.4.3.415>. (pages 9 and 24)
- TorchVision maintainers and contributors. Torchvision: Pytorch’s computer vision library. <https://github.com/pytorch/vision>, 2016. (pages 24 and 26)
- Mitsuru Matsuura and Satoshi Hara. Active model selection: A variance minimization approach. In *NeurIPS 2023 Workshop on Adaptive Experimental Design and Active Learning in the Real World*, 2023. URL <https://openreview.net/forum?id=vBwftUDTtz>. (page 21)
- Joshua Maynez, Priyanka Agrawal, and Sebastian Gehrmann. Benchmarking large language model capabilities for conditional generation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9194–9213, 2023. (page 21)
- Wes McKinney. Data structures for statistical computing in python. *SciPy 2010*, 2010. doi: 10.25080/Majora-92bf1922-00a. URL <https://doi.org/10.25080/Majora-92bf1922-00a>. (page 26)
- Muhammad Naeem and Sohail Asghar. KEGG Metabolic Relation Network (Directed). UCI Machine Learning Repository, 2011a. DOI: <https://doi.org/10.24432/C5CK52>. (page 26)
- Muhammad Naeem and Sohail Asghar. KEGG Metabolic Reaction Network (Undirected). UCI Machine Learning Repository, 2011b. DOI: <https://doi.org/10.24432/C5G609>. (page 26)
- Phuc Nguyen, Deva Ramanan, and Charless Fowlkes. Active testing: An efficient and robust framework for estimating accuracy. In *International Conference on Machine Learning*, pages 3759–3768. PMLR, 2018. (page 12)
- Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023. (page 24)
- The pandas development team. pandas-dev/pandas: Pandas, February 2020. URL <https://doi.org/10.5281/zenodo.3509134>. (page 26)
- F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011. (pages 9, 23, 24, 26, 27, and 28)
- Jo Plested, Musa Phiri, and Tom Gedeon. Deep transfer learning for image classification: a survey. *Artificial Intelligence Review*, 2026. (page 24)
- Felipe Maia Polo, Lucas Weber, Leshem Choshen, Yuekai Sun, Gongjun Xu, and Mikhail Yurochkin. tinybenchmarks: evaluating llms with fewer examples. *arXiv preprint arXiv:2402.14992*, 2024. (page 21)
- Alvaro Prat, Leo Zhang, Charlotte M Deane, Yee Whye Teh, and Garrett M Morris. Sigmadock: Untwisting molecular docking with fragment-based se (3) diffusion. *arXiv preprint arXiv:2511.04854*, 2025. (page 1)
- Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 8748–8763, 2021. URL <https://proceedings.mlr.press/v139/radford21a.html>. (pages 1, 9, 24, 25, and 26)
- Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 11 2019. URL <https://arxiv.org/abs/1908.10084>. (pages 9, 25, 26, and 27)

- Hippolyt Ritter, Aleksandar Botev, and David Barber. A scalable laplace approximation for neural networks. In *International Conference on Learning Representations*, 2018. URL <https://openreview.net/forum?id=Skdvd2xAZ>. (page 24)
- James M. Robins, Andrea Rotnitzky, and Lue Ping Zhao. Estimation of regression coefficients when some regressors are not always observed. *Journal of the American Statistical Association*, 89(427):846–866, 1994. ISSN 01621459, 1537274X. URL <http://www.jstor.org/stable/2290910>. (page 22)
- Pablo Romeu-Guallart and Francisco Zamora-Martinez. SML2010. UCI Machine Learning Repository, 2014. DOI: <https://doi.org/10.24432/C5RS3S>. (page 26)
- Gayathri Saranathan, Mahammad Parwez Alam, James Lim, Suparna Bhattacharya, Soon Yee Wong, Martin Foltin, and Cong Xu. DELE: Data efficient LLM evaluation. In *ICLR 2024 Workshop on Navigating and Addressing Data Problems for Foundation Models*, 2024. URL <https://openreview.net/forum?id=I8bsxPWLNF>. (page 21)
- Christoph Sawade, Niels Landwehr, Steffen Bickel, and Tobias Scheffer. Active risk estimation. In *Proceedings of the 27th International Conference on International Conference on Machine Learning, ICML’10*, page 951–958, Madison, WI, USA, 2010a. Omnipress. ISBN 9781605589077. (pages 1 and 12)
- Christoph Sawade, Niels Landwehr, and Tobias Scheffer. Active estimation of f-measures. In J. Lafferty, C. Williams, J. Shawe-Taylor, R. Zemel, and A. Culotta, editors, *Advances in Neural Information Processing Systems*, volume 23. Curran Associates, Inc., 2010b. URL [https://proceedings.neurips.cc/paper\\_files/paper/2010/file/d7a728a67d909e714c0774e22cb806f2-Paper.pdf](https://proceedings.neurips.cc/paper_files/paper/2010/file/d7a728a67d909e714c0774e22cb806f2-Paper.pdf). (page 1)
- Burr Settles. Active learning literature survey. Computer Sciences Technical Report 1648, University of Wisconsin–Madison, 2009. URL <http://axon.cs.byu.edu/~martinez/classes/778/Papers/settles.activelearning.pdf>. (page 1)
- Maria-Eleni Sfyraiki and Jun-Kun Wang. Revisiting active sequential prediction-powered mean estimation. *arXiv preprint arXiv:2604.18569*, 2026. (page 21)
- Sanbao Su, Xin Li, Thang Doan, Sima Behpour, Wenbin He, Liang Gou, Fei Miao, and Liu Ren. Metaat: Active testing for label-efficient evaluation of dense recognition tasks. In *European Conference on Computer Vision*, pages 325–342. Springer, 2024. (page 21)
- Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, et al. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023. (page 1)
- Michael E. Tipping. Sparse bayesian learning and the relevance vector machine. *J. Mach. Learn. Res.*, 1: 211–244, September 2001. ISSN 1532-4435. doi: 10.1162/15324430152748236. URL <https://doi.org/10.1162/15324430152748236>. (page 23)
- Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023. (page 1)
- Pauli Virtanen, Ralf Gommers, Travis E Oliphant, Matt Haberland, Tyler Reddy, David Cournapeau, Evgeni Burovski, Pearu Peterson, Warren Weckesser, Jonathan Bright, et al. Scipy 1.0: fundamental algorithms for scientific computing in python. *Nature methods*, 17(3):261–272, 2020. (page 26)
- Rajan Vivek, Kawin Ethayarajh, Diyi Yang, and Douwe Kiela. Anchor points: Benchmarking models with much fewer examples. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1576–1601, 2024. (page 21)

- Shuohang Wang, Yang Liu, Yichong Xu, Chenguang Zhu, and Michael Zeng. Want to reduce labeling cost? gpt-3 can help. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 4195–4205, 2021. (page 1)
- Ross Wightman. PyTorch Image Models. <https://github.com/huggingface/pytorch-image-models>, 2019. (page 26)
- Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45. Association for Computational Linguistics, 2020. doi: 10.18653/v1/2020.emnlp-demos.6. URL <https://aclanthology.org/2020.emnlp-demos.6>. (page 26)
- Emine Yilmaz, Peter Hayes, Raza Habib, Jordan Burgess, and David Barber. Sample efficient model evaluation. *arXiv preprint arXiv:2109.12043*, 2021. (pages 1 and 12)
- Dayou Yu, Weishi Shi, and Qi Yu. Actively testing your model while it learns: Realizing label-efficient learning in practice. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, editors, *Advances in Neural Information Processing Systems*, volume 36, pages 31404–31427. Curran Associates, Inc., 2023. URL [https://proceedings.neurips.cc/paper\\_files/paper/2023/file/63ef323523f3be8b58ed9277cc747485-Paper-Conference.pdf](https://proceedings.neurips.cc/paper_files/paper/2023/file/63ef323523f3be8b58ed9277cc747485-Paper-Conference.pdf). (pages 1 and 21)
- Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 11975–11986, 2023. (page 1)
- Tijana Zrnica and Emmanuel J Candès. Active statistical inference. *arXiv preprint arXiv:2403.03208*, 2024a. (pages 21 and 33)
- Tijana Zrnica and Emmanuel J Candès. Cross-prediction-powered inference. *Proceedings of the National Academy of Sciences*, 121(15):e2322083121, 2024b. (page 21)

# Appendix Contents

---

<b>A</b>	<b>Extended Related Work</b>	<b>21</b>
<b>B</b>	<b>Further Details on Experimental Setup</b>	<b>23</b>
B.1	Section 5.2 . . . . .	23
B.2	Section 5.3 . . . . .	25
B.3	Section 5.4 . . . . .	25
B.4	Computational Resources and Licenses . . . . .	25
<b>C</b>	<b>Additional Results</b>	<b>27</b>
C.1	Empirical Verification of $\lambda^\dagger$ . . . . .	27
C.2	Additional Ablations . . . . .	27
C.3	Additional Metrics . . . . .	29
C.4	Additional Comparisons . . . . .	31
<b>D</b>	<b>Additional Details</b>	<b>34</b>
D.1	On the Choice of the Surrogate Score in Section 2.1 . . . . .	34
D.2	Simplified Expressions for the Proposals of LURE and PPAT . . . . .	34
D.3	Extended Background . . . . .	35
<b>E</b>	<b>Mathematical Results</b>	<b>37</b>
E.1	Definitions . . . . .	37
E.2	Generalising LURE with Triangular-Arrays . . . . .	37
E.3	Finite-Sample Unbiasedness and Variance . . . . .	40
E.4	Myopic Oracle Proposals . . . . .	46
E.5	Asymptotics . . . . .	48

## A Extended Related Work

---

Here, we provide a more detailed discussion of related work.

**Applications and extensions of active testing.** Active testing has recently been extended to a range of settings. In the LLM evaluation setting, Huang et al. (2026) proposed *AcTracer* which partitions the test pool via hidden-state representations of the target LLM and performs multi-stage sampling using online variance estimates and confidence traces. Su et al. (2024) addressed dense computer vision tasks such as segmentation and object detection. Yu et al. (2023) incorporated active testing into model training, and Ashury-Tahan et al. (2024); Matsuura and Hara (2023) studied active testing for model selection.

**Benchmark subsampling.** A related but distinct line of work reduces evaluation cost by constructing smaller test sets from existing *labelled* LLM benchmarks. Maynez et al. (2023) showed that uniform subsampling can preserve stable model rankings in some settings, and Polo et al. (2024); Saranathan et al. (2024); Vivek et al. (2024) developed more sophisticated dataset-reduction methods. These approaches are complementary to ours in their goal of reducing evaluation cost, but address a different setting: they assume an already labelled test set and aim to support comparisons across many models, whereas we study the problem of acquiring new labels to estimate the risk of a given model on a fixed unlabelled test pool.

**Active measurement.** Also related to our work is *Active Measurement* (AM, (Hamilton et al., 2025)), which also builds on LURE and combines model predictions with importance sampling without replacement, but targets scientific measurements (e.g., total bird counts) rather than a model’s test risk. Our framework uses predictions in fundamentally different ways: AM uses AI predictions to define the proposal itself and iteratively refines the predictor with acquired labels, whereas PPAT uses a fixed proxy as a *control variate* that residualises the loss, with the proposal constructed separately from a surrogate tailored to this residualised objective. Moreover, we note that AM assumes their proxy model can be updated as more labels are acquired, which is less general than our setup<sup>5</sup>. In principle, AM’s iterative proposal refinement and weighting schemes could also be combined with PPAT’s control-variate estimator.

**Active statistical inference.** *Active statistical inference* (ASI, Zrnic and Candès (2024a)) considers adaptive label sampling for statistical inference, proposing batch and sequential methods that yield valid confidence intervals for  $M$ -estimation problems. Recent work revisits the sequential mean-estimation setting and provides non-asymptotic guarantees for online policies that choose the probability of querying each ground-truth label (Sfyraki and Wang, 2026). While risk estimation is a particular instance of  $M$ -estimation, their framework is not framed around model evaluation, and their sequential formulation is closer to online selective labelling than to pool-based acquisition: labels are acquired point-by-point from a stream rather than selected from the full remaining pool. This differs from our setting, where we focus on pool-based risk estimation and use black-box predictions as control variates within an unbiased active-testing estimator. We compare with ASI in §C.4.3 and find that we can outperform their approach in their “active” setting while achieving similar performance in the “batch” setting.

**Prediction-powered inference and classical augmented estimators.** *Prediction-powered inference* (PPI, Angelopoulos et al. (2023a)) was introduced to construct tighter confidence intervals in the semi-supervised setting by leveraging machine learning predictions, and has since been extended in several directions: PPI++ (Angelopoulos et al., 2023b) provides a computationally efficient, loss-based formulation with a power-tuning parameter  $\lambda$ ; Stratified PPI (Fisch et al., 2024) improves efficiency via data stratification; and Cross-PPI (Zrnic and Candès, 2024b) trains the predictor within the same prediction-powered pipeline. PPI is also related to classical augmented estimators: in survey sampling, regression estimators use auxiliary

---

<sup>5</sup>Indeed, our approach and theoretical results naturally transfer to the setting where our proxy model can also be updated as more labels are acquired.

variables to improve the efficiency of the estimator (Cochran, 1977); in causal inference, augmented inverse probability weighted estimators combine inverse-probability weighting with an outcome-model correction and are doubly robust when either the propensity model or the outcome model is correctly specified (Robins et al., 1994; Bang and Robins, 2005).

Our work is related to these works in spirit but differs in both objective and setting: rather than semi-supervised inference with i.i.d. labelled and unlabelled data, we study active risk estimation over a finite test pool, using black-box predictions as a control variate inside an importance-weighted active-testing estimator.

## B Further Details on Experimental Setup

---

Here, we provide further details for the experiments in §5.

### B.1 Section 5.2

Here, we provide details for the experiments in §5.2.

#### B.1.1 Regression Experiments

**Datasets.** Below, we provide further details on the UCI datasets used. We use  $n$  to denote the full dataset size and  $d$  to denote the number of features.

- **Keggundirected** ( $n = 63,608$ ,  $d = 27$ ): A biology dataset derived from KEGG metabolic pathways represented as undirected reaction networks. The inputs are graph-level descriptors of pathway topology.
- **Keggdirected** ( $n = 48,827$ ,  $d = 20$ ): A related KEGG dataset where metabolic pathways are represented as directed relation networks. The features summarise structural properties of these directed pathway graphs.
- **Bike** ( $n = 17,379$ ,  $d = 17$ ): Hourly bike-rental data from the Capital Bikeshare system in 2011–2012. The covariates include weather, seasonal, and calendar information.
- **Sml** ( $n = 4,137$ ,  $d = 26$ ): A smart-home monitoring dataset collected over approximately 40 days in a domotic house. The inputs include indoor, outdoor, weather, humidity, lighting, and actuator measurements.

We load each of the above datasets from the `uci_datasets` repository. For each dataset, we randomly sample a fixed training set of 250 points which we use to train the model that we wish to evaluate. The remainder of the dataset is used as the test pool.

**Models.** We choose  $f$  to be a Gaussian process regression model with a Radial Basis Function (RBF) kernel. We optimise its hyperparameters using the marginal likelihood, with the lengthscale initialised at 1.0. We use the `Scikit-Learn` implementation (Pedregosa et al., 2011).

For our surrogate model, we use a Bayesian linear regression model (Tipping, 2001) with the default hyperparameters from `Scikit-Learn` (Pedregosa et al., 2011) and ablate with different choices of surrogate models in §C.2.2. We again use the `Scikit-Learn` implementation.

For the choice of  $g$  for **PPAT**, we use the **TabPFN-2.5** (Grinsztajn et al., 2025) foundation model and ablate with different choices of  $g$  in §C.2.1. To make predictions on our test pool, we use in-context learning (Grinsztajn et al., 2025), where we use our training data as the in-context data.

**Baselines.** We compare **PPAT** with **LURE** (Kossen et al., 2021), **ASE** Kossen et al. (2022), and **Random** (sampling points randomly without replacement and using the unweighted estimator in (2)). All approaches use the same surrogate model. For **PPAT**, we compare with the three different choices of  $\lambda$  described in §3.3. To compute the acquisition scores for **ASE**, we use 100 Monte Carlo samples for each point: 10 samples from the posterior on the model parameters, and 10 samples from the model likelihood. See §D.3.1 for more details about the **ASE** acquisition function.

For **PPAT** and **LURE**, we compute the acquisition scores using the simplified formulas discussed in §D.2, which can be readily computed using the mean prediction and predictive variance output by the surrogate model. For the **PPAT** plug-in estimate of  $\lambda$ ,  $\hat{\lambda}_M$ , we initialise it at 0.5 and update it after every 100 acquired points using (14).

**Active testing setup and metrics.** We take our loss function to be the squared error  $\ell$ . We use a budget of  $M = 500$  points and update our surrogate model after each newly acquired point. Each experiment is run for 1000 trials. Following Kossen et al. (2021, 2022), we report the median squared error. The mean error is presented in §C.3.1. Here, the error is defined as the difference between our estimate of the risk and the true risk on the test pool.

### B.1.2 Classification Experiments

**Datasets.** Below, we provide further details about the classification dataset used. We provide the size of the train and test sets used as well as the resolution of the original images. Note that for Tiny-ImageNet we use the validation set as the test set, this is because the original test set is *unlabelled*.

- CIFAR-10 (Krizhevsky (2009), 50,000 train, 10,000 test;  $32 \times 32 \times 3$ ): A natural-image classification dataset with 10 object classes. Each class contains 6,000 images.
- CIFAR-100 (Krizhevsky (2009), 50,000 train, 10,000 test;  $32 \times 32 \times 3$ ): A more fine-grained version of CIFAR-10 with 100 object classes. Each class contains 600 images, making the classification task harder.
- Tiny-ImageNet (Le and Yang (2015), 100,000 train, 10,000 test;  $64 \times 64 \times 3$ ): A reduced version of ImageNet with 200 classes. It contains higher-resolution images and substantially more classes than CIFAR-10 and CIFAR-100.

We load each dataset from the `torchvision` (maintainers and contributors, 2016) library. For each dataset, we respect the original splits and use the provided training set and test set, where the test set is treated as the test pool and the training set is used to train  $f$ .

**Models.** In line with standard practice in modern image-classification pipelines (Chen et al., 2020; Grill et al., 2020; Radford et al., 2021; Oquab et al., 2023; Plested et al., 2026), we use a fixed pretrained image encoder to map each input  $\mathbf{x}$  to an embedding  $e(\mathbf{x})$ , and train lightweight prediction models on top of these embeddings. Here, the encoder  $e$  is a map  $e : \mathbb{R}^{d_1} \rightarrow \mathbb{R}^{d_2}$ , where  $d_1$  is the size of the input space,  $d_2$  is the size of the embeddings and  $d_2 \ll d_1$ . In all of our classification experiments, we use embeddings from a CLIP image encoder (Radford et al., 2021). Specifically, we use the ViT-B-32 encoder with model card `openai/clip-vit-base-patch32` from the CLIP library.

The model  $f$  whose risk we evaluate is a linear classifier trained on the CLIP embeddings. This corresponds to the standard linear probing setup, where a linear prediction head is trained on top of a frozen representation to evaluate the usefulness of the learned features (Alain and Bengio, 2016; Kornblith et al., 2019; Chen et al., 2020). For this, we use the logistic regression model class from `Scikit-Learn` (Pedregosa et al., 2011), where we set the maximum number of iterations to 300 and keep the remaining hyperparameters at their default values.

For the active-testing surrogate, it is important to use a model with well-calibrated predictive uncertainties, since our acquisition rules depend on expectations under the surrogate predictive distribution  $\pi_m(\cdot | \mathbf{x})$ . We therefore use a Laplace-approximated Bayesian neural network (MacKay, 1992; Daxberger et al., 2021), implemented using the `laplace-redux` library. The surrogate is an MLP trained on the CLIP embeddings, with three hidden layers of width 128 for CIFAR-10 and CIFAR-100, and five hidden layers of width 128 for Tiny-ImageNet. We train the network for 200 epochs using Adam (Kingma and Ba, 2014) with batch size 64, learning rate  $10^{-3}$ , and weight decay  $10^{-4}$ . We then fit a last-layer Laplace approximation with a Kronecker-factored structure (Ritter et al., 2018), optimising the prior precision by maximising the marginal likelihood. For prediction, we use the linearised predictive distribution induced by the Laplace posterior, which gives an approximate Gaussian distribution over the latent network outputs around the trained weights (Immer et al., 2021). We then obtain class probabilities by Monte Carlo through the softmax link function using 100 samples. We again ablate with different surrogates in §C.2.2.

For the choice of  $g$ , we use the zero-shot predictions of a pretrained CLIP (Radford et al., 2021) model, specifically the ViT-L-14 variant from the `sentence-transformers` library (Reimers and Gurevych, 2019) with model card ViT-L-14::datacomp\_xl\_s13b\_b90k. We again ablate with different choices of  $g$  in §C.2.1.

**Baselines.** All our baselines use the same surrogate model and follow the same setup as for the regression datasets. To compute the acquisition scores for ASE, we use 100 samples from the posterior on the model parameters. Note that for classification problems we do not need to approximate the expectation over  $Y$  as  $y$  is discrete.

**Active testing setup and metrics.** We take our loss to be cross-entropy. The rest of our setup is the same as for the regression problems, with the exception that we do not update our surrogate model, in line with the image-classification experiments in (Kossen et al., 2021, 2022).

## B.2 Section 5.3

Here, we provide details for the experiments in Figs. 3 and 4.

**The choice of the acquisition strategy.** The regression and classification experiments in Fig. 4 follow the same setup as detailed in §B.1.1 and B.1.2. The PPAT Acq. + LURE approach uses the PPAT proposal (13) and the LURE estimator (4); the PPAT Acq. + Empirical approach again uses the same PPAT proposal but with the unweighted estimator (2); PPAT uses its standard estimator and acquisition rule. We use  $\lambda = 1$  for all the approaches.

**The choice of the estimator.** The regression and classification experiments in Fig. 3 follow the same setup as detailed in §B.1.1 and B.1.2. The LURE Acq. + PPI approach uses the LURE proposal and the PPI estimator (8)<sup>6</sup>; the Random + PPI approach uses random sampling in combination with the PPI estimator; PPAT uses its standard estimator and acquisition rule. We use  $\lambda = 1$  for all the approaches.

## B.3 Section 5.4

Here, we provide details for the coverage experiments in §5.4 (Fig. 5). The regression and classification experiments follow the same setup as detailed in §B.1.1 and B.1.2. We use an error level of  $\delta = 0.1$  and compute the confidence intervals using (23), noting that Random is a special case of LURE with  $V_m = 1, \forall m$ . Coverage is computed with respect to the true test risk on the full pool for each labelling budget. We report the fraction of the 1000 trials for which the confidence interval contains this full-pool test risk.

## B.4 Computational Resources and Licenses

All our experiments were run on a single NVIDIA H100 80GB GPU. Below, we provide the licenses for the packages and datasets used in this work.

---

<sup>6</sup>Note that we slightly abuse terminology here as (8) is not the PPI estimator described in (Angelopoulos et al., 2023a), but instead the unbiased, importance weighted version of it. For random sampling, the two coincide.

Table 1: Software packages used in our experiments.

Project	Citation	License	URL
NumPy	Harris et al. (2020)	BSD	<a href="https://numpy.org">https://numpy.org</a>
SciPy	Virtanen et al. (2020)	BSD	<a href="https://scipy.org">https://scipy.org</a>
scikit-learn	Pedregosa et al. (2011)	BSD	<a href="https://scikit-learn.org">https://scikit-learn.org</a>
Matplotlib	Hunter (2007)	PSF-based	<a href="https://matplotlib.org">https://matplotlib.org</a>
pandas	pandas development team (2020); McKinney (2010)	BSD	<a href="https://pandas.pydata.org">https://pandas.pydata.org</a>
uci_datasets	Evans (2022)	MIT	<a href="https://github.com/treforevans/uci_datasets">https://github.com/treforevans/uci_datasets</a>
CLIP	Radford et al. (2021)	MIT	<a href="https://github.com/openai/CLIP">https://github.com/openai/CLIP</a>
sentence-transformers	Reimers and Gurevych (2019)	Apache-2.0	<a href="https://www.sbert.net">https://www.sbert.net</a>
torchvision	maintainers and contributors (2016)	BSD	<a href="https://pytorch.org/vision">https://pytorch.org/vision</a>
laplace-redux	Daxberger et al. (2021)	MIT	<a href="https://github.com/runame/laplace-redux">https://github.com/runame/laplace-redux</a>
wandb	Biewald (2020)	MIT	<a href="https://wandb.ai">https://wandb.ai</a>
transformers	Wolf et al. (2020)	Apache-2.0	<a href="https://huggingface.co/docs/transformers">https://huggingface.co/docs/transformers</a>
timm	Wightman (2019)	Apache-2.0	<a href="https://huggingface.co/docs/timm">https://huggingface.co/docs/timm</a>

Table 2: Datasets used in our experiments.

Dataset	Citation	License	URL
CIFAR-10	Krizhevsky (2009)	Not specified	<a href="https://www.cs.toronto.edu/~kriz/cifar.html">https://www.cs.toronto.edu/~kriz/cifar.html</a>
CIFAR-100	Krizhevsky (2009)	Not specified	<a href="https://www.cs.toronto.edu/~kriz/cifar.html">https://www.cs.toronto.edu/~kriz/cifar.html</a>
Tiny-ImageNet	Le and Yang (2015)	Not specified	<a href="http://cs231n.stanford.edu/tiny-imagenet-200.zip">http://cs231n.stanford.edu/tiny-imagenet-200.zip</a>
keggdirected	Naeem and Asghar (2011a)	CC BY 4.0	<a href="https://doi.org/10.24432/C5CK52">https://doi.org/10.24432/C5CK52</a>
keggundirected	Naeem and Asghar (2011b)	CC BY 4.0	<a href="https://doi.org/10.24432/C5G609">https://doi.org/10.24432/C5G609</a>
bike	Fanaee-T (2013)	CC BY 4.0	<a href="https://doi.org/10.24432/C5W894">https://doi.org/10.24432/C5W894</a>
sml	Romeu-Guallart and Zamora-Martinez (2014)	CC BY 4.0	<a href="https://doi.org/10.24432/C5RS3S">https://doi.org/10.24432/C5RS3S</a>

## C Additional Results

### C.1 Empirical Verification of $\lambda^\dagger$

Here, we empirically validate the suitability of our target  $\lambda^\dagger$ . To do this, we compare **PPAT** run with the fixed value of  $\lambda^\dagger$ , where  $\lambda^\dagger$  is computed on the full pool, with **PPAT** run with  $\lambda \in \{0.1, 0.25, 0.5, 0.75, 1.0, 1.5, 2.5, 5.0\}$ . For the experiments, we use one regression dataset (**Keggdirected**) and one classification dataset (**CIFAR-10**), and follow the same setup in §B.1.1 and §B.1.2.

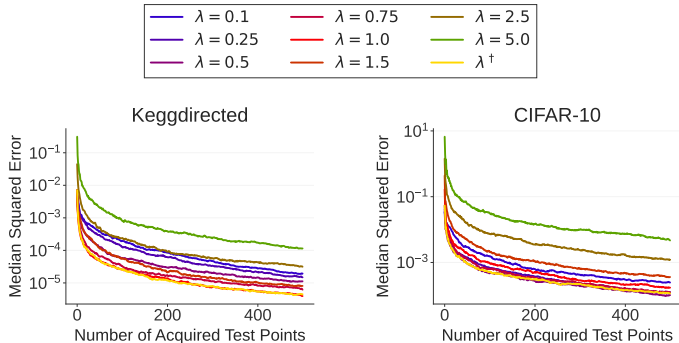


Figure 6: Plot of **PPAT** for a range of fixed  $\lambda$  values and  $\lambda^\dagger$  for the datasets **Keggdirected** and **CIFAR-10**.  $\lambda^\dagger$  is computed on the full pool. The plots report the median squared error over 1000 trials.

Fig. 6 plots the median squared error across 1000 trials for all the different values of  $\lambda$  described previously as well as  $\lambda^\dagger$ . As we can see,  $\lambda^\dagger$  achieves the same median squared error as the best  $\lambda$  for both datasets. Besides the discussion in §3.3, this further supports the use of  $\lambda^\dagger$  as a suitable  $\lambda$  to target instead of the intractable  $\lambda^*$ .

### C.2 Additional Ablations

Here, we ablate with different proxy models and different surrogate models.

#### C.2.1 Different Proxies

To ablate with different proxy models, we use one regression dataset (**Keggdirected**) and one classification dataset (**CIFAR-10**), and follow the same setup as in §B.1.1 and B.1.2. For the regression dataset, we compare with the **TabPFN-2.0** foundation model, where we make predictions on the pool using in-context learning with the context set being our training set in the same way as before. For the classification case, we compare with the **ViT-B-32** and **ViT-B-16** variants of the **CLIP** model from the **sentence-transformers** library<sup>7</sup> (Reimers and Gurevych, 2019).

From Fig. 7, we see that across both datasets, **PPAT** remains robust to different choices of the proxy model. Indeed, for all proxy choices, **PPAT** still retains the lowest median squared error for all choices of  $\lambda$ .

#### C.2.2 Different Surrogate Models

To ablate with different surrogate models, we use one regression dataset (**Keggdirected**) and one classification dataset (**CIFAR-10**), and follow the same setup as in §B.1.1 and B.1.2. For the regression dataset, we use a random forest regression model from **Scikit-Learn** (Pedregosa et al., 2011) with 300 trees, with the

<sup>7</sup>Their corresponding model cards are `ViT-B-32::datacomp_xl_s13b_b90k` and `ViT-B-16::datacomp_xl_s13b_b90k`

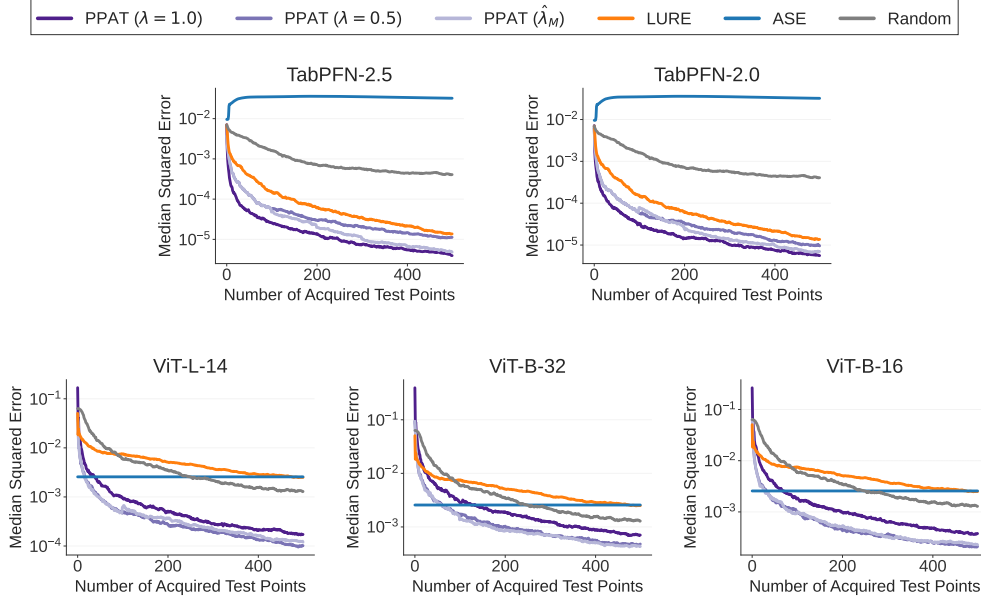


Figure 7: Ablation with different proxy models comparing **PPAT** with **Random**, **ASE**, **LURE**. **Top row**: results on the **Keggdirected** dataset with **TabPFN-2.5** and **TabPFN-2.0** as the proxies. **Bottom row**: results on the **CIFAR-10** dataset with the **ViT-L-14**, **ViT-B-32**, **ViT-B-16** CLIP models as the proxies. Plots show the median squared error across 1000 trials.

remainder of the hyperparameters at their default values. Similarly, for the classification dataset, we use a random forest classification model from **Scikit-Learn** (Pedregosa et al., 2011) with 500 trees, with the remainder of the hyperparameters at their default values.

From Fig. 8, we see that across both datasets **PPAT** remains robust to different choices of surrogate models. Indeed, for all surrogate choices, **PPAT** still retains the lowest median squared error for all choices of  $\lambda$ , except for the random forest model on **CIFAR-10** where  $\lambda = 1$  performs similarly to **LURE**.

Moreover, on the **Keggdirected** dataset, **ASE** outperforms **Random** when using the random forest surrogate and improves as more labels are acquired. This suggests that the poor behaviour of **ASE** with the Bayesian linear model surrogate is not only due to active sampling itself, but also due to surrogate misspecification: because **ASE** estimates the risk through the surrogate predictive model, errors in this model can translate directly into bias in the risk estimate. The random forest surrogate appears to approximate the losses on this tabular regression task more accurately, reducing this bias enough for the additional acquired labels to improve the estimate. On **CIFAR-10**, however, **ASE** performs substantially worse than **Random** when using the random forest surrogate. This suggests that the random forest is not sufficiently expressive for this CLIP-embedding classification task, so the bias introduced by the **ASE** estimate dominates.

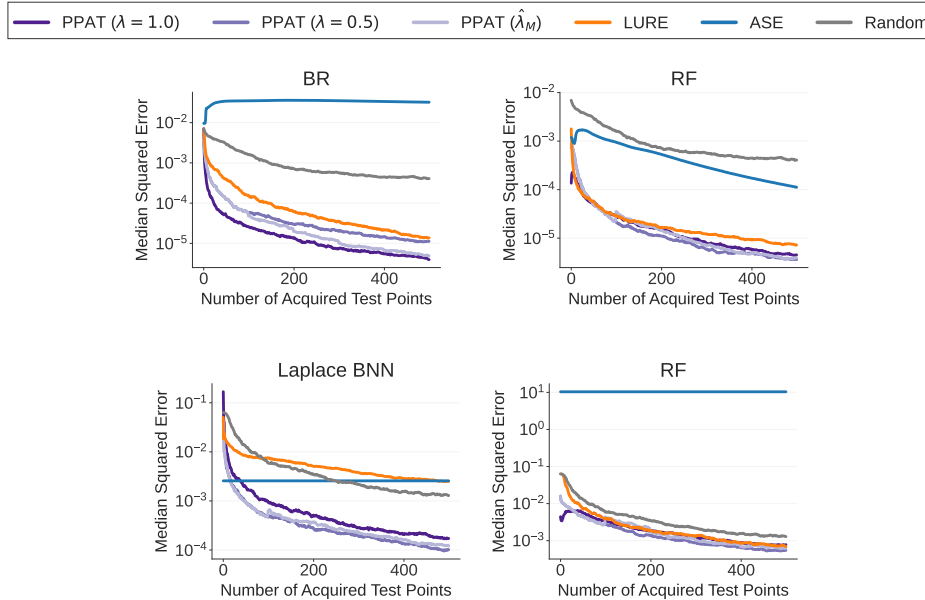


Figure 8: Ablation with different surrogate models comparing **PPAT** with **Random**, **ASE**, **LURE**. **Top row**: results on the **Keggdirected** dataset with a Bayesian linear regression (BR) and random forest (RF) model as the surrogate. **Bottom row**: results on the **CIFAR-10** dataset with a Laplace BNN and random forest model as the surrogate. Plots show the median squared error across 1000 trials. Note that **ASE** results in **biased** estimates of the risk.

### C.3 Additional Metrics

Here, we present additional metrics for the experiments in §5.

#### C.3.1 Mean Error

Figs. 9, 10, 11, 12 show the mean error (bias) for the experiments in Figs. 1, 2, 3, and 4 respectively. The mean error here is the mean difference between the estimate of the risk and the true test risk on the full pool.

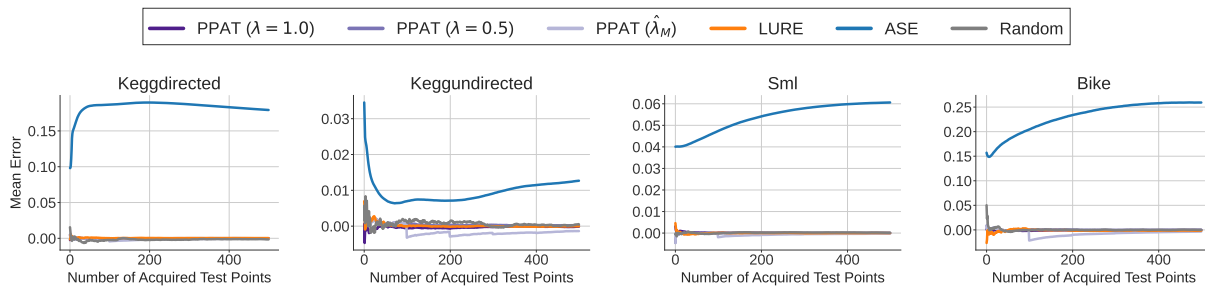


Figure 9: Experiments on UCI datasets comparing **PPAT** with **Random**, **ASE**, **LURE**. Plots show the mean error (bias) across 1000 trials. Note that **ASE** results in **biased** estimates of the risk.

Across Figs. 9 and 10, **ASE** exhibits persistent non-zero mean error, as expected: **ASE** estimates the risk through the surrogate predictive model, and any surrogate misspecification can therefore translate directly

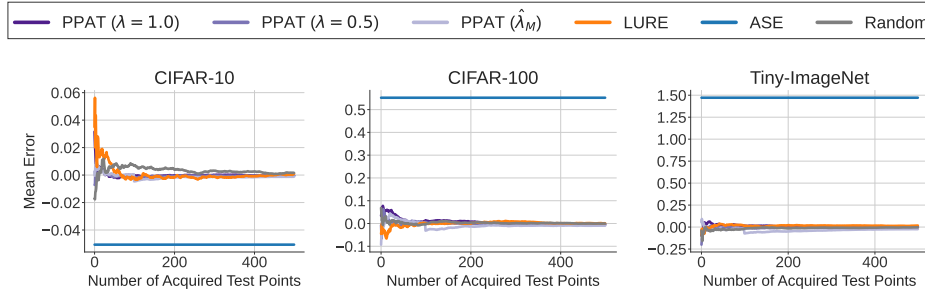


Figure 10: Experiments on classification datasets comparing **PPAT** with **Random**, **ASE**, **LURE**. Plots show the mean error (bias) across 1000 trials. Note that **ASE** results in **biased** estimates of the risk.

into bias in the resulting risk estimate. The plug-in version of **PPAT** can also introduce finite-sample bias because  $\lambda$  is estimated from the actively acquired labels. However, this bias decreases as more labels are acquired, as expected from our theoretical results.

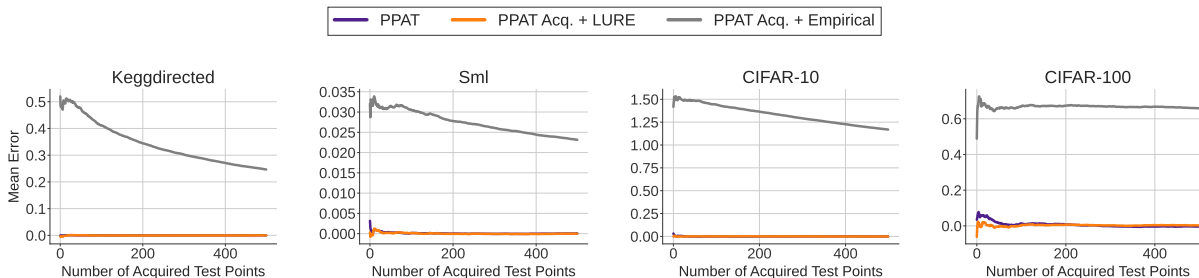


Figure 11: Experiments on UCI and classification datasets studying the influence of our estimator. We compare **PPAT** with  $\lambda = 1$  to using the **PPAT** acquisition strategy,  $Q^{\text{PPAT}}$ , but with the **LURE** estimate of the risk (**PPAT Acq. + LURE**) and the empirical estimate of the risk (**PPAT Acq. + Empirical**). Plots show the mean error (bias) across 1000 trials. Note that **PPAT Acq. + Empirical** results in **biased** estimates of the risk as the data is not randomly sampled.

In Fig. 11, **PPAT Acq. + Empirical** is biased for the expected reason: the data are selected using a non-uniform active proposal, but the estimator does not apply the **LURE** importance-weighting correction. Moreover, the ablations **PPAT Acq. + LURE** in Fig. 11 and **LURE Acq. + PPI** in Fig. 12, show a bias of approximately zero, which is expected as the estimators are still using the **LURE** importance-weighting correction.

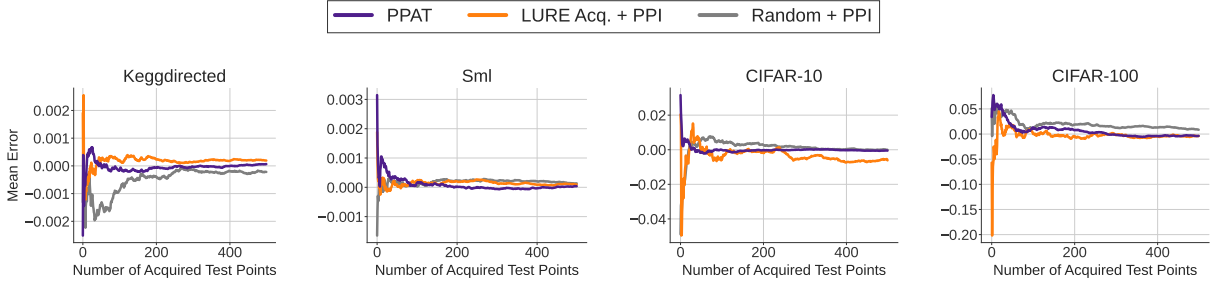


Figure 12: Experiments on UCI and classification datasets studying the influence of our acquisition strategy. We compare **PPAT** with  $\lambda = 1$  to using the PPI estimator with the LURE acquisition strategy (**LURE Acq. + PPI**) and random sampling (**Random + PPI**). Plots show the mean error (bias) across 1000 trials.

### C.3.2 Confidence Interval Widths

Fig. 13 shows that the faster attainment of the target coverage level in Fig. 5 is not simply due to overly conservative confidence intervals. Across representative datasets, we see that **PPAT** achieves lower mean widths than competing approaches, while reaching the desired coverage level at a similar or substantially faster rate. The main exception is **CIFAR-100**, where **PPAT** with  $\lambda = 0.5$  performs similarly to **LURE** in terms of mean width. Overall, this suggests that the prediction-powered control variate improves not only point estimation, but also (frequentist) uncertainty quantification.

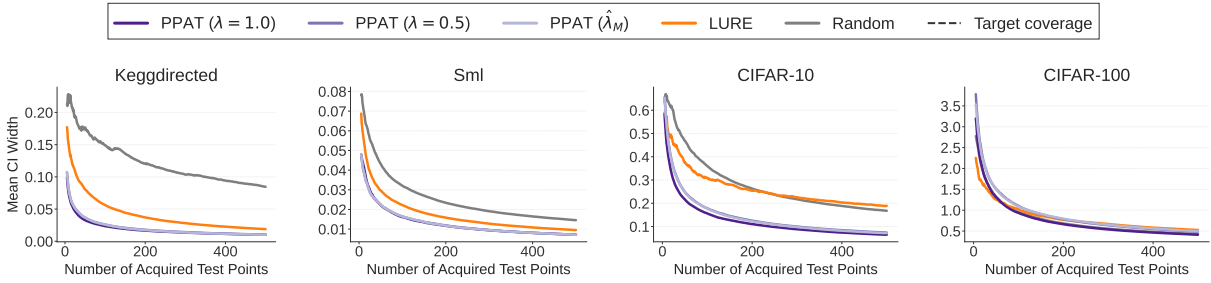


Figure 13: Representative experiments on UCI and classification datasets for the mean widths of the confidence intervals of **PPAT**, **Random**, **LURE**. Plots show the widths of asymptotic confidence intervals constructed with  $\delta = 0.1$  across 1000 trials. The **ASE** baseline is excluded as it does **not** provide asymptotically valid confidence intervals.

## C.4 Additional Comparisons

Here, we present a broader set of comparative analyses against **PPAT**.

### C.4.1 Further Comparisons with Prediction-Powered Inference

Fig. 14 extends Fig. 4 by additionally including the standard **Random** and **LURE** baselines, shown as dashed lines. Across all datasets, **Random + PPI** outperforms both **Random** and **LURE**, further demonstrating that prediction-powered inference provides a strong baseline. Nevertheless, **PPAT** consistently achieves lower median squared error. This shows that the gains of **PPAT** do not come only from using a PPI-like estimator, but also from adaptively sampling points in a way that is tailored to reducing the variance of that estimator.

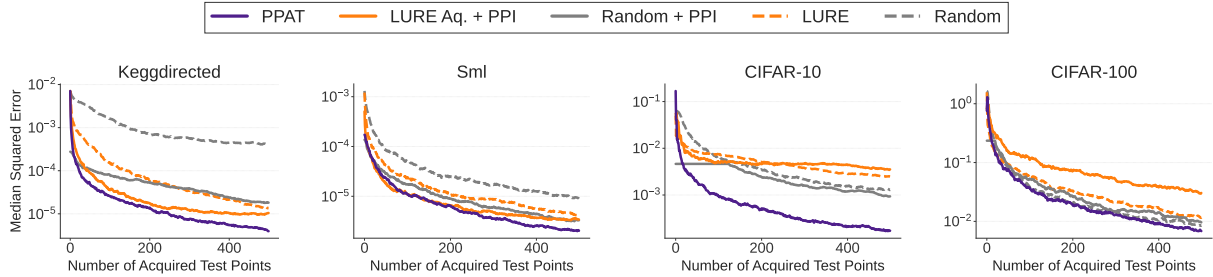


Figure 14: Experiments on UCI and classification datasets studying the influence of our acquisition strategy. We compare **PPAT** with  $\lambda = 1$  to using the PPI estimator with the LURE acquisition strategy (**LURE Acq. + PPI**, —) and random sampling (**Random + PPI**, —), as well as the standard **LURE** (---) and **Random** (---) baselines. Plots show the median squared error across 1000 trials.

#### C.4.2 Comparison with Active Surrogate Estimators Using the Proxy

In §5, we compared **PPAT** with the original formulation of **ASE**. Here, we consider a related baseline that uses the proxy predictions directly. Specifically, since the proxy  $g$  is evaluated on the full test pool, we form the proxy losses  $\tilde{\ell}_i$  and estimate the risk by their full-pool average. We refer to this baseline as **Proxy**. This tests whether the initial proxy model is already accurate enough to replace the unknown labels, rather than using the proxy only as a control variate as in **PPAT**.

This baseline is related to **ASE**, but is not directly comparable to the original **ASE** approach. In our setting, the proxy  $g$  can be a fixed black-box predictor: even if its predictions are available on the entire pool, we may not have access to posterior samples, ensemble members, or other measures of epistemic uncertainty needed to compute the XWED acquisition score. Like **ASE**, however, **Proxy** generally yields a biased estimate of the risk and does not inherit the unbiasedness guarantee of **PPAT**.

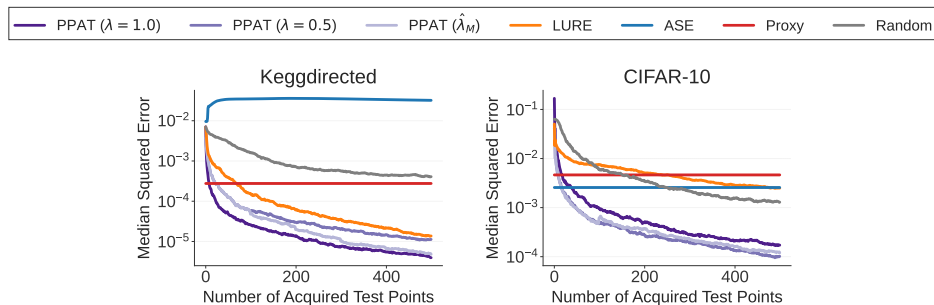


Figure 15: Experiments on **Keggdirected** and **CIFAR-10** comparing **PPAT** with **Random**, **ASE**, **Proxy**, **LURE**. Plots show the median squared error across 1000 trials. Note that **ASE** and **Proxy** result in **biased** estimates of the risk.

We report this comparison on **Keggdirected** and **CIFAR-10**, representing one regression dataset and one classification dataset. We use the same setup as in Figs. 1 and 2 which is described in detail in §B.1.1 and §B.1.2 respectively.

Fig. 15 shows that **PPAT** continues to outperform the **Proxy** baseline for all choices of  $\lambda$  on both datasets. This indicates that, although the proxy predictions contain useful information, directly treating them as labels is not sufficient to match the performance of our approach. Interestingly, the **Proxy** baseline performs better

than [ASE](#) on the regression dataset, but worse than [ASE](#) on the classification dataset. A likely explanation is that, in the regression setting, the [TabPFN-2.5](#) proxy makes predictions using in-context learning with the task-specific training set, making its proxy losses relatively informative for the test pool. In contrast, in the classification setting, the [CLIP](#) proxy predictions are zero-shot and are not adapted to the specific downstream classification task, so the resulting proxy losses may be less aligned with the true losses.

### C.4.3 Comparison with Active Statistical Inference

Here, we compare [PPAT](#) with Active Statistical Inference ([ASI](#), [Zrnic and Candès \(2024a\)](#)). Like before, we use one regression ([Keggdirected](#)) and one classification dataset ([CIFAR-10](#)). For a fair comparison with [PPAT](#), we use the “batch” setting of [ASI](#) for [CIFAR-10](#) and the “active” setting of [ASI](#) for [Keggdirected](#), where we use the default hyperparameters from the paper for both<sup>8</sup>. Note that for the “active” setting of [ASI](#), the proxy model is updated periodically as we collect more labels, whereas for us the proxy model is kept fixed.

Table 3: Median squared error, mean confidence-interval width, and mean coverage on [Keggdirected](#) and [CIFAR-10](#) over 1000 trials, reported to two significant figures.  $\pm$  indicates one standard error; standard errors below  $1 \times 10^{-3}$  are reported as 0.00. Note that [ASE](#) does not provide confidence intervals, so its width and coverage results are omitted.

Method	Keggdirected			CIFAR-10		
	Med. squared error	Mean width	Mean coverage	Med. squared error	Mean width	Mean coverage
<a href="#">ASE</a> ( <a href="#">Kossen et al., 2022</a> )	$3.2 \times 10^{-2} \pm 0.00$	—	—	$2.6 \times 10^{-3} \pm 0.00$	—	—
<a href="#">LURE</a> ( <a href="#">Kossen et al., 2021</a> )	$1.4 \times 10^{-5} \pm 0.00$	$1.9 \times 10^{-2} \pm 0.00$	$0.89 \pm 0.0099$	$2.5 \times 10^{-3} \pm 0.00$	$1.9 \times 10^{-1} \pm 0.0047$	$0.68 \pm 0.015$
Random	$4.1 \times 10^{-4} \pm 0.00$	$7.2 \times 10^{-1} \pm 0.013$	$0.084 \pm 0.0018$	$1.3 \times 10^{-3} \pm 0.00$	$1.7 \times 10^{-1} \pm 0.0011$	$0.83 \pm 0.010$
<a href="#">ASI</a> ( <a href="#">Zrnic and Candès, 2024a</a> )	$4.5 \times 10^{-5} \pm 0.00$	$4.9 \times 10^{-2} \pm 0.00$	$0.99 \pm 0.0024$	$9.8 \times 10^{-5} \pm 0.00$	$6.5 \times 10^{-2} \pm 0.0010$	$0.98 \pm 0.0042$
<a href="#">PPAT</a> ( $\lambda = 1$ )	$4.0 \times 10^{-6} \pm 0.00$	$1.0 \times 10^{-2} \pm 0.00$	$0.90 \pm 0.0096$	$1.7 \times 10^{-4} \pm 0.00$	$6.5 \times 10^{-2} \pm 0.00$	$0.90 \pm 0.0097$
<a href="#">PPAT</a> ( $\lambda = 0.5$ )	$1.1 \times 10^{-5} \pm 0.00$	$1.0 \times 10^{-2} \pm 0.00$	$0.89 \pm 0.010$	$1.0 \times 10^{-4} \pm 0.00$	$7.4 \times 10^{-2} \pm 0.00$	$0.90 \pm 0.0097$
<a href="#">PPAT</a> ( $\lambda = \hat{\lambda}_M$ )	$4.8 \times 10^{-6} \pm 0.00$	$1.0 \times 10^{-2} \pm 0.00$	$0.88 \pm 0.010$	$1.2 \times 10^{-4} \pm 0.00$	$7.1 \times 10^{-2} \pm 0.00$	$0.89 \pm 0.010$

Following the original [ASI](#) paper, we report the final mean coverage and widths of the confidence intervals of both approaches, as well as the median squared error. From Table 3 we see that on [Keggdirected](#), [PPAT](#) improves over [ASI](#) in both risk estimation and confidence-interval width while still attaining the desired coverage level: each of the three [PPAT](#) variants attains a lower median squared error and a substantially narrower mean interval width than [ASI](#), whose intervals are moreover conservative relative to the nominal level. We believe this advantage stems from the pool-based nature of our acquisition strategy: because [PPAT](#) conditions on the full test pool, each acquisition is guided not only by the previously acquired labels but also by the remaining pool, allowing it to use more information when deciding which points to query. By contrast, the active variant of [ASI](#) used here acquires labels in a more streaming manner. On [CIFAR-10](#), however, [ASI](#) performs comparably to, or slightly better than, [PPAT](#) in terms of median squared error and mean interval width, while [PPAT](#) attains coverage closer to the nominal level. This is unsurprising, since the batch variant of [ASI](#) used here is closely related to our approach: both are ultimately built on importance-weighted estimates of the risk but with different proposals.

<sup>8</sup>This includes the mixing hyperparameter  $\tau$  and the finetuning batch size. We did not find any noticeable gains from further tuning of the hyperparameters.

## D Additional Details

### D.1 On the Choice of the Surrogate Score in Section 2.1

Here, we discuss the choice of the surrogate score in §2.1. This also explains our choice in §3.2.

The myopic oracle proposal of §2.1 is given by

$$Q_m^*(i) \propto |\ell_i|. \quad (17)$$

Because the true label  $y_i$  (and hence  $\ell_i$ ) is unknown for  $i \in S_m$ ,  $|\ell_i|$  must be replaced by a surrogate score computed from the surrogate predictive distribution  $\pi_m(\cdot | \mathbf{x}_i)$ . There are two natural candidates:

$$a_m^{\text{abs}}(i) := \mathbb{E}_{Y \sim \pi_m(\cdot | \mathbf{x}_i)} [|\mathcal{L}(f(\mathbf{x}_i), Y)|], \quad (18)$$

$$a_m(i) := \sqrt{\mathbb{E}_{Y \sim \pi_m(\cdot | \mathbf{x}_i)} [\mathcal{L}(f(\mathbf{x}_i), Y)^2]}. \quad (19)$$

The score  $a_m^{\text{abs}}$  is obtained by applying  $\pi_m$  directly to the oracle expression  $|\ell_i|$ ; this is the score that one might expect on first reading. The score  $a_m$  in (19) is the one we use. The two scores coincide only when  $\mathcal{L}(f(\mathbf{x}_i), Y)$  is constant under  $\pi_m(\cdot | \mathbf{x}_i)$ ; otherwise, by Jensen’s inequality applied to  $|\cdot|$ ,

$$a_m^{\text{abs}}(i) \leq a_m(i),$$

with strict inequality whenever the surrogate has nontrivial predictive uncertainty about the loss. Below, we discuss why  $a_m$  is the more principled surrogate of the oracle.

Firstly, note that the oracle proposal  $Q_m^*$  in (17) is derived by minimising the conditional sampling variance of LURE, which is upper-bounded by the second-moment objective

$$\mathbb{E} \left[ \left( \frac{\ell_{I_m}}{N Q_m(I_m)} \right)^2 \mid I_{1:m-1} \right] = \frac{1}{N^2} \sum_{i \in S_m} \frac{\ell_i^2}{Q_m(i)}. \quad (20)$$

See §E.4 for further details. Now, the objective on the right-hand side is *quadratic* in the loss: the per-point contribution is  $\ell_i^2/Q_m(i)$ , not  $|\ell_i|/Q_m(i)$ . The closed-form oracle  $Q_m^*(i) \propto |\ell_i|$  arises only after solving this quadratic problem in closed form via Lagrange multipliers; the absolute value is a consequence of the optimisation, not of the objective.

When  $\ell_i$  is unknown and a surrogate  $\pi_m$  is introduced, the right quantity to substitute into (20) is therefore the predictive expectation of  $\ell_i^2$ , i.e. the second moment, not a point estimate of  $|\ell_i|$ . Concretely, replacing  $\ell_i^2$  by  $\mathbb{E}_{\pi_m(\cdot | \mathbf{x}_i)} [\mathcal{L}(f(\mathbf{x}_i), Y)^2]$  in (20) and re-solving the same minimisation problem yields exactly

$$Q_m^{\text{surr}}(i) \propto \sqrt{\mathbb{E}_{\pi_m(\cdot | \mathbf{x}_i)} [\mathcal{L}(f(\mathbf{x}_i), Y)^2]} = a_m(i).$$

In other words,  $a_m$  is the surrogate score induced by the *variance objective itself*, not by an after-the-fact substitution into the closed-form oracle.

### D.2 Simplified Expressions for the Proposals of LURE and PPAT

Here, we describe the proposals used for LURE and PPAT in our experiments. We discuss the case for the MSE loss – the case for the cross-entropy loss follows similarly.

**LURE.** Let  $f_i := f(\mathbf{x}_i)$ ,  $\mu_{m,i} := \mathbb{E}_{\pi_m(\cdot|\mathbf{x}_i)}[y]$  and  $\sigma_{m,i}^2 := \text{Var}_{\pi_m(\cdot|\mathbf{x}_i)}(y)$ . For the MSE loss, we follow Kossen et al. (2021) and use the proposal

$$Q_m(i) \propto \mathbb{E}_{\pi_m(\cdot|\mathbf{x}_i)}[(f_i - y)^2] = \text{Var}_{\pi_m(\cdot|\mathbf{x}_i)}(y) + \left(\mathbb{E}_{\pi_m(\cdot|\mathbf{x}_i)}[y] - f_i\right)^2 = \sigma_{m,i}^2 + (\mu_{m,i} - f_i)^2.$$

Note that this is not the same as the proposal  $Q_m(i) \propto \sqrt{\mathbb{E}_{\pi_m(\cdot|\mathbf{x}_i)}[(f_i - y)^4]}$  described in §2.1 and §D.1; since we know that our loss is nonnegative, we can instead use the simpler  $Q_m(i) \propto \mathbb{E}_{\pi_m(\cdot|\mathbf{x}_i)}[(f_i - y)^2]$ .

From the above, we can see that under the MSE loss the LURE proposal has the standard bias–variance decomposition form: it favours points for which either the surrogate is uncertain about the label, or the surrogate mean disagrees strongly with the prediction of the model being evaluated.

**PPAT.** Let  $c_i = \tilde{\ell}_i - \tilde{R}$ , and  $\tilde{\ell}_i = (f_i - g(\mathbf{x}_i))^2$ . The proposal  $a_{m,\lambda}$  (12) admits the following decomposition

$$\mathbb{E}_{\pi_m(\cdot|\mathbf{x}_i)}\left[\left(\mathcal{L}(f(\mathbf{x}_i), Y) - \lambda c_i\right)^2\right] = \text{Var}_{\pi_m(\cdot|\mathbf{x}_i)}(\mathcal{L}(f(\mathbf{x}_i), Y)) + \left(\mathbb{E}_{\pi_m(\cdot|\mathbf{x}_i)}[\mathcal{L}(f(\mathbf{x}_i), Y)] - \lambda c_i\right)^2. \quad (21)$$

This shows that the PPAT proposal favours points for which either the surrogate is uncertain about the loss, or the surrogate mean loss is far from the proxy term  $\lambda c_i$ .

Under the MSE loss, (21) becomes

$$\mathbb{E}_{\pi_m(\cdot|\mathbf{x}_i)}\left[\left((f_i - Y)^2 - \lambda c_i\right)^2\right] = \text{Var}_{\pi_m(\cdot|\mathbf{x}_i)}((f_i - Y)^2) + \left(\mathbb{E}_{\pi_m(\cdot|\mathbf{x}_i)}[(f_i - Y)^2] - \lambda c_i\right)^2.$$

Now, using the standard bias–variance decomposition, the second term simplifies to

$$\left(\sigma_{m,i}^2 + (\mu_{m,i} - f_i)^2 - \lambda c_i\right)^2.$$

To simplify the first term, let  $a_{m,i} = f_i - \mu_{m,i}$  and  $z = Y - \mu_{m,i}$  so that  $\mathbb{E}[z] = 0$ ,  $\text{Var}(z) = \sigma_{m,i}^2$ , and  $(f_i - Y)^2 = (a_{m,i} - z)^2$ . Then we have:

$$\text{Var}((f_i - Y)^2) = \text{Var}(z^2 - 2a_{m,i}z) = \left(\mathbb{E}[z^4] - \sigma_{m,i}^4\right) + 4a_{m,i}^2\sigma_{m,i}^2 - 4a_{m,i}\mathbb{E}[z^3].$$

Now, assuming the surrogate predictive distribution is Gaussian<sup>9</sup>, i.e.  $Y | \mathbf{x}_i \sim \mathcal{N}(\mu_{m,i}, \sigma_{m,i}^2)$ , then  $\mathbb{E}[z^3] = 0$  and  $\mathbb{E}[z^4] = 3\sigma_{m,i}^4$ , giving

$$\text{Var}_{\pi_m(\cdot|\mathbf{x}_i)}((f_i - Y)^2) = 2\sigma_{m,i}^4 + 4\sigma_{m,i}^2(\mu_{m,i} - f_i)^2.$$

Therefore, under the MSE loss and a Gaussian surrogate, the PPAT proposal simplifies to

$$Q_{m,\lambda}^{\text{PPAT}}(i) \propto \sqrt{2\sigma_{m,i}^4 + 4\sigma_{m,i}^2(\mu_{m,i} - f_i)^2 + \left(\sigma_{m,i}^2 + (\mu_{m,i} - f_i)^2 - \lambda c_i\right)^2}.$$

## D.3 Extended Background

### D.3.1 Active Testing with Active Surrogate Estimators

Here, we provide further details on the XWED acquisition function used in Kossen et al. (2022).

At acquisition round  $m$ , let  $\pi_m(\cdot | \mathbf{x})$  denote the surrogate predictive distribution used in the main text. When this surrogate is Bayesian, ensemble–based, or otherwise represented as a mixture, we write  $\phi_m$  for

<sup>9</sup>This is the case for our experiments where we use a Bayesian linear regression model with Gaussian priors as the surrogate. It is also the case for a large family of models such as Gaussian process regression with a Gaussian likelihood and Bayesian linear regression with conjugate Gaussian priors.

the fitted quantities that determine the surrogate at round  $m$  – for example, posterior hyperparameters, variational parameters, or the trained ensemble – and let  $\Theta$  denote the latent component of the surrogate predictive model. With this notation, the round- $m$  surrogate predictive distribution can be written as the posterior predictive mixture

$$\pi_m(Y | \mathbf{x}) \equiv \pi_m(Y | \mathbf{x}; \phi_m) = \mathbb{E}_{\Theta \sim \pi_m(\cdot; \phi_m)} [\pi_m(Y | \mathbf{x}, \Theta; \phi_m)].$$

The XWED score is then given by

$$\begin{aligned} \text{XWED}(\mathbf{x}) &= \mathbb{E}_{Y \sim \pi_m(\cdot | \mathbf{x})} [-\mathcal{L}(f(\mathbf{x}), Y) \log \pi_m(Y | \mathbf{x})] \\ &\quad - \mathbb{E}_{\Theta \sim \pi_m(\cdot; \phi_m)} [\mathbb{E}_{Y \sim \pi_m(\cdot | \mathbf{x}, \Theta; \phi_m)} [-\mathcal{L}(f(\mathbf{x}), Y) \log \pi_m(Y | \mathbf{x}, \Theta; \phi_m)]] . \end{aligned}$$

with the following acquisition rule at round  $m$ :

$$i_m = \arg \max_{i \in S_m} \text{XWED}(\mathbf{x}_i).$$

## E Mathematical Results

Here, we present the full statements and proofs for the results in the main body of our paper. Since the main body states results in a simplified form for readability, we first introduce the definitions needed to state these results formally in §E.1. All subsequent results in this section make use of these definitions.

We note that our finite-sample unbiasedness, variance, and  $L^2$ -consistency arguments (§E.2–E.5.1) closely follow the analysis of Farquhar et al. (2021, Appendix B.4–B.7). The main difference is that we state the arguments conditionally on a fixed finite test pool and for general real-valued arrays  $\xi_{M,i}$ , which generalises the results in Farquhar et al. (2021) to include both standard LURE ( $\xi_{M,i} = \ell_i$ ) and PPAT ( $\xi_{M,i} = \ell_i - \lambda(\tilde{\ell}_i - \tilde{R}_M)$ ). In addition to this, §E.5.2 provides a new asymptotic normality result for the resulting LURE-style estimators.

### E.1 Definitions

We first introduce the definition of an *active proposal process* (APP), which formalises the adaptive sampling proposals considered throughout the paper.

**Definition E.1** (Active proposal process, APP). Let  $H_0$  denote all information available before querying the true labels. This includes the inputs  $\mathbf{x}_{1:N}$ , model predictions, or other side information, but it does not include the unqueried true labels.

An admissible active proposal process (APP)  $Q_{1:M} = (Q_1, \dots, Q_M)$  is generated by a collection of measurable maps  $\eta_1, \dots, \eta_M$  such that, at step  $m$ , the map  $\eta_m$  takes as input the observed history  $H_0, (I_1, y_{I_1}), \dots, (I_{m-1}, y_{I_{m-1}})$  and returns a probability distribution on  $S_m$  with full support. The proposal used at step  $m$  is denoted by

$$Q_m = \eta_m(H_0, (I_1, y_{I_1}), \dots, (I_{m-1}, y_{I_{m-1}})),$$

and satisfies

$$Q_m(i) = 0 \quad \text{for } i \notin S_m, \quad \sum_{i \in S_m} Q_m(i) = 1, \quad Q_m(i) > 0 \quad \text{for } i \in S_m.$$

The next queried index is sampled according to

$$\Pr(I_m = i \mid H_0, I_1, y_{I_1}, \dots, I_{m-1}, y_{I_{m-1}}) = Q_m(i), \quad i \in S_m.$$

We now formalise the definition of a surrogate model introduced in §2.1.

**Definition E.2** (Surrogate model). Let  $\mathcal{F}_m = \sigma(\{I_1, \dots, I_m\})$ . At round  $m$ , a surrogate model is an  $\mathcal{F}_{m-1}$ -measurable Markov kernel  $\pi_m(\cdot \mid x)$  on  $\mathcal{Y}$ . That is, for each input  $x$ ,  $\pi_m(\cdot \mid x)$  is a predictive distribution over the unknown label, constructed using only information available before round  $m$ .

Typical examples include Bayesian neural networks, random forests and Gaussian processes.

### E.2 Generalising LURE with Triangular-Arrays

We begin by introducing a general triangular-array formulation that captures both the standard LURE estimator and the PPAT estimator as special cases, and we establish two algebraic identities – a telescoping representation in terms of running averages and a normalisation identity for the finite-pool weights – that underpin all subsequent results.

Let  $(\mathbf{x}_1, \mathbf{x}_2, \dots)$  be a sequence of inputs, with associated labels  $(y_1, y_2, \dots)$ . Moreover, let  $(N_1, N_2, \dots)$  be a sequence of test pool sizes, with  $N_M/M \rightarrow \alpha$  as  $M \rightarrow \infty$  for some  $\alpha > 1$ , and let  $(\xi_{M,i})_{M \geq 1, i=1, \dots, N_M}$  be a triangular array of reals, with  $\xi_{M,i} = h_M(\mathbf{x}_{1:N_M}, y_i)$  for some function  $h_M : \mathcal{X}^{N_M} \times \mathcal{Y} \rightarrow \mathbb{R}$ . Define  $R_M$  as

$$R_M := \frac{1}{N_M} \sum_{i=1}^{N_M} \xi_{M,i}.$$

**Remark E.3.** In this paper, we consider either  $\xi_{M,i} = \ell_i$  (standard LURE), where  $\ell_i = \mathcal{L}(f(\mathbf{x}_i), y_i)$ , or  $\xi_{M,i} = \ell_i - \lambda(\tilde{\ell}_i - \tilde{R}_M)$  (PPAT), where  $\tilde{\ell}_i = \mathcal{L}(f(\mathbf{x}_i), g(\mathbf{x}_i))$  and  $\tilde{R}_M = \frac{1}{N_M} \sum_{i=1}^{N_M} \tilde{\ell}_i$ . Note that in both cases,  $R_M = \frac{1}{N_M} \sum_{i=1}^{N_M} \ell_i$ .

Consider the LURE estimator of  $R_M$  given by

$$\hat{R}_M = \frac{1}{M} \sum_{m=1}^M V_{M,m} \xi_{M, I_{M,m}},$$

with  $I_{M,m} \mid I_{M,1:m-1} \sim Q_{M,m}$ , where, for each  $M$ ,  $Q_{M,1}, \dots, Q_{M,M}$  is a sequence of active proposals in the sense of Def. E.1, and

$$V_{M,m} = 1 + \frac{N_M - M}{N_M - m} \left( \frac{1}{(N_M - m + 1) Q_{M,m}(I_{M,m})} - 1 \right). \quad (22)$$

Throughout this section, for each  $M$ , we treat the finite test pool  $(\mathbf{x}_1, y_1), \dots, (\mathbf{x}_{N_M}, y_{N_M})$  as fixed. All probabilities, expectations, variances, and asymptotic statements are with respect to the randomisation of the active sampling procedure only. Moreover, we write  $N$  for  $N_M$  whenever the dependence on  $M$  is clear and let  $S_{M,m} = [N] \setminus \{I_{M,1}, \dots, I_{M,m-1}\}$  be the set of unqueried indices before round  $m$ , and  $\mathcal{F}_{M,m} = \sigma(I_{M,1}, \dots, I_{M,m})$ .

We first state a telescoping representation of LURE in terms of running averages.

**Lemma E.4** (Alternative representation of LURE). *For each  $M$ , define*

$$\gamma_{M,m} = \frac{N(N-M)}{(N-m)(N-m+1)}$$

and

$$A_{M,m} = \frac{1}{N} \left\{ \frac{\xi_{M, I_{M,m}}}{Q_{M,m}(I_{M,m})} + \sum_{t=1}^{m-1} \xi_{M, I_{M,t}} \right\}.$$

Then

$$\hat{R}_M = \frac{1}{M} \sum_{m=1}^M \gamma_{M,m} A_{M,m}.$$

*Proof.* We expand

$$\sum_{m=1}^M \gamma_{M,m} A_{M,m} = \sum_{m=1}^M \gamma_{M,m} \left[ \frac{\xi_{M, I_{M,m}}}{N Q_{M,m}(I_{M,m})} + \frac{1}{N} \sum_{t=1}^{m-1} \xi_{M, I_{M,t}} \right].$$

Collect the coefficient of a fixed sampled value  $\xi_{M,I_{M,t}}$ , where  $t \in \{1, \dots, M\}$ . This coefficient is

$$\frac{\gamma_{M,t}}{NQ_{M,t}(I_{M,t})} + \sum_{m=t+1}^M \frac{\gamma_{M,m}}{N}.$$

The first term is given by

$$\frac{\gamma_{M,t}}{NQ_{M,t}(I_{M,t})} = \frac{N-M}{(N-t)(N-t+1)Q_{M,t}(I_{M,t})},$$

and for the second term we have

$$\sum_{m=t+1}^M \frac{\gamma_{M,m}}{N} = \sum_{m=t+1}^M \frac{N-M}{(N-m)(N-m+1)}.$$

Now, using

$$\frac{1}{k(k+1)} = \frac{1}{k} - \frac{1}{k+1},$$

and setting  $k = N - m$ , this telescopes to:

$$\sum_{m=t+1}^M \frac{N-M}{(N-m)(N-m+1)} = (N-M) \sum_{k=N-M}^{N-t-1} \frac{1}{k(k+1)} = 1 - \frac{N-M}{N-t}.$$

The coefficient of  $\xi_{M,I_{M,t}}$  is therefore given by:

$$\begin{aligned} & 1 - \frac{N-M}{N-t} + \frac{N-M}{(N-t)(N-t+1)Q_{M,t}(I_{M,t})}. \\ & = 1 + \frac{N-M}{N-t} \left\{ \frac{1}{(N-t+1)Q_{M,t}(I_{M,t})} - 1 \right\} \\ & = V_{M,t}. \end{aligned}$$

Finally, we have

$$\sum_{m=1}^M \gamma_{M,m} A_{M,m} = \sum_{t=1}^M V_{M,t} \xi_{M,I_{M,t}},$$

and dividing by  $M$  gives the desired result. □

We now state a normalisation identity for the finite-pool weights.

**Lemma E.5** (Telescoping identity). *For each  $M$ , we have  $\sum_{m=1}^M \gamma_{M,m} = M$ .*

*Proof.* Using the definition of  $\gamma_{M,m}$ ,

$$\sum_{m=1}^M \gamma_{M,m} = N(N-M) \sum_{m=1}^M \frac{1}{(N-m)(N-m+1)}.$$

Setting  $k = N - m$ , we have:

$$\sum_{m=1}^M \frac{1}{(N-m)(N-m+1)} = \sum_{k=N-M}^{N-1} \frac{1}{k(k+1)}.$$

Now, using the identity

$$\frac{1}{k(k+1)} = \frac{1}{k} - \frac{1}{k+1},$$

the sum telescopes:

$$\sum_{k=N-M}^{N-1} \frac{1}{k(k+1)} = \frac{1}{N-M} - \frac{1}{N}.$$

Therefore

$$\sum_{m=1}^M \gamma_{M,m} = N(N-M) \left( \frac{1}{N-M} - \frac{1}{N} \right) = M.$$

□

### E.3 Finite-Sample Unbiasedness and Variance

Using the running-average representation of Lemma E.4, we now establish two finite-sample properties of our general formulation of LURE: that it is unbiased for the test-pool risk under any APP, and that its variance admits a clean decomposition into a sum of conditional variances of the importance-weighted increments. Using this, we then prove Prop. 3.1 from the main text

**Proposition E.6** (Unbiasedness and finite-sample variance). *Assume that we have an APP in the sense of Def. E.1. Then, for each  $m = 1, \dots, M$ ,  $\mathbb{E}[A_{M,m} \mid \mathcal{F}_{M,m-1}] = R_M$ , and, consequently,  $\mathbb{E}[A_{M,m}] = R_M$ . Moreover, we have that*

$$\text{Var}(A_{M,m}) = \frac{1}{N^2} \mathbb{E} \left[ \text{Var} \left( \frac{\xi_{M,I_{M,m}}}{Q_{M,m}(I_{M,m})} \mid \mathcal{F}_{M,m-1} \right) \right],$$

and for  $m \neq k$ , we have  $\text{Cov}(A_{M,m}, A_{M,k}) = 0$ . It therefore follows that

$$\mathbb{E}[\widehat{R}_M] = R_M$$

and

$$\text{Var}(\widehat{R}_M) = \frac{1}{M^2} \sum_{m=1}^M \gamma_{M,m}^2 \text{Var}(A_{M,m}).$$

Equivalently,

$$\text{Var}(\widehat{R}_M) = \frac{1}{M^2 N^2} \sum_{m=1}^M \gamma_{M,m}^2 \mathbb{E} \left[ \text{Var} \left( \frac{\xi_{M,I_{M,m}}}{Q_{M,m}(I_{M,m})} \mid \mathcal{F}_{M,m-1} \right) \right].$$

*Proof.* Fix  $m \in 1, \dots, M$ . Conditional on  $\mathcal{F}_{M,m-1}$ , the set  $S_{M,m}$ , the proposal  $Q_{M,m}$ , and the previously sampled values  $\xi_{M,I_{M,1}}, \dots, \xi_{M,I_{M,m-1}}$  are fixed. Moreover, by admissibility we know that

$$\mathbb{P}(I_{M,m} = i \mid \mathcal{F}_{M,m-1}) = Q_{M,m}(i), \quad i \in S_{M,m}.$$

Therefore, the conditional expectation of the importance-weighted sampled value is

$$\begin{aligned}\mathbb{E}\left[\frac{\xi_{M,I_{M,m}}}{NQ_{M,m}(I_{M,m})} \middle| \mathcal{F}_{M,m-1}\right] &= \sum_{i \in S_{M,m}} Q_{M,m}(i) \frac{\xi_{M,i}}{NQ_{M,m}(i)} \\ &= \frac{1}{N} \sum_{i \in S_{M,m}} \xi_{M,i}.\end{aligned}$$

Adding the contribution from the previously sampled indices gives

$$\begin{aligned}\mathbb{E}[A_{M,m} \mid \mathcal{F}_{M,m-1}] &= \frac{1}{N} \sum_{i \in S_{M,m}} \xi_{M,i} + \frac{1}{N} \sum_{t=1}^{m-1} \xi_{M,I_{M,t}} \\ &= \frac{1}{N} \sum_{i=1}^N \xi_{M,i} \\ &= R_M.\end{aligned}$$

Taking expectations then gives  $\mathbb{E}[A_{M,m}] = R_M$ . Now, noting that

$$A_{M,m} = \frac{1}{N} \frac{\xi_{M,I_{M,m}}}{Q_{M,m}(I_{M,m})} + \frac{1}{N} \sum_{t=1}^{m-1} \xi_{M,I_{M,t}},$$

where the second term is  $\mathcal{F}_{M,m-1}$ -measurable, we have

$$\text{Var}(A_{M,m} \mid \mathcal{F}_{M,m-1}) = \frac{1}{N^2} \text{Var}\left(\frac{\xi_{M,I_{M,m}}}{Q_{M,m}(I_{M,m})} \middle| \mathcal{F}_{M,m-1}\right).$$

Since  $\mathbb{E}[A_{M,m} \mid \mathcal{F}_{M,m-1}] = R_M$  is deterministic, the law of total variance then yields the desired variance formula:

$$\begin{aligned}\text{Var}(A_{M,m}) &= \mathbb{E}[\text{Var}(A_{M,m} \mid \mathcal{F}_{M,m-1})] \\ &= \frac{1}{N^2} \mathbb{E}\left[\text{Var}\left(\frac{\xi_{M,I_{M,m}}}{Q_{M,m}(I_{M,m})} \middle| \mathcal{F}_{M,m-1}\right)\right]\end{aligned}$$

Suppose now that  $m < k$ . Since  $A_{M,m}$  is  $\mathcal{F}_{M,k-1}$ -measurable, we have

$$\begin{aligned}\text{Cov}(A_{M,m}, A_{M,k}) &= \mathbb{E}[(A_{M,m} - R_M)(A_{M,k} - R_M)] \\ &= \mathbb{E}[(A_{M,m} - R_M)\mathbb{E}[A_{M,k} - R_M \mid \mathcal{F}_{M,k-1}]] \\ &= 0.\end{aligned}$$

The case  $k < m$  follows by symmetry.

Finally, using Lemma E.4 and E.5 gives:

$$\begin{aligned}\mathbb{E}[\widehat{R}_M] &= \frac{1}{M} \sum_{m=1}^M \gamma_{M,m} \mathbb{E}[A_{M,m}] \\ &= \frac{R_M}{M} \sum_{m=1}^M \gamma_{M,m} \\ &= R_M.\end{aligned}$$

The variance formula follows similarly by using Lemma E.4 again and the fact that the  $A_{M,m}$ 's are pairwise uncorrelated.  $\square$

Prop. 3.1 follows as an immediate corollary to the above proposition. We restate the proposition more formally below.

**Corollary E.7** (Proposition 3.1). *Consider the setup of §E.2 and fix any  $\lambda \in \mathbb{R}$ . Then, for any APP,*

$$\mathbb{E}\left[\widehat{R}_M^{\text{PPAT}}(\lambda)\right] = R_M.$$

*Proof.* This follows immediately from Prop. E.6. Indeed, fix  $\lambda \in \mathbb{R}$  and apply Prop. E.6 with the triangular array

$$\xi_{M,i} = \ell_i - \lambda c_{M,i} = \ell_i - \lambda(\tilde{\ell}_i - \tilde{R}_M).$$

The corresponding LURE estimator is exactly

$$\widehat{R}_M^{\text{PPAT}}(\lambda) = \frac{1}{M} \sum_{m=1}^M V_{M,m} \xi_{M,I_{M,m}}.$$

Moreover, the finite-pool mean of this array is

$$\frac{1}{N} \sum_{i=1}^N \xi_{M,i} = \frac{1}{N} \sum_{i=1}^N \ell_i - \lambda \frac{1}{N} \sum_{i=1}^N (\tilde{\ell}_i - \tilde{R}_M) = R_M,$$

since  $\frac{1}{N} \sum_{i=1}^N (\tilde{\ell}_i - \tilde{R}_M) = 0$ . Prop. E.6 therefore gives

$$\mathbb{E}\left[\widehat{R}_M^{\text{PPAT}}(\lambda)\right] = R_M,$$

as claimed. □

We now more formally state and prove Prop. 3.2 from the main text.

**Proposition E.8** (Proposition 3.2). *Consider the setup of §E.2 and fix an APP  $Q_{M,1:M}$ . Let  $c_{M,i} = \tilde{\ell}_i - \tilde{R}_M$  and define*

$$\widehat{R}_M^{\text{LURE}} = \frac{1}{M} \sum_{m=1}^M V_{M,m} \ell_{I_{M,m}}, \quad \widehat{C}_M^{\text{LURE}} = \frac{1}{M} \sum_{m=1}^M V_{M,m} c_{M,I_{M,m}}.$$

*Then, for every  $\lambda \in \mathbb{R}$ ,*

$$\text{Var}(\widehat{R}_M^{\text{PPAT}}(\lambda)) = \text{Var}(\widehat{R}_M^{\text{LURE}}) - 2\lambda \text{Cov}(\widehat{R}_M^{\text{LURE}}, \widehat{C}_M^{\text{LURE}}) + \lambda^2 \text{Var}(\widehat{C}_M^{\text{LURE}}).$$

*Thus,  $\text{Var}(\widehat{R}_M^{\text{PPAT}}(\lambda)) < \text{Var}(\widehat{R}_M^{\text{LURE}})$  whenever  $\lambda \in (\min\{0, 2\lambda^*\}, \max\{0, 2\lambda^*\})$  and  $\text{Var}(\widehat{C}_M^{\text{LURE}}) > 0$ , where*

$$\lambda^* = \frac{\text{Cov}(\widehat{R}_M^{\text{LURE}}, \widehat{C}_M^{\text{LURE}})}{\text{Var}(\widehat{C}_M^{\text{LURE}})}.$$

*In particular,  $\lambda^*$  minimises the variance of  $\widehat{R}_M^{\text{PPAT}}(\lambda)$ .*

*Proof.* Fix  $\lambda \in \mathbb{R}$ . By Remark E.3,  $\widehat{R}_M^{\text{PPAT}}(\lambda)$  is the LURE estimator associated with the residualised array  $\zeta_{M,i}(\lambda) = \ell_i - \lambda c_{M,i}$ . That is,

$$\begin{aligned}\widehat{R}_M^{\text{PPAT}}(\lambda) &= \frac{1}{M} \sum_{m=1}^M V_{M,m} \zeta_{M,I_{M,m}}(\lambda) \\ &= \frac{1}{M} \sum_{m=1}^M V_{M,m} (\ell_{I_{M,m}} - \lambda c_{M,I_{M,m}}) \\ &= \widehat{R}_M^{\text{LURE}} - \lambda \widehat{C}_M^{\text{LURE}}.\end{aligned}$$

Now, by the properties of variance we have

$$\text{Var}(\widehat{R}_M^{\text{PPAT}}(\lambda)) = \text{Var}(\widehat{R}_M^{\text{LURE}}) - 2\lambda \text{Cov}(\widehat{R}_M^{\text{LURE}}, \widehat{C}_M^{\text{LURE}}) + \lambda^2 \text{Var}(\widehat{C}_M^{\text{LURE}}),$$

as claimed.

For the second part, write  $V_C := \text{Var}(\widehat{C}_M^{\text{LURE}}) > 0$  and  $b := \text{Cov}(\widehat{R}_M^{\text{LURE}}, \widehat{C}_M^{\text{LURE}})$ , and set

$$\phi(\lambda) := \text{Var}(\widehat{R}_M^{\text{PPAT}}(\lambda)) - \text{Var}(\widehat{R}_M^{\text{LURE}}) = V_C \lambda^2 - 2b \lambda = V_C \lambda(\lambda - 2\lambda^*),$$

where  $\lambda^* = b/V_C$ . This is a strictly convex parabola with roots 0 and  $2\lambda^*$ , so  $\phi(\lambda) < 0$  if and only if  $\lambda$  lies strictly between these roots, i.e.  $\lambda \in (\min\{0, 2\lambda^*\}, \max\{0, 2\lambda^*\})$ ; this is the stated variance–reduction condition. Examining the derivatives of  $\phi$  then gives  $\lambda^*$  as the minimiser of  $\text{Var}(\widehat{R}_M^{\text{PPAT}}(\lambda))$ , as required.  $\square$

### E.3.1 PPAT Variance Bound

Here, we formally state and prove the variance bound of Prop. 3.4. Throughout we fix the budget  $M$  and the finite test pool, and write  $N$  for  $N_M$ . Recall from Remark E.3 that, for a fixed  $\lambda \in \mathbb{R}$ , the PPAT estimator is exactly the LURE estimator associated with the triangular array  $\zeta_{M,i}(\lambda) = \ell_i - \lambda c_{M,i}$ . That is, by the running–average representation of Lemma E.4,

$$\widehat{R}_M^{\text{PPAT}}(\lambda) = \frac{1}{M} \sum_{m=1}^M V_{M,m} \zeta_{M,I_{M,m}}(\lambda) = \frac{1}{M} \sum_{m=1}^M \gamma_{M,m} A_{M,m},$$

where  $\gamma_{M,m}$  and  $A_{M,m}$  are as defined there for the array  $\zeta_{M,\cdot}(\lambda)$ .

**Proposition E.9** (Variance bound for PPAT). *Suppose we are given an APP satisfying a uniform overlap condition: there exists  $\beta > 0$  such that*

$$Q_{M,m}(i) \geq \frac{\beta}{N}, \quad m = 1, \dots, M, \quad i \in S_{M,m}.$$

*Then, for every  $\lambda \in \mathbb{R}$ ,*

$$\text{Var}(\widehat{R}_M^{\text{PPAT}}(\lambda)) \leq \frac{1}{\beta M} \frac{N}{N - M + 1} \left( \frac{1}{N} \sum_{i=1}^N (\ell_i - \lambda c_{M,i})^2 \right).$$

*Whenever  $\frac{1}{N} \sum_{i=1}^N c_{M,i}^2 > 0$ , the right–hand side is minimised over  $\lambda$  by*

$$\lambda^\dagger = \frac{N^{-1} \sum_{i=1}^N \ell_i c_{M,i}}{N^{-1} \sum_{i=1}^N c_{M,i}^2}.$$

*Proof.* Fix  $\lambda \in \mathbb{R}$ , abbreviate  $\zeta_i := \zeta_{M,i}(\lambda) = \ell_i - \lambda c_{M,i}$ , and set  $D(\lambda) := \frac{1}{N} \sum_{i=1}^N \zeta_i^2 = \frac{1}{N} \sum_{i=1}^N (\ell_i - \lambda c_{M,i})^2$ . By Prop. E.6, the running averages  $A_{M,1}, \dots, A_{M,M}$  are pairwise uncorrelated and satisfy

$$\begin{aligned} \mathbb{E}[A_{M,m} \mid \mathcal{F}_{M,m-1}] &= R_M, \\ \text{Var}(\widehat{R}_M^{\text{PPAT}}(\lambda)) &= \frac{1}{M^2} \sum_{m=1}^M \gamma_{M,m}^2 \text{Var}(A_{M,m}), \\ \text{Var}(A_{M,m}) &= \frac{1}{N^2} \mathbb{E} \left[ \text{Var} \left( \frac{\zeta_{I_{M,m}}}{Q_{M,m}(I_{M,m})} \mid \mathcal{F}_{M,m-1} \right) \right]. \end{aligned}$$

We first bound  $\text{Var}(A_{M,m})$  uniformly in  $m$ . For this, fix an arbitrary  $m$  and condition on  $\mathcal{F}_{M,m-1}$ . Bounding the conditional variance by the conditional second moment and evaluating the latter gives

$$\begin{aligned} \text{Var} \left( \frac{\zeta_{I_{M,m}}}{Q_{M,m}(I_{M,m})} \mid \mathcal{F}_{M,m-1} \right) &\leq \mathbb{E} \left[ \frac{\zeta_{I_{M,m}}^2}{Q_{M,m}(I_{M,m})^2} \mid \mathcal{F}_{M,m-1} \right] \\ &= \sum_{i \in S_{M,m}} Q_{M,m}(i) \frac{\zeta_i^2}{Q_{M,m}(i)^2} \\ &= \sum_{i \in S_{M,m}} \frac{\zeta_i^2}{Q_{M,m}(i)}. \end{aligned}$$

Now, the overlap condition implies  $1/Q_{M,m}(i) \leq N/\beta$  for every  $i \in S_{M,m}$ . As the summands are nonnegative and  $S_{M,m} \subseteq \{1, \dots, N\}$ ,

$$\sum_{i \in S_{M,m}} \frac{\zeta_i^2}{Q_{M,m}(i)} \leq \frac{N}{\beta} \sum_{i \in S_{M,m}} \zeta_i^2 \leq \frac{N}{\beta} \sum_{i=1}^N \zeta_i^2.$$

The right-hand side is deterministic, so taking expectations and dividing by  $N^2$  yields

$$\text{Var}(A_{M,m}) \leq \frac{1}{N^2} \cdot \frac{N}{\beta} \sum_{i=1}^N \zeta_i^2 = \frac{1}{\beta} \left( \frac{1}{N} \sum_{i=1}^N \zeta_i^2 \right) = \frac{D(\lambda)}{\beta},$$

a bound independent of  $m$ .

We now aggregate over the different rounds. The weights  $\gamma_{M,m} = \frac{N(N-M)}{(N-m)(N-m+1)}$  are positive, and since  $(N-m)(N-m+1)$  is decreasing in  $m$  on  $\{1, \dots, M\}$ ,  $\gamma_{M,m}$  is increasing in  $m$ ; its maximum is therefore attained at  $m = M$ ,

$$\max_{1 \leq m \leq M} \gamma_{M,m} = \gamma_{M,M} = \frac{N(N-M)}{(N-M)(N-M+1)} = \frac{N}{N-M+1}.$$

Using  $\gamma_{M,m}^2 \leq \gamma_{M,m} \cdot \max_{1 \leq k \leq M} \gamma_{M,k}$  together with the identity  $\sum_{m=1}^M \gamma_{M,m} = M$  of Lemma E.5, gives

$$\sum_{m=1}^M \gamma_{M,m}^2 \leq \left( \max_{1 \leq k \leq M} \gamma_{M,k} \right) \sum_{m=1}^M \gamma_{M,m} = \frac{N}{N-M+1} M.$$

Combining this with the variance decomposition above then gives

$$\begin{aligned} \text{Var}(\widehat{R}_M^{\text{PPAT}}(\lambda)) &= \frac{1}{M^2} \sum_{m=1}^M \gamma_{M,m}^2 \text{Var}(A_{M,m}) \leq \frac{1}{M^2} \cdot \frac{N}{N-M+1} M \cdot \frac{D(\lambda)}{\beta} \\ &= \frac{1}{\beta M} \frac{N}{N-M+1} D(\lambda), \end{aligned}$$

as claimed.

Finally, only  $D(\lambda)$  depends on  $\lambda$ . Expanding  $D(\lambda)$  gives

$$D(\lambda) = \frac{1}{N} \sum_{i=1}^N \ell_i^2 - 2\lambda \frac{1}{N} \sum_{i=1}^N \ell_i c_{M,i} + \lambda^2 \frac{1}{N} \sum_{i=1}^N c_{M,i}^2,$$

a quadratic in  $\lambda$  with leading coefficient  $\frac{1}{N} \sum_i c_{M,i}^2$ . When  $\frac{1}{N} \sum_i c_{M,i}^2 > 0$  it is strictly convex, and its derivative vanishes at the unique point

$$\lambda^\dagger = \frac{N^{-1} \sum_{i=1}^N \ell_i c_{M,i}}{N^{-1} \sum_{i=1}^N c_{M,i}^2}.$$

As the constant multiplying  $D(\lambda)$  is independent of  $\lambda$ ,  $\lambda^\dagger$  minimises the upper bound.  $\square$

### E.3.2 Tightness of the PPAT Variance Bound

Here, we discuss when the bound in Prop. E.9 is tight and, consequently, when we can expect  $\lambda^\dagger$  to serve as a suitable alternative to  $\lambda^*$ . In particular, we show that the tightness of the bound depends critically on how small  $M/N$  is and how close the proposal is to being uniform.

We begin by noting that the bound is derived by chaining four separate inequalities to relax the exact variance. By tracking when each of these relaxations holds as an equality, we can better understand its overall tightness. As before, write  $\zeta_i := \zeta_{M,i}(\lambda) = \ell_i - \lambda c_{M,i}$  for the residualised loss,  $R_M = N^{-1} \sum_i \zeta_i$  for its pool mean, and  $\sigma_\zeta^2 := N^{-1} \sum_i (\zeta_i - R_M)^2$  for its pool variance, so that  $D(\lambda) = \sigma_\zeta^2 + R_M^2$ . The proof passes through the following four relaxations:

- (i) *Second-moment relaxation.* We bound the conditional variance of the round- $m$  increment by its conditional second moment, which discards the (nonnegative) squared mean of the *remaining* residuals,  $N^{-2} (\sum_{i \in S_{M,m}} \zeta_i)^2$ .
- (ii) *Overlap relaxation.* We replace  $1/Q_{M,m}(i)$  by its worst-case value  $N/\beta$  based on the uniform overlap condition. This is exact only where the proposal sits exactly at the overlap floor,  $Q_{M,m}(i) = \beta/N$ , for every point with  $\zeta_i \neq 0$ . The more concentrated (non-uniform) the proposal is, the looser this step becomes.
- (iii) *Pool-extension relaxation.* We enlarge the sum over the remaining pool  $S_{M,m}$  to a sum over the full pool  $\{1, \dots, N\}$ , discarding the squared residuals of the  $m - 1$  points already queried. This discarded mass is a vanishing fraction of the total whenever the number of acquired labels is small relative to the pool,  $M \ll N$ .
- (iv) *Round-weight aggregation.* Finally, we bound  $\sum_m \gamma_{M,m}^2$  by  $(\max_m \gamma_{M,m}) \sum_m \gamma_{M,m} = (\max_m \gamma_{M,m}) M$ . Since  $\gamma_{M,m}$  is *strictly increasing* in  $m$ , this step is strict for any  $M \geq 2$ : the later rounds contribute disproportionately more to the bound than to a hypothetical “equal-weight” aggregation. The size of this gap is controlled by the spread of the weights across rounds,

$$\frac{\gamma_{M,M}}{\gamma_{M,1}} = \frac{N(N-1)}{(N-M)(N-M+1)} \rightarrow 1 \quad \text{as } M/N \rightarrow 0,$$

i.e., when only a small fraction of the pool is queried, the weights  $\gamma_{M,1}, \dots, \gamma_{M,M}$  are all  $\approx 1$ , and this relaxation becomes negligible.

It's not difficult to see that the above relaxations are governed by two separate sets of conditions. Firstly, the relaxations (ii)-(iv) are all controlled by the same two knobs – *how small  $M/N$  is* and *how close the proposal is to uniform* – and become negligible together as  $M/N \rightarrow 0$  under a near-uniform proposal. Secondly, relaxation (i) depends only on whether the residualised losses are “centred”, i.e. on the ratio  $R_M^2/\sigma_\zeta^2$  between the squared pool mean and the pool variance of  $\zeta_i$ .

In the settings that motivate our approach, we expect relaxation (i) to be negligible. Its cost is controlled solely by the ratio  $R_M^2/\sigma_\zeta^2$ , which is small precisely when the residualised losses are highly variable across the pool relative to their average magnitude,  $\sigma_\zeta^2 \gg R_M^2$ . This is exactly the high-variance regime in which active testing is beneficial. Thus, we expect our bound to be tight when the budget is a small fraction of the pool ( $M \ll N$ ) and our proposal is close to uniform. In particular, we note that  $\lambda^*$  coincides with  $\lambda^\dagger$  when our proposal is uniform.

## E.4 Myopic Oracle Proposals

Here, we consider choosing proposals that minimise the variance of our estimator. As minimising the full-horizon variance is intractable, we instead target a tractable *myopic* proxy. We state our result for the general triangular array  $\xi_{M,i}$  considered above; Prop. 3.3 in the main text then follows immediately by taking  $\xi_{M,i} = \ell_i - \lambda(\tilde{\ell}_i - \tilde{R})$ .

Recall from §E.2-E.3 that, for a general triangular array  $\xi_{M,i}$ , the estimator  $\hat{R}_M = \frac{1}{M} \sum_{m=1}^M V_{M,m} \xi_{M,I_{M,m}}$  admits the running-average representation of Lemma E.4 and the variance decomposition

$$\text{Var}(\hat{R}_M) = \frac{1}{M^2} \sum_{m=1}^M \gamma_{M,m}^2 \text{Var}(A_{M,m})$$

of Prop. E.6. Since the sum  $\sum_{t=1}^{m-1} \xi_{M,I_{M,t}}$  in  $A_{M,m}$  is  $\mathcal{F}_{M,m-1}$ -measurable and  $\gamma_{M,m}$  is a deterministic constant, the only part of the round- $m$  contribution that depends on the proposal  $Q_{M,m}$  is the conditional variance of the importance-weighted increment,

$$\text{Var}(A_{M,m} \mid \mathcal{F}_{M,m-1}) = \frac{1}{N_M^2} \text{Var}\left(\frac{\xi_{M,I_{M,m}}}{Q_{M,m}(I_{M,m})} \mid \mathcal{F}_{M,m-1}\right).$$

The full-horizon variance-minimising proposal is intractable, because the choice of  $Q_{M,m}$  also shapes the laws of all later pools  $S_{M,m+1}, \dots$ . Following §3.2, we therefore adopt a *myopic* objective: at each round  $m$ , conditionally on the history  $\mathcal{F}_{M,m-1}$ , we choose the proposal minimising this conditional variance. The next proposition solves this problem in closed form for an arbitrary array.

**Proposition E.10** (Myopic oracle proposal). *Fix  $M \geq 1$  and a round  $m \in \{1, \dots, M\}$ , and condition on  $\mathcal{F}_{M,m-1}$ , so that the remaining pool  $S_{M,m}$  and the values  $(\xi_{M,i})_{i \in S_{M,m}}$  are fixed. Assume  $\sum_{j \in S_{M,m}} |\xi_{M,j}| > 0$ . Then, among all proposals supported on  $S_{M,m}$ , the conditional variance*

$$\mathcal{V}(Q_{M,m}) := \text{Var}\left(\frac{\xi_{M,I_{M,m}}}{N_M Q_{M,m}(I_{M,m})} \mid \mathcal{F}_{M,m-1}\right)$$

*is minimised by*

$$Q_{M,m}^*(i) = \frac{|\xi_{M,i}|}{\sum_{j \in S_{M,m}} |\xi_{M,j}|},$$

*with minimal value*

$$\mathcal{V}(Q_{M,m}^*) = \frac{1}{N_M^2} \left[ \left( \sum_{i \in S_{M,m}} |\xi_{M,i}| \right)^2 - \left( \sum_{i \in S_{M,m}} \xi_{M,i} \right)^2 \right] \geq 0.$$

*Proof.* Condition throughout on  $\mathcal{F}_{M,m-1}$ . To lighten notation, let  $S = S_{M,m}$ ,  $\xi_i = \xi_{M,i}$ ,  $Q = Q_{M,m}$ ,  $I = I_{M,m}$  and  $N = N_M$ . Set  $W := \xi_I / (N Q(I))$ .

Firstly, note that the conditional mean of  $W$  does not depend on  $Q$ . Indeed,

$$\mathbb{E}[W \mid \mathcal{F}_{M,m-1}] = \sum_{i \in S} Q(i) \frac{\xi_i}{N Q(i)} = \frac{1}{N} \sum_{i \in S} \xi_i.$$

Now, because  $\mathcal{V}(Q) = \mathbb{E}[W^2 \mid \mathcal{F}_{M,m-1}] - (\mathbb{E}[W \mid \mathcal{F}_{M,m-1}])^2$  and the second term is constant in  $Q$ , minimising  $\mathcal{V}(Q)$  is equivalent to minimising the conditional second moment

$$\mathbb{E}[W^2 \mid \mathcal{F}_{M,m-1}] = \sum_{i \in S} Q(i) \frac{\xi_i^2}{N^2 Q(i)^2} = \frac{1}{N^2} \sum_{i \in S} \frac{\xi_i^2}{Q(i)}.$$

Applying the Cauchy–Schwarz inequality with  $a_i = |\xi_i| / \sqrt{Q(i)}$  and  $b_i = \sqrt{Q(i)}$  then gives

$$\left( \sum_{i \in S} |\xi_i| \right)^2 = \left( \sum_{i \in S} a_i b_i \right)^2 \leq \left( \sum_{i \in S} a_i^2 \right) \left( \sum_{i \in S} b_i^2 \right) = \left( \sum_{i \in S} \frac{\xi_i^2}{Q(i)} \right) \underbrace{\sum_{i \in S} Q(i)}_{=1}.$$

Hence  $\sum_{i \in S} \xi_i^2 / Q(i) \geq (\sum_{i \in S} |\xi_i|)^2$ . Since  $b_i > 0$ , equality holds iff  $a_i = c b_i$  for some constant  $c$ , i.e.  $Q(i) = |\xi_i| / c$  for all  $i \in S$ . Noting that  $\sum_{i \in S} Q(i) = 1$ ,  $c$  is forced to be  $c = \sum_{j \in S} |\xi_j|$ , giving

$$Q^*(i) = \frac{|\xi_i|}{\sum_{j \in S} |\xi_j|}, \quad i \in S,$$

which is a valid distribution because  $\sum_{j \in S} |\xi_j| > 0$ .

Now, substituting  $Q^*$  into the second moment yields  $\frac{1}{N^2} (\sum_{i \in S} |\xi_i|)^2$ ; subtracting the constant squared mean  $\frac{1}{N^2} (\sum_{i \in S} \xi_i)^2$  gives the stated minimal value of  $\mathcal{V}$ , which is non-negative by the triangle inequality  $|\sum_i \xi_i| \leq \sum_i |\xi_i|$ .  $\square$

Prop. 3.3 follows as an immediate corollary to Prop. E.10.

**Corollary E.11** (Proposition 3.3). *Consider the setup of §E.2 and fix  $\lambda \in \mathbb{R}$ . Take the PPAT array*

$$\xi_{M,i} = \ell_i - \lambda(\tilde{\ell}_i - \tilde{R}_M)$$

*of Remark E.3, where  $\tilde{\ell}_i = L(f(x_i), g(x_i))$  and  $\tilde{R}_M = \frac{1}{N_M} \sum_{i=1}^{N_M} \tilde{\ell}_i$ . Fix a round  $m \in \{1, \dots, M\}$  and condition on  $\mathcal{F}_{M,m-1}$  and assume  $\sum_{j \in S_{M,m}} |\ell_j - \lambda(\tilde{\ell}_j - \tilde{R}_M)| > 0$ . Then, among all proposals supported on  $S_{M,m}$ , the myopic proxy*

$$\mathcal{V}(Q_{M,m}) := \text{Var} \left( \frac{\ell_{I_{M,m}} - \lambda(\tilde{\ell}_{I_{M,m}} - \tilde{R}_M)}{N_M Q_{M,m}(I_{M,m})} \mid \mathcal{F}_{M,m-1} \right)$$

*is minimised by*

$$Q_{M,m,\lambda}^*(i) = \frac{|\ell_i - \lambda(\tilde{\ell}_i - \tilde{R}_M)|}{\sum_{j \in S_{M,m}} |\ell_j - \lambda(\tilde{\ell}_j - \tilde{R}_M)|},$$

*This is exactly the myopic oracle proposal of Proposition 3.3.*

*Proof.* Fix  $\lambda \in \mathbb{R}$ . This is an application of Prop. E.10 to the triangular array  $\xi_{M,i} = \ell_i - \lambda(\tilde{\ell}_i - \tilde{R}_M)$  of Remark E.3. With this array,  $\xi_{M,I_{M,m}} = \ell_{I_{M,m}} - \lambda(\tilde{\ell}_{I_{M,m}} - \tilde{R}_M)$ , so the conditional variance  $\mathcal{V}(Q_{M,m})$  of Prop. E.10 is precisely the residualised myopic proxy in the statement, and the condition  $\sum_{j \in S_{M,m}} |\xi_{M,j}| > 0$  holds by assumption. Prop. E.10 therefore yields that, among all proposals supported on  $S_{M,m}$ , the minimiser is

$$Q_{M,m}^*(i) = \frac{|\xi_{M,i}|}{\sum_{j \in S_{M,m}} |\xi_{M,j}|} = \frac{|\ell_i - \lambda(\tilde{\ell}_i - \tilde{R}_M)|}{\sum_{j \in S_{M,m}} |\ell_j - \lambda(\tilde{\ell}_j - \tilde{R}_M)|} =: Q_{M,m,\lambda}^*(i),$$

which is the stated proposal.  $\square$

Note that setting  $\lambda = 0$  in Corr. E.11 recovers the LURE oracle  $Q_m^*(i) \propto |\ell_i|$  of §2.1.

## E.5 Asymptotics

We now establish the asymptotic guarantees underlying the confidence intervals used in the main text. We first work with our general triangular-array version of the LURE estimator, proving consistency, asymptotic normality, and the validity of a plug-in variance estimate for constructing confidence intervals. We then specialise these results to PPAT with a fixed value of  $\lambda$ , before extending the argument to the plug-in choice  $\hat{\lambda}_M$ . We conclude by discussing the main assumptions required for these asymptotic results.

### E.5.1 Consistency of the General LURE Estimator

Here we establish that, under a uniform overlap condition on our proposals and a mild second-moment-like condition on the finite-pool array, the general LURE estimator is  $L^2$ -consistent for the test-pool risk; the proof proceeds by bounding the conditional variance of each running average using the overlap condition and aggregating across rounds.

**Theorem E.12** ( $L^2$ -consistency). *Assume that  $N_M/M \rightarrow \alpha > 1$ . Assume also that there exists  $\beta > 0$  such that, for all sufficiently large  $M$ ,*

$$Q_{M,m}(i) \geq \frac{\beta}{N_M}, \quad m = 1, \dots, M, \quad i \in S_{M,m}.$$

*Then, for every  $m = 1, \dots, M$ ,*

$$\text{Var}(A_{M,m}) \leq \frac{1}{\beta N_M} \sum_{i=1}^{N_M} \xi_{M,i}^2 = \frac{1}{\beta} \left( \frac{1}{N_M} \sum_{i=1}^{N_M} \xi_{M,i}^2 \right).$$

*If, in addition  $\frac{1}{N_M} \sum_{i=1}^{N_M} \xi_{M,i}^2 = O(1)$ , then*

$$\mathbb{E}[(\hat{R}_M - R_M)^2] \rightarrow 0.$$

*More generally, the same conclusion holds if  $\frac{1}{N_M} \sum_{i=1}^{N_M} \xi_{M,i}^2 = o(M)$ .*

*Proof.* To simplify our notation, we write  $N = N_M$ . From the proof of Prop. E.6, we have

$$\text{Var}(A_{M,m}) = \mathbb{E}[\text{Var}(A_{M,m} \mid \mathcal{F}_{M,m-1})].$$

Moreover,

$$\text{Var}(A_{M,m} \mid \mathcal{F}_{M,m-1}) \leq \mathbb{E} \left[ \left( \frac{\xi_{M,I_{M,m}}}{NQ_{M,m}(I_{M,m})} \right)^2 \middle| \mathcal{F}_{M,m-1} \right].$$

This gives:

$$\begin{aligned} \text{Var}(A_{M,m} \mid \mathcal{F}_{M,m-1}) &\leq \sum_{i \in S_{M,m}} Q_{M,m}(i) \frac{\xi_{M,i}^2}{N^2 Q_{M,m}(i)^2} \\ &= \frac{1}{N^2} \sum_{i \in S_{M,m}} \frac{\xi_{M,i}^2}{Q_{M,m}(i)}. \end{aligned}$$

Now, using the overlap condition we get  $\frac{1}{Q_{M,m}(i)} \leq \frac{N}{\beta}$  and consequently:

$$\text{Var}(A_{M,m} \mid \mathcal{F}_{M,m-1}) \leq \frac{1}{N^2} \frac{N}{\beta} \sum_{i \in S_{M,m}} \xi_{M,i}^2 \leq \frac{1}{\beta N} \sum_{i=1}^N \xi_{M,i}^2.$$

Taking expectations then gives

$$\text{Var}(A_{M,m}) \leq \frac{1}{\beta N} \sum_{i=1}^N \xi_{M,i}^2.$$

Moreover, since  $N_M/M \rightarrow \alpha > 1$ , the sequence  $\max_{1 \leq m \leq M} \gamma_{M,m}$  is bounded. Indeed,

$$\gamma_{M,m} = \frac{N(N-M)}{(N-m)(N-m+1)} \leq \frac{N}{N-M+1},$$

and

$$\frac{N}{N-M+1} \rightarrow \frac{\alpha}{\alpha-1} < \infty.$$

Finally, using Prop. E.6, there exists a constant  $C_\gamma < \infty$  such that, for all sufficiently large  $M$ ,

$$\begin{aligned} \mathbb{E}[(\widehat{R}_M - R_M)^2] &= \text{Var}(\widehat{R}_M) \\ &= \frac{1}{M^2} \sum_{m=1}^M \gamma_{M,m}^2 \text{Var}(A_{M,m}) \\ &\leq \frac{1}{M^2} \sum_{m=1}^M C_\gamma^2 \frac{1}{\beta N} \sum_{i=1}^N \xi_{M,i}^2 \\ &= \frac{C_\gamma^2}{\beta M} \left( \frac{1}{N} \sum_{i=1}^N \xi_{M,i}^2 \right). \end{aligned}$$

If  $N^{-1} \sum_i \xi_{M,i}^2 = O(1)$ , the right-hand side is  $O(M^{-1})$ , and hence tends to zero. More generally, it tends to zero whenever  $N^{-1} \sum_i \xi_{M,i}^2 = o(M)$ .  $\square$

### E.5.2 Asymptotic Normality and Confidence Intervals for the General LURE Estimator

Building on the martingale structure exposed in §E.3, we now establish a new central limit theorem for  $\sqrt{M}(\widehat{R}_M - R_M)$  under the uniform overlap condition and a  $(2 + \delta)$ -moment condition, and we show that the asymptotic variance can be consistently estimated from the queried labels alone, yielding the asymptotically valid confidence intervals reported in §4. Our result is obtained by applying a martingale central limit theorem for triangular arrays (Billingsley, 1995, Theorem 35.12).

**Theorem E.13** (Martingale CLT for LURE). *Assume  $N_M/M \rightarrow \alpha > 1$ . Assume that our APP satisfies a uniform overlap condition:*

$$Q_{M,m}(i) \geq \frac{\beta}{N_M}, \quad m = 1, \dots, M, \quad i \in S_{M,m},$$

for all sufficiently large  $M$ . For  $m = 1, \dots, M$ , define  $s_{M,m}^2 = \text{Var}(A_{M,m} \mid \mathcal{F}_{M,m-1})$  and  $\sigma_M^2 = \frac{1}{M} \sum_{m=1}^M \gamma_{M,m}^2 s_{M,m}^2$ . Assume that  $\sigma_M^2 \rightarrow_p \sigma^2$  for some  $0 < \sigma^2 < \infty$ . Assume also that, for some  $\delta > 0$ ,

$$\frac{1}{N_M} \sum_{i=1}^{N_M} |\xi_{M,i}|^{2+\delta} = O(1).$$

Then

$$\sqrt{M}(\widehat{R}_M - R_M) \Rightarrow N(0, \sigma^2).$$

*Proof.* Firstly, define the martingale–difference array  $\Delta_{M,m} = \frac{\gamma_{M,m}}{\sqrt{M}}(A_{M,m} - R_M)$ . By Proposition E.6, we have

$$\mathbb{E}[A_{M,m} - R_M \mid \mathcal{F}_{M,m-1}] = 0,$$

and thus

$$\mathbb{E}[\Delta_{M,m} \mid \mathcal{F}_{M,m-1}] = 0.$$

Now, using Lemma E.4 and Lemma E.5, we get

$$\begin{aligned} \sqrt{M}(\widehat{R}_M - R_M) &= \sqrt{M} \left[ \frac{1}{M} \sum_{m=1}^M \gamma_{M,m} A_{M,m} - R_M \right] \\ &= \frac{1}{\sqrt{M}} \sum_{m=1}^M \gamma_{M,m} (A_{M,m} - R_M) \\ &= \sum_{m=1}^M \Delta_{M,m}. \end{aligned}$$

Here, the predictable quadratic variation is given by

$$\sum_{m=1}^M \mathbb{E}[\Delta_{M,m}^2 \mid \mathcal{F}_{M,m-1}] = \frac{1}{M} \sum_{m=1}^M \gamma_{M,m}^2 s_{M,m}^2 = \sigma_M^2.$$

Now, by assumption, we have  $\sigma_M^2 \rightarrow_p \sigma^2$ . It remains to verify the conditional Lindeberg condition.

Let  $p = 2 + \delta$ . As in the proof of Thm. E.12, the sequence  $\max_{1 \leq m \leq M} \gamma_{M,m}$  is bounded. Moreover,

$$A_{M,m} - R_M = \frac{\xi_{M,I_{M,m}}}{NQ_{M,m}(I_{M,m})} - \frac{1}{N} \sum_{i \in S_{M,m}} \xi_{M,i}.$$

Using  $|a - b|^p \leq 2^{p-1}(|a|^p + |b|^p)$ , we have

$$\begin{aligned} \mathbb{E}[|A_{M,m} - R_M|^p \mid \mathcal{F}_{M,m-1}] &\leq 2^{p-1} \mathbb{E} \left[ \left| \frac{\xi_{M,I_{M,m}}}{NQ_{M,m}(I_{M,m})} \right|^p \mid \mathcal{F}_{M,m-1} \right] \\ &\quad + 2^{p-1} \left| \frac{1}{N} \sum_{i \in S_{M,m}} \xi_{M,i} \right|^p. \end{aligned}$$

For the first term, our overlap condition gives

$$\begin{aligned} \mathbb{E} \left[ \left| \frac{\xi_{M,I_{M,m}}}{NQ_{M,m}(I_{M,m})} \right|^p \mid \mathcal{F}_{M,m-1} \right] &= \frac{1}{N^p} \sum_{i \in S_{M,m}} \frac{|\xi_{M,i}|^p}{Q_{M,m}(i)^{p-1}} \\ &\leq \frac{1}{N^p} \left( \frac{N}{\beta} \right)^{p-1} \sum_{i \in S_{M,m}} |\xi_{M,i}|^p \\ &\leq \beta^{1-p} \left( \frac{1}{N} \sum_{i=1}^N |\xi_{M,i}|^p \right). \end{aligned}$$

For the second term, we have

$$\left| \frac{1}{N} \sum_{i \in S_{M,m}} \xi_{M,i} \right|^p \leq \left( \frac{1}{N} \sum_{i=1}^N |\xi_{M,i}| \right)^p \leq \frac{1}{N} \sum_{i=1}^N |\xi_{M,i}|^p.$$

We therefore have that there exists a constant  $C_{p,\beta} < \infty$  such that

$$\mathbb{E}[|A_{M,m} - R_M|^p \mid \mathcal{F}_{M,m-1}] \leq C_{p,\beta} \left( \frac{1}{N} \sum_{i=1}^N |\xi_{M,i}|^p \right).$$

Consequently, for another finite constant  $C < \infty$ , we have

$$\begin{aligned} \sum_{m=1}^M \mathbb{E}[|\Delta_{M,m}|^p \mid \mathcal{F}_{M,m-1}] &= \sum_{m=1}^M \frac{\gamma_{M,m}^p}{M^{p/2}} \mathbb{E}[|A_{M,m} - R_M|^p \mid \mathcal{F}_{M,m-1}] \\ &\leq CM^{1-p/2} \left( \frac{1}{N} \sum_{i=1}^N |\xi_{M,i}|^p \right) \\ &= CM^{-\delta/2} \left( \frac{1}{N} \sum_{i=1}^N |\xi_{M,i}|^{2+\delta} \right) \\ &\rightarrow 0. \end{aligned}$$

Hence, for every  $\varepsilon > 0$ ,

$$\sum_{m=1}^M \mathbb{E}[\Delta_{M,m}^2 \mathbf{1}\{|\Delta_{M,m}| > \varepsilon\} \mid \mathcal{F}_{M,m-1}] \leq \varepsilon^{-\delta} \sum_{m=1}^M \mathbb{E}[|\Delta_{M,m}|^{2+\delta} \mid \mathcal{F}_{M,m-1}] \rightarrow 0.$$

This is the conditional Lindeberg condition.

We can now apply the martingale central limit theorem for triangular arrays: if a martingale–difference array has predictable quadratic variation converging in probability to  $\sigma^2$  and satisfies the conditional Lindeberg condition, then its sum converges in distribution to  $N(0, \sigma^2)$ . Concretely, we have

$$\sum_{m=1}^M \Delta_{M,m} \Rightarrow N(0, \sigma^2).$$

Using  $\sum_{m=1}^M \Delta_{M,m} = \sqrt{M}(\widehat{R}_M - R_M)$ , then gives our desired result.  $\square$

**Remark E.14** (Explicit form of the predictable variance). The conditional variance  $s_{M,m}^2$  can be written explicitly as

$$s_{M,m}^2 = \text{Var} \left( \frac{\xi_{M,I_{M,m}}}{NQ_{M,m}(I_{M,m})} \middle| \mathcal{F}_{M,m-1} \right).$$

Equivalently,

$$s_{M,m}^2 = \frac{1}{N^2} \left[ \sum_{i \in S_{M,m}} \frac{\xi_{M,i}^2}{Q_{M,m}(i)} - \left( \sum_{i \in S_{M,m}} \xi_{M,i} \right)^2 \right].$$

Thus, the asymptotic variance in Thm. E.13 is the limit of

$$\sigma_M^2 = \frac{1}{M} \sum_{m=1}^M \gamma_{M,m}^2 \frac{1}{N^2} \left[ \sum_{i \in S_{M,m}} \frac{\xi_{M,i}^2}{Q_{M,m}(i)} - \left( \sum_{i \in S_{M,m}} \xi_{M,i} \right)^2 \right].$$

**Theorem E.15** (Consistent estimation of the CLT variance and studentized CLT). *Assume the conditions of Thm. E.13. In addition, assume  $\frac{1}{N_M} \sum_{i=1}^{N_M} |\xi_{M,i}|^4 = O(1)$ . Moreover, define*

$$\widehat{\sigma}_M^2 = \frac{1}{M} \sum_{m=1}^M \gamma_{M,m}^2 (A_{M,m} - \widehat{R}_M)^2.$$

Then  $\widehat{\sigma}_M^2 \rightarrow_p \sigma^2$ . Consequently, we have

$$\frac{\sqrt{M}(\widehat{R}_M - R_M)}{\widehat{\sigma}_M} \Rightarrow N(0, 1).$$

An approximate  $(1 - \eta)$ -level confidence interval for  $R_M$  is given by:

$$\widehat{R}_M \pm \Phi^{-1}(1 - \eta/2) \frac{\widehat{\sigma}_M}{\sqrt{M}}, \tag{23}$$

where  $\Phi$  is the CDF of the standard normal distribution.

*Proof.* Define the infeasible realised quadratic variation

$$\widetilde{\sigma}_M^2 := \frac{1}{M} \sum_{m=1}^M \gamma_{M,m}^2 (A_{M,m} - R_M)^2.$$

We first show that  $\tilde{\sigma}_M^2 - \sigma_M^2 \rightarrow_p 0$ . Let

$$Y_{M,m} = \frac{\gamma_{M,m}^2}{M} [(A_{M,m} - R_M)^2 - s_{M,m}^2].$$

Then  $\mathbb{E}[Y_{M,m} | \mathcal{F}_{M,m-1}] = 0$ . Thus  $\sum_{m=1}^M Y_{M,m}$  is a martingale. Since  $\max_m \gamma_{M,m}$  is bounded and, by the same overlap argument used in Thm. E.13,

$$\mathbb{E}[|A_{M,m} - R_M|^4 | \mathcal{F}_{M,m-1}] \leq C \left( \frac{1}{N} \sum_{i=1}^N |\xi_{M,i}|^4 \right),$$

we have

$$\mathbb{E} [(\tilde{\sigma}_M^2 - \sigma_M^2)^2] \leq \frac{C}{M} \left( \frac{1}{N} \sum_{i=1}^N |\xi_{M,i}|^4 \right) \rightarrow 0.$$

Consequently,  $\tilde{\sigma}_M^2 - \sigma_M^2 \rightarrow_p 0$ . Now, since  $\sigma_M^2 \rightarrow_p \sigma^2$ , it therefore follows that  $\tilde{\sigma}_M^2 \rightarrow_p \sigma^2$ .

It now remains to replace  $R_M$  by  $\hat{R}_M$ . Let  $D_{M,m} = A_{M,m} - R_M$ . Then  $A_{M,m} - \hat{R}_M = D_{M,m} - (\hat{R}_M - R_M)$  and, consequently,

$$\tilde{\sigma}_M^2 - \hat{\sigma}_M^2 = -2(\hat{R}_M - R_M) \frac{1}{M} \sum_{m=1}^M \gamma_{M,m}^2 D_{M,m} + (\hat{R}_M - R_M)^2 \frac{1}{M} \sum_{m=1}^M \gamma_{M,m}^2.$$

By Thm. E.12<sup>10</sup>, we have  $\hat{R}_M - R_M = o_p(1)$ . Note that we also have:

$$\frac{1}{M} \sum_{m=1}^M \gamma_{M,m}^2 = O(1), \tag{24}$$

$$\frac{1}{M} \sum_{m=1}^M \gamma_{M,m}^2 D_{M,m} = O_p(1), \tag{25}$$

where (24) follows from the boundedness of  $\gamma_{M,m}$ , and (25) follows from  $\tilde{\sigma}_M^2 \rightarrow_p \sigma^2$  and an application of the Cauchy-Schwarz inequality:

$$\left| \frac{1}{M} \sum_{m=1}^M \gamma_{M,m}^2 D_{M,m} \right| \leq \left( \frac{1}{M} \sum_{m=1}^M \gamma_{M,m}^2 \right)^{1/2} \left( \frac{1}{M} \sum_{m=1}^M \gamma_{M,m}^2 D_{M,m}^2 \right)^{1/2}.$$

Consequently, we have  $\hat{\sigma}_M^2 - \tilde{\sigma}_M^2 = o_p(1)$ . Combining this with  $\tilde{\sigma}_M^2 \rightarrow_p \sigma^2$  then yields:

$$\hat{\sigma}_M^2 \rightarrow_p \sigma^2.$$

The studentised CLT follows from Thm. E.13 and Slutsky's theorem. □

### E.5.3 Fixed- $\lambda$ PPAT

Here, we prove Thm. 4.1 for the fixed- $\lambda$  case; the plug-in case is treated in §E.5.4. This follows as an immediate corollary to the theorems in §E.5.1-E.5.2. The key observation is that, for any fixed  $\lambda \in \mathbb{R}$ , PPAT is exactly the general LURE estimator of §E.2 applied to the residualised triangular array

$$\zeta_{M,i}(\lambda) = \ell_i - \lambda c_{M,i}, \quad c_{M,i} = \tilde{\ell}_i - \tilde{R}_M,$$

<sup>10</sup>This applies here as the fourth moment condition implies the second moment condition.

whose finite-pool mean is the test-pool risk  $R_M$ ; the consistency, central-limit, and variance-estimation results then transfer directly once the moment conditions on this array are translated into separate conditions on the true and proxy losses.

Recall from Remark E.14 that, for the array  $\zeta_{M,\cdot}(\lambda)$ , the predictable quadratic variation  $\sigma_M^2$  of Thm. E.13 takes the explicit form

$$\sigma_M^2(\lambda) = \frac{1}{M} \sum_{m=1}^M \gamma_{M,m}^2 \frac{1}{N_M^2} \left[ \sum_{i \in S_{M,m}} \frac{\zeta_{M,i}(\lambda)^2}{Q_{M,m}(i)} - \left( \sum_{i \in S_{M,m}} \zeta_{M,i}(\lambda) \right)^2 \right]. \quad (26)$$

**Corollary E.16** (Fixed- $\lambda$  PPAT, Theorem 4.1). *Consider the setup of §E.2 and fix  $\lambda \in \mathbb{R}$ . Assume that the pool ratio satisfies  $N_M/M \rightarrow \alpha > 1$  and that our APP satisfies a uniform overlap condition: there exists  $\beta > 0$  such that, for all sufficiently large  $M$ ,*

$$Q_{M,m}(i) \geq \frac{\beta}{N_M}, \quad m = 1, \dots, M, \quad i \in S_{M,m}.$$

Then the following hold.

1. (Consistency.) Assume  $\frac{1}{N_M} \sum_{i=1}^{N_M} \ell_i^2 = O(1)$  and  $\frac{1}{N_M} \sum_{i=1}^{N_M} \tilde{\ell}_i^2 = O(1)$ , then

$$\widehat{R}_M^{\text{PPAT}}(\lambda) - R_M \xrightarrow{p} 0.$$

2. (Asymptotic normality and confidence intervals.) Assume, additionally,  $\frac{1}{N_M} \sum_{i=1}^{N_M} \ell_i^4 = O(1)$  and  $\frac{1}{N_M} \sum_{i=1}^{N_M} \tilde{\ell}_i^4 = O(1)$ , and the predictable quadratic variation (26) satisfies  $\sigma_M^2(\lambda) \xrightarrow{p} \sigma_{\text{PPAT}}^2(\lambda)$  for some  $\sigma_{\text{PPAT}}^2(\lambda) \in (0, \infty)$ , then

$$\sqrt{M} \{ \widehat{R}_M^{\text{PPAT}}(\lambda) - R_M \} \Rightarrow \mathcal{N}(0, \sigma_{\text{PPAT}}^2(\lambda)).$$

Moreover, writing  $A_{M,m}(\lambda)$  for the running average of Lemma E.4 applied to the array  $\zeta_{M,\cdot}(\lambda)$ , i.e.

$$A_{M,m}(\lambda) = \frac{1}{N_M} \left( \frac{\zeta_{M,I_{M,m}}(\lambda)}{Q_{M,m}(I_{M,m})} + \sum_{t=1}^{m-1} \zeta_{M,I_{M,t}}(\lambda) \right),$$

the plug-in estimator

$$\widehat{\sigma}_M^2(\lambda) = \frac{1}{M} \sum_{m=1}^M \gamma_{M,m}^2 (A_{M,m}(\lambda) - \widehat{R}_M^{\text{PPAT}}(\lambda))^2$$

satisfies  $\widehat{\sigma}_M^2(\lambda) \xrightarrow{p} \sigma_{\text{PPAT}}^2(\lambda)$ , and hence

$$\frac{\sqrt{M} \{ \widehat{R}_M^{\text{PPAT}}(\lambda) - R_M \}}{\widehat{\sigma}_M(\lambda)} \Rightarrow \mathcal{N}(0, 1).$$

In particular,

$$\widehat{R}_M^{\text{PPAT}}(\lambda) \pm \Phi^{-1}\left(1 - \frac{\eta}{2}\right) \frac{\widehat{\sigma}_M(\lambda)}{\sqrt{M}}$$

is an asymptotic  $(1 - \eta)$ -level confidence interval for  $R_M$ , where  $\Phi$  is the CDF for the standard normal distribution.

*Proof.* By Remark E.3, for any fixed  $\lambda \in \mathbb{R}$  the PPAT estimator  $\widehat{R}_M^{\text{PPAT}}(\lambda)$  is exactly the generic LURE estimator of §E.2 applied to the array  $\xi_{M,i} = \zeta_{M,i}(\lambda) = \ell_i - \lambda c_{M,i}$ . Since  $c_{M,i} = \tilde{\ell}_i - \tilde{R}_M$  has zero pool average,  $\frac{1}{N_M} \sum_{i=1}^{N_M} c_{M,i} = 0$ , the finite-pool mean of this array is

$$\frac{1}{N_M} \sum_{i=1}^{N_M} \zeta_{M,i}(\lambda) = \frac{1}{N_M} \sum_{i=1}^{N_M} \ell_i - \lambda \frac{1}{N_M} \sum_{i=1}^{N_M} c_{M,i} = R_M.$$

Hence the quantity “ $R_M$ ” appearing in Thms. E.12, E.13 and E.15 for this array is precisely the test-pool risk, and it only remains to translate the moment conditions on  $\xi_{M,i}$  into conditions on  $\ell_i$  and  $\tilde{\ell}_i$ .

It remains to translate the moment assumptions on  $\xi_{M,i}$  into the corresponding assumptions on  $\ell_i$  and  $\tilde{\ell}_i$ . Fix  $p \geq 1$ . Since  $c_{M,i} = \tilde{\ell}_i - \tilde{R}_M$  and  $\tilde{R}_M$  is the pool average of the  $\tilde{\ell}_i$ , Jensen’s inequality gives  $|\tilde{R}_M|^p \leq \frac{1}{N_M} \sum_j |\tilde{\ell}_j|^p$ , so that, using  $|a - b|^p \leq 2^{p-1}(|a|^p + |b|^p)$ ,

$$\frac{1}{N_M} \sum_{i=1}^{N_M} |c_{M,i}|^p \leq 2^{p-1} \left( \frac{1}{N_M} \sum_{i=1}^{N_M} |\tilde{\ell}_i|^p + |\tilde{R}_M|^p \right) \leq C_p \frac{1}{N_M} \sum_{i=1}^{N_M} |\tilde{\ell}_i|^p,$$

for a constant  $C_p < \infty$ . As  $\lambda$  is fixed, applying the same inequality to  $\zeta_{M,i}(\lambda) = \ell_i - \lambda c_{M,i}$  yields

$$\frac{1}{N_M} \sum_{i=1}^{N_M} |\zeta_{M,i}(\lambda)|^p \leq C_{p,\lambda} \left( \frac{1}{N_M} \sum_{i=1}^{N_M} |\ell_i|^p + \frac{1}{N_M} \sum_{i=1}^{N_M} |\tilde{\ell}_i|^p \right), \quad (27)$$

for a constant  $C_{p,\lambda} < \infty$ .

We first establish consistency. Taking  $p = 2$  in (27), the assumptions  $\frac{1}{N_M} \sum \ell_i^2 = O(1)$  and  $\frac{1}{N_M} \sum \tilde{\ell}_i^2 = O(1)$  imply  $\frac{1}{N_M} \sum \xi_{M,i}^2 = O(1)$ . Thm. E.12 then gives  $\mathbb{E}[(\widehat{R}_M^{\text{PPAT}}(\lambda) - R_M)^2] \rightarrow 0$ , and  $L^2$  convergence implies  $\widehat{R}_M^{\text{PPAT}}(\lambda) - R_M \xrightarrow{p} 0$ .

We now turn to the central limit theorem. Taking  $p = 4$  in (27), the fourth-moment assumptions give  $\frac{1}{N_M} \sum |\xi_{M,i}|^4 = O(1)$ , and in particular the  $(2+\delta)$ -moment condition  $\frac{1}{N_M} \sum |\xi_{M,i}|^{2+\delta} = O(1)$  with  $\delta = 2$ . The assumed stabilisation  $\sigma_M^2(\lambda) \xrightarrow{p} \sigma_{\text{PPAT}}^2(\lambda) \in (0, \infty)$  is exactly the predictable-quadratic-variation condition of Thm. E.13 for the array  $\zeta_{M,\cdot}(\lambda)$ . Thm. E.13 therefore yields

$$\sqrt{M} \{ \widehat{R}_M^{\text{PPAT}}(\lambda) - R_M \} \Rightarrow \mathcal{N}(0, \sigma_{\text{PPAT}}^2(\lambda)).$$

Finally, we address estimation of the asymptotic variance and the asymptotic confidence interval. The bound  $\frac{1}{N_M} \sum |\xi_{M,i}|^4 = O(1)$  established above is precisely the additional condition required by Thm. E.15. Applying that theorem to the array  $\zeta_{M,\cdot}(\lambda)$  shows that the plug-in estimator is consistent,  $\widehat{\sigma}_M^2(\lambda) \xrightarrow{p} \sigma_{\text{PPAT}}^2(\lambda)$ , and, by Slutsky’s theorem, gives both the studentised limit and the asymptotic  $(1 - \eta)$ -level confidence interval for  $R_M$ .  $\square$

**Remark E.17.** Setting  $\lambda = 0$  gives  $\zeta_{M,i}(0) = \ell_i$  and  $\widehat{R}_M^{\text{PPAT}}(0) = \widehat{R}_M^{\text{LURE}}$ , so Corr. E.16 specialises to the corresponding consistency, asymptotic normality, and confidence-interval statements for LURE.

#### E.5.4 Plug-in PPAT

Finally, we extend the results in §E.5.3 for the plug-in PPAT estimator, where the tuning parameter  $\lambda$  is replaced by the data-driven estimate  $\hat{\lambda}_M$  in §3.3. The proof compares the plug-in estimator with the PPAT estimator that uses the finite-pool oracle coefficient  $\lambda_M^\dagger$ . We show that replacing  $\lambda_M^\dagger$  by  $\hat{\lambda}_M$  changes the estimator only by  $o_p(M^{-1/2})$ . Thus, at the  $\sqrt{M}$  rate, the plug-in estimator has the same asymptotic behaviour as this oracle estimator: it is consistent, satisfies the same central limit theorem, and yields asymptotically valid confidence intervals when the variance is estimated from the data.

To simplify our notation, for any given triangular array  $a_{M,1:N_M}$  we will write

$$\bar{a}_M = \frac{1}{N_M} \sum_{i=1}^{N_M} a_{M,i}, \quad \hat{\mu}_M(a) = \frac{1}{M} \sum_{m=1}^M V_{M,m} a_{M,I_{M,m}}.$$

In this way,  $\hat{\mu}_M(a)$  is the LURE estimator applied to the array  $a$ . As before, define

$$c_{M,i} = \tilde{\ell}_i - \tilde{R}_M, \quad \tilde{R}_M = \frac{1}{N_M} \sum_{i=1}^{N_M} \tilde{\ell}_i,$$

and note that  $\bar{c}_M = 0$ . Moreover, let

$$D_M = \frac{1}{N_M} \sum_{i=1}^{N_M} c_{M,i}^2, \quad G_M = \frac{1}{N_M} \sum_{i=1}^{N_M} \ell_i c_{M,i},$$

such that the oracle finite-pool coefficient in §3.3 can be written as  $\lambda_M^\dagger = \frac{G_M}{D_M}$ , whenever  $D_M > 0$ .

As discussed in §3.3,  $c_{M,i}$  is known for all test points so  $D_M$  can be computed exactly. On the other hand, we estimate  $G_M$  via LURE,

$$\hat{G}_M = \hat{\mu}_M(\ell c) = \frac{1}{M} \sum_{m=1}^M V_{M,m} \ell_{I_{M,m}} c_{M,I_{M,m}},$$

and set  $\hat{\lambda}_M = \frac{\hat{G}_M}{D_M}$ . Using the notation defined earlier, the plug-in PPAT estimator is

$$\hat{R}_M^{\text{PPAT}}(\hat{\lambda}_M) = \hat{\mu}_M(\ell) - \hat{\lambda}_M \hat{\mu}_M(c).$$

We now formally state and prove the consistency and asymptotic normality of the plug-in PPAT estimator. We prove the confidence interval result as a corollary in Corr. E.19.

**Theorem E.18** (Consistency and asymptotic normality of plug-in PPAT, Theorem 4.1). *Assume that the pool ratio satisfies  $N_M/M \rightarrow \alpha > 1$  and that the APP satisfies a uniform overlap condition: there exists  $\beta > 0$  such that, for all sufficiently large  $M$ ,*

$$Q_{M,m}(i) \geq \frac{\beta}{N_M}, \quad m = 1, \dots, M, \quad i \in S_{M,m}.$$

*Assume also that there exists  $d > 0$  such that, for all sufficiently large  $M$ ,*

$$D_M = \frac{1}{N_M} \sum_{i=1}^{N_M} c_{M,i}^2 \geq d.$$

*Then the following hold.*

1. (Consistency.) Assume

$$\frac{1}{N_M} \sum_{i=1}^{N_M} \ell_i^2 = O(1), \quad \frac{1}{N_M} \sum_{i=1}^{N_M} c_{M,i}^2 = O(1), \quad \frac{1}{N_M} \sum_{i=1}^{N_M} \ell_i^2 c_{M,i}^2 = O(1).$$

Then

$$\widehat{\lambda}_M - \lambda_M^\dagger = O_p(M^{-1/2}) \quad \text{and} \quad \widehat{R}_M^{\text{PPAT}}(\widehat{\lambda}_M) - R_M = O_p(M^{-1/2}).$$

In particular, both quantities are  $o_p(1)$ , so  $\widehat{R}_M^{\text{PPAT}}(\widehat{\lambda}_M) - R_M \rightarrow_p 0$  and  $\widehat{\lambda}_M - \lambda_M^\dagger \rightarrow_p 0$ .

2. (Asymptotic normality.) Assume, in addition, that for some  $\delta > 0$ ,

$$\frac{1}{N_M} \sum_{i=1}^{N_M} |\ell_i|^{2+\delta} = O(1), \quad \frac{1}{N_M} \sum_{i=1}^{N_M} |c_{M,i}|^{2+\delta} = O(1),$$

and define the oracle residualised array  $z_{M,i}^\dagger = \ell_i - \lambda_M^\dagger c_{M,i}$ , with running averages and predictable variances

$$A_{M,m}(z^\dagger) = \frac{1}{N_M} \left\{ \frac{z_{M,I_{M,m}}^\dagger}{Q_{M,m}(I_{M,m})} + \sum_{t=1}^{m-1} z_{M,I_{M,t}}^\dagger \right\},$$

$$s_{M,m}^2(z^\dagger) = \text{Var}(A_{M,m}(z^\dagger) \mid \mathcal{F}_{M,m-1}).$$

If the predictable quadratic variation stabilises,

$$\sigma_{M,\dagger}^2 := \frac{1}{M} \sum_{m=1}^M \gamma_{M,m}^2 s_{M,m}^2(z^\dagger) \rightarrow_p \sigma_\dagger^2$$

for some  $0 < \sigma_\dagger^2 < \infty$ , then

$$\sqrt{M} \left\{ \widehat{R}_M^{\text{PPAT}}(\widehat{\lambda}_M) - R_M \right\} \Rightarrow N(0, \sigma_\dagger^2).$$

*Proof.* Throughout we apply the  $L^2$ -consistency result (Thm. E.12) and the martingale central limit theorem for LURE (Thm. E.13) to fixed finite-pool arrays.

We first bound  $\widehat{\lambda}_M - \lambda_M^\dagger$ . Since

$$\widehat{\lambda}_M - \lambda_M^\dagger = \frac{\widehat{G}_M - G_M}{D_M}$$

and  $D_M \geq d > 0$ , it suffices to control  $\widehat{G}_M - G_M$ . Here  $\widehat{G}_M = \widehat{\mu}_M(\ell c)$  is the LURE estimator of the finite-pool mean  $G_M$  of the array  $g_{M,i} := \ell_i c_{M,i}$ , which satisfies

$$\frac{1}{N_M} \sum_{i=1}^{N_M} g_{M,i}^2 = \frac{1}{N_M} \sum_{i=1}^{N_M} \ell_i^2 c_{M,i}^2 = O(1).$$

Thm. E.12 applied to  $g_{M,i}$  therefore gives  $\widehat{G}_M - G_M = O_p(M^{-1/2})$ , and hence

$$\widehat{\lambda}_M - \lambda_M^\dagger = O_p(M^{-1/2}).$$

Similarly, the centred proxy array  $c_{M,i}$  has zero finite-pool mean,  $\bar{c}_M = 0$ , and satisfies  $\frac{1}{N_M} \sum_i c_{M,i}^2 = O(1)$ , so Thm. E.12 applied to  $c_{M,i}$  gives

$$\widehat{\mu}_M(c) = \widehat{\mu}_M(c) - \bar{c}_M = O_p(M^{-1/2}).$$

We next describe the properties of the oracle residualised array  $z_{M,i}^\dagger = \ell_i - \lambda_M^\dagger c_{M,i}$  needed for our results. First,  $\lambda_M^\dagger$  is bounded: by Cauchy–Schwarz,

$$|G_M| = \left| \frac{1}{N_M} \sum_{i=1}^{N_M} \ell_i c_{M,i} \right| \leq \left( \frac{1}{N_M} \sum_{i=1}^{N_M} \ell_i^2 \right)^{1/2} \left( \frac{1}{N_M} \sum_{i=1}^{N_M} c_{M,i}^2 \right)^{1/2},$$

so that

$$|\lambda_M^\dagger| = \frac{|G_M|}{D_M} \leq \left( \frac{N_M^{-1} \sum_i \ell_i^2}{D_M} \right)^{1/2} = O(1),$$

using  $D_M \geq d > 0$  and  $N_M^{-1} \sum_i \ell_i^2 = O(1)$ . Second, since  $\bar{c}_M = 0$ , the array  $z^\dagger$  has finite-pool mean

$$\frac{1}{N_M} \sum_{i=1}^{N_M} z_{M,i}^\dagger = \frac{1}{N_M} \sum_{i=1}^{N_M} \ell_i - \lambda_M^\dagger \frac{1}{N_M} \sum_{i=1}^{N_M} c_{M,i} = R_M,$$

and, using  $(a - b)^2 \leq 2a^2 + 2b^2$  together with the boundedness of  $\lambda_M^\dagger$ ,

$$\frac{1}{N_M} \sum_{i=1}^{N_M} (z_{M,i}^\dagger)^2 \leq 2 \frac{1}{N_M} \sum_{i=1}^{N_M} \ell_i^2 + 2(\lambda_M^\dagger)^2 \frac{1}{N_M} \sum_{i=1}^{N_M} c_{M,i}^2 = O(1).$$

Thm. E.12 applied to the fixed array  $z^\dagger$  therefore gives

$$\widehat{\mu}_M(z^\dagger) - R_M = O_p(M^{-1/2}).$$

We now prove consistency. Writing  $\widehat{R}_M^{\text{PPAT}}(\widehat{\lambda}_M) = \widehat{\mu}_M(\ell) - \widehat{\lambda}_M \widehat{\mu}_M(c)$  and using the linearity of  $\widehat{\mu}_M$ ,

$$\begin{aligned} \widehat{R}_M^{\text{PPAT}}(\widehat{\lambda}_M) - R_M &= \widehat{\mu}_M(\ell) - \widehat{\lambda}_M \widehat{\mu}_M(c) - R_M \\ &= \widehat{\mu}_M(\ell) - \lambda_M^\dagger \widehat{\mu}_M(c) - R_M - (\widehat{\lambda}_M - \lambda_M^\dagger) \widehat{\mu}_M(c) \\ &= (\widehat{\mu}_M(z^\dagger) - R_M) - (\widehat{\lambda}_M - \lambda_M^\dagger) \widehat{\mu}_M(c). \end{aligned}$$

The first term is  $O_p(M^{-1/2})$  by the previous step, while the second term is

$$(\widehat{\lambda}_M - \lambda_M^\dagger) \widehat{\mu}_M(c) = O_p(M^{-1/2}) O_p(M^{-1/2}) = O_p(M^{-1}).$$

Therefore

$$\widehat{R}_M^{\text{PPAT}}(\widehat{\lambda}_M) - R_M = O_p(M^{-1/2}),$$

which proves part 1.

It remains to prove asymptotic normality. Since  $|\lambda_M^\dagger| = O(1)$ , the additional  $(2 + \delta)$ -moment assumptions control the corresponding moment of  $z^\dagger$ : by  $|a - b|^{2+\delta} \leq 2^{1+\delta}(|a|^{2+\delta} + |b|^{2+\delta})$ ,

$$\frac{1}{N_M} \sum_{i=1}^{N_M} |z_{M,i}^\dagger|^{2+\delta} \leq C \left[ \frac{1}{N_M} \sum_{i=1}^{N_M} |\ell_i|^{2+\delta} + |\lambda_M^\dagger|^{2+\delta} \frac{1}{N_M} \sum_{i=1}^{N_M} |c_{M,i}|^{2+\delta} \right] = O(1)$$

for a finite constant  $C$ . The fixed array  $z^\dagger$  thus satisfies the hypotheses of the LURE central limit theorem (Thm. E.13), and, using the assumed variance stabilisation  $\frac{1}{M} \sum_{m=1}^M \gamma_{M,m}^2 s_{M,m}^2(z^\dagger) \rightarrow_p \sigma_\dagger^2$ , we obtain

$$\sqrt{M} \{\widehat{\mu}_M(z^\dagger) - R_M\} \Rightarrow N(0, \sigma_\dagger^2).$$

Finally, we show that replacing  $\lambda_M^\dagger$  by  $\widehat{\lambda}_M$  is first-order negligible. From the decomposition above,

$$\sqrt{M} \left\{ \widehat{R}_M^{\text{PPAT}}(\widehat{\lambda}_M) - \widehat{\mu}_M(z^\dagger) \right\} = -(\widehat{\lambda}_M - \lambda_M^\dagger) \sqrt{M} \widehat{\mu}_M(c).$$

We have shown  $\widehat{\lambda}_M - \lambda_M^\dagger = O_p(M^{-1/2}) = o_p(1)$ , while  $\sqrt{M} \widehat{\mu}_M(c) = O_p(1)$ ; hence the right-hand side is  $o_p(1)$ , that is,

$$\sqrt{M} \left\{ \widehat{R}_M^{\text{PPAT}}(\widehat{\lambda}_M) - \widehat{\mu}_M(z^\dagger) \right\} = o_p(1).$$

Combining this asymptotic equivalence with the central limit theorem for  $\widehat{\mu}_M(z^\dagger)$  and applying Slutsky's theorem yields

$$\sqrt{M} \left\{ \widehat{R}_M^{\text{PPAT}}(\widehat{\lambda}_M) - R_M \right\} \Rightarrow N(0, \sigma_\dagger^2),$$

which proves part 2.  $\square$

The preceding theorem gives a limiting normal distribution with generally unknown variance. We now show that this variance can be replaced by an estimate computed from the acquired labels, while still yielding asymptotically valid confidence intervals. This allows us to obtain approximate confidence intervals for our plug-in PPAT estimator.

**Corollary E.19** (Studentised plug-in PPAT CLT). *Assume the conditions of Thm. E.18, and in addition that*

$$\frac{1}{N_M} \sum_{i=1}^{N_M} |\ell_i|^4 = O(1), \quad \frac{1}{N_M} \sum_{i=1}^{N_M} |c_{M,i}|^4 = O(1).$$

For  $\lambda \in \mathbb{R}$ , define the residualised array  $z_{M,i}(\lambda) = \ell_i - \lambda c_{M,i}$  with running average

$$A_{M,m}(\lambda) = \frac{1}{N_M} \left\{ \frac{z_{M,I_{M,m}}(\lambda)}{Q_{M,m}(I_{M,m})} + \sum_{t=1}^{m-1} z_{M,I_{M,t}}(\lambda) \right\},$$

and let

$$\widehat{\sigma}_M^2(\widehat{\lambda}_M) = \frac{1}{M} \sum_{m=1}^M \gamma_{M,m}^2 \left\{ A_{M,m}(\widehat{\lambda}_M) - \widehat{R}_M^{\text{PPAT}}(\widehat{\lambda}_M) \right\}^2.$$

Then

$$\widehat{\sigma}_M^2(\widehat{\lambda}_M) \rightarrow_p \sigma_\dagger^2,$$

and hence

$$\frac{\sqrt{M} \left\{ \widehat{R}_M^{\text{PPAT}}(\widehat{\lambda}_M) - R_M \right\}}{\widehat{\sigma}_M(\widehat{\lambda}_M)} \Rightarrow N(0, 1).$$

*Proof.* Recall from Thm. E.18 the oracle residualised array  $z_{M,i}^\dagger = \ell_i - \lambda_M^\dagger c_{M,i}$ , together with the facts  $|\lambda_M^\dagger| = O(1)$  and  $\widehat{\lambda}_M - \lambda_M^\dagger = o_p(1)$ . Evaluated at the deterministic oracle coefficient,  $\widehat{\sigma}_M^2(\lambda_M^\dagger)$  is precisely

the LURE variance estimator for the fixed array  $z^\dagger$ . The variance-estimation result of Thm. E.15 therefore applies: its moment condition holds because  $|\lambda_M^\dagger| = O(1)$  and the fourth-moment assumptions give

$$\frac{1}{N_M} \sum_{i=1}^{N_M} |z_{M,i}^\dagger|^4 = O(1),$$

This gives  $\widehat{\sigma}_M^2(\lambda_M^\dagger) \rightarrow_p \sigma_\dagger^2$ .

It remains to replace  $\lambda_M^\dagger$  with  $\widehat{\lambda}_M$ . Both  $A_{M,m}(\lambda)$  and  $\widehat{R}_M^{\text{PPAT}}(\lambda) = \widehat{\mu}_M(\ell) - \lambda \widehat{\mu}_M(c)$  are affine in  $\lambda$ , so

$$A_{M,m}(\widehat{\lambda}_M) - \widehat{R}_M^{\text{PPAT}}(\widehat{\lambda}_M) = \left\{ A_{M,m}(\lambda_M^\dagger) - \widehat{R}_M^{\text{PPAT}}(\lambda_M^\dagger) \right\} - (\widehat{\lambda}_M - \lambda_M^\dagger) \{A_{M,m}(c) - \widehat{\mu}_M(c)\},$$

where  $A_{M,m}(c)$  is the running average of the array  $c_{M,\cdot}$ . Substituting this identity into  $\widehat{\sigma}_M^2(\widehat{\lambda}_M)$  and expanding, every term other than  $\widehat{\sigma}_M^2(\lambda_M^\dagger)$  carries a factor of  $\widehat{\lambda}_M - \lambda_M^\dagger$ . Since  $\widehat{\lambda}_M - \lambda_M^\dagger = o_p(1)$ , while  $\widehat{\sigma}_M^2(\lambda_M^\dagger)$  and the empirical quadratic average

$$\frac{1}{M} \sum_{m=1}^M \gamma_{M,m}^2 \{A_{M,m}(c) - \widehat{\mu}_M(c)\}^2$$

are both  $O_p(1)$  (the latter by the fourth-moment condition on  $c$ ), these terms are  $o_p(1)$ . Hence

$$\widehat{\sigma}_M^2(\widehat{\lambda}_M) - \widehat{\sigma}_M^2(\lambda_M^\dagger) = o_p(1), \quad \text{and therefore} \quad \widehat{\sigma}_M^2(\widehat{\lambda}_M) \rightarrow_p \sigma_\dagger^2.$$

The studentised CLT then follows from Thm. E.18 and Slutsky's theorem.  $\square$

### E.5.5 Discussion of Assumptions

Thms. E.12, E.13 and E.15 rely on a number of assumptions, which we now examine one at a time. Broadly, they fall into four groups: the finite-population asymptotic regime, two conditions on the active proposal (admissibility and uniform overlap), a family of moment conditions on the loss array, and the stabilisation of the predictable quadratic variation. As we argue below, the first three are either standard, directly enforceable, or mild conditions on the fixed test pool that involve no distributional assumptions; only the last is genuinely substantive, which we verify through our empirical results. Throughout, recall that these results are stated for the general triangular array  $\xi_{M,i}$  and specialise to LURE ( $\xi_{M,i} = \ell_i$ ) and PPAT ( $\xi_{M,i} = \zeta_{M,i}(\lambda) = \ell_i - \lambda(\tilde{\ell}_i - \tilde{R}_M)$ ).

**Finite-population regime:** All three results are stated in the finite-population asymptotic regime, in which the pool size  $N_M$  and the budget  $M$  grow together with  $N_M/M \rightarrow \alpha$  for some  $\alpha > 1$ . This is the standard regime for without-replacement sampling from a finite population (Farquhar et al., 2021), and it simply says that we label a fixed fraction  $M/N_M \rightarrow 1/\alpha \in (0, 1)$  of the pool. The requirement  $\alpha > 1$  rules out the degenerate case in which we label (essentially) the entire pool.

**Uniform overlap.** This is a positivity (overlap) condition requiring that, at every round, each remaining point retains at least a fixed fraction of uniform sampling mass. It is the finite-population analogue of the overlap/positivity assumption that is ubiquitous in importance sampling, inverse-probability weighting, and off-policy evaluation, and it plays exactly the same role here: it keeps the importance weights bounded. Crucially, this condition is not merely plausible but is enforced by construction in our experiments, as described in §3.2.

**Moment conditions on the loss array.** The three results require control of, respectively, the second,  $(2 + \delta)$ -th, and fourth finite-pool moments of the array:

$$\underbrace{\frac{1}{N_M} \sum_{i=1}^{N_M} \xi_{M,i}^2 = O(1)}_{\text{consistency (Thm. E.12)}, \quad \underbrace{\frac{1}{N_M} \sum_{i=1}^{N_M} |\xi_{M,i}|^{2+\delta} = O(1)}_{\text{normality (Thm. E.13)}, \quad \underbrace{\frac{1}{N_M} \sum_{i=1}^{N_M} |\xi_{M,i}|^4 = O(1)}_{\text{variance/CIs (Thm. E.15)}.$$

The single most important point about these conditions is that they are statements about the empirical moments of a fixed, deterministic collection of losses: because we work conditionally on the test pool, they involve no distributional or i.i.d. assumption on the data whatsoever. They simply require that the average magnitude of the (residualised) losses does not blow up as the pool grows.

For PPAT, the array  $\xi_{M,i} = \zeta_{M,i}(\lambda)$  is a linear combination of the true loss  $\ell_i$  and the centred proxy loss  $c_{M,i} = \tilde{\ell}_i - \tilde{R}_M$ . Since centring does not increase moments beyond a constant factor, each of the above conditions reduces to the same bounded-moment requirement on the true losses  $\ell_i$  and the proxy losses  $\tilde{\ell}_i$  separately – for instance,  $N_M^{-1} \sum_i \ell_i^4 = O(1)$  and  $N_M^{-1} \sum_i \tilde{\ell}_i^4 = O(1)$  for the fourth-moment condition. These are easy to satisfy. For any *bounded* loss – for example the 0–1 loss, or any loss on a compact label space – all finite-pool moments are automatically bounded and the conditions hold trivially. For *unbounded* losses such as squared error or cross-entropy, they hold provided the losses do not develop pool-growing heavy tails, i.e. provided no vanishingly small subset of points carries losses large enough to dominate the average  $p$ -th power; in practise, this is a weak requirement.

**Stabilisation of the predictable quadratic variation.** The asymptotic normality and variance-estimation results (Thms. E.13 and E.15) additionally require that the predictable quadratic variation  $\sigma_M^2 = \frac{1}{M} \sum_{m=1}^M \gamma_{M,m}^2 s_{M,m}^2$  converge in probability to a finite, strictly positive constant  $\sigma^2$ . This is the one genuinely non-trivial assumption, which we now discuss below.

Firstly, we note that it is the standard stabilisation condition for a martingale CLT (Hall et al., 1980, Chapter 3). The centred, rescaled error  $\sqrt{M}(\hat{R}_M - R_M)$  is a sum of martingale differences, and the fluctuations of such a sum are governed by its accumulated conditional variance  $\sigma_M^2$ ; the CLT holds precisely when this accumulated variance stabilises.

Secondly, the finiteness of  $\sigma^2$  is essentially free: combining the overlap bound with the second-moment condition gives  $s_{M,m}^2 \leq \beta^{-1} N_M^{-1} \sum_i \xi_{M,i}^2$ , and since  $\gamma_{M,m}$  is uniformly bounded,  $\sigma_M^2$  is uniformly bounded above. Moreover, the positivity of  $\sigma^2 > 0$  asks that the estimator have genuinely non-degenerate fluctuations at the  $\sqrt{M}$  scale. It fails only in degenerate situations, for example if the residualised losses became asymptotically constant across the pool. The convergence requirement, however, is more substantive and generally does not follow from pool-level moment bounds alone, except in simpler cases such as non-adaptive proposals or random sampling without replacement. While it is difficult to establish this convergence for arbitrary fully adaptive proposals, our coverage experiments in §5.4 provide indirect empirical support for it: the PPAT intervals attain the nominal coverage level, and do so at least as quickly as, and often faster than, the competing methods. Such behaviour would be difficult to achieve if  $\sigma_M^2$  failed to stabilise. We also note that while the random and LURE baselines sometimes fall short of nominal coverage at the budgets considered, we expect this to improve with an increased labelling budget.