Diagnostic Tool for Out-of-Sample Model Evaluation

Anonymous authors
Paper under double-blind review

Abstract

Assessment of model fitness is a key part of machine learning. The standard paradigm of model evaluation is analysis of the average loss over future data. This is often explicit in model fitting, where we select models that minimize the average loss over training data as a surrogate, but comes with limited theoretical guarantees. In this paper, we consider the problem of characterizing a batch of out-of-sample losses of a model using a calibration data set. We provide finite-sample limits on the out-of-sample losses that are statistically valid under quite general conditions and propose a diagonistic tool that is simple to compute and interpret. Several numerical experiments are presented to show how the proposed method quantifies the impact of distribution shifts, aids the analysis of regression, and enables model selection as well as hyperparameter tuning.

1 Introduction

Fitting a model to data is a central task in machine learning, signal processing, statistics and other areas (Bishop, 2006; Hastie et al., 2009; Söderström & Stoica, 2001; Kay, 1993; Fitzmaurice et al., 2011). A fitted model f can be assessed by considering a loss function $\ell(\cdot)$ that evaluates the model on future data points. This is called *out-of-sample* analysis, since it considers data points beyond those in the training data sample.

Classical statistical models are often assessed by verifying assumption validity through statistical or graphical tools, a process called model diagnostics, diagnostic checks, or regression diagnostics (Ruppert et al., 2003; Belsley et al., 1980). The term has also been established in a wider sense for general checks to verify a model fitness for given data, see e.g. Casper et al. (2022); Zhang et al. (2022). We propose a diagnostic tool that helps evaluating the model performance on out-of-sample data.

In this paper, we consider the problem of characterizing m out-of-sample losses of a model f using a calibration data set. The choice of m will depend on the application. In online learning, $m = 1$ is reasonable since the model may only be valid for a single prediction. In problems with low data-volumes, such as a predictive health-care model to be used on e.g. 30 future patients with a rare desease, $m = 30$ may be appropriate. For models in high volume applications, such as e-commerce, the limit $m \to \infty$ may be relevant. We derive finite-sample *limits* on the out-of-sample losses that are statistically valid under quite general conditions and for any of the choices of m. An illustration is provided in Fig. 1, where we fit a model for predicting house prices to training data and evaluate its absolute prediction error on a calibration data set $\mathcal{D}$. The average loss on $\mathcal{D}$ is a form of 'cross-validation' and is indicated in the figure. The figure also illustrates the proposed diagnostic statistic: an upper bound $\ell_\alpha^\beta(\mathcal{D})$ on the β-fraction of m yet unobserved prediction errors that holds with confidence level $1 - \alpha$. By computing this statistical limit $\ell_\alpha^\beta(\mathcal{D})$ as function of α, we quantify how probable different out-of-sample losses are for the model f. Since this quantification is valid for any size of $\mathcal{D}$, the limit can be employed as a diagnostic tool also in cases where calibration data is scarce or costly to obtain. Moreover, as the validity does not depend on the distribution of the data used to train f, the limit can also be used to analyze the severity of distribution shifts, as we will illustrate in the numerical experiments below.

The rest of the paper is organized as follows. We first formalize the general problem of interest, then propose a measure $\ell_\alpha^\beta(\mathcal{D})$ which we refer to as the level-α limit (LAL) on the β-fraction of m out-of-sample losses and

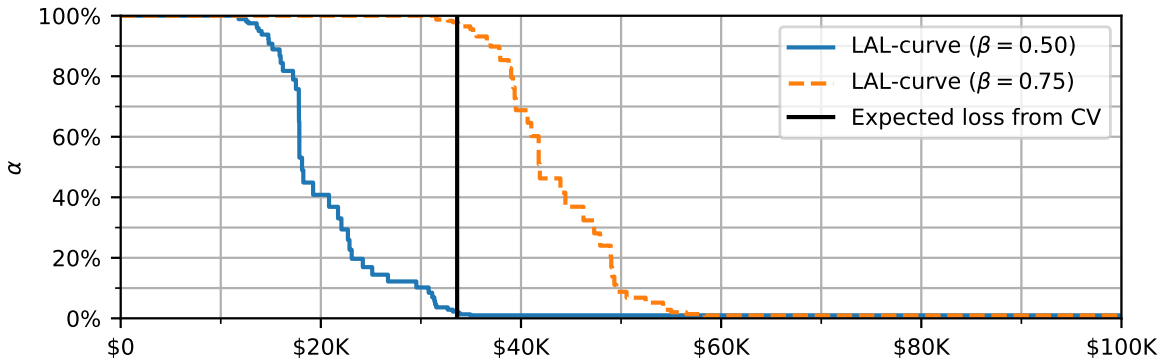


Figure 1: Model error diagonistic. Consider a predictive model $f(X)$ of house price Y in an area described by a feature vector X. How will this model perform in m future price predictions? We evaluate the model using the prediction error $\ell(X, Y) = |Y - f(X)|$ as the chosen loss function. The cross-validation method (CV) evaluates the model by estimating the expected out-of-sample loss, which in this case is circa \$36K (indicated by black vertical line). Being an estimate of an expectation, it provides limited information about what individual loss values might be observed. Suppose the model will be used in $m = 100$ prediction instances, and we want to bound at least $\beta = 50\%$ of these future losses. Then the LAL-curve infers that with probability at least 80% (level $\alpha = 20\%$), this fraction of prediction errors will not exceed \$25K (see blue solid curve) For a stronger guarantees, e.g. bounding at least $\beta = 75\%$ of the future losses, the LAL-curve indicates that prediction errors of \$48K must be tolerated at level $\alpha = 20\%$ (see orange dashed curve). Full details for this experiment are provided in Sec. 4.1.

prove its statistical guarantees. This is followed by a series of numerical experiments that demonstrate the utility of LAL. In the closing discussion, the proposed method is related to existing literature.

2 Problem Statement

Let f denote a model fitted to data samples $\mathcal{D}_0$ drawn from a distribution p_0. We aim to quantify the performance of f on m out-of-sample data points $\{Z_1, \ldots, Z_m\}$ drawn from distribution p, which may differ from p_0. The performance is quantified using any real-valued loss function $\ell(z)$ the user wants. We let upper case letters denote random variables, e.g., Z, and let the lower case version, z, represent their realization.

Example 1 (Density Estimation using a Gaussian model)**.** *A data point is a vector $z \in \mathbb{R}^K$ and we consider a Gaussian density model $f(z) = \mathcal{N}\left(z\,;\,\widehat{\mu}\,,\,\widehat{\Sigma}\right)$, with a fitted mean $\widehat{\mu}$ and covariance matrix $\widehat{\Sigma}$. A common loss function is the negative log-likelihood*

$$\ell(z) = -2\ln f(z) = (z - \widehat{\mu})^{\mathsf{T}}\widehat{\Sigma}^{-1}(z - \widehat{\mu}) + \ln|\widehat{\Sigma}|,$$

ignoring the constant and scaling by a factor 2.

Example 2 (Regression)**.** *A data point is a pair of features x and a label y, i.e., $z = (x, y)$. The model $f(x)$ is any estimate of the conditional expectation function $\mathbb{E}[y|x]$. Example models include Gaussian process regressors, random forest regressors, and neural networks. A common loss function is the squared-error loss $\ell(x, y) = |y - f(x)|^2$.*

We assume that we have access to a *calibration data set* $\mathcal{D} = \{Z_1^c, \ldots, Z_n^c\}$ with n samples drawn randomly from p, such that the combined data set of $n+m$ samples is exchangeable, i.e., the joint density of the random vector $(Z_1^c, \ldots, Z_n^c, Z_1, \ldots, Z_m)$ is invariant under relabeling of the data points. While the calibration data set $\{Z_1^c, \ldots, Z_n^c\}$ is available to use in computations, the out-of-sample data $\{Z_1, \ldots, Z_m\}$ is future, yet unseen data. Exchangeability includes the common independent and identically distributed (iid) data assumption as a special case. An example of exchangeable *non*-iid data is sampling without replacement from a finite population.

The problem we consider is to characterize m unknown *out-of-sample losses*

$$\{\ell(Z_1), \ldots, \ell(Z_m)\} \tag{1}$$

of the model f. Specifically, we want to bound a fraction $\beta \in (0,1)$ of the m losses with a confidence $1-\alpha$. That is, we want to find a statistical limit $\ell_\alpha^\beta(\mathcal{D})$ on how large future losses we are likely to observe for the model:

$$\mathbb{P}\Big[\text{at least a fraction } \beta \text{ of losses } \ell(Z_i) \text{ respects } \ell(Z_i) \leq \ell_\alpha^\beta(\mathcal{D})\Big] \geq 1-\alpha. \tag{2}$$

We will call such a value $\ell_\alpha^\beta(\mathcal{D})$ the *level-α limit* for the β-fraction of m losses, or a LAL for short.

We call the graph of $\ell_\alpha^\beta(\mathcal{D})$ versus α a *LAL-curve*, as illustrated in, e.g., Fig. 1. By plotting the value $\ell_\alpha^\beta(\mathcal{D})$ on the horizontal axis, we can visualize the tail behaviour of the out-of-sample losses for f in a transparent manner. Thus given a valid LAL, we propose to use the LAL-curve as a diagnostic tool for model evaluation.

3 Method

For notational convenience, let $L_i = \ell(Z_i)$ and arrange the m out-of-sample losses (1) in increasing order: $L_{(1)} \leq L_{(2)} \leq \cdots \leq L_{(m)}$. The criterion (2) can be expressed compactly as

$$\boxed{\mathbb{P}\left[L_{(\lceil m\beta \rceil)} > \ell_\alpha^\beta(\mathcal{D})\right] \leq \alpha} \tag{3}$$

and our objective is to find a limit $\ell_\alpha^\beta(\mathcal{D})$ for any given $\beta \in (0,1]$ and $\alpha \in (0,1)$.

For the calibration data $\mathcal{D}$, let $L_i^c = \ell(Z_i^c)$ and define the set of losses $\{L_1^c, \ldots, L_n^c\}$ and order statistics $L_{(1)}^c \leq L_{(2)}^c \leq \cdots \leq L_{(n)}^c$. We also define the special cases $L_{(0)}^c$ and $L_{(n+1)}^c$ to be the infimum and supremum of the support of the distribution of $\{L_i^c\}_{i=1}^m$, possibly $\pm\infty$.

3.1 General LAL expression

The following result holds generically; it assumes two or more continuously distributed random variables don't take exactly the same value. The probability of such an event is zero. To simplify the reading, we will skip the assertions of 'almost surely' and 'assuming no ties'.

Theorem 1. *Let $\ell_\alpha^\beta(\mathcal{D}) = L_{(k^\star)}^c$, where*

$$k^\star = \min\left\{k \in \{1, \ldots, n+1\} \mid a(k) \leq \alpha\right\},$$

$$a(k) = \sum_{j=k}^{n+1} \frac{\binom{n-j+m-\lceil m\beta \rceil}{n-j}\binom{j+\lceil m\beta \rceil-1}{j}}{\binom{n+m}{m}} \tag{4}$$

This is a valid LAL, satisfying (3).

Proof. Define the set of calibration losses $\mathcal{L}^c = \{L_1^c, \ldots, L_n^c\}$ and out-of-sample losses $\mathcal{L} = \{L_1, \ldots, L_m\}$.

Consider the case when $\mathcal{L}^c \cup \mathcal{L}$ has a continuous joint distribution, so there are no ties. In that case

$$\mathbb{P}\left[L_{(j)}^c \leq L_{(i)} < L_{(j+1)}^c\right] = \frac{\binom{n-j+m-i}{n-j}\binom{j+i-1}{j}}{\binom{n+m}{m}} \tag{5}$$

for $1 \leq i \leq m$, $1 \leq j \leq n$, which is proven by a combinatorial argument. First take all $n+m$ loss values in $\mathcal{L}^c \cup \mathcal{L}$ and place them in sorted order on the real number line. This exact list of sorted values may have come from either of the $\binom{n+m}{m}$ equally probable partitions of $\mathcal{L}^c \cup \mathcal{L}$ into $\mathcal{L}^c$ and $\mathcal{L}$, due to exchangeability. How many of these partitions fulfil $L_{(j)}^c \leq L_{(i)} < L_{(j+1)}^c$? Among all $n-j+m-i$ lossvalues greater than $L_{(i)}$, $n-j$ must belong to $\mathcal{L}^c$, and there are $\binom{n-j+m-i}{n-j}$ equally probable outcomes that satisfy this requirement,

due to exchangeability. There are similarly $\binom{j+i-1}{j}$ equally probable outcomes that places j values of $\mathcal{L}^c$ and $i-1$ values from $\mathcal{L}$ to the left of $L_{(i)}$. These combinations are independent, showing (5).

Because $\mathbb{P}\left[L_{(i)} > L^c_{(k)}\right] = \sum_{j=k}^{n} \mathbb{P}\left[L^c_{(j)} \leq L_{(i)} < L^c_{(j+1)}\right]$, we have that

$$\mathbb{P}\left[L_{(\lceil m\beta \rceil)} > \ell^{\beta}_{\alpha}(\mathcal{D})\right] = a(k^\star) \leq \alpha$$

which verifies the theorem for continuous variables. Next, we prove the result for discrete random variables. Let F be the joint cdf for $\mathcal{L}^c \cup \mathcal{L}$, with where each loss takes values in a finite or countable set $V = \{v_i\}$ of real numbers. Since we may get ties with non-zero probability, the preceding analysis fails. Construct instead a random vector $(\lambda_1, \ldots, \lambda_{n+m})$ with cdf F^* such that

$$F^*(l_1, \ldots, l_{n+m}) \geq F(l_1, \ldots, l_{n+m}) \text{ always}$$
$$F^*(l_1, \ldots, l_{n+m}) = F(l_1, \ldots, l_{n+m}) \text{ if } (l_1, \ldots, l_{n+m}) \in V^{n+m}$$

Define further a set of random variables $\overline{\lambda}_i = \min\{v \in V | v \geq \lambda_i\}$ for all i. Now $(\overline{\lambda}_1, \ldots, \overline{\lambda}_{n+m})$ is equal to $(L^c_1 \ldots L^c_n, L_1, \ldots, L_m)$ in distribution. Also, $(\overline{\lambda}_{(i)} > \overline{\lambda}_{(j)}) \Rightarrow (\lambda_{(i)} > \lambda_{(j)})$ for all i, j. Together, this means $\mathbb{P}\left[L_{(\lceil \beta m \rceil)} > L^c_{(k^\star)}\right] = \mathbb{P}\left[\overline{\lambda}_{(\lceil \beta m \rceil + n)} > \overline{\lambda}_{(k^\star)}\right] \leq \mathbb{P}\left[\lambda_{(\lceil \beta m \rceil + n)} > \lambda_{(k^\star)}\right] = a(k^\star)$. where we have used the result on continuous random variables on F^* to compute $k^\star$. □

Remark 1. *The definition of the* LAL, (2)*, only demands that the limit level is at least* $1-\alpha$*. However, the more conservative* $\ell^{\beta}_{\alpha}(\mathcal{D})$ *is, the larger is the excess coverage. From the proof, we see that if the joint set of losses* $(L^c_1 \ldots L^c_n, L_1, \ldots, L_m)$ *has no ties, one may compute the* exact *coverage* $1 - a(k^\star)$*. Thus when the method is conservative, it is transparently so.*

Remark 2. *The proof technique above is inspired by Fligner & Wolfe (1976) which proves the result for the special case of iid data. It is noteworthy that when generalizing from iid to exchangeable data, we keep the same level of precision.*

3.2 LAL for a single out-of-sample data point

When $m = 1$, the LAL takes a very simple closed form. By deriving it from basic principles rather than using Thm. 1, we also obtain a tightness guarantee.

Theorem 2. *For a single out-of-sample data point (m=1), a* LAL *can be constructed as* $\ell^1_{\alpha}(\mathcal{D}) = L^c_{(k^\star)}$*, where* $k^\star = \lceil (n+1)(1-\alpha) \rceil$*. For continuous data distributions, the almost sure out-of-sample loss guarantee is*

$$\alpha - \frac{1}{1+n} \leq \mathbb{P}\left[L_1 > \ell^1_{\alpha}(\mathcal{D})\right] \leq \alpha \tag{6}$$

For discrete data distributions, only the upper bound in (6) *can be guaranteed.*

Proof. When $(L^c_1, \ldots, L^c_n, L_1)$ are continuous, the values are almost surely unique, and therefore there are $\binom{n+1}{1} = \frac{1}{n+1}$ equally likely ways to select which one is L_1. Only one such selection obeys $L^c_{(j)} \leq L_1 \leq L^c_{(j+1)}$. Therefore,

$$\mathbb{P}[L^c(j) \leq L_1 \leq L^c_{(j+1)}] = \frac{1}{1+n} \text{ and } \mathbb{P}[L_1 \leq L^c_{(k^\star)}] = \frac{k^\star}{1+n}$$

Because $(1-\alpha) \leq \lceil (n+1)(1-\alpha) \rceil / (n+1) \leq (1-\alpha) + 1/(n+1)$ we compute

$$\alpha - \frac{1}{1+n} \leq \mathbb{P}[L_1 > L^c_{(k^\star)}] \leq \alpha$$

When $(L^c_1, \ldots, L^c_n, L_1)$ are discrete and ties are possible, we can still prove the upper bound. Consider rank of R of L_1, i.e., put $\{L^c_1, \ldots, L^c_n, L_1\}$ in nondecreasing order, and let R denote the position of L_1. When there are ties, position the tied values in a uniformly random way. By construction R is uniformly distributed over $1, \ldots, (n+1)$ and

$$\mathbb{P}[L_1 \leq L_{(k^\star)}] \geq \mathbb{P}[R \geq k^\star] = \frac{k^\star}{1+n} \geq \alpha$$

□

Remark 3. *For iid data and $m = 1$, the LAL-curve approaches the complementary cdf of L_1, i.e., $1 - F$. This facilitates the interpretation of the LAL-curve as a quantile point estimate. To see this, let $\widehat{F}_n^{-1}$ denote the empirical quantile function of L_1 based on the losses $L_1^c, \dots, L_n^c$. Then the LAL of Thm. 2 can equivalently be defined as*

$$\ell_\alpha^\beta(\mathcal{D}) = \begin{cases} \widehat{F}_n^{-1}\left(\frac{n+1}{n}(1-\alpha)\right) & \text{if } \frac{n+1}{n}(1-\alpha) \in (0,1) \\ L_{(n+1)}^c & \text{else} \end{cases} \tag{7}$$

If L has a bounded and connected range, $\widehat{F}_n^{-1}$ converges uniformly to F^{-1} (Bogoya et al., 2016), and $\ell_\alpha^\beta(\mathcal{D}) \to F^{-1}(1-\alpha)$. Plotting α on the vertical axis against $F^{-1}(1-\alpha)$ is identical to plotting the complementary cdf $1 - F(\ell)$ on the vertical axis against ℓ, so in this case, the LAL-curve converges to the graph of the complementary cdf.

3.3 LAL for large out-of-sample data sets

If the number of out-of-sample data points is very large, we may use a limit argument and let $m \to \infty$ in on Thm. 1.

Corollary 1. *Let $\mathrm{BIN}^{-1}(\cdot; n, \beta)$ denote the quantile function of a binomial distribution with parameters (n, β). For an infinite sequence of exchangeable losses $(L_1^c \dots L_n^c, L_1, \dots)$, let $\ell_\alpha^\beta(\mathcal{D}) = L_{(k^\star)}^c$ with*

$$k^\star = 1 + \mathrm{BIN}^{-1}(1-\alpha; n, \beta).$$

This LAL satisfies a limit form of (3):

$$\lim_{m\to\infty} \mathbb{P}\left[L_{(\lceil m\beta \rceil)} > \ell_\alpha^\beta(\mathcal{D})\right] \le \alpha \tag{8}$$

Proof. The result follows by expanding the binomial coefficients with Stirling's formula and taking the limit for $m \to \infty$. See the appendix for details. □

Starting from Cor. 1, we can find a different interpretation of the LAL when $m \to \infty$.

Remark 4. *Consider the case of iid data, where L_i^c and L_j have a common cdf F for all i, j. Let F^{-1} be the quantile function, and $\widehat{F}_m^{-1}$ be the empirical quantile function based on $\{L_1, \dots, L_m\}$.*

As $m \to \infty$, we have that $L_{(\lceil m\beta \rceil)} = \widehat{F}_m^{-1}(\beta) \to F^{-1}(\beta)$. Therefore, (8) simplifies to $\mathbb{P}\left[F^{-1}(\beta) > \ell_\alpha^\beta(\mathcal{D})\right] \le \alpha$, which gives another interpretation of the LAL as the boundary of a confidence interval $C_\alpha^\beta(\mathcal{L}^c) := (-\infty, \ell_\alpha^\beta(\mathcal{D})]$. This interval satisfies $\mathbb{P}\left[F^{-1}(\beta) \notin C_\alpha^\beta(\mathcal{L}^c)\right] \le \alpha$ and is thus a valid confidence interval for the β-quantile of L_i for all i.

This intuition is also useful in the non-iid case as $m \to \infty$. The De Finetti theorem (Kingman, 1978) states that if $(L_1^c \dots L_n^c, L_1, L_2 \dots)$ forms an infinite sequence of exchangeable random variables, there is an auxiliary random variable ζ such that all the conditional variables $(L_i^c|\zeta)$ and $(L_j|\zeta)$ are iid with cdf F_ζ. Therefore, the interpretation of the LAL as the boundary of a confidence interval is useful in the exchangeable-but-not-iid data setting, even if its interpretation must be handled with more caution.

4 Experiments

This section presents examples of how the LAL can be applied to analyzing the out-of-sample loss of a model or family of models. Code to reproduce all experiments can be found at (link omitted under review).

4.1 Study of asymptotics

This experiment illustrates the limit for different m corresponding to Thm. 1, Cor. 1 and Thm. 2. The data set consist of California housing prices from the 1990 census (Kelley Pace & Barry, 1997), covering 20 640 housing blocks.

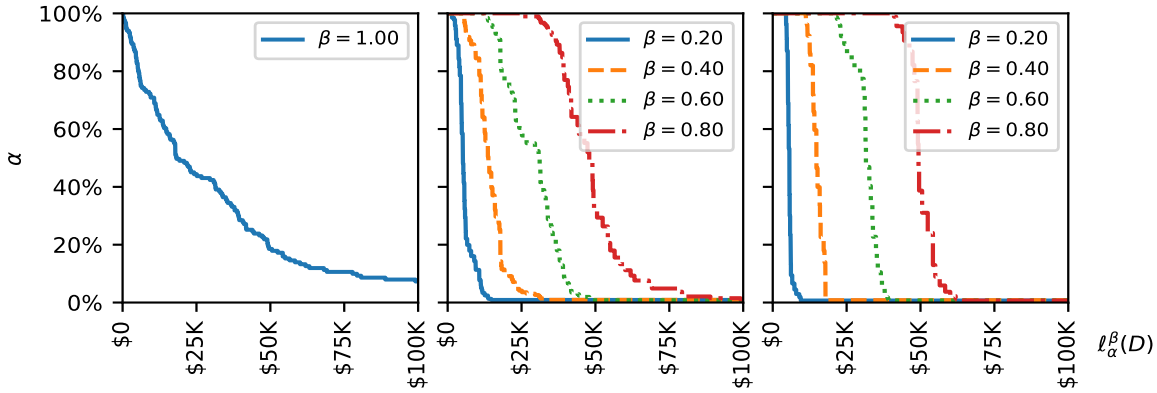


Figure 2: LAL-curves. From left to right, $m = 1$ computed with Thm. 2, $m = 30$ computed with Thm. 1 and $m = \infty$ computed with Cor. 1. The curves assure that a fraction β of out-of-sample losses will not exceed $\ell_\alpha^\beta(\mathcal{D})$, with at least probability $1 - \alpha$. For example, we can see that among the $m = 30$ next samples, at least 80% of them will have prediction losses less than $\$70,000$, with a confidence of 90%.

Each data point $z = (x, y)$ represents a single city block. The label $y \in \mathbb{R}$ is the median house value in the block, and the feature vector $x \in \mathbb{R}^8$ consists of continuous variables such as block coordinates, median house ages and tenant median income. The training data set $\mathcal{D}_0$ has $n_0 = 15\ 000$ sampled without replacement. The calibration data set $\mathcal{D}$ has $n = 150$ and is sampled without replacement from the remaining data.

The model is a regression for the logarithm of the median house value, operating on standardized features and labels, and uses a random Fourier feature basis (Rahimi & Recht, 2008). Let K random Fourier functions with bandwidth b be stacked in a vector ϕ. The model is $f(x) = \exp[\phi(x)^\mathsf{T}\hat{\theta}]$, where $\hat{\theta}$ is found by L2-regularized least squares regression. The hyperparameters (number of basis functions K, bandwidth b and regularization strength λ) are tuned by five-fold cross-validation on the training data. The loss function is the absolute error expressed in dollars

$$\ell(x, y) = |y - f(x)|.$$

We now turn to limiting m out-of-sample losses, where $m \in \{1, 30, \infty\}$. LAL-curves were drawn with varying β-fractions, shown in Fig. 1 and Fig. 2. The empirical risk, defined as the average loss on the calibration data, is also calculated, and the calibration losses are presented as a histogram.

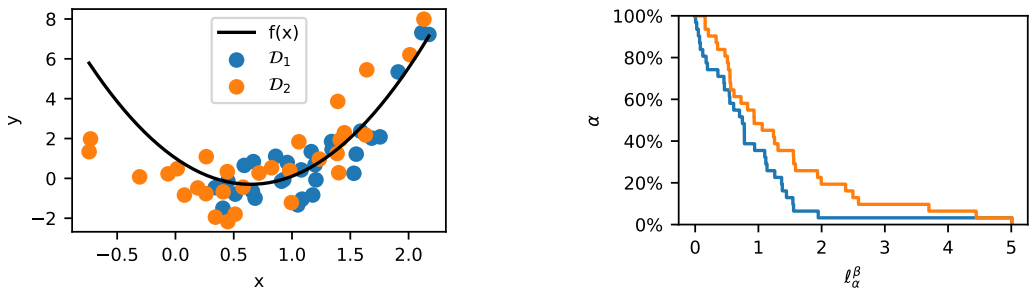
Averaging the losses over the calibration data gives an unbiased point estimate of the expected out-of-sample loss for f but it, by itself, lacks statistical guarantees. While it estimates the mean out-of-sample prediction error to be around \$35 000, the LAL-curve for a single new prediction error ($m = 1$) in Fig. 2 shows that it may be close to \$80 000 (for $\alpha \approx 10\%$).

If we now consider a batch of $m = 30$ predictions, the LAL-curve for $m = 30$ in Fig. 2 informs us that $\beta = 80\%$ of them will have errors smaller than \$70 000 (for $\alpha \approx 10\%$). The number of data points blocks not analyzed are 5 490, so we may also consider the limiting case $m \to \infty$. The LAL curve now tells us that $\beta = 80\%$ of prediction errors will be smaller than \$60 000 (for $\alpha \approx 10\%$).

By comparing the LAL-curves in Fig. 2, we learn how the out-of-sample batch size m affects the tail behavior of the LAL at a fixed β. The LAL-curve is slowly decaying for $m = 1$. For $m = 30$, the decay is faster. For $m = \infty$ the decay is even more abrupt. Some intuition can be gained for iid data. As m increases, the variance of $L_{(\lceil m\beta \rceil)}$ decreases (it is asymptotically normal and $\sqrt{m}$-consistent, see e.g. (Vaart, 1998, Cor. 21.5)), and the bound in (3) can be made tighter.

4.2 Distribution shift analysis

This experiment illustrates the analysis of distribution shifts (Quiñonero-Candela et al., 2009), also known as concept drift, using the LAL-curve. The experiment also verifies Thm. 2 numerically. The availability of a calibration data $\mathcal{D}$ invites the notion of refitting or fine-tuning, to adapt to potential distribution shift (Lee



(a) Fitted model and calibration data

(b) LAL-curves, $m = 1$, $\beta = 1$

Figure 3: Quadratic regression model $f(x)$ evaluated using absolute prediction error loss function $\ell(x, y) = |y - f(x)|$. (a) A fitted model and two different calibration data sets $\mathcal{D}_1 \overset{iid}{\sim} p_1$ (identical to training data distribution) and $\mathcal{D}_2 \overset{iid}{\sim} p_2$ (shifted distribution). (b) LAL-curves under distributions p_1 and p_2. A single out-of-sample loss exceeds the limit $\ell_\alpha^\beta(\mathcal{D})$ by a probability of at most α, as given by the curve. (See also Thm. 2).

et al., 2023). This is not always possible, e.g. when calibration data is limited (too small n). It may still be desirable to evaluate the model performance on out-of-distribution data. Since the LAL is valid for any n, it will always be an available diagnostic tool to analyze the extent to which distribution shift affects the model performance.

Lu et al. (2018) proposes a taxonomy for distribution shift detection methods. In this taxonomy, the LAL would qualify as an *error rate-based* method, as it is based on a loss function. Example methods in this group are the Drift Detection Method (DDM) (Gama et al., 2004) and Black Box Shift Estimation (BBSE) (Lipton et al., 2018) which detects distribution shift by monitoring when classifier performance changes with statistical significance. In contrast, the LAL does focus on distributions of out-of-sample losses, enabling analysis of *how/shift severity* for distribution shifts (again using the Lu et al. (2014) taxonomy). Other methods for *how*, focus on statistical distances between feature distributions (Rabanser et al., 2019). While such methods do not need labels, they do not measure shifts in the way that matters for the model, i.e., the loss.

Let $Z_i = (X_i, Y_i)$, with real valued X_i and Y_i, and generate data according to

$$X_i \sim \mathcal{N}\left(\mu\,,\,\sigma^2\right) \tag{9}$$

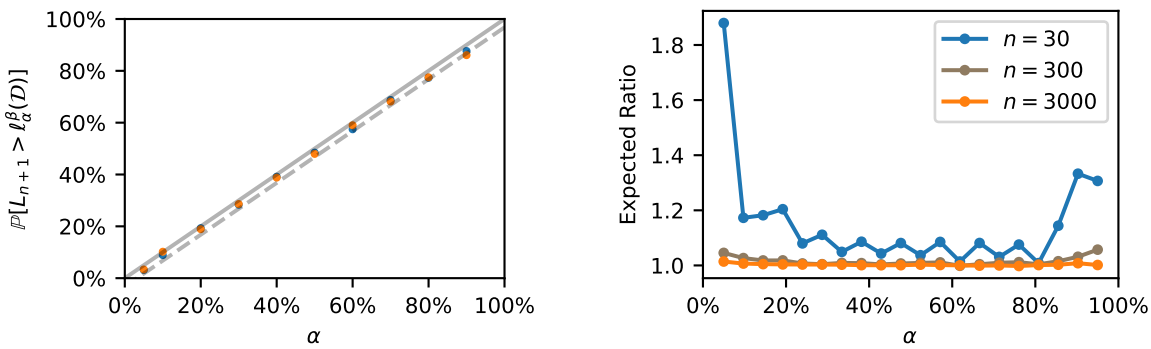
$$Y_i | X_i = x \sim \mathcal{N}\left(x(x-1)(x+1)\,,\,1\right) \tag{10}$$

The training data set $\mathcal{D}_0$ was created with $n_0 = 100$, $\mu = 1$, $\sigma = 0.5$. A quadratic regression model was used, since it approximates the conditional expectation function well over the training data; $f(x) = \widehat{\theta}_0 + \widehat{\theta}_1 x + \widehat{\theta}_2 x^2$ was fitted to $\mathcal{D}_0$ via least-squares. See Fig. 3a. We analyze the case of out-of-sample batch size $m = 1$. The performance of the model is evaluated using the absolute error loss:

$$\ell(x, y) = |y - f(x)| \tag{11}$$

The out-of-sample losses are quantified for two different data distributions. In the first case, $\mu = 1$, $\sigma = 0.5$, so there is no distribution shift, and we use a calibration data set $\mathcal{D}_1$ for which $n = 30$. In the second case, $\mu = 0.75$, $\sigma = 0.75$, resulting in a shift in X and we use a calibration data set $\mathcal{D}_2$ also having $n = 30$. The LAL-curves in both cases are presented in Fig. 3b, which reveal significantly larger out-of-sample losses for the distribution p_2 from which $\mathcal{D}_2$ was drawn. The 5%-tail losses are nearly twice as large for the shifted distribution.

Under the same experimental setup, we verified the tightness result of Thm. 2. Using 2 000 Monte Carlo runs, the empirical coverage was computed for different α, and plotted in Fig. 4a. Moreover, Rem. 3 states the convergence of the LAL to the quantile function $F^{-1}(1-\alpha)$ as n increases. This is illustrated in Fig. 4b.



(a) Empirical coverage

(b) Convergence of the LAL to quantile function

Figure 4: The finite-sample guarantee of Thm. 2 is verified in (a), which illustrates both the upper bound (solid line) and lower bound (dashed line). (b) Illustration of the connection between LAL and the quantile function (Rem. 3). The expected value of the ratio $\ell_\alpha^\beta(\mathcal{D})/F^{-1}(1-\alpha)$ is evaluated for different values of n with data drawn from the shifted distribution p_2. As n increases, the expected ratio approaches 1, from above.

The quantile function was numerically approximated using 10^5 samples. The expected LAL is computed using 2 000 Monte Carlo runs with data sampled from p_2. We see that the LAL approaches the quantile function as n increases.

4.3 Classification error analysis

This experiment shows that the proposed methodology can be applied to classification models as well. We will also consider how adversarial distribution shifts manifest in the LAL-curve. We use the Palmer Penguin data set, popularized by Horst et al. (2020). The 333 complete record data points are pairs $z_i = (x_i, y_i)$ of features $x_i \in \mathbb{R}^7$ and labels $y_i \in \{1, 2, 3\}$. The labels y_i are categorical, encoding the penguin species. The features x_i are vectors of (Island, Bill Length, Bill Depth, Flipper Length, Body Mass, Sex, Year), a mixture of categorical and integer-valued features.

We use a training data set $\mathcal{D}_0$ of $n_0 = 150$, leaving 183 samples for calibration. Using $\mathcal{D}_0$, we fit $f(x)$ via multinomial logistic regression using L2-regularization and cross-validation. The model output is a three-dimensional vector $f(x) = [f_1(x), f_3(x), f_3(x)]^\mathsf{T}$ approximating the conditional probabilities, so that $f_i(x)$ approximates $\mathbb{P}[Y = i|X = x]$. The model is to be evaluated on calibration data using the misclassification probability loss

$$\ell(x, y) = 1 - f_y(x) \tag{12}$$

The out-of-sample batch size was set to $m = 1$.

Two different calibration data sets $\mathcal{D}_1$ and $\mathcal{D}_2$ of sample size $n = 50$ were constructed. We sampled from the 183 held out data points without replacement. For $\mathcal{D}_1$, the initial probability of selecting a sample was uniform over the data. For $\mathcal{D}_2$, the initial probability of selecting a sample was proportional to $\ell(x, y)$, effectively an adversarial reweighting of samples. The LAL-curves of Thm. 2 (set $m = 1$ and β arbitrary) are presented in Fig. 5, showing that LAL-curves can be applied to classification models. The fact that one curve comes from an adversarial calibration sample is manifest by larger LAL for every confidence α compared to the non-adversarial calibration sample.

4.4 Regression error analysis

This experiment shows how alternative loss functions can be used to analyze the asymmetry of errors in regression problems. We use the UCI Airfoil data set (Dua & Graff, 2017). The task is to predict a label

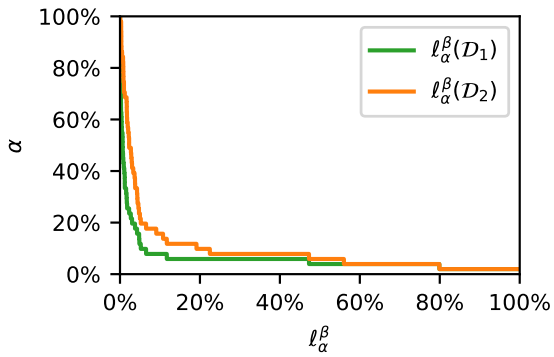


Figure 5: LAL-curves ($m = 1$) for the classification error analysis. The loss function indicates the certainty of misclassification – a loss of 80% means that the model assigned 80% probability to the wrong labels. Data set $\mathcal{D}_1$, is exchangeable with the training data. Data set $\mathcal{D}_2$ is adversarially sampled, presenting notably larger LAL.

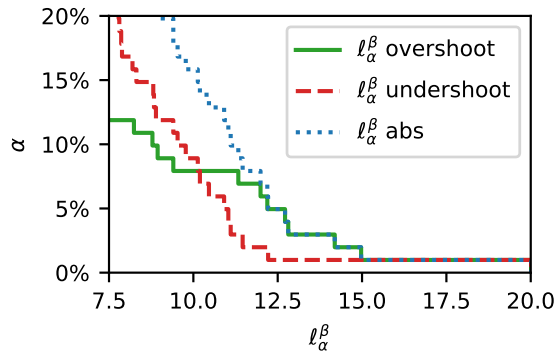


Figure 6: LAL-curves ($m = 1$) for the regression error analysis. Using different loss functions, all in units of dB, gives deeper insight in the model fit. Looking at α around $2 - 8\%$, using loss for under- and overshoot, we see that the model is more likely to overshoot out-of-sample data than it is to undershoot.

$y \in \mathbb{R}$ representing the sound measured in dB. The features $x \in \mathbb{R}^5$ Each data point is a feature-label pair $z_i = (x_i, y_i)$.

The calibration data $\mathcal{D}$ is constructed by weighted sampling of $n = 100$ samples. The probability to draw a data point (x_i, y_i) is proportional to $\exp(\begin{bmatrix}1 & 0 & 0 & 0 & -1\end{bmatrix} x_i)$, making data points with high frequency and small displacement more likely to sample, similar to the distribution shift experiments in Tibshirani et al. (2020, sec. 2.2). The remaining 1 403 data points constitute the training data set $\mathcal{D}_0$.

The model $f(x)$ uses a spline basis that is fit using least-squares with L2-regularization and cross-validation. To study the asymmetry of prediction errors, we compute the LAL-curves for a subsequent experiment ($m = 1$) using three different loss functions,

$$\text{overshoot loss} \quad \ell(x, y) = \max(0, f(x) - y) \tag{13}$$

$$\text{undershoot loss} \quad \ell(x, y) = \max(0, y - f(x)) \tag{14}$$

$$\text{absolute loss} \quad \ell(x, y) = |y - f(x)| \tag{15}$$

which are all in units of dB to enable a physical interpretation. The results are shown in Fig. 6, where we observe that f produces more severe overshoots than undershoots. The chosen loss functions (13-15) differs from the loss function used for model fitting (squared-error), illustrating the freedom that an analyst has to characterize the performance of f.

4.5 Model comparison

This experiment is concerned with model selection. The data used is the monthly number of earthquakes worldwide with magnitude ≥ 5 between 2012 and 2022 (USGS, 2022). There are 120 data points $z_1, \ldots, z_{120}$, with $z_i \in \{0, 1, 2, \ldots\}$ These are randomly split into 100 and 20 data points, forming $\mathcal{D}_0$ and $\mathcal{D}$. We learn two models of the number of earthquakes per month z, $\mathbb{P}(Z = z)$, using the maximum likelihood method: a Poisson model $f^{\text{Poisson}}(z)$ and a Negative Binomial model $f^{\text{NegBin}}(z)$. The models are evaluated using the negative log-likelihood loss

$$\ell(z) = -\log f(z) \tag{16}$$

and Fig. 7 presents their respective LAL-curves for a subsequent earthquake, i.e., $m = 1$. From this result we can conclude that the simpler one-parameter Poisson model produces much larger out-of-sample losses than the two-parameter Negative Binomial model.

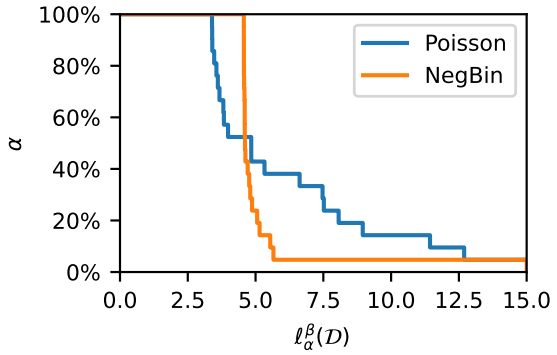


Figure 7: Comparison of models for earthquake statistics, using LAL curves and $m = 1$, $\beta = 1$. The loss function used is the negative log-likelihood of data under the fitted model. By considering how it handles the $\alpha = 20\%$ most difficult-to-fit data points, we see that the Negative Binomial model provides a better fit than the Poisson model.

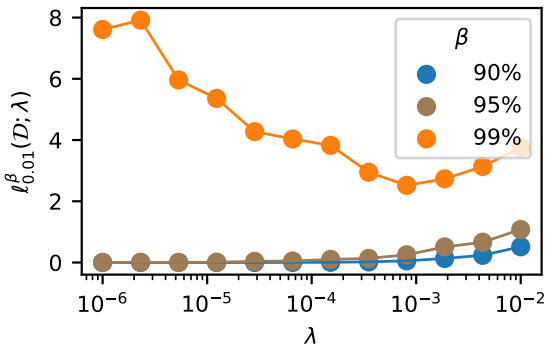


Figure 8: Relation between a regularization parameter λ in neural network model $f^\lambda(x)$, and the corresponding LAL for $m \to \infty$, varying β, and $(1-\alpha) = 99\%$ confidence. Appropriate regularization improves the performance in the tail of the loss distribution ($\beta = 99\%$), with an optimum around $\lambda \approx 10^{-3}$. The bulk of the losses ($\beta = 95\%$) are not reduced by regularization; their LAL increase with λ.

We compare the LAL analysis with a common model selection metric, the Akaike Information Criterion (AIC) (Ding et al., 2018). It evaluates the models as $\text{AIC}^{\text{Poisson}} = 1\,926.34$ and $\text{AIC}^{\text{NegBin}} = 1\,015.63$. One may also compare models by the average loss on the calibration data, yielding 5.69/4.78 nats for the Poisson/NegBin. Both these model selection metrics favor the NegBin model, in agreement with the LAL-curve. Whereas the model selection metrics report a single number, the LAL-curve presents a more nuanced picture, showing the reduction in tail losses.

4.6 Hyper-parameter tuning

This experiment shows how the LAL-curve can be used for tuning a hyperparameter when training a neural network.

The data set is the MNIST handwritten digits (Bottou et al., 1994). The training data set $\mathcal{D}_0$ has $n_0 = 6 \cdot 10^4$ data points, and the calibration data set $\mathcal{D}$ has $n = 10^4$ data points. The features x_i are images of handwritten digits, and the labels y_i are integers `0-9`. Data points are tuples $z_i = (x_i, y_i)$.

We construct a family of models $f^\lambda(x) = \text{NN}(x; \theta^\lambda)$. The function $\text{NN}(x; \theta^\lambda)$ consists of a dense feed-forward neural network with 3 hidden layers of 250 units each, parameters θ^λ, ReLU activations and softmax output. The parameter θ^λ is learned by minimizing the cross entropy, using the Adam optimizer with an L2-regularization parameter λ (aka. 'weight decay' in the deep learning literature). The optimization was run for 100 epochs, employing a batch size of 1024 and learning rate 0.01. The output of the network $f^\lambda(x)$ is a 10-dimensional vector, where the ith component approximates $\mathbb{P}[y = i|x]$.

We evaluate the model using the negative log-likelihood loss

$$\ell(x, y) = -\log f_y^\lambda(x) = -\log \left[\text{NN}(x; \theta^\lambda)\right]_y \tag{17}$$

Since deep learning methods may be deployed without retraining over a large number of predictions, we consider the case $m \to \infty$ (Cor. 1) and study the LAL for a β-fraction of out-of-sample losses. The results in Fig. 8 show that small regularization initially reduces losses for outliers without increasing loss on nominal samples.

5 Discussion

This section elaborates on connections to related fields.

5.1 Model evaluation and selection techniques

Cross-validatory out-of-sample expected loss estimation is arguably the most common approach to model evaluation in machine learning (Arlot & Celisse, 2010; Stone, 1974; Bishop, 2006; Hastie et al., 2009). The idea is that model performance is measured by expected loss on out-of-sample data (called the *risk*), and cross-validation estimates this quantity. In its most common form, a fraction of the training data is held out from model fitting. The average loss of the model is computed on the held out data, forming an estimate of the risk. Repeated splitting and refitting of the model (k-fold cross-validation) can be used to estimate the bias and variance of such estimates. Model evaluation with LAL shifts focus to evaluating model performance on probabalistic bounds on out-of-sample losses. This is significant when the out-of-sample loss distribution is multimodal, skewed or otherwise not well described by its mean and variance alone.

Evaluating models with respect to the risk may be done without cross-validation. Statistical learning theories, such as the VC-theory (Vapnik, 1991), provide asymptotic bounds on the risk under certain assumptions on the models and the data. Similarly, M-estimation (Vaart, 1998) provides another asymptotically valid method of fitting parametric models, quantifying the convergence of the average loss on training data to the risk. If the losses are bounded and the samples are iid or constructed as sampling without replacement, non-parametric results for inferring the risk appear to be promising (Waudby-Smith & Ramdas, 2020), improving on both the Hoeffding inequality and the Empirical Bernstein bounds (Audibert et al., 2009; Maurer & Pontil, 2009). For unbounded losses and non-iid exchangeable data, the inferential problem of estimating the risk remain open. LAL-curves provide a nonparemetric and nonasymptotic way to evaluate model performance by not using the risk as the quantity of comparison.

In this work, we have computed $\ell_\alpha^\beta(\mathcal{D})$ at given confidence levels α. Conversely, one can interpret $\ell_\alpha^\beta(\mathcal{D})$ as the boundary value for rejection of the null hypothesis that model predictions are exchangeable, at level α. Others have used hypothesis testing for model evaluation. Posterior predictive checks (Gelman et al., 2013; Rubin, 1984), and the data consistency criterion, (Lindholm et al., 2019) rely on exchangeability between observed data and data generated by the model to this end. Those methods are used to test whether a model is compatible with data, and reports a p-value for the test. A LAL-based analysis acknowledges that models are always misspecified and instead quantifies how well a model performs.

An analyst may wish to not only diagnose and compare models, but also decide which model in a set of candidates to use. This is called *model selection* (Ding et al., 2018). Two principal goals motivate the model selection algorithm: 1) model selection for inference, or 2) model selection for prediction. The first goal has been used to motivate well known information criteria such as the AIC (Stoica & Selen, 2004). Information critera has also been combined with hypothesis testing to assess if a certain model or model class can be said to be significantly closer to the true one than other candidates (Vuong, 1989). Regarding the second goal, Ding et al. (2018) posits that the 'best' model is the one with smallest expected out-of-sample loss, which may be estimated by cross-validation. By fixing α at any desired level, the LAL can be used to rank models in fashion similar to cross-validatory risk estimation. However, by employing the full LAL-curve as indicated in Sec. 4.5, a better informed decision can be taken by balancing typical versus outlier performance.

The Value-at-Risk (VaR) is a risk measurement used in the finance industry. For a random variable L describing finanical loss, the value-at-risk at level $1-\alpha$ can be defined as (Jorion, 2007, Eq 5.1)

$$\mathrm{VaR}_{1-\alpha}(L) = \inf\{x : \mathbb{P}[L > x] \leq \alpha\}$$

Note that x is not a random variable in the above expression. Comparing with (3) and fixing $m = 1$, we find that the VaR is the smallest LAL, conditional on the calibration data $\mathcal{D}$. Any VaR constitute a valid LAL for $m = 1$, but the converse is not true. Many VaR estimation methods rely on parametric assumptions or are only asymptotically valid. (Christoffersen, 2009; Alemany et al., 2013) The relaxed definition of LAL in contrast to VaR enables the distribution-free methodology presented in this article. For its use in e.g. financial risk measurements, VaR has been further developed into several variants such as Conditional VaR,

Tilted VaR, Entropic VaR, and more. (Rockafellar & Uryasev, 2002; Li et al., 2021) This line of reasearch may be beneficial for the LAL as a diagnostic tool, and constitute a direction for further research, but some of the requirements on financial risk metrics may not necessarily carry over to all model performance metrics.

5.2 Non-parametric statistics and conformal prediction

For ease of exposition, the LAL has primarily been discussed as a point value $\ell_\alpha^\beta(\mathcal{D})$. One can also consider it as the boundary point of the interval $(-\infty, \ell_\alpha^\beta(\mathcal{D})]$. As such, it can be understood as a variant of a non-parametric tolerance interval (Thm. 1), a prediction interval (Thm. 2) or a confidence interval (Rem. 4). See for instance Vardeman (1992) for a comparison between the different statistical intervals.

Fligner & Wolfe (1976) show how to construct non-asymptotic, non-parametric prediction intervals for quantiles of future data, and is a methodological precursor for the results in this paper.

Considering LAL a confidence interval of a quantile for iid data, there are other results with exact coverage (Zieliński & Zieliński, 2005), whereas the formula presented in this article is sometimes conservative. One could use that method with LAL to get exact guarantee. Such intervals are constructed via extra randomization and become harder to interpret. The LAL-curves in specific have no exact counterpart. We have therefore chosen to avoid this construction.

The theory of nonparametric prediction intervals also forms a foundation for conformal prediction. This field focuses on producing *prediction sets* for the output of any predictive model f, see e.g. Vovk et al. (2005) or Angelopoulos & Bates (2021) for introductions to the field. We will clarify the connection between conformal prediction and the LAL, in the case of split-conformal inference. The general case is essentially identical.

Consider a set of exchangeable random vectors $\{W_i\}_{i=1}^{M+1}$ taking values in $\mathcal{W}$. We wish to produce a prediction set $\mathcal{C}_\alpha\left(\{W_i\}_{i=1}^{M}\right)$ so that

$$\mathbb{P}[W_{M+1} \in \mathcal{C}_\alpha\left(\{W_i\}_{i=1}^{M}\right)] \geq 1-\alpha$$

To this end, define a real-valued *nonconformity score* $A : W_i \mapsto A(W_i) = A_i$, with the semantic that a large value means W_i does not conform to the general data set. Since the score is real valued, one can employ similar principles as in Thm. 2 to define a prediction interval $\mathcal{A}_\alpha$ such that

$$\mathbb{P}[A_{M+1} \in \mathcal{A}_\alpha\left(\{A_i\}_{i=1}^{M}\right)] \geq 1-\alpha$$

By letting $\mathcal{C}_\alpha$ be the inverse image of $\mathcal{A}_\alpha\left(\{A_i\}\right)$ under A, i.e.,

$$\mathcal{C}_\alpha\left(\{W_i\}\right) = \left\{w | A(w) \in \mathcal{A}_\alpha\left(\{A_i\}_{i=1}^{M}\right)\right\},$$

we ensure the desired coverage. Conformal prediction methodology is thus largely centered on finding suitable nonconformity scores A that are computationally tractable and handle various inference targets and data distributions (Angelopoulos & Bates, 2021).

More recently (Lei et al., 2018, Thm. 2.1), an upper bound on the coverage rate was derived,

$$1-\alpha+\frac{1}{M+1} \geq \mathbb{P}[W_{M+1} \in \mathcal{C}_\alpha\left(\{W_i\}\right)] \geq 1-\alpha,$$

that holds whenever the non-conformity scores are almost surely unique. The line of reasoning is similar but not identical to Thm. 2.

6 Conclusion

We have proposed the level-α loss (LAL) curve as a diagonistic tool for out-of-sample analysis of a model f. The method requires specifying a loss function of interest and the access to a calibration data set. In return it provides finite-sample guarantees about the probability of a batch of out-of-sample losses exceeding a certain threshold. The LAL is simple to compute and easy to interpret. A series of numerical experiments have been presented to show its usefulness in regression error analysis, distribution shift analysis, model selection and hyper-parameter tuning. We anticipate that there are many other areas of applications for this methodology.

References

Ramon Alemany, Catalina Bolancé, and Montserrat Guillén. A nonparametric approach to calculating value-at-risk. *Insurance: Mathematics and Economics*, 52(2):255–262, 2013. ISSN 0167-6687. doi: https://doi.org/10.1016/j.insmatheco.2012.12.008. URL https://www.sciencedirect.com/science/article/pii/S0167668713000048.

Anastasios N. Angelopoulos and Stephen Bates. A gentle introduction to conformal prediction and distribution-free uncertainty quantification, 2021. URL https://arxiv.org/abs/2107.07511.

Sylvain Arlot and Alain Celisse. A survey of cross-validation procedures for model selection. *Statistics Surveys*, 4, Jan 2010. ISSN 1935-7516. doi: 10.1214/09-SS054. URL https://projecteuclid.org/journals/statistics-surveys/volume-4/issue-none/A-survey-of-cross-validation-procedures-for-model-selection/10.1214/09-SS054.full.

Jean-Yves Audibert, Rémi Munos, and Csaba Szepesvári. Exploration–exploitation tradeoff using variance estimates in multi-armed bandits. *Theoretical Computer Science*, 410(19):1876–1902, April 2009. ISSN 03043975. doi: 10.1016/j.tcs.2009.01.016. URL https://linkinghub.elsevier.com/retrieve/pii/S030439750900067X.

David A. Belsley, Edwin Kuh, and Roy E. Welsch. *Regression Diagnostics: Identifying Influential Data and Sources of Collinearity*. Wiley Series in Probability and Statistics. Wiley, 1 edition, Jun 1980. ISBN 978-0-471-05856-4. doi: 10.1002/0471725153. URL https://onlinelibrary.wiley.com/doi/book/10.1002/0471725153.

Christopher M. Bishop. *Pattern recognition and machine learning*. Information science and statistics. Springer, New York, 2006. ISBN 978-0-387-31073-2.

J. M. Bogoya, A. Böttcher, and E. A. Maximenko. From convergence in distribution to uniform convergence. *Boletín de la Sociedad Matemática Mexicana*, 22(2):695–710, October 2016. ISSN 1405-213X, 2296-4495. doi: 10.1007/s40590-016-0105-y. URL http://link.springer.com/10.1007/s40590-016-0105-y.

Léon Bottou, Corinna Cortes, John S. Denker, Harris Drucker, Isabelle Guyon, Lawrence D. Jackel, Yann Le Cun, Urs A. Muller, Eduard Säckinger, Patrice Simard, and Vladimir Vapnik. Comparison of classifier methods: a case study in handwritten digit recognition. In *Proceedings of the 12th IAPR International Conference on Pattern Recognition, Conference B: Computer Vision & Image Processing.*, volume 2, pp. 77–82, Jerusalem, October 1994. IEEE. URL http://leon.bottou.org/papers/bottou-cortes-94.

Stephen Casper, Kaivalya Hariharan, and Dylan Hadfield-Menell. Diagnostics for deep neural networks with automated copy/paste attacks. In *NeurIPS ML Safety Workshop*, 2022. URL https://openreview.net/forum?id=l-kqvueSRp7.

Peter Christoffersen. Value–at–risk models. In *Handbook of financial time series*, pp. 753–766. Springer, 2009.

Jie Ding, Vahid Tarokh, and Yuhong Yang. Model selection techniques: An overview. *IEEE Signal Processing Magazine*, 35(6):16–34, 2018. doi: 10.1109/MSP.2018.2867638. URL https://doi.org/10.1109/MSP.2018.2867638.

Dheeru Dua and Casey Graff. UCI Machine Learning Repository, 2017. URL http://archive.ics.uci.edu/ml.

Garrett M. Fitzmaurice, Nan M. Laird, and James H. Ware. *Applied longitudinal analysis*. Wiley series in probability and statistics. Wiley, Hoboken, N.J, 2nd ed edition, 2011. ISBN 978-0-470-38027-7.

Michael A. Fligner and Douglas A. Wolfe. Some Applications of Sample Analogues to the Probability Integral Transformation and a Coverage Property. *The American Statistician*, 30(2):78, May 1976. ISSN 00031305. doi: 10.2307/2683799. URL https://www.jstor.org/stable/2683799?origin=crossref.

João Gama, Pedro Medas, Gladys Castillo, and Pedro Rodrigues. Learning with drift detection. In Ana L. C. Bazzan and Sofiane Labidi (eds.), *Advances in Artificial Intelligence – SBIA 2004*, volume 3171 of *Lecture Notes in Computer Science*, pp. 286–295. Springer Berlin Heidelberg, Berlin, Heidelberg, 2004. ISBN 978-3-540-23237-7. doi: 10.1007/978-3-540-28645-5_29. URL http://link.springer.com/10.1007/978-3-540-28645-5_29.

Andrew Gelman, John B Carlin, Hal S Stern, David B Dunson, Aki Vehtari, and Donald B Rubin. *Bayesian Data Analysis, Third Edition.* CRC Press, Hoboken, 2013. ISBN 978-1-4398-9820-8.

Trevor Hastie, Robert Tibshirani, and Jerome H. Friedman. *The elements of statistical learning: data mining, inference, and prediction.* Springer series in statistics. Springer, New York, NY, 2nd edition, 2009. ISBN 978-0-387-84857-0.

Allison Marie Horst, Alison Presmanes Hill, and Kristen B Gorman. palmerpenguins: Palmer Archipelago (Antarctica) penguin data, 2020. URL https://allisonhorst.github.io/palmerpenguins/.

Philippe Jorion. *Value at risk: the new benchmark for managing financial risk.* McGraw-Hill, New York, 3rd ed edition, 2007. ISBN 978-0-07-173692-3.

Steven M. Kay. *Fundamentals of statistical signal processing.* Prentice Hall signal processing series. Prentice-Hall PTR, Englewood Cliffs, N.J, 1993. ISBN 978-0-13-345711-7 978-0-13-504135-2 978-0-13-280803-3.

R. Kelley Pace and Ronald Barry. Sparse spatial autoregressions. *Statistics & Probability Letters*, 33(3): 291–297, May 1997. ISSN 01677152. doi: 10.1016/S0167-7152(96)00140-X. URL https://linkinghub.elsevier.com/retrieve/pii/S016771529600140X.

J. F. C. Kingman. Uses of exchangeability. *The Annals of Probability*, 6(2), Apr 1978. ISSN 0091-1798. doi: 10.1214/aop/1176995566. URL https://projecteuclid.org/journals/annals-of-probability/volume-6/issue-2/Uses-of-Exchangeability/10.1214/aop/1176995566.full.

Yoonho Lee, Annie S Chen, Fahim Tajwar, Ananya Kumar, Huaxiu Yao, Percy Liang, and Chelsea Finn. Surgical fine-tuning improves adaptation to distribution shifts. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=APuPRxjHvZ.

Jing Lei, Max G'Sell, Alessandro Rinaldo, Ryan J. Tibshirani, and Larry Wasserman. Distribution-Free Predictive Inference for Regression. *Journal of the American Statistical Association*, 113(523):1094–1111, July 2018. ISSN 0162-1459, 1537-274X. doi: 10.1080/01621459.2017.1307116. URL https://www.tandfonline.com/doi/full/10.1080/01621459.2017.1307116.

Tian Li, Ahmad Beirami, Maziar Sanjabi, and Virginia Smith. Tilted empirical risk minimization, Mar 2021. URL http://arxiv.org/abs/2007.01162. arXiv:2007.01162 [cs, math, stat].

Andreas Lindholm, Dave Zachariah, Petre Stoica, and Thomas B. Schon. Data Consistency Approach to Model Validation. *IEEE Access*, 7:59788–59796, 2019. ISSN 2169-3536. doi: 10.1109/ACCESS.2019.2915109. URL https://ieeexplore.ieee.org/document/8708204/.

Zachary Lipton, Yu-Xiang Wang, and Alexander Smola. Detecting and correcting for label shift with black box predictors. In Jennifer Dy and Andreas Krause (eds.), *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 3122–3130. PMLR, 10–15 Jul 2018. URL https://proceedings.mlr.press/v80/lipton18a.html.

Jie Lu, Anjin Liu, Fan Dong, Feng Gu, Joao Gama, and Guangquan Zhang. Learning under Concept Drift: A Review. *IEEE Transactions on Knowledge and Data Engineering*, pp. 1–1, 2018. ISSN 1041-4347, 1558-2191, 2326-3865. doi: 10.1109/TKDE.2018.2876857. URL http://arxiv.org/abs/2004.05785. arXiv:2004.05785.

Ning Lu, Guangquan Zhang, and Jie Lu. Concept drift detection via competence models. *Artificial Intelligence*, 209:11–28, April 2014. ISSN 00043702. doi: 10.1016/j.artint.2014.01.001. URL https://linkinghub.elsevier.com/retrieve/pii/S0004370214000034.

Andreas Maurer and Massimiliano Pontil. Empirical Bernstein Bounds and Sample-Variance Penalization. In *COLT 2009 Proceedings*, Montreal, Quebec, Canada, 2009. URL https://www.cs.mcgill.ca/~colt2009/papers/012.pdf.

Joaquin Quiñonero-Candela, Masashi Sugyiama, Anton Schwaighofer, and Neil D. Lawrence (eds.). *Dataset shift in machine learning*. Neural information processing series. MIT Press, Cambridge, Mass, 2009. ISBN 978-0-262-17005-5.

Stephan Rabanser, Stephan Günnemann, and Zachary Lipton. Failing Loudly: An Empirical Study of Methods for Detecting Dataset Shift. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d' Alché-Buc, E. Fox, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL https://proceedings.neurips.cc/paper/2019/file/846c260d715e5b854ffad5f70a516c88-Paper.pdf.

Ali Rahimi and Benjamin Recht. Weighted sums of random kitchen sinks: Replacing minimization with randomization in learning. In D. Koller, D. Schuurmans, Y. Bengio, and L. Bottou (eds.), *Advances in Neural Information Processing Systems*, volume 21. Curran Associates, Inc., 2008. URL https://proceedings.neurips.cc/paper/2008/file/0efe32849d230d7f53049ddc4a4b0c60-Paper.pdf.

Herbert Robbins. A remark on stirling's formula. *The American Mathematical Monthly*, 62(1):26–29, 1955. ISSN 00029890, 19300972. URL http://www.jstor.org/stable/2308012.

R.Tyrrell Rockafellar and Stanislav Uryasev. Conditional value-at-risk for general loss distributions. *Journal of Banking & Finance*, 26(7):1443–1471, 2002. ISSN 0378-4266. doi: https://doi.org/10.1016/S0378-4266(02)00271-6. URL https://www.sciencedirect.com/science/article/pii/S0378426602002716.

Donald B. Rubin. Bayesianly Justifiable and Relevant Frequency Calculations for the Applied Statistician. *The Annals of Statistics*, 12(4):1151–1172, December 1984. ISSN 0090-5364. doi: 10.1214/aos/1176346785. URL http://projecteuclid.org/euclid.aos/1176346785.

David Ruppert, M. P. Wand, and R. J. Carroll. *Semiparametric Regression*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 2003. ISBN 978-0-511-75545-3.

P. Stoica and Y. Selen. Model-order selection: a review of information criterion rules. *IEEE Signal Processing Magazine*, 21(4):36–47, Jul 2004. ISSN 1558-0792. doi: 10.1109/MSP.2004.1311138.

M. Stone. Cross-Validatory Choice and Assessment of Statistical Predictions. *Journal of the Royal Statistical Society: Series B (Methodological)*, 36(2):111–133, January 1974. ISSN 00359246. doi: 10.1111/j.2517-6161.1974.tb00994.x. URL https://onlinelibrary.wiley.com/doi/10.1111/j.2517-6161.1974.tb00994.x.

Torsten Söderström and Petre Stoica. *System identification*. Prentice Hall international series in systems and control engineering. Prentice-Hall, New York, NY, reprint edition, 2001. ISBN 978-0-13-881236-2.

Ryan J. Tibshirani, Rina Foygel Barber, Emmanuel J. Candès, and Aaditya Ramdas. Conformal Prediction Under Covariate Shift. *arXiv:1904.06019 [stat]*, July 2020. URL http://arxiv.org/abs/1904.06019.

USGS. Earthquake Catalog, 2022. URL https://earthquake.usgs.gov/fdsnws/event/1/query.csv?starttime=2012-01-01%2000:00:00&endtime=2022-01-01%2000:00:00&minmagnitude=5&orderby=time.

A. W. van der Vaart. *Asymptotic Statistics*. Cambridge University Press, 1 edition, October 1998. doi: 10.1017/CBO9780511802256. URL https://www.cambridge.org/core/product/identifier/9780511802256/type/book.

V. Vapnik. Principles of Risk Minimization for Learning Theory. In J. Moody, S. Hanson, and R. P. Lippmann (eds.), *Advances in Neural Information Processing Systems*, volume 4. Morgan-Kaufmann, 1991. URL https://proceedings.neurips.cc/paper/1991/file/ff4d5fbbafdf976cfdc032e3bde78de5-Paper.pdf.

Stephen B. Vardeman. What about the other intervals? *The American Statistician*, 46(3):193, Aug 1992. ISSN 00031305. doi: 10.2307/2685212. URL https://www.jstor.org/stable/2685212?origin=crossref.

Vladimir Vovk, A. Gammerman, and Glenn Shafer. *Algorithmic learning in a random world*. Springer, New York, 2005. URL https://doi.org/10.1007/b106715.

Quang H. Vuong. Likelihood ratio tests for model selection and non-nested hypotheses. *Econometrica*, 57(2): 307, Mar 1989. ISSN 00129682. doi: 10.2307/1912557. URL https://www.jstor.org/stable/1912557?origin=crossref.

Ian Waudby-Smith and Aaditya Ramdas. Estimating means of bounded random variables by betting, 2020. URL https://arxiv.org/abs/2010.09686.

Yuhui Zhang, Jeff Z HaoChen, Shih-Cheng Huang, Kuan-Chieh Wang, James Zou, and Serena Yeung. Drml: Diagnosing and rectifying vision models using language. In *NeurIPS ML Safety Workshop*, 2022. URL https://openreview.net/forum?id=losu6IAaPeB.

Ryszard Zieliński and Wojciech Zieliński. Best exact nonparametric confidence intervals for quantiles. *Statistics*, 39(1):67–71, 2005. URL https://doi.org/10.1080/02331880412331329854.

A Details for Proof of Corollary 1

We must first show that

$$\lim_{m\to\infty} \frac{\binom{n-j+m-\lceil m\beta\rceil}{n-j}\binom{j+\lceil m\beta\rceil-1}{j}}{\binom{n+m}{m}} = \binom{n}{j}\beta^j(1-\beta)^{n-j}$$

By definition of binomial coefficient, this is

$$\lim_{m\to\infty} \frac{(n-j+m-\lceil m\beta\rceil)!(j+\lceil m\beta\rceil-1)!m!n!}{(n-j)!(m-\lceil m\beta\rceil)!j!(\lceil m\beta\rceil-1)!(n+m)!}$$

Rearrangement of factors gives

$$\binom{n}{j}\lim_{m\to\infty} \frac{(n-j+m-\lceil m\beta\rceil)!(j+\lceil m\beta\rceil-1)!m!}{(m-\lceil m\beta\rceil)!(\lceil m\beta\rceil-1)!(n+m)!}$$

So we must now show that

$$\lim_{m\to\infty} \underbrace{\frac{(n-j+m-\lceil m\beta\rceil)!(j+\lceil m\beta\rceil-1)!m!}{(m-\lceil m\beta\rceil)!(\lceil m\beta\rceil-1)!(n+m)!}}_{=:H(m)} = \beta^j(1-\beta)^{n-j}$$

By the upper and lower bounds in Stirling's formula (Robbins, 1955).

$$\sqrt{2\pi n}\ \left(\frac{n}{e}\right)^n e^{\frac{1}{12n+1}} < n! \leq \sqrt{2\pi n}\ \left(\frac{n}{e}\right)^n e^{\frac{1}{12n}}$$

Let $\delta \coloneqq \lceil m\beta \rceil - m\beta$. We introduce

$$
\begin{aligned}
h(m,Q) = &(2\pi(n-j+m(1-\beta)-\delta))^{1/2} && \left(\frac{n-j+m(1-\beta)-\delta}{e}\right)^{n-j+m(1-\beta)-\delta} && \exp\left[\frac{1}{12(n-j+m(1-\beta)-\delta)+Q}\right] \\
&\times (2\pi(j+m\beta+\delta-1))^{1/2} && \left(\frac{j+m\beta+\delta-1}{e}\right)^{j+m\beta+\delta-1} && \exp\left[\frac{1}{12(j+m\beta+\delta-1)+Q}\right] \\
&\times (2\pi m)^{1/2} && \left(\frac{m}{e}\right)^{m} && \exp\left[\frac{1}{12m+Q}\right] \\
&\times (2\pi(m(1-\beta)-\delta))^{-1/2} && \left(\frac{e}{m(1-\beta)-\delta}\right)^{m(1-\beta)-\delta} && \exp\left[\frac{-1}{12(m(1-\beta)-\delta)+(1-Q)}\right] \\
&\times (2\pi(m\beta-1+\delta))^{-1/2} && \left(\frac{e}{m\beta-1+\delta}\right)^{m\beta-1+\delta} && \exp\left[\frac{-1}{12(m\beta-1+\delta)+(1-Q)}\right] \\
&\times (2\pi(n+m))^{-1/2} && \left(\frac{e}{n+m}\right)^{n+m} && \exp\left[\frac{-1}{12(n+m)+(1-Q)}\right]
\end{aligned}
$$

allowing us to state succinctly

$$h(m,1) < H(m) \leq h(m,0)$$

The proof is complete if we can show that $\lim_{m\to\infty} h(m,Q) = \beta^j(1-\beta)^{n-j}$. To see this we rearrange the factors. We will also use little-oh notation, i.e. $f(m) = o(g(m))$ iff $\lim_{m\to\infty} |f(m)|/g(m) = 0$. When taking limits, we use that $0 \leq \delta < 1$ for all m.

$$
\begin{aligned}
h(m,Q) = &\underbrace{\left(\frac{2^3\pi^3}{2^3\pi^3}\frac{(n-j+m(1-\beta)-\delta)(j+m\beta+\delta-1)m}{(m(1-\beta)-\delta)(m\beta-1+\delta)(n+m)}\right)^{1/2}}_{=1+o(1)} \\
&\times \underbrace{e^{n-j+m(1-\beta)-\delta+j+m\beta+\delta-1+m-m(1-\beta)+\delta-m\beta+1-\delta-n-m)}}_{=1} \\
&\times \underbrace{\left(\frac{n-j+m(1-\beta)-\delta}{m(1-\beta)-\delta}\right)^{m(1-\beta)-\delta}}_{=\exp(n-j)+o(1)} \underbrace{\left(\frac{j+m\beta+\delta-1}{m\beta+\delta-1}\right)^{m\beta+\delta-1}}_{=\exp(j)+o(1)} \underbrace{\left(\frac{m}{n+m}\right)^{m}}_{=\exp(-n)+o(1)} \\
&\times \underbrace{(n-j+m(1-\beta)-\delta)^{n-j}}_{=m^{n-j}(1-\beta)^{n-j}+o(m^{n-j})} \underbrace{(j+\beta m+\delta-1)^{j}}_{=m^j\beta^j+o(m^j)} \underbrace{(n+m)^{-n}}_{=m^{-n}+o(m^{-n})} \\
&\times \underbrace{\exp\left[(12(n-j+m(1-\beta)-\delta)+Q)^{-1}+(12(j+m\beta+\delta-1)+Q)^{-1}+(12m+Q)^{-1}\right]}_{=1+o(1)} \\
&\times \underbrace{\exp\left[-(12(m(1-\beta)-\delta)+(1-Q))^{-1}-(12(m\beta-1+\delta)+(1-Q))^{-1}-(12(n+m)+(1-Q))^{-1}\right]}_{=1+o(1)}
\end{aligned}
$$

By calculus of little-oh notation we find

$$h(m,Q) = \exp(n-j)\exp(j)\exp(-n)m^{n-j}m^{j}m^{-n}\beta^j(1-\beta)^{n-j} + o(1)$$

which in turn means

$$\lim_{m\to\infty} h(m,Q) = \beta^j(1-\beta)^{n-j}.$$

We have thus shown that the limit of (4) is

$$\lim_{m\to\infty} a(k) = \sum_{j=k}^{n} \binom{n}{j} \beta^j(1-\beta)^{n-j},$$

and by recognizing the binomial cumulative distribution function (and denoting it with the symbol $\mathrm{BIN}(\cdot; n, \beta)$) we see

$$\lim_{m\to\infty} a(k) = 1 - \mathrm{BIN}(k-1; n, \beta).$$

Further algebra shows that

$$k^\star = \min\left\{k \in \{1, \ldots, n+1\} \middle| \lim_{m\to\infty} a(k) \leq \alpha\right\} = 1 + \mathrm{BIN}^{-1}(1-\alpha; n, \beta)$$

B Extended Distribution Shift experiment

Continuing on the experiment on distribution shift, from Sec. 4.2. We keep $n = 30$, and also fix the random seed across runs to make the comparison clear. We generate calibration data sets $\mathcal{D}$ for three scenarios:

Scenario 1, Increasing mean
The mean μ shift from 0.5 (no shift) to 3 (large shift), and the standard deviation $\sigma = 0.5$ (no shift).

Scenario 2, Increasnig variance
The mean $\mu = 0.5$ (no shift), and the standard deviation shifts from $\sigma = 0.5$ (no shift) to 2 (large shift)

Scenario 3, Decreasing variance
The mean $\mu = 0.5$ (no shift), and the standard deviation shifts from $\sigma = 0.5$ (no shift) to 0.05 (large shift)

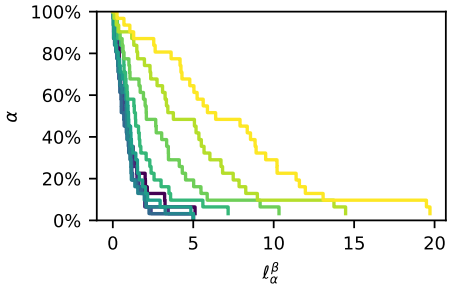
These three series of calibration-sets have their LAL-curves plotted; see Fig. 9.

The results of Scenario 1 show that as the mean shift increase, model performance generally deteriorate. One detail should be noted for moderate mean-shifts. Some of the outlier data points with $x < 0.5$ contributed to a large LAL when there is no shift, but with increasing shift, these data points are moved towards 0.5, where model performance is generally better. This effect means that the LAL decreases somewhat, as can be observed for $\alpha \approx 10\%$. When the mean-shift increases, outliers with $x > 0.5$ produce larger losses, which is observed through the LAL-curve moving to the right. A large mean-shift moves most of the calibration set to x where model performance is worse, as indicated by the whole LAL-curve shifting to the right, not only for selected α.

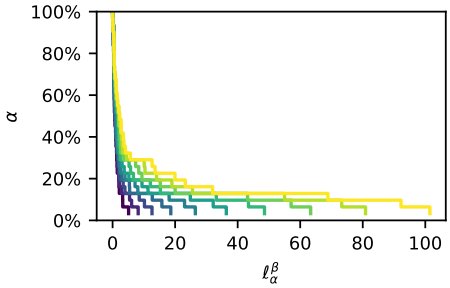
The results of Scenario 2 show that the tail of the LAL-curve increase with feature variance, indicating that model performance on outliers get worse. Inliers are not as much affected.

The results of Scenario 3 show that the LAL-curve moves to the left with decreasing feature variance, indicating smaller losses. This means that the data concentrates in regions of feature space where the model performs well. We do observe some distribution shift, but this model still performs well on the data.

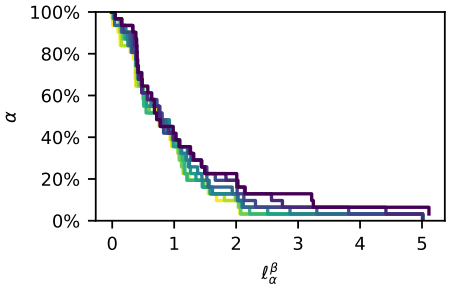
Taken together, this indicate that the LAL can indicate distribution shift, that it only detects shifts relevant for model performance, and that it indicates if the shift incurs model degradation generally, or only on outliers.



(a) LAL curves for distribution shift in Scenario 1, Increasing mean



(b) LAL curves for distribution shift in Scenario 2, Increasnig variance



(c) LAL curves for distribution shift in Scenario 3, Decreasing variance

Figure 9: LAL-curves are drawn for various calibration sets $\mathcal{D}$ under distributions shift, as described in Appendix B. Dark blue line = no shift. Yellow line = large shift.