Towards Error Measures which Influence a Learners Inductive Bias to the Ground Truth

A. I. Parkes
University of Southampton
A.I.Parkes@soton.ac.uk

A. J. Sobey
University of Southampton
The Alan Turing Institute
ajs502@soton.ac.uk

D. A. Hudson
University of Southampton
dominic@soton.ac.uk

Abstract

Artificial intelligence is applied in a range of sectors, and is relied upon for decisions requiring a high level of trust. For regression methods, trust is increased if they approximate the true input-output relationships and perform accurately outside the bounds of the training data. But often performance off-test-set is poor, especially when data is sparse. This is because the conditional average, which in many scenarios is a good approximation of the ‘ground truth’, is only modelled with conventional Minkowski-r error measures when the data set adheres to restrictive assumptions, with many real data sets violating these. To combat this there are several methods that use prior knowledge to approximate the ‘ground truth’. However, prior knowledge is not always available, and this paper investigates how error measures affect the ability for a regression method to model the ‘ground truth’ in these scenarios. Current error measures are shown to create an unhelpful bias and a new error measure is derived which does not exhibit this behaviour. This is tested on 36 representative data sets with different characteristics, showing that it is more consistent in determining the ‘ground truth’ and in giving improved predictions in regions beyond the range of the training data.

Keywords function approximation · explainable AI · error measures · inductive bias · regression

1 The Importance of Correctly Interpreting the Ground Truth

The artificial intelligence industry is expected to grow to $40 billion dollars by 2025 (Orbis Research 2018), revolutionising sectors from manufacturing to healthcare. Increasing the use of artificial intelligence for applications where dependability, safety or security are of concern requires trustworthy and explainable methods to facilitate human understanding and decision making. Regression problems, where relationships between the inputs and continuous output(s) are identified, comprise a large portion of the cutting edge of the machine learning field (Li et al. 2020). However, none of the commonly used performance error measures for regression problems assess whether the input-output relationships are modelled accurately; methods reporting low prediction error do not necessarily approximate the ground truth. If it cannot be verified that these methods accurately approximate the ground truth of their data sets then they cannot be trusted to perform in many real-world applications.

There are many situations where regression is required but there is not a full understanding of the input-output relationships, these range from engineering (Parkes et al. 2018) (Richmond et al. 2020) and financial (De Prado 2018) to medical applications (Sidey-Gibbons and Sidey-Gibbons 2019) (Wang et al. 2010). This can be due to a lack of domain knowledge as well as sparse or unrepresentative data sets, because data can be expensive to gather in large quantities or because the phenomenon of interest is rare. Machine learning methods like neural networks have large modelling flexibility, given sufficient data they can find patterns in problems where complexity prohibits the explicit programming of a systems precise physical nature. But they are known to produce physically inconsistent results and cannot generalise to off test sets, so have therefore not modelled the ground truth of the system (Willard et al. 2020).

Due to the flexibility of machine learning regression methods, any number of arbitrary patterns could be modelled which provide a good approximation to the fundamental relationship over a limited input-output domain, despite these functions having no similarities outside of this range; this effect is illustrated for two functions in Figure 1. This
Error Measures with a Bias to the Ground Truth
demonstrates the common issue of poor extrapolation for current machine learning techniques. For example, a method trained on the $x$ domain between $0 - 10$, with limited or poor quality data for the $x$ domain above 10, would have no way to discern which of the two curves was the true $x - y$ relationship. Both curves would produce similar error profiles for $x < 10$, using current error measures.

![Figure 1: Example of two curves tracking similar patterns in the x range from 0-10 but diverging above x values of 10.](image)

Therefore it cannot be assumed that regression methods producing low Minkowski-$r$ error measures, such as the Mean Squared or Mean Absolute Errors, model the ground truth accurately. Even methods to balance the bias-variance trade-off, such as regularisation and cross-validation, act to minimise the error or to fix overfitting but don’t account for the ground truth. In fact, there are restrictive assumptions about a data set that must be met to ensure that a regression method approximates the conditional average of the data set, the average output value conditioned on each input in turn (Bishop 1995). One of the methods with the least restrictive assumptions are neural networks, and these assumptions are discussed further in Section 2, where it is noted that the number of applications which violate these assumptions is larger than is apparent in the standard literature.

Even if there is certainty that the regression producing minimal loss approximates the conditional averages of the data set, the conditional averages do not necessarily approximate the fundamental relationships which generated the data set. This can be caused by sparse areas of data which can be affected by the ‘law of small numbers’ creating misleading averages (Tversky and Kahneman 1971). Additionally, the interaction of noise and high convexity or concavity in input-output relationships creates a gap between the conditional averages of a data set and the underlying relationships which generated them, a direct result from Jensen’s inequality (Jensen 1906), discussed in more detail in Section 3.

To address this problem a number of methods exist to improve a method’s approximations of the ground truth. These methods include biasing the architecture of learning methods to known relationships (Park and Park 2019) (Anderson et al. 2019) (Zhang et al. 2018); using physics-guided initialization (Read et al. 2019) (Sultan et al. 2018); and adding a ‘distance from the ground truth’ measure to the loss function (Karpatne et al. 2018). They can be used to produce Reduced Order Models (Lucia et al. 2004) or improve the prediction from the physical model alone. These have been shown to produce better extrapolation predictions as their approximated patterns are robust to sparse regions of data (Willard et al. 2020). Methods also exist to mitigate poor inductive biases or learning biases, where a method tends towards error minima which produce trained methods that do not generalise off-test set, by generating extra data (Abusitta et al. 2020). However, as the discussed methods either train models so that they have an inbuilt bias to the known input-output relationships, or generate extra data based on an understanding of the system being modelled, knowledge of these relationships is required. There is currently no way to measure the extent to which a method approximates the ground truth if it is not already known.

All regression methods require a loss function to produce predictions on a test set. Traditional maximum likelihood regression approaches such as generalized linear models, support vector regressions, and other kernel regressions require an intelligent selection of multiple model parameters (Nelder and Wedderburn 1972) and (Drucker et al. 1996). These parameter choices are often based on variable distributions, variable dependence and noise characteristics (Murphy 2012) which are not necessarily known. Bayesian methods, including Gaussian processes and Bayesian neural networks, provide a solution to not knowing the best parameter value, or basis function, to use in a model by using...
’non-informative priors’ (Bernardo 1979), and integrating over all feasible values to identify the optimal parameters for any given application (Lampinen and Vehtari 2001). However, in addition to the increased computational complexity this causes (Green et al. 2015), applications of the ‘no free lunch’ theorem state that if you make no assumptions concerning the target, then you have no assurances about how well you generalise (MacKay 1992), so if there is no prior knowledge to build into a Bayesian method and ‘non-informative priors’ are used exclusively, then the model will have limited practical uses. In addition, despite being able to perform inference without reference to a loss function, Gaussian processes still require a loss function to identify the optimal prediction for a testing set (Rasmussen and Williams 2006). They use the same Minkowski-r family of error measures and therefore have the same problems determining the conditional average.

This paper focuses on the situation where there is a partial, or no, understanding of the input-output relationships in a data set, so current approaches to fitting to the ground truth can not be used. A new error measure to assess how accurately the ground truth is mapped by a regression is proposed and is shown to be effective. The interaction between the conditional averages and the ground truth is also formalised for these data sets. It highlights that in these situations minimising traditional error measures cannot guarantee an accurate approximation of the ground truth, and discusses potential causes for the poor inductive bias produced by these error measures. The error measures developed in this study can be applied to any regression method, as they all require the use of a loss function to make predictions. Neural networks are employed as the exemplary method in this study as they are used prolifically in the machine learning literature due to their flexibility and ease of implementation.

2 Current Regression Error Measures

Nearly all common regression error measures are point-based and can be defined as in (Botchkarev 2018):

\[ G_{i=1,\ldots,n}(N[D(y_i, \hat{y}_i)]) \]

where \(D\) is some distance measure between the two points \(\hat{y}_i\) the predicted target variable(s) and \(y_i\) from the measured data set, \(N\) is some normalisation method, and \(G\) some aggregation method. Other error measures include the \(R^2\) or Pearson correlation coefficient but interpreting model performance from correlation measures is more difficult than interpreting point-based error measures (Taylor 1990), as they provide limited intuition about model behaviour for a single prediction or use in ‘real-life’ scenarios.

The most popular error measures used for regression are the Minkowski-r distance measures (Hanson and Burr 1987), where the distance measure is the Euclidean distance to some power \(r\) and the aggregation is the mean,

\[ M = \frac{1}{r} \sum_i N[|y_i - \hat{y}_i|^r], r \in \mathbb{R}. \]

In these error measures higher \(r\) values increase the penalty for large deviations, and smaller \(r\) values reduce the influence of outliers in feature space during learning. When the regression output only has one dimension, these measures equate to Mean Squared Error for \(r = 2\) and Mean Absolute Error for \(r = 1\). Other variants can be produced by varying the normalisation method; for example, Mean Absolute Percentage Error is produced using a normalisation with the target variable value. Theoretically, minimising the Mean Squared Error of a method creates an approximation of the conditional mean of the data set (Bishop 1995), and a method which minimises Mean Absolute Error should approximate the conditional median of the data, as this is less affected by outliers, however, these proofs rely heavily on several assumptions. The assumptions vary slightly for different methods, one of the least restrictive sets is that for neural networks:

(i) the datapoints are independent;
(ii) the distribution of the target variable is to be deterministic of the input with Gaussian noise, e.g. \(y = \phi(x) + \epsilon\) where \(\phi\) depends only on input variable \(x\), and \(\epsilon \sim N(0, \sigma^2)\);
(iii) the standard deviation of noise, \(\sigma\), is not dependent on the input \(x\);
(iv) and the data set and neural network must be sufficiently large.

These are not trivial assumptions; requiring that the standard deviation of the noise is not dependent on the input values implies that only homoscedastic data sets can be modelled. In many applications these assumptions do not hold, although meaningful results can still be produced by applying neural networks with Mean Squared Error or Mean Absolute Error error measure there is no certainty that the trained networks approximate the conditional average of the
Error Measures with a Bias to the Ground Truth

data set. For example, if noise in the target variable is not Gaussian, then the results cannot distinguish between the true
distribution and any other distribution with the same mean and variance. It is suggested that this is the reason regression
networks can model a data set with low errors, but model arbitrary patterns rather than the conditional average of the
data set.

This paper specifically investigates data sets where at least one $\phi$, the input-output relationship for a single input feature,
is a non-linear function. So there is some curvature, either concave or convex, in the system being modelled. For
applications modelling physical causality in a system a conditional average is likely to approximate the ground truth
of the data set, although depending on the characteristics of input-output relationship $\phi$ and any noise the choice of
conditional average may be important. By synthesising results from (Jensen 1906) and (MacGillivray 1981) we have
that, for data sets where $\phi$ has a larger curvature and higher standard deviations of noise, the conditional median will
approximate the ground truth closer than the conditional mean compared to data sets with a lower curvature and smaller
standard deviation of noise, where the mean and median will tend to the same value as the conditional distributions
have less of a skew.

However, the conditional average does not necessarily approximate the ground truth for all data sets. For example if the
relationship between inputs and outputs are a relation and not a function, there are several valid values for the outputs,
illustrated in Figure 2. This means the conditional average of the data set can hold no useful information about the data
at all, as the average of several solutions is not necessarily itself a solution.

The use of Minkowski-r error measures, most commonly Mean Squared Error and Mean Absolute Error, on data sets
which do not adhere to the above assumptions leads to poor performance outside areas of dense data as the ground truth
has not been approximated. This reduces the level of trust a user has for a model and inhibits our understanding of why
certain predictions are correct and some are incorrect. This lack of reproducibility is a growing problem in the field,
(Hutson 2018) and our lack of understanding means that these models are unusable in some domains (Voosen 2017). It
also reduces the transferability of trained methods as the causal relationships are not identified by these models.

3 Determination of a New Error Measure: Mean Fit to Median

To derive an error measure which quantifies how accurately a regression method models the ground truth in a data set,
an approximation of the isolated input-output relationships is required. Therefore, the proposed new measure calculates
a proxy for the relationship between each input variable and the output, the conditional average independent of other
input variables. It measures the distance between the approximation of these relationships and the proxy curves.

First, these relationships are formalised for any data set; for clarity the principles of the new error measure are explained
using a two dimensional example, where input $x$ is used to predict output $y$, and this is expanded to include scenarios
with multi-dimensional inputs for the formal statement of the error measure. In this initial reduced set-up, let \( \phi \) be the relationship between input variable \( x \) and output variable \( y \), such that \( y = \phi(x) \). For simplicity, the only curves \( \phi \) discussed are either convex or concave across their entire domain, although it is suggested that the arguments extend to a piecewise curve.

From this data set we have the distributions \( X \) and \( Y(= \phi(X)) \) shown on the axis of Figure 3a and \( E[X] \) and \( E[Y](= E[\phi(X)]) \) which are defined as the expectation, or mean, of the distributions. Assuming these distributions are sufficiently continuous, a naive proxy for \( \phi \) can be derived by estimating the conditional averages of the data set. To produce this the \( X \) domain must first be divided into \( n \) equally spaced bins \( (X_1, X_2, \ldots, X_n) \), Figure 3b. Then \( Y(= \phi(X)) \) is partitioned based on the data points’ corresponding \( x \) values, producing \( (Y_1, Y_2, \ldots, Y_n) \). \( Y_i \) is therefore the set of all the observed output values for the input values in \( X_i \). Assuming a sufficiently large set of partitions \( n \), the mean of all \( Y_i \) and the midpoints of \( X_i \) create an initial proxy for the conditional average: \( \text{mid}(X_i), E[Y_i] \) for \( i = 1, 2, \ldots, n \).

![Figure 3: Illustration of (A) expectation of X and Y distribution, (B) partitioning X distribution to create proxy curve.](image)

However, from Jensen’s inequality we have that \( \phi(E[X]) \leq E[\phi(X)] \) for convex \( \phi \) (Jensen 1906). This is illustrated in Figure 4a where the function value at the average of the input variable distribution \( (\phi(E[X])) \) is less than the average of the distribution of the function values of the inputs \( (E[\phi(X)]) \). This means that for any \( \phi \) which is not a straight line, a gap between \( \phi(E[X]) \) and \( E[\phi(X)] \) will exist and hence the mean of \( Y_i \) will not approximate \( \phi(X_i) \) for these applications. Formal bounds on this inequality exist (Gao et al. 2018), but cannot be applied without prior knowledge of \( \phi \).

We know that \( \text{med}(\phi(X)) \leq E[\phi(X)] \) for unimodal right-skewed distributions from basic statistics (MacGillivray 1981), so the median of each \( Y_i \) (denoted \( \text{med}(\phi(Y_i)) \)) could produce better estimates of \( \phi \) under certain constraints. If \( \phi \) is non-linear then \( X_i \) and \( Y_i \) will have a skew and, as continuity is already assumed, these subsets of the distributions will be unimodal. Jensen’s inequality also holds for medians; \( \phi(\text{med}(X)) \leq \text{med}(\phi(X)) \) for convex \( \phi \) (Merkle 2005). Figure 4b. So although there will also be a gap between \( \text{med}(\phi(X)) \) and the true curve \( \phi \) , \( \text{med}(Y_i) \) will produce a closer approximation than \( E[Y_i] \). The proposed measure is therefore the ‘Mean Fit to Median’ which is the mean of the distance between the proxy points for the conditional median and the isolated input-output relationships predicted by a machine learning regression method.

Here the notation is expanded to include multiple input variables, so that there are \( m \) independent input variables \( x^j \) such that \( y = \phi(x^1, x^2, \ldots, x^m) \). The distribution of each input variable is partitioned as above, formalised this produces

\[
\text{proxy curves} = \{ \text{med}(Y^j_i) | i = 1, 2, \ldots, n, j = 1, 2, \ldots, m \},
\]

where \( Y^j_i \) is the set of \( y \) values corresponding to the set \( X^j_i \).

\[^1\text{And } \phi(E[X]) \geq E[\phi(X)] \text{ for concave } \phi.\]
Error Measures with a Bias to the Ground Truth

Figure 4: Illustrations of Jensen’s inequality $\phi(E[X]) \leq E[\phi(X)]$ for convex function $\phi$ where A) the transformed mean of a distribution $\phi(E[X])$ is less than the mean of the transformed distribution $E[\phi(X)]$ (Jensen 1906) and B) where the gap caused by this inequality is smaller for the median of the distributions than for the mean. (Merkle 2005).

The distance between these proxy curves and the input-output relationships learnt during training is calculated to produce the Mean Fit to Median Error value. The input-output relationships learnt by a trained regression method, are produced by probing a trained method to predict output values for each input variable in turn by using inputs, cycling from the minimum to maximum observed values of each input, with all other inputs at the median value. Explicitly, if $P$ denotes the representations learnt from a model such that $P(x^1, x^2, \ldots, x^m)$ is the predicted output value for inputs $x^1, x^2, \ldots, x^m$. Then for input variable $j$, the predicted input-output relationship is approximated by the set

$$\text{learnt input-output curves} = \{p_{X_i}^j | i = 1, 2, \ldots, n, j = 1, 2, \ldots, m\},$$

where $p_{X_i}^j = P\left(\text{mid}(X_i^j), \text{med}(X^k) \text{ for } k \neq j\right)$ with $\text{mid}(X)$ denoting the midpoint of the set $X$. This allows the Mean Fit to Median Error measure to be written as

$$\text{Mean Fit to Median} = \frac{1}{m} \sum_{j=1}^{m} \frac{1}{n} \sum_{i=1}^{n} \left|\text{med}(Y_i^j) - p_{X_i}^j\right|.$$  \hspace{1cm} (1)

It is only possible to calculate this error measure after normal training has terminated, as the input-output relationships learnt by the method are required to perform the calculation. This means that currently it is only possible to use the ‘Fit to Median’ measure as a post-training evaluation method. However, it can be used to identify the method with the best ground truth approximation from a set of trained methods. Due to the extension of Jensen’s inequality to medians (Merkle 2005) the Fit to Median does not generalise to solve inverse problems, as illustrated in Figure 2.

4 Generating Artificial Data Sets

For many real regression applications a full understanding of the system and its uncertainties is not available. This means that benchmarking to what extent a trained method replicates the true relationships between inputs and outputs is not possible. To allow this to be accurately measured artificial data sets are used, with a range of fully defined variable relationships and differing levels of uncertainty, with the aim of approximating the complexity and characteristics of common regression problems. The data is also generated with the intention of violating assumptions (ii)-(iv), discussed in Section 2.

This data generation process is formalised as follows: the input variables are generated such that $X \sim N_6(\mu, \Sigma)$, with $\mu \in \{(\mu_0, \ldots, \mu_5)|\mu \in \mathbb{R}\}$ and for initial benchmarking $\Sigma$, the covariance matrix, is diagonal with a view to
Error Measures with a Bias to the Ground Truth

investigate scenarios where input variables are not independent in a further study. In practice, \( \mu_i \) is randomly generated as \( \mu_i \sim \text{unif}(0, 10) \) from numpy random number modules and the number of observations generated is 1,000,000. To generate one output \( y \) the following procedure is used;

\[
y = \phi(x_0, x_1, \ldots, x_5) = \phi(X) = W \cdot F(X),
\]

where \( W \in \{(w_0, \ldots, w_5)|w_i \in [0,1], \sum w_i = 1\} \) and,

\[
F = (\tilde{f}_0, \ldots, \tilde{f}_5).
\]

Where noise in the \( y \) direction is introduced as

\[
\tilde{f}(x) = f(x) + \epsilon,
\]

with

\[
\epsilon \sim N(0, \sigma^2),
\]

and \( \sigma \in \mathbb{R} \) not dependent on \( x \). The isolated input-output functions, \( f_i \), are represented as

\[
f \in \mathbb{P}_5[x],
\]

where \( \mathbb{P}_n[x] = \{p_0 + p_1 x + p_2 x^2 + \cdots + p_n x^n | p_i \in \mathbb{R}\} \). For each data set the degree of the polynomials \( f_i \) are randomly chosen from the range specified, for example \( \mathbb{P}_{a-b} \) indicates the degrees of \( f_i \) are chosen integers from the range from \( a - b \) inclusive. Then the coefficients are randomly selected uniformly from the interval \((-10, 10)\).

This produces a data set for a regression where the known relationships between input and output variables are \( W \cdot F \) or \( w_i \cdot f_i \) \( \forall i \), which are all functions and will be denoted \( f_i \), and is the isolated relationship between the input \( i \) and the output. A range of curvatures of \( f_i \); noise in the \( y \) direction; and covariance matrices for generating \( X \) are tested, Table 1 to identify the types of data where the error measure is effective.

| \( \sigma \) | \( \mathbb{P}_{0-1} \) | \( \mathbb{P}_{2-3} \) | \( \mathbb{P}_{4-5} \) |
|-------------|-----------------|-----------------|-----------------|
| 0.1         | a               | b               | c               |
| 1           | d               | e               | f               |
| 10          | g               | h               | i               |
| 100         | j               | k               | l               |

Table 1: Varieties of dataset generated for trialling the error measure, Illustrated in Figure 5

The data generated for this study aims to replicate a scenario where a single target variable is dependent on 6 independent input variables, with their own specific relationship to the target. Each input variable has an arbitrarily generated polynomial relationship, \( f_i \), with noise independent of the other input variables, the target, \( y \), is the weighted average of each \( f_i \).

For data sets with non-linear polynomials this creates non-Gaussian noise in the \( y \) distribution, violating assumption (ii). The reason for violating this assumption is that many applications require modelling of systems with non-Gaussian noise in the target variable. Although the noise introduced in these artificial data sets has the same Gaussian distribution for each input variable, their introduction at the input variable level is sufficient to provide non-Gaussian noise in the output variable. The interaction between input-level noise also produces heteroscedastic data sets, which violates assumption (iii).

The noise distributions are illustrated by the scatter plot in Figure 5c and 5d, the combination of noise from the other 5 inputs creates an increased spread below the concave \( x_5 - y \) relationship, explained by Jensen’s inequality. This skewed conditional distribution of the output variable means that often the conditional median approximates the ground truth closer than the conditional mean.

Input variables are generated as normal distributions around an arbitrary mean, so there is not sufficient data across the entirety of each \( x_i \) domain, this violates assumption (iv). The shading of boxplots in Figure 5a, 5c and 5e illustrates the normal distribution of \( x_i \) values in one of the generated data sets, where a mean of 9.25 and standard deviation of 1, has created ranges of sparse data for \( x_i < 6 \) and \( x_i > 13 \).

This methodology produces a data set which is not equivalent to any specific data set, but allows validation of the different regression error measures on a data set that can be controlled whilst replicating the main features from real data. However, the artificial input values all have a normal distribution with equal mean and median values whereas
Error Measures with a Bias to the Ground Truth

![Diagram](image)

Figure 5: Illustrations of different data sets generated for this study: A) boxplots showing the spread of y values for each interval of \(x_0\) along with \(f_0\) from a data set of type a from Table 1; shading of each boxplot shows the quantity of datapoints in any given interval, illustrating the sparse areas of data created at the tails of the \(x_1\) distribution; B) scatter plot of \(x_0\) to \(y\) for a data set of type a showing \(f_0\); C) boxplots showing the spread of y values for each interval of \(x_3\) along with \(f_3\) from a data set of type h; D) scatter plot of \(x_3\) to \(y\) for a data set of type h showing \(f_3\); E) boxplots showing the spread of y values for each interval of \(x_5\) along with \(f_5\) from a data set of type l; F) scatter plot of \(x_5\) to \(y\) for a data set of type l showing \(f_5\).

Many real-life situations include variables which show more complex distributions. Arbitrary regression data sets from the UCI Machine Learning Repository (Dua and Graff 2017), as well as privately available data sets, were analysed to contextualise the artificial data sets used in this study. The input-output relationships in real data sets, such as wind turbine power, fish toxicity and housing price regression data, are seen to be similar to those used to generate the artificial data sets.
5 Assessment of the Mean Fit to Median Error Measure

Three separate data sets are tested for each type of data set in Table 1 with different randomly generated parameters. For each of these data sets, 1,500 neural networks are trained using normal backpropagation of random sizes ranging from (1,1) to (3,1000) and a loss function based on the Mean Absolute Error, full hyperparameters are provided in Table 2. After training, each network is tested with the newly derived Mean Fit to Median Error Measure, as well as Mean Absolute Error and Mean Squared Error, as these are the most popular error measures used for regression (Fildes and Goodwin 2007) (McCarthy et al. 2006). The Mean Absolute Error is used as the Minkowski metric to compare to because the curvature of the input-output relationships means the conditional median is closer to the ground truth than the conditional mean (Bishop 1995), although the Mean Square Error was also trialled due to its prolific use in the field but showed similar behaviour to the Mean Absolute Error.

| Hyperparameter           | Value or set |
|-------------------------|--------------|
| Number of hidden layers  | [1,3]        |
| Number of neurons in each hidden layer | [1,1000] |
| Number of epochs        | 20           |
| Batch Size              | 50           |
| Early Stopping Patience | 5            |
| Error function          | Mean Absolute Error |
| Learning rule           | AdaMax (Kingma and Ba 2014) |
| Activation Function     | ReLU         |
| Regulariser             | None         |
| Initialiser             | Random Normal ($\mu = 0, \sigma = 0.1$) |

Table 2: Selected Hyperparameters

Since the data sets are all artificial, the underlying relationships within them are known explicitly, therefore it is possible to assess how well each network fits the ground truth of the data set. This measure is called ‘Mean Fit to the Ground Truth’ which is derived similarly to the Mean Fit To Median with the difference that the input-output function $\phi$ is used to generate the data set,

\[
\text{Mean Fit to Ground Truth} = \frac{1}{m} \sum_{j=1}^{m} \frac{1}{n} \sum_{i=1}^{n} |\phi\left(\text{mid}(X^j_i), \text{med}(X^k_i) \text{ for } k \neq j\right) - p_j|.
\]  

This involves cycling from the minimum to maximum observed values, taking the true value of $\phi$ for the midpoint of each partition in the input domain, with all of the other inputs held at the median value. It is only possible to calculate this measure for artificial data sets.

The errors from 1,500 separate network runs are calculated and each error measure is normalised independently as the data sets are artificial, therefore the absolute magnitude of an error measure has limited meaning. The error measures are compared to assess if the Mean Fit to Median measure approximates the Mean Fit to the Ground Truth of the data set better than the Mean Absolute Error or Mean Squared Error. If there is a correlation between the Mean Fit to the Median and the Mean Fit to the Ground Truth then the Mean Fit to the Median can be used as a measure that acts as a proxy for how well a network approximates the ground truth of these data sets.

The relationships between the Mean Absolute Error and the Mean Fit to the Ground Truth show a high variation across the different data sets, nearly all relationships involve bifurcations, although this becomes less prominent as the degree of the polynomials in the data set decreases. Figure 6a shows an almost perpendicular bifurcation at 0.4 normalised Mean Absolute Error whereas Figures 7a and 8a have a more irregular pattern but still show prominent bifurcations. The cause of these patterns is discussed in Section 6. Contrastingly, the Mean Fit to the Median and the Mean Fit to the Ground Truth clearly approach more of a one-to-one relationship. This is quantified using $R^2$ from a linear regression performed on the normalised data, the lines are shown on Figures 6, 7 and 8.

For a data set with entirely linear $f_i$ and a low standard deviation of noise, there is almost total agreement between the Mean Fit to the Median and Mean Fit to the Ground Truth, Figure 6c. This correlation lessens as the standard deviation of noise increases and as the $f_i$ increases in degree, or curvature. Figure 7c shows results from a data set with increased curvature but low noise, compared to Figure 6c, and although the correlation is still prominent, the $R^2$ has decreased from 0.98 to 0.85. The data sets with polynomial $f_i$ of degrees 4-5 and high noise levels are hard to model, requiring at
Error Measures with a Bias to the Ground Truth

Figure 6: 1,500 networks assessed on the data set of **type d**, where the data is fit with a linear regression a) the Mean Absolute Error against Mean Fit To Ground Truth with $R^2$ of 0.0036 and b) the Mean Fit To Median against Mean Fit To Ground Truth with $R^2$ of 0.9801.

Figure 7: 1,500 networks assessed on the data set of **type b**, where the data is fit with a linear regression a) the Mean Absolute Error against Mean Fit To Ground Truth with $R^2$ of 0.0016 and b) the Mean Fit To Median against Mean Fit To Ground Truth with $R^2$ of 0.8464.

At least three hidden layers in a neural network to approximate accurately. The results from this data set show the lowest
Error Measures with a Bias to the Ground Truth

Figure 8: 1,500 networks assessed on the data set of type 1, where the data is fit with a linear regression a) the Mean Absolute Error against Mean Fit To Ground Truth with $R^2$ of 0.2704 and b) the Mean Fit To Median against Mean Fit To Ground Truth with $R^2$ of 0.5184.

Figure 9: Heatmaps illustrating the average $R^2$ between error measures and the Mean Fit To Ground Truth from the different types of data set, stipulated in Table 1 a) the average $R^2$ of the traditional, Mean Absolute Error, error measure against the Mean Fit To Ground Truth and b) the average $R^2$ of the proposed, Mean Fit To Median, error measure against the Mean Fit To Ground Truth.
correlation between Mean Fit to the Median and Mean Fit to the Ground Truth with an $R^2$ of 0.52, but this is still significantly higher than the correlation between Mean Absolute Error and Mean Fit to the Ground Truth.

This decrease in correlation between Mean Fit to the Median and Mean Fit to the Ground Truth as more noise and curvature is introduced to the data sets can be explained by Jensen’s inequality. The increase in noise and curvature of $f_i$ create a larger skew in the conditional distributions of the output variable, this skew causes the mean and median of the conditional distribution to move away from the ground truth or $\phi(X_i)$. If the median is not an accurate representation of the ground truth then the Mean Fit to the Median will not be as effective as it directly measures the distance of a prediction from the conditional median. However, the Mean Fit to the Median approximates the Mean Fit to the Ground Truth better than the Mean Absolute Error for all analysed data sets. Networks with low Mean Fit to the Median therefore replicate the ground truth better than networks with low Mean Absolute Error.

The $R^2$ values from all of the regressions on the 36 data sets are collated and this is illustrated in Figure 9a and 9b. Comparing the average $R^2$ value of the linear regressions performed for each different type of data set, shows where the Mean Fit to the Median is most effective. All of the $R^2$ values for the Mean Absolute Error are lower, 0.02-0.57, than their Mean Fit to the Median counterparts, 0.39-0.99, from Figure 9a where a higher $R^2$ indicates a tighter positive correlation between the error measure and the Mean Fit to the Ground Truth of the data set. For data sets with lower degree polynomial relationships the $R^2$ values for the Mean Fit to the Median are particularly high, approaching 1, Figure 9b. However, data sets with these characteristics should be simple enough to analyse that advanced machine learning techniques, such as neural networks, do not produce any further insight than simpler analysis techniques.

All of the $R^2$ values for the Mean Fit to the Median are higher than the Mean Absolute Error values, meaning the Fit to Median is better at assessing how well a network models than ground truth than the Mean Absolute Error. This increase in $R^2$ ranges from 0.29-0.96, with an average of 0.6, across all 36 data sets. Choosing networks with a low Mean Fit to the Median increases the likelihood of a better model of the underlying relationships within a data set, when compared to relying on the most commonly used error measures.

### 6 Investigation into the Poor Bias from Minkowski-r Error Measures

The causes of the bifurcations in the relationship between Mean Absolute Error and the Mean Fit to the Ground Truth are investigated. A major factor affecting the error of a trained network is its size and shape. The networks trained had a random number of layers and neurons, with a maximum size of (3,1000) and therefore the bifurcations imply that some networks trained with Mean Absolute Error have a bias to the ground truth and that some have a bias to a pattern that is not the ground truth, negatively affecting their ability to replicate this relationship. One of the assumptions required for minimum Mean Absolute Error to model the conditional median of a data set is the use of a sufficiently large neural network. As randomly sized networks are used in this study this assumption is not necessarily satisfied. Therefore, to evaluate the effect of the size of a network on the error, the complexity, or number of connections, are indicated by colour in Figure 10.

There is no clear relationship between the size of the network and error for either the Mean Absolute Error, Figure 10a, the Mean Fit to the Median or the Mean Fit to the Ground Truth, Figure 10b. This suggests that the bifurcations are not caused by over or under-fitting of the networks. Although some relationship between the network size and error would be expected when using a large range of network sizes, the maximum number of layers trialled is only 3 and as all of the networks employ early stopping it is suggested this procedure decreases the chances of networks overfitting.

It is confirmed that, on the whole, networks with low Mean Fit To Median also have a low Mean Absolute Error but not the lowest observed Mean Absolute Error, so fitting to the ground truth of a data set does not produce the lowest point-to-point accuracy but does provide an acceptable level. For all 36 trialled data sets, 70% show lower Mean Absolute Error as Mean Fit To Median lowers, this means the Mean Fit To Median error measure can be used alongside Mean Absolute Error to produce networks which do not compromise the point-to-point accuracy of prediction significantly, but do increase the fit to the ground truth of the data set.

The characteristics of the data sets, their distributions and underlying relationships, are investigated as possible causes of the bifurcations. The normalised difference between the Mean Fit to the Median and ‘Fit to Mean’ error measures are indicated by data point colour in Figure 11. The ‘Fit to Mean’ error measure is derived similarly to the Mean Fit to the Median except the proxy curve is $E[\phi(X)]$ as opposed to $med(\phi(X))$ in Equation 1.

The networks with the lowest Mean Absolute Error approximate relationships which are closer to the mean of the data than the median of the data and hence have a positive difference between Mean Fit to the Median and ‘Fit to Mean’. These lighter points in Figure 11a form the bottom tail of the bifurcation and have a higher Mean Fit to the Ground Truth error than the main body. This suggests that as networks more closely fit to the conditional mean of a data set, this approximation stops fitting to the ground truth. As the difference between Mean Fit to the Median and ‘Fit to Mean’
Error Measures with a Bias to the Ground Truth

Figure 10: 1,500 networks assessed on the data set of **type e** with a) Mean Absolute Error against Mean Fit To Ground Truth and b) Mean Fit To Median against Mean Fit To Ground Truth. Where complexity, or number of connections in each network, is indicated by the colour of each point and shows no pattern for either plot.

Figure 11: 1,500 networks assessed on the data set of **type l** with a) Mean Absolute Error against Mean Fit To Ground Truth and b) Mean Fit To Median against Mean Fit To Ground Truth. The difference between the Mean Fit To Median and ‘Fit to Mean’ of each network is indicated by their colour, with networks where the ‘Fit to Mean’ is lower coloured lighter and networks where the Mean Fit To Median is lower coloured darker.

becomes negative the networks fits closer to the conditional median of the data set than the mean, indicated by the darker data points. The networks which have the lowest Mean Fit to the Median, comparative to the ‘Fit to Mean’, diverge away from the main body as the upper part of the bifurcation as the Mean Absolute Error from these networks is high.
Therefore, the bifurcations noted for these data sets may be caused by the bias created by training with the Mean Absolute Error. The minimum Mean Absolute Error does not guarantee a fit of the conditional median as shown in previous sections, in this particular data set the networks with lowest Mean Absolute Error are closest to the conditional mean of the data set. For the data sets used in this study the conditional median is a better approximation of the ground truth than the conditional mean, hence the lower tail in Figure 11a which contains networks which fit closely to the conditional mean but do not closely fit to the ground truth. This shows an unhelpful inductive bias is created by the use of the Mean Absolute Error.

New error measures are required to improve regression approximations of the ground truth of a data set, these will need to take into account the origins of the unhelpful bias created by Minkowski-r error measures. As neural networks display low neuroplasticity, the bias will be decided early on in the training process which makes it difficult to overcome using conventional error measures. This suggests that the bifurcation of networks into helpful or unhelpful bias originates from inherent stochasticity in the training of a neural network. Explicitly this is a combination of the initialisation and the order the data set is shown to the network in the first epoch, as well as any stochastic elements introduced by learning rules.

7 Conclusion

Regression methods can report low conventional error measures by mapping arbitrary patterns between the inputs and outputs. In many applications, such as where the data is non-linear or sparse, methods trained on these error measures demonstrate a poor approximation of the ground truth, which reduces trust in the method. This paper develops a error measure, the Mean Fit To Median error, which improves the approximation of the conditional average for applications where no understanding of the input-output relationships exist but where a better modelling of the ground truth would provide a greater understanding of the problem and allow for extrapolation. The error measure is robust across a larger range of data sets than the traditional error measures and can be used in a regression method to better approximate the ground truth, in order to provide more explainable machine learning. The Mean Fit To Median error, which measures how far from a proxy of the conditional median predictions are, is compared to the traditional Minkowski-r error measures on 36 different data sets and is shown to have a stronger correlation to the Fit to Ground Truth, with an average increase in $R^2$ of 0.6 compared to the Mean Absolute Error. Methods with low Mean Fit To Median error model the ground truth of a data set more reliably than those with a low Mean Absolute Error, by creating a bias to the ground truth. Selecting trained methods with a low Fit To Median Error, in combination with the use of traditional loss functions, increases the likelihood that the relationships between the inputs and the outputs of a regression problem are accurately mapped, even when there is no understanding of what these relationships are.

Acknowledgments

The authors thank Shell Shipping and Maritime for funding this work. The work was also kindly supported by the Lloyds Register Foundation. The authors acknowledge the use of the IRIDIS High Performance Computing Facility, and associated support services at the University of Southampton, in the completion of this work.

References

Abusitta, A., E. Aimeur, and O. A. Wahab (2020). “Generative Adversarial Networks for mitigating biases in machine learning systems”. In: European Conference on Artificial Intelligence.
Anderson, B., T. S. Hy, and R. Kondor (2019). “Cormorant: covariant molecular neural networks”. In: Advances in Neural Information Processing Systems. Vol. 32, pp. 14537–14546.
Bernardo, J. M. (1979). “Reference posterior distributions for Bayesian inference”. In: Journal of the Royal Statistical Society: Series B (Methodological) 41.2, pp. 113–128.
Bishop, C. (1995). “Neural Networks for Pattern Recognition”. In: Oxford University Press. Chap. 6, pp. 194–225.
Botchkarev, A. (2018). “Performance metrics (error measures) in machine learning regression, forecasting and prognostics: Properties and typology”. In: arXiv preprint arXiv:1809.03006.
De Prado, M. L. (2018). Advances in financial machine learning. John Wiley & Sons.
Drucker, H. et al. (1996). “Support vector regression machines”. In: Advances in Neural Information Processing systems 9, pp. 155–161.
Dua, D. and C. Graff (2017). UCI Machine Learning Repository. URL: http://archive.ics.uci.edu/ml.
Fieldes, R. and P. Goodwin (2007). “Against your better judgment? How organizations can improve their use of management judgment in forecasting”. In: Interfaces 37.6, pp. 570–576.
Gao, X., M. Sitharam, and A. Roitberg (2018). “Bounds on the Jensen gap, and implications for mean-concentrated distributions”. In: *arXiv:1712.05267v4*.

Green, P. J. et al. (2015). “Bayesian computation: a summary of the current state, and samples backwards and forwards”. In: *Statistics and Computing* 25.4, pp. 835–862.

Hanson, S. and D. Burr (1987). “Minkowski-r back-propagation: learning in connectionist models with non-euclidian error signals.” In: *Neural Information Processing Systems (NIPS 1987)*.

Hutson, M. (2018). *Artificial intelligence faces reproducibility crisis*.

Jensen, J. (1906). “Sur les fonctions convexes et les inégalités entre les valeurs moyennes”. In: *Acta Math.* 30, pp. 175–193.

Karpatne, A. et al. (2018). “Physics-guided neural networks (PGNN): an application in lake temperature modelling”. In: *arXiv:1710.11431v2*.

Kingma, D. P. and J. Ba (2014). “Adam: a method for stochastic optimization”. In: *arXiv preprint arXiv:1412.6980*.

Lampinen, J. and A. Vehtari (2001). “Bayesian approach for neural networks—review and case studies”. In: *Neural networks* 14.3, pp. 257–274.

Li, S. et al. (2020). “Scalable Gaussian Process regression networks”. In: *International Joint Conference on Artificial Intelligence (IJCAI)*.

Lucia, D. J., P. S. Beran, and W. A. Silva (2004). “Reduced-order modeling: new approaches for computational physics”. In: *Progress in Aerospace Sciences* 40.1-2, pp. 51–117.

MacGillivray, H. (1981). “The mean, median, mode inequality and skewness for a class of densities”. In: *Australian Journal of Statistics* 23, pp. 247–250.

MacKay, D. J. C. (1992). “A practical Bayesian framework for backpropagation networks”. In: *Neural Computation* 4.3, pp. 448–472.

McCarthy, Teresa M et al. (2006). “The evolution of sales forecasting management: a 20-year longitudinal study of forecasting practices”. In: *Journal of Forecasting* 25.5, pp. 303–324.

Merkle, M. (2005). “Jensen’s inequality for medians”. In: *Statistics and Probability Letters* 71, pp. 277–281.

Murphy, K. P. (2012). *Machine learning: a probabilistic perspective*. MIT press.

Nelder, J. A. and R. W. M. Wedderburn (1972). “Generalized linear models”. In: *Journal of the Royal Statistical Society. Series A (General)* 135.3, pp. 370–384. ISSN: 00359238.

Orbis Research (2018). *Machine Learning Market*. (Visited on 03/07/2019).

Park, J. and J. Park (2019). “Physics-induced graph neural network: an application to wind-farm power estimation”. In: *Energy* 187, p. 115883.

Parkes, A. I., A. J. Sobey, and D. A. Hudson (2018). “Physics-based shaft power prediction for large merchant ships using neural networks”. In: *Ocean Engineering* 166, pp. 92–104.

Rasmussen, C. E. and C. K. I. Williams (2006). *Gaussian Processes for machine learning*. The MIT Press.

Read, J. S. et al. (2019). “Process-guided deep learning predictions of lake water temperature”. In: *Water Resources Research* 55.11, pp. 9173–9190.

Richmond, M. et al. (2020). “Stochastic assessment of aerodynamics within offshore wind farms based on machine-learning”. In: *Renewable Energy* 161, pp. 650–661.

Sidey-Gibbons, J. A. M. and C. J. Sidey-Gibbons (2019). “Machine learning in medicine: a practical introduction”. In: *BMC medical research methodology* 19.1, p. 64.

Sultan, M. M., H. K. Wayment-Steele, and V . S. Pande (2018). “Transferable neural networks for enhanced sampling of protein dynamics”. In: *Journal of chemical theory and computation* 14.4, pp. 1887–1894.

Taylor, R. (1990). “Interpretation of the correlation coefficient: a basic review”. In: *Journal of Diagnostic Medical Sonography* 6.1, pp. 35–39.

Tversky, A. and D. Kahneman (1971). “Belief in the law of small numbers”. In: *Psychological Bulletin* 76 (2), pp. 105–110.

Voosen, P. (2017). *The AI detectives*. American Association for the Advancement of Science.

Wang, Y . et al. (2010). “High-dimensional pattern regression using machine learning: from medical images to continuous clinical variables”. In: *Neuroimage* 50.4, pp. 1519–1535.

Willard, J. et al. (2020). *Integrating physics-based modeling with machine learning: a survey*. arXiv:2003.04919

Zhang, L. et al. (2018). “End-to-end symmetry preserving inter-atomic potential energy model for finite and extended systems”. In: *Advances in Neural Information Processing Systems*. Vol. 31, pp. 4436–4446.