Mathematics for Computer Science Project 2

1 Task Description

The project requires completing one of the following two tasks, and when submitting, include the project code and a report. Various methods such as non-linear regression or deep learning can be employed.

1.1 Code Implementation (4 point)

The code should consist of at least three parts:

- data preprocessing (1 point);
- model construction (2 point);
- performance evaluation (1 point).

1.2 Report Writing (6 point)

The report should include:

- Description and analysis of the problem definition (1 point);
- Introduction to the data preprocessing process (1 point);
- Detailed explanation of the methods used and implementation process (3 point);
- Test results (1 point).

2 Task 1 - PAs-Modeling

2.1 Introduction

Behavioral modeling of power amplifiers (PAs) is aimed at better understanding the relationship between the input signals and output signals of PAs and enabling better digital predistortion. The input signal and output signal of a PA is represented as $X = [..., x_{n-m}, x_{n-m+1}, ..., x_n, ...]$ and $Y = [..., y_{n-m}, y_{n-m+1}, ..., y_n, ...]$, respectively. An illustrative example is shown in Fig.1, in which the behavioral modeling of the PA is to approximate the function f.

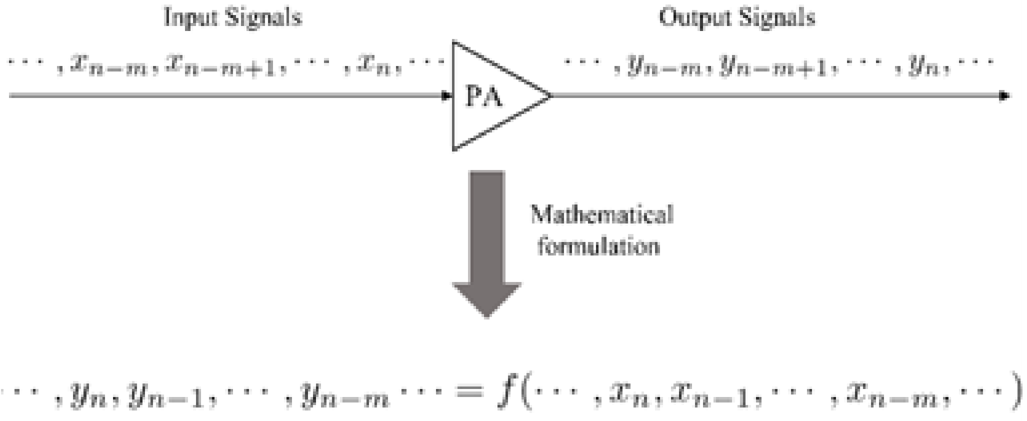
Input Signals
$\cdots, x_{n-m}, x_{n-m+1}, \cdots, x_n, \cdots$
PA
Output Signals
$\cdots, y_{n-m}, y_{n-m+1}, \cdots, y_n, \cdots$
Mathematical formulation
$\cdots, y_n, y_{n-1}, \cdots, y_{n-m} \cdots = f(\cdots, x_n, x_{n-1}, \cdots, x_{n-m}, \cdots)$

Figure 1: Diagram of the relationship between the input signals and output signals of a PA. The function f represents the black box of the PA.

With memory effects and non-linearity, a real PA can be represented as:

$$y_n = f(x_n, x_{n-1}, ..., x_{n-m})$$

where y_n is determined by $x_n, x_{n-1}, ..., x_{n-m}$, and m is the PA's memory depth.

2.2 Dataset

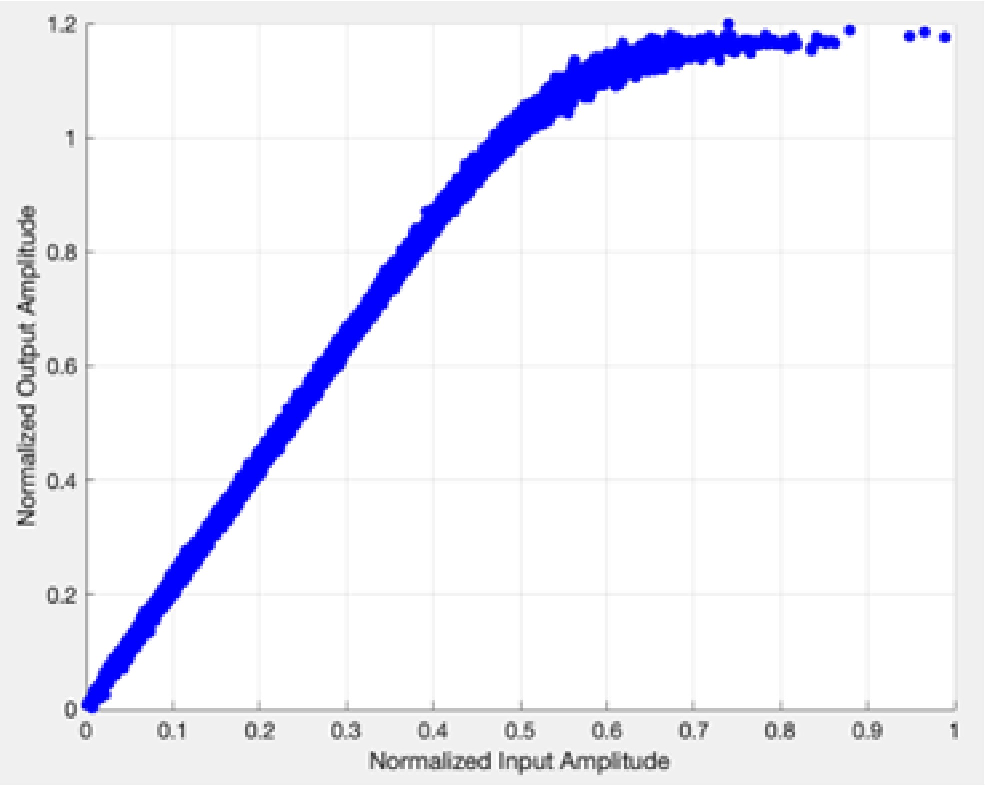

Figure 2: The Normalized Amplitude/amplitude (AM/AM) response of the PA.

The normalized amplitude/amplitude (AM/AM) response of the PA is shown in Fig. 2. The spread in the AM/AM response suggests that the PA has memory effects. We collect the dataset and request you to conduct behavioral modeling, namely capturing the relationship between the input signals and output signals of the PA.

In the PA dataset, ’PA_data_test.mat’ refers to the test set, ’PA_data_train.mat’ refers to the training set, and in the mat file, ’paInput’ represents the input data, and ’paOutput’ represents the output data, datas are all one-dimensional complex arrays.

```
dataset/
├── task1/
│   ├── PA_data_test.mat
│   └── PA_data_train.mat
└── task2/
    └── ......
```

2.3 Evaluation metric

The performance is evaluated by the Normalized Mean Square Error (NMSE) between the predicted signals and the actual PA output signals. Since the signal matrix is a complex matrix, when calculating NMSE, it is necessary to split the real and imaginary parts into real numbers, calculate them separately, and then sum them up.The calculation formula is as follows:

$$NMSE = 10lg(\frac{\sum_{n=1}^{N}(I_{out}(n) - \hat{I}_{out}(n))^2 + (Q_{out}(n) - \hat{Q}_{out}(n))^2}{\sum_{n=1}^{N}(I_{out}(n))^2 + (Q_{out}(n))^2})$$

where $\hat{I}_{out}(n)$ represents the real part of the predicted value, $\hat{Q}_{out}(n)$ represents the imaginary part of the predicted value, $I_{out}(n)$ represents the real part of the actual value, and $Q_{out}(n)$ represents the imaginary part of the actual value.

3 Task 2 - CSI-Prediction

3.1 Introduction

Channel prediction refers to predicting the future channel state based on past Channel State Information (CSI) data and other relevant information in wireless communication systems. The main purpose of this task is to anticipate the trend of channel variations, thereby taking corresponding adjustment measures in communication systems to maximize communication performance and resource utilization efficiency.

Here, we will use Channel Frequency Response (CFR) as the prediction data. CFR, which represents the frequency attenuation and phase delay of the channel at different frequencies, is commonly used to describe the characteristics of the channel in the frequency domain. The CFR at time t is a complex vector $h_t \in C^{n_s}$, where n_s represents the number of frequencies. Therefore, the channel prediction task can be expressed as:

$$(h_{t+1}, h_{t+2}, ..., h_{t+l}) = F(h_{t-r+1}, h_{t-r+2}, ..., h_t),$$

where r is the length of historical CFR used for prediction, and l is the length of future CFR to be predicted, indicating using the past r time points' CFR to predict the future l time points' CFR.

3.2 Dataset

The dataset consists of real-world data from a scenario with drone. The feature is CFR, represented as a complex matrix with two dimensions: time and frequency, i.e. $CFR \in C^{n_t \times n_s}$, where n_t represents the length of the time series, and n_s represents the number of frequencies.

In the CFR dataset, 'CFR_test.mat' refers to the test set, 'CFR_train_(0,1,2,3).mat' refers to the training set, and in the mat file, 'CFR' represents the CFR data, and CFR are all two-dimensional complex arrays. The frequency dimension is 28-dimensional.

```
dataset/
│
├─ task1/
│   │
│   └─ ......
│
└─ task2/
    │
    ├─ CFR_test.mat
    ├─ CFR_train_0.mat
    ├─ CFR_train_1.mat
    ├─ CFR_train_2.mat
    └─ CFR_train_3.mat
```

3.3 Evaluation metric

The performance is evaluated by the Normalized Mean Square Error (NMSE) between the predicted CFR and the actual CFR. Since the CFR matrix is a complex matrix, when calculating NMSE, it is necessary to split the real and imaginary parts into real numbers, calculate them separately, and then sum them up. The calculation formula is as follows:

$$NMSE = 10lg(\frac{\sum_{n=1}^{N}(I_{out}(n) - \hat{I}_{out}(n))^2 + (Q_{out}(n) - \hat{Q}_{out}(n))^2}{\sum_{n=1}^{N}(I_{out}(n))^2 + (Q_{out}(n))^2})$$

where $\hat{I}_{out}(n)$ represents the real part of the predicted value, $\hat{Q}_{out}(n)$ represents the imaginary part of the predicted value, $I_{out}(n)$ represents the real part of the actual value, and $Q_{out}(n)$ represents the imaginary part of the actual value.

4 Hint

- Since the dataset consists of complex numbers, most neural network architectures or time series analysis algorithms cannot handle complex numbers. Therefore, it is possible to split the complex numbers into real and imaginary parts for separately processing. Similarly, when calculating NMSE, the real and imaginary parts are calculated separately;

- NMSE is not used as a scoring metric. As long as the method used is reasonable and the implementation process is correct, there is no need to focus too much on the effectiveness of NMSE;

- In Task1, the data is only one-dimensional, leaning towards traditional algorithms such as non-linear regression. In Task2, due to its higher dimensionality, the data may be more suitable for deep learning algorithms.
- In Task 1, the objective is to predict one output data using a certain length (m) of input data. In Task 2, the objective is to predict a certain length (l) of future data using a certain length (r) of historical data. The lengths of the input and output windows (m,l,r) can be determined by yourself.

5 Submission

Submit a single zip file named StudentID_StudentName.zip that includes:

1. The Python code file(s) in a separate folder named 'code' .
2. The PDF report (named as StudentID_StudentName.pdf).

```
StudentID_StudentName.zip
|
|__ code/
|   |
|   |__ ......
|
|__ StudentID_StudentName.pdf
```