

0.1 Datasets

0.1.1 Gaussian Process

The 2D Gaussian Process is the natural extension of the 1D Gaussian Process described in subsection 0.1.1 where we use the squared exponential kernel. We continue to use the same range of lengthscale across both input dimensions as the 1D Gaussian Process.

The following plots show some samples from the GP dataset.

TODO

Add plots of GP dataset

0.1.2 Sawtooth

The 2D Sawtooth dataset is the natural extension of the 1D Sawtooth dataset described in subsection 0.1.2. We continue to use the same period T and noise n across both input dimensions as the 1D Sawtooth dataset.

The following plots show some samples from the Sawtooth dataset.

TODO

Add plots of Sawtooth dataset

0.1.3 Restricted Sawtooth

By accident when generating the 2D Sawtooth dataset, we ended up restricted the ‘direction of travel’ of the sawtooth function to the line of $x_1 = x_2$ or $x_1 = -x_2$. This was not intentional but when training both models on this dataset, we found very interesting results. As the models only learn a subset of the ‘full sawtooth’ function, we can see how well the models can generalize to samples from the full sawtooth function.

The following plots show some samples from the Restricted Sawtooth dataset.

0.2 Post or Pre Relative Attention Function

TODO

Highlight MLP

In our original formulation of the TETNP (section 0.3) we pass the matrix of differences (Δ) between x values through a function F to apply non-linearities then add it to the dot product attention Equation 0.3.3, whilst this performs well we can also consider applying this non-linearity after combining the dot product attention and the relative attention, this method is called the ‘Post Relative Attention Function’.

$$\text{Attention}(\mathbf{Q}, \mathbf{K}, \mathbf{V}, \mathbf{X}) = \text{softmax}(\mathbf{E}) \mathbf{V} \quad (0.2.1)$$

$$\text{Pre: } \mathbf{E}_{ij} = \mathbf{D}_{ij} + \text{MLP}(\mathbf{\Delta}_{ij}) \quad (0.2.2)$$

$$\text{Post: } \mathbf{E}_{ij} = \text{MLP}(\text{cat}[\mathbf{D}_{ij}, \mathbf{\Delta}_{ij}]) \quad (0.2.3)$$

Where

$$\mathbf{D}_{ij} = \mathbf{q}_i \cdot \mathbf{k}_j / \sqrt{d_k} \quad \mathbf{\Delta}_{ij} = \mathbf{x}_i - \mathbf{x}_j \quad (0.2.4)$$

We will investigate the performance of the TETNP with the ‘Post Relative Attention Function’ compared to the original ‘Pre Relative Attention Function’.

TODO

Add Results

0.3 ConvNP vs TNP

0.3.1 Gaussian Process

0.3.2 Sawtooth

0.3.3 Rotational Equivariance