Neural Network Model for Imprecise Regression with Interval-Valued Dependent Variables

A summary and analysis

Ali Zindari 7048032

Based on the paper by

Krasymyr Tretiak, Georg Schollmeyer, and Scott Ferson

August 13, 2025

Overview of the Talk

1. Problem and Motivation
2. Basic Notation and Background
3. Related Works
4. Method and Contribution
5. Experiments
6. Summary

Problem and Motivation

- Real-world Data / measurements are **imprecise**: inaccurate measurement, lack of knowledge, wrong understanding of how the world works!
- This uncertainty is called epistemic uncertainty.
- Theoretically, it can be reduced by having more data.

In this work, the focus is on regression problems and dealing with uncertainty when we have continuous real-valued output.

Regression and Uncertainty

- Usually, in regression the output / prediction is a single number in $y \in \mathbb{R}$.
- To capture uncertainty, the easiest way is to output an interval $[y^{\min}, y^{\max}]$ instead.
- You see this in real-world products in form of ($\pm$) for weight or power etc.
- It is common to assume uniform distribution over values inside the interval (should we?).

Problem

An infinitely wide interval definitely contains the true value but is useless!!

This works proposes an efficient method to get tighter intervals!

Basic Notation and Definitions

An interval can be denoted as:

$$[x] = [\underline{x}, \overline{x}] = \{x \in \mathbb{R} \mid \underline{x} \leq x \leq \overline{x}\} \tag{1}$$

For intervals $[a] = [\underline{a}, \overline{a}]$, $[b] = [\underline{b}, \overline{b}]$ basic operations are:

$$\begin{aligned} [a] + [b] &= [\underline{a} + \underline{b},\ \overline{a} + \overline{b}] \\ [a] - [b] &= [\underline{a} - \overline{b},\ \overline{a} - \underline{b}] \end{aligned} \tag{2}$$

- The second operation is problematic! assume $a = b = [1, 2]$. Then we have:

$$[a] - [b] = [1, 2] - [1, 2] = [-1, 1]$$

Dependency Problem

Naive replacement of floating point computations leads to unnecessary wide intervals.

Precise Regression

The paper a linear dependence between input and output parameterized by $\mathbf{w}^\star$ with an additive gaussian noise:

$$y = \mathbf{w}_\star^\top \mathbf{x} + \epsilon \tag{3}$$

where $\mathbf{x} \in \mathbb{R}^m$ is m dimensional input feature sampled from some distribution $\mathcal{P}_X$ and $y \in \mathbb{R}$ is the true label. The MSE loss over n data pair $\{(\mathbf{x}_i, y_i)\}_{i=1,\dots,n}$ in matrix form is defined as:

$$\hat{\mathbf{w}} = \arg\min_{\mathbf{w} \in \mathbb{R}^m} \frac{1}{2n} \|\mathbf{y} - X\mathbf{w}\|^2 \tag{4}$$

Assuming X is full rank we have the unique solution of:

$$\hat{\mathbf{w}} = (X^\top X)^{-1} X^\top \mathbf{y} \tag{5}$$

Imprecise Regression

In imprecise regression, we assume labels y_is are intervals $y_i = [\underline{y_i}, \overline{y_i}]$. This means the parameters of our model are intervals $\mathbf{w} = [\underline{\mathbf{w}}, \overline{\mathbf{w}}]$.

Two main approaches

For predictor $\mathbf{h} = X[\mathbf{w}]$ and interval labels $[\mathbf{y}]$ there are two main approaches:

$$L_{\text{INN}} = \frac{1}{2}\|[\mathbf{h}] - [\mathbf{y}]\|^2 \tag{6}$$

$$L_{\text{ISH}} = \frac{1}{2}\|\underline{\mathbf{h}} - \underline{\mathbf{y}}\|^2 + \frac{1}{2}\|\overline{\mathbf{h}} - \overline{\mathbf{y}}\|^2 \tag{7}$$

The loss (6) involves interval operation which makes it prone to *dependency* problem.

Dependency Problem

Let's take a closer look at Equation (6):

$$L_{\text{INN}} = \frac{1}{2}\|[\mathbf{h}] - [\mathbf{y}]\|^2 = \left\| \begin{pmatrix} [\mathbf{w}^\top]\mathbf{x}_1 - [y_1] \\ [\mathbf{w}^\top]\mathbf{x}_2 - [y_2] \\ \vdots \\ [\mathbf{w}^\top]\mathbf{x}_n - [y_n] \end{pmatrix} \right\|^2 = \frac{1}{2}\sum_{i=1}^{n} \left(\underbrace{[\mathbf{w}^\top]\mathbf{x}_i - [y_i]}_{\text{interval operation}} \right)^2$$

- Since we have interval subtraction, we face *dependency* problem.
- Suppose you predict all intervals correctly such that $[\mathbf{w}^\top]\mathbf{x}_i = [y_i] \ \forall i \in [n]$,

$$L_{\text{INN}} = \frac{1}{2}\sum_{i=1}^{n} ([y_i] - [y_i])^2 \gg 0$$

- We keep updating weights which is unnecessary and norm of weights gets bigger.

Related Works

- Center Method (CM) (see [1])
  - Considers both data and labels are interval.
  - Takes the mid-point of interval and uses regular methods to solve the problems.
  - Scalable and efficient (e.g. use of neural networks).
  - Allows for interval input.
  - Ignores all the important information in the intervals.
- Center and Range Method (CRM and CCRM) (see [2, 3])
  - Trains two independent models for lower and upper bound of interval.
  - CCRM adds a constraint to make sure upper bound is bigger or equal to lower bound.
  - Allows for interval input.
  - Scalable and efficient (e.g. use of neural networks).

Sharp Collection Region (SCR)

What is SCR?

- The **Sharp Collection Region (SCR)** [4] is the set of all *precise* models f such that:

$$f(\mathbf{x}_i) \in [\underline{y}_i, \overline{y}_i] \quad \forall i$$

- It captures the full range of parameter values that are compatible with the interval-valued data.

Simple example:

$$x = 2, \quad [y] = [3, 5], \quad f(x) = wx$$

$$\Rightarrow \quad f(2) = 2w \in [3, 5] \quad \Rightarrow \quad w \in [1.5, 2.5]$$

SCR: all models $f_w(x) = wx$ with $w \in [1.5, 2.5]$

Gradient Descent

Recall the loss function for linear regression:

$$L(\mathbf{w}_k) = \frac{1}{2}\|X[\mathbf{w}_k] - [\mathbf{y}]\|^2$$

The update rule for Gradient Descent (GD) with stepsize γ is:

$$\mathbf{w}_{k+1} = \mathbf{w}_k - \gamma \nabla L(\mathbf{w}_k) = \mathbf{w}_k - \gamma X^\top X[\mathbf{w}_k] + \gamma X^\top [\mathbf{y}]$$

Which comes from the fact that:

$$\nabla L(w_k) = X^\top (X[\mathbf{w}_k] - [\mathbf{y}])$$

- $\mathbf{w}_{k+1}$ is linear in $\mathbf{w}_k$.
- As a result, after k steps of GD, $\mathbf{w}_k$ is a linear function of $\mathbf{w}_0$ and $\mathbf{y}$.

Method

We know that $\mathbf{w}_k$ is some linear function $f(\cdot)$ of $\mathbf{w}_0, \mathbf{y}$.

$$\mathbf{w}_k = f(\mathbf{w}_0, \mathbf{y})$$

The rough idea is to linearly approximate the radius of error.

Mean Value Theorem
Let $f : \mathbb{R}^d \to \mathbb{R}$ be differentiable, and let $\mathbf{a}, \mathbf{b} \in \mathbb{R}^d$. Then there exists $\mathbf{c} \in (\mathbf{a}, \mathbf{b})$ such that:

$$f(\mathbf{b}) = f(\mathbf{a}) + \nabla f(\mathbf{c})^\top (\mathbf{b} - \mathbf{a})$$

Now, we extend the above definition to interval-valued function:

$$f([\mathbf{x}]) = f(\text{mid}[\mathbf{x}]) + \mathbf{J}_\mathbf{w}([\mathbf{x}])([\mathbf{x}] - \text{mid}[\mathbf{x}]) \tag{8}$$

If f is linear, its gradient is constant and one can choose any $c \in (a, b)$.

Method

For linear regression problems with interval labels, Equation (8) becomes:

$$[\mathbf{w}_{k+1}(\mathbf{w}_0, \mathbf{y})] = \mathbf{w}_{k+1}\big(\operatorname{mid}(\mathbf{w}_0), \operatorname{mid}(\mathbf{y})\big) + \begin{pmatrix} \frac{\partial \mathbf{w}_{k+1}}{\partial \mathbf{w}_0} \\ \frac{\partial \mathbf{w}_{k+1}}{\partial \mathbf{y}} \end{pmatrix}^\top \begin{pmatrix} [\mathbf{w}_0] - \operatorname{mid}(\mathbf{w}_0) \\ [\mathbf{y}] - \operatorname{mid}(\mathbf{y}) \end{pmatrix} \tag{9}$$

$$\frac{\partial \mathbf{w}_{k+1}}{\partial \mathbf{w}_0} = \frac{\partial \mathbf{w}_{k+1}}{\partial \mathbf{w}_k} \frac{\partial \mathbf{w}_k}{\partial \mathbf{w}_{k-1}} \cdots \frac{\partial \mathbf{w}_1}{\partial \mathbf{w}_0} = (I - \gamma X^\top X)^k$$

$$\frac{\partial \mathbf{w}_{k+1}}{\partial \mathbf{y}} = \gamma X^\top$$

- $\mathbf{w}_{k+1}$ is obtained by running GD on mid-point labels.
- Note: $\mathbf{w}_{k+1} = \mathbf{w}_k - \gamma X^\top X \mathbf{w}_k + \gamma X^\top \mathbf{y}$ when $\mathbf{y} = \operatorname{mid}([\mathbf{y}])$.
- $[\mathbf{y}] - \operatorname{mid}(\mathbf{y})$ gives a symmetric interval centered at zero.

Intuition

- We run GD on mid-points and compute the error once in the end.
- $\frac{\partial \mathbf{w}_{k+1}}{\partial \mathbf{y}}$ measures how our parameter changes with a slight change in $\mathbf{y}$.
- Since model is linear, the error scales linear and is constant over time since $\frac{\partial \mathbf{w}_{k+1}}{\partial \mathbf{y}} = \gamma X^\top$.
- One can assume $\mathbf{w}_0$ is initialized at zero and the only uncertainty comes from $\mathbf{y}$.

$$\text{uncertainty} = \frac{\partial \mathbf{w}_{k+1}}{\partial \mathbf{y}}([\mathbf{y}] - \text{mid}(\mathbf{y})) = \gamma X^\top([\mathbf{y}] - \text{mid}(\mathbf{y}))$$

$$\implies [\mathbf{w}_{k+1}(\mathbf{w}_0, \mathbf{y})] = \mathbf{w}_{k+1}(\text{mid}(\mathbf{w}_0), \text{mid}(\mathbf{y})) \pm \frac{1}{2}\gamma X^\top([\mathbf{y}] - \text{mid}(\mathbf{y}))$$

Experiments: wine quality dataset

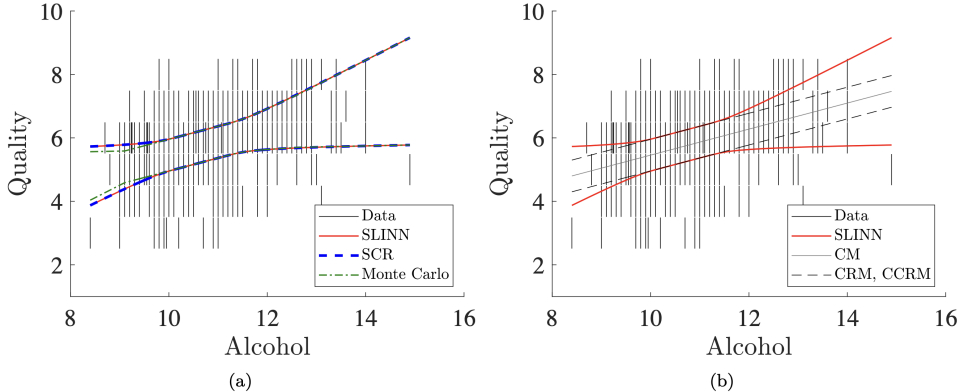

Figure 3: Linear regression example: (a) expectation band obtained from the SLINN (bounded in red lines), from the SCR method (dashed blue lines), Monte Carlo exploration (green dash-dotted lines); (b) the CRM and CCRM methods (dashed black lines) and the CM method produced the solid gray line.

Experiments: wine quality dataset

Table 2: Summary of results from the selected model. The values are interval coefficients from the linear regression.

	SLINN	SCR	Monte Carlo
(intercept)	[-2.61, 7.48]	[-2.61 ,7.48]	[-1.68, 6.56]
`Volatile acidity`	[-56.27, 20.55]	[-56.27, 20.56]	[-46.56, 10.85]
`Citric acid`	[-37.93, 40.21]	[-37.93, 40.21]	[-29.85, 32.12]
`Chlorides`	[-138.01, 73.44]	[-138.18, 73.53]	[-107.86, 43.21]
`Sulphates`	[-18.78, 50.20]	[-18.78, 50.20]	[-14.63, 46.06]
`Alcohol`	[-1.30, 10.73]	[-1.30, 10.73]	[-0.96,10.39]

Experiments: housing price

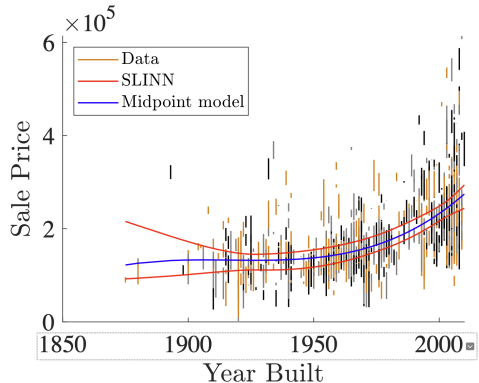

Figure 4: Nonlinear regression example. Expectation band (in red) obtained by fitting a 3rd-order polynomial to interval data and a 3rd-order polynomial fitted to midpoints of the data set.

Strengths

- **If we believe that this choice of loss function is justified,** we know we should not use naive GD for linear regression.

Weaknesses

- The proposed loss is not clearly motivated or theoretically justified. The paper assumes it is preferable to other interval-based losses without comparing or proving its benefits.
- The method lacks both theoretical grounding (e.g., no convergence guarantees or optimality results) and practical applicability (only tested on toy problems).
- It is unclear whether the optimization procedure converges to a meaningful solution. Also the paper does not formally define what a "good" solution means in this imprecise setting.
- The claim of producing tighter output intervals is not supported by analysis, intuition, or proof. The role of the mean value form is stated but not convincingly explained or justified.
- There is no discussion on how to initial $\mathbf{w}_0$ and what are its implications?

Weaknesses

- The experiments are very minimal and toy-level. For example, using linear regression on a single-feature model (alcohol content).
- The method cannot be extended to neural networks or high dimensional problems. It is not then clear why should we even need to be imprecise for a linear regression with single feature?
- A large portion of the paper is dedicated to trivial derivations (e.g., multiple steps of GD, feature squaring) instead of elaborating on the main contribution.
- Two pages on momentum extension which is completely unnecessary for linear regression.

Thanks for your Time!

References I

[1] Lynne Billard and Edwin Diday. “Regression analysis for interval-valued data”. In: *Data analysis, classification, and related methods*. Springer, 2000, pp. 369–374.

[2] Francisco de AT de Carvalho, Eufrasio de A Lima Neto, and Camilo P Tenorio. “A new method to fit a linear regression model for interval-valued data”. In: *Annual conference on artificial intelligence*. Springer. 2004, pp. 295–306.

[3] Eufrásio de A Lima Neto and Francisco de AT De Carvalho. “Constrained linear regression models for symbolic interval-valued variables”. In: *Computational Statistics & Data Analysis* 54.2 (2010), pp. 333–347.

[4] Georg Schollmeyer and Thomas Augustin. “Statistical modeling under partial identification: Distinguishing three types of identification regions in regression analysis with interval data”. In: *International Journal of Approximate Reasoning* 56 (2015), pp. 224–248.