Project Milestone: A Simple Parameter Inference with Transformers

Adam Kleman
30 November 2025

List of read articles:

- Vaswani et al. – Attention is All You Need https://arxiv.org/abs/1706.03762
- Jet: A Modern Transformer-Based Normalizing Flow https://arxiv.org/html/2412.15129v1
- Neural Spline Flows https://arxiv.org/abs/1906.04032
- Flow-based Conformal Prediction for Multi-dimensional Time Series https://arxiv.org/abs/2502.05709
- Normalizing Flows for Probabilistic Modeling and Inference https://arxiv.org/abs/1912.02762

Task overview:

The goal of this project is to predict the amplitude A_i and frequency ω_i of the periodic function $y_i(A_i, \omega_i, t) = A_i \cdot \sin(\omega_i \cdot t)$, where the input data consist of vectors $V_i(A_i, \omega_i)$ representing different time discretizations. The project focuses on searching for optimal hyperparameters, comparing a classical encoder-only transformer architecture using a regression head or a flow-based head for frequency prediction. Another goal is to examine the influence of dataset noise, dataset size, and time discretization on the stability and performance of the model.

Current progress:

At the beginning of the project, an initial encoder-only transformer model was implemented together with a synthetic dataset parameterized by amplitude, frequency, and time discretization. All data generation and experiments were performed with a fixed random seed to ensure full reproducibility. For a dataset of size N=1000, different frequency intervals – specifically [0.0,10.0], [0.1,10.0], and [0.5,10.0] – were compared to evaluate their effect on model behavior.

Suitable time discretization values were also tested, with $t_{disc} \in \{10, 20, 30, 40, 50, 100\}$. Training length was subsequently examined, and experiments showed that stable performance is achieved at approximately 160 epochs.

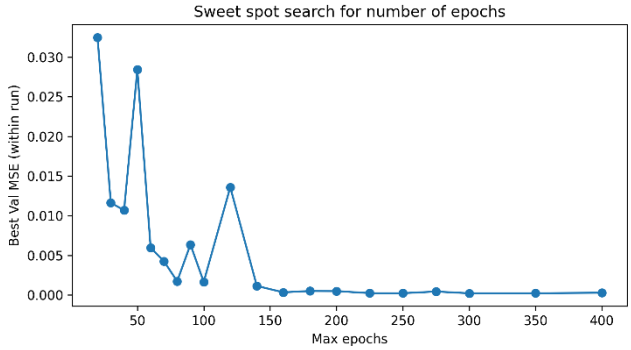

Fig. 1: Optimized number of sufficient epochs for training.

After selecting the interval ω ∈ [0.5, 10], time discretization t_{disc}= 100, and 160 epochs, a slight improvement in model performance was observed when the dataset was expanded. Therefore, datasets of size N ∈ {1000, 2500, 5000, 7500, 10000} were tested together with $t_{disc} \in \{60, 70, 85, 100, 125, 150\}$.

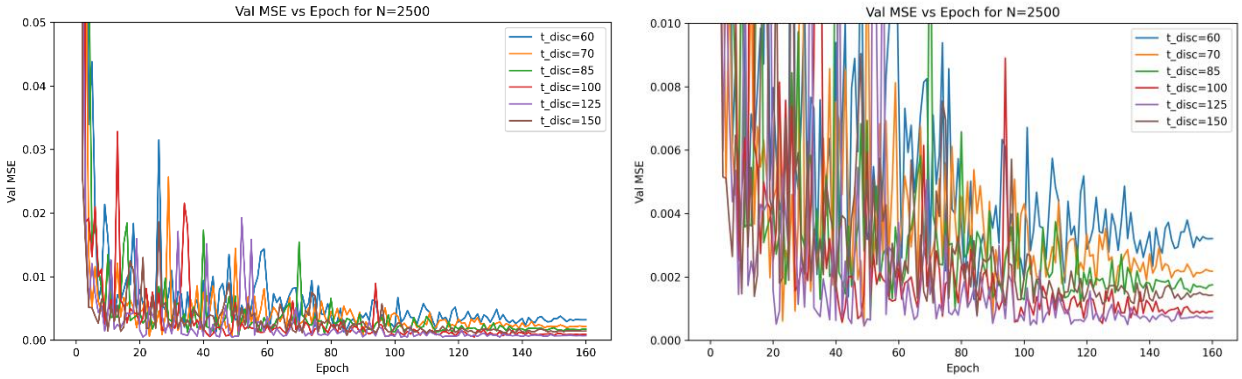

Fig. 2: Examples of different time discretizations and corresponding loss curves for N=1000 (left: slightly zoomed, right: fully zoomed).

In the next phase, systematic hyperparameter search was performed for N=1000 and t_disc=100. All combinations of parameter values ***d_model*** ∈ {32, 64, 128, 256}, ***n_heads*** ∈ {1, 2, 4, 8}, ***num_layers*** ∈ {1, 2, 3, 4}, and ***dim_f*** ∈ {64, 128, 256, 512} were evaluated. I chose the fifth model from the table since it achieved similar performance to larger configurations but used significantly fewer parameters, making it a more efficient choice for further experiments.

Tab. 1: Top 5 models with performance at validation MSE.

d_model	n_heads	num_layers	dim_f	validation MSE
256	8	1	256	0.000172
256	8	1	64	0.000190
256	8	1	128	0.000204
256	8	1	512	0.000229
64	2	2	64	0.000239

Gaussian noise of the form $V_i = \sin(\omega_i\, t) + \mathcal{N}(\mu, \sigma)$ was then added to the dataset to analyze model behavior under more challenging conditions.

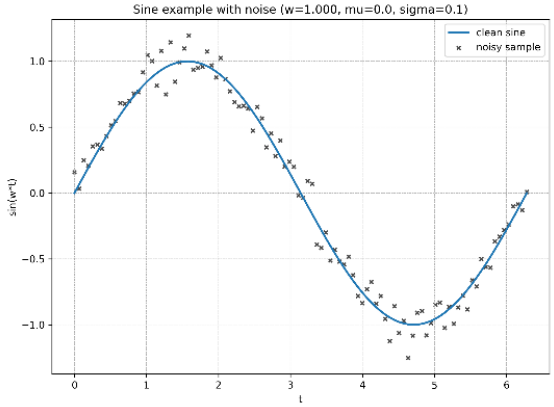

Fig. 3: Example of ideal wave without noise and created V_i with gaussian noise

The transformer architecture was extended by replacing the standard regression head with a flow-based head, allowing the model to learn the full probability distribution of the target parameter instead of predicting a single deterministic value. During inference, the model generates multiple samples from the learned distribution, and their mean is used as the final point estimate.

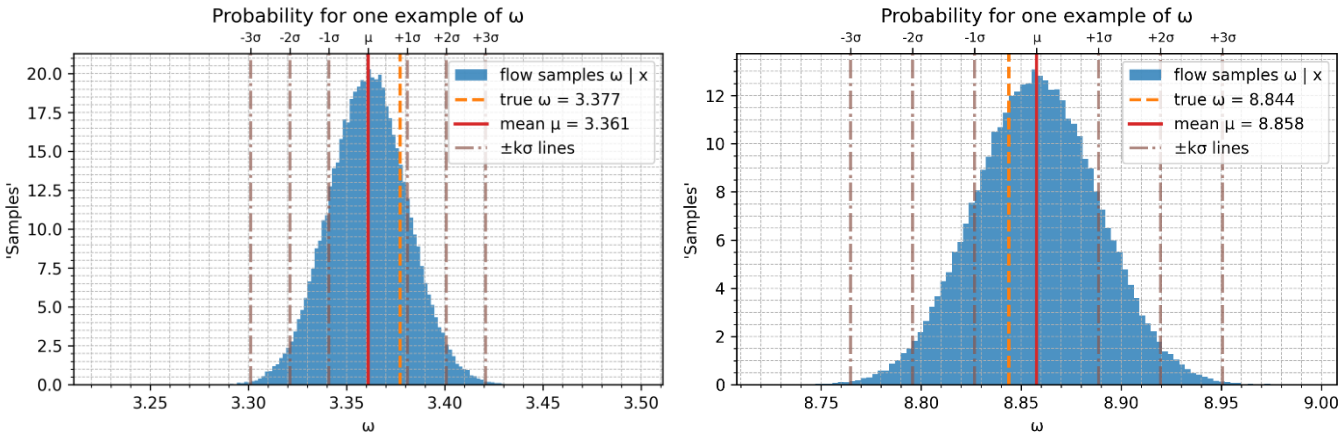

Fig. 4: Examples of predicted probability density functions for several input signals.

I noticed that for larger values of ω, the model often produces wider predicted distributions, which suggests lower confidence in the estimate compared to smaller frequencies. In the scatter plot, there also appears to be a pattern where higher frequencies tend to result in larger prediction errors.

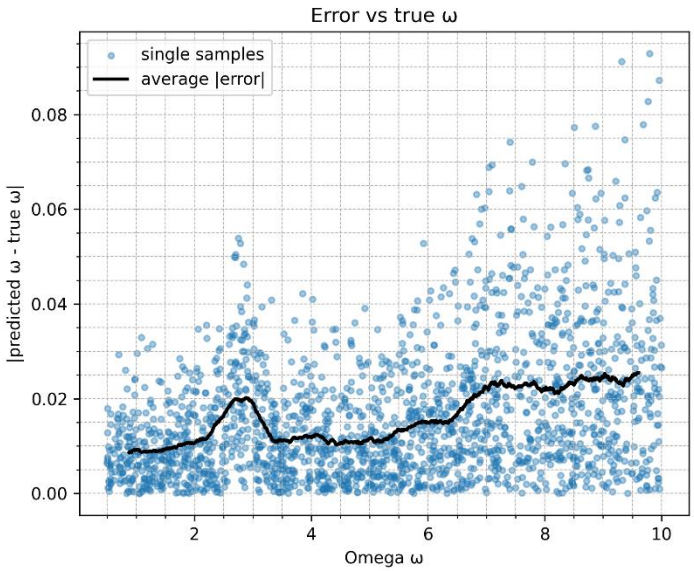

Fig. 5: Relationship between the frequency ω_i and the absolute difference between the predicted and true value.

Currently the work continues with studying the model's ability to predict two frequencies ω_i and ω_j from a mixture signal of the form $y(\omega_i, t) = \sin(\omega_i t) + \sin(\omega_j t)$, which introduces a more challenging inference problem.

Repository link: https://gitlab.fit.cvut.cz/klemaada/mvi-sp or https://github.com/y4h1ko/mvi-sp