Neural Diffusion Processes

Vincent Dutordoir [1 2] **Alan Saul** [2] **Zoubin Ghahramani** [1 3] **Fergus Simpson** [2]

Abstract

Neural network approaches for meta-learning distributions over functions have desirable properties such as increased flexibility and a reduced complexity of inference. Building on the successes of denoising diffusion models for generative modelling, we propose Neural Diffusion Processes (NDPs), a novel approach that learns to sample from a rich distribution over functions through its finite marginals. By introducing a custom attention block we are able to incorporate properties of stochastic processes, such as exchangeability, directly into the NDP's architecture. We empirically show that NDPs can capture functional distributions close to the true Bayesian posterior, demonstrating that they can successfully emulate the behaviour of Gaussian processes and surpass the performance of neural processes. NDPs enable a variety of downstream tasks, including regression, implicit hyperparameter marginalisation, non-Gaussian posterior prediction and global optimisation.

1. Introduction

Gaussian processes (GPs) offer a powerful framework for defining distributions over functions (Rasmussen & Williams, 2006). It is an appealing framework because Bayes rule allows one to reason consistently about the predictive distribution, allowing the model to be data efficient. However, for many problems, GPs are not an appropriate prior. Consider, for example, a function that has a single discontinuity at some unknown location. This is one classic example of a distribution of functions that cannot be expressed in terms of a GP (Neal, 1998).

One popular approach to these problems is to abandon GPs, in favour of Neural Network-based generative models. Successful methods include the meta-learning approaches of Neural Processes (NPs) (Garnelo et al., 2018b;a), and VAE-based models (Mishra et al., 2020; Fortuin et al., 2020). By leveraging a large number of small datasets during training, they are able to transfer knowledge across datasets at prediction time. Using Neural Networks (NNs) is appealing since most of the computational effort is expended during the training process, while the task of prediction becomes more straightforward. A further major advantage of a NN-based approach is that they are not restricted by the Gaussianity.

In this work, we strive to enhance the capabilities of NN-based generative models for meta-learning function by extending probabilistic denoising diffusion models (Sohl-Dickstein et al., 2015; Song & Ermon, 2020; Ho et al., 2020). Diffusion models have demonstrated superior performance over existing methods in generating images (Nichol et al., 2021; Ramesh et al., 2022), molecular structures (Xu et al., 2022; Hoogeboom et al., 2022), point clouds (Luo & Hu, 2021), and audio signal data (Kong et al., 2020). The central challenge we address is the Bayesian inference of functions, a fundamentally different task not previously approached by diffusion models. Although we acknowledge that several recent works have emerged since our initial preprint, providing complementary insights on extending diffusion models to function spaces (Phillips et al., 2022; Kerrigan et al., 2022; Lim et al., 2023; Bond-Taylor & Willcocks, 2023; Pidstrigach et al., 2023; Franzese et al., 2023).

Contributions We propose a novel generative model, the Neural Diffusion Process (NDP), which defines a probabilistic model over functions via their finite marginals. NDPs generalise diffusion models to stochastic processes by allowing the indexing of the function's marginals onto which the model diffuses. We take particular care to enforce properties of stochastic processes, including exchangeability, facilitating the training process. These properties are enforced using a novel *bi-dimensional* attention block, which guarantees equivariance over the ordering of the input dimensionality and the sequence (i.e., datapoints). From the experiments, we conclude: firstly, NDPs are an improvement over existing NN-based generative models for functions such as Neural Processes (NPs). Secondly, NDPs are an attractive alternative to GPs for specifying appropriate (i.e., non-Gaussian) priors over functions. Finally, we present a novel global optimisation method using NDPs.

[1]Department of Engineering, University of Cambridge, Cambridge, UK [2]Secondmind, Cambridge, UK [3]Google DeepMind. Correspondence to: Vincent Dutordoir <vd309@cam.ac.uk>.

Proceedings of the 40^{th} International Conference on Machine Learning, Honolulu, Hawaii, USA. PMLR 202, 2023.

2. Background

The aim of this section is to provide an overview of the key concepts used throughout the manuscript.

2.1. Gaussian Processes

A Gaussian Process (GP) $f : \mathbb{R}^D \to \mathbb{R}$ is a stochastic process such that, for any finite collection of points $x_1, ..., x_n \in \mathbb{R}^D$ the random vector $(f_1, \ldots, f_n)$ with $f_i = f(x_i)$, follows a multivariate normal distribution (Rasmussen & Williams, 2006). For a dataset $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^n$, where values are corrupted by Gaussian noise $y_i = f(x_i) + \eta$, GPs offer exact inference of the posterior $p(\boldsymbol{y}^* \mid \mathcal{D})$. This leads to data-efficient learning and accurate uncertainty estimations as illustrated in Fig. 1a.

GPs are stochastic processes (SPs) which satisfy the Kolmogorov Extension Theorem (KET). KET states that all finite-dimensional marginal distributions p are consistent with each other under permutation (exchangeability) and marginalisation. Let $n \in \mathbb{N}$ and π be a permutation of $\{1, \ldots, n\}$, then the following holds for the GP's joint:

$$p(f_1, \ldots, f_n) = p(f_{\pi(1)}, \ldots, f_{\pi(n)}), \text{ and} \tag{1}$$

$$p(f_1) = \int p(f_1, f_2, \ldots, f_n)\, \mathrm{d}f_2 \ldots \mathrm{d}f_n. \tag{2}$$

Despite these favourable properties, GPs are plagued by several limitations. Firstly, encoding prior assumptions through analytical covariance functions can be extremely difficult, especially in higher dimensions (Wilson et al., 2016; Simpson et al., 2021a; Liu et al., 2020). Secondly, by definition, GPs assume a multivariate *Gaussian* distribution for each finite collect of predictions —limiting the set of functions it can model (Neal, 1998).

2.2. Neural Processes and Meta-Learning Functions

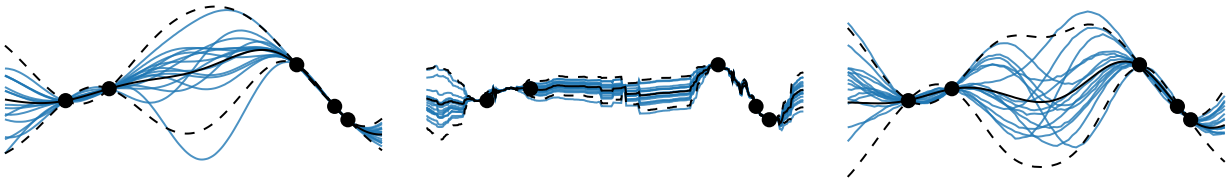
Neural Process (NP) models were introduced as a flexible alternative to GPs, utilising an encoder-decoder architecture to meta-learn dataset distributions in an amortised fashion (Garnelo et al., 2018b). While being beneficial in many respects, they nonetheless present several limitations: Conditional NPs assume independence amongst all function values in the predictive posterior, leading to a lack of correlated function samples at test time (Garnelo et al., 2018a). Latent NPs tackle this but have non-tractable likelihoods, leading to crude approximations during inference and performance restrictions. Attentive NPs improve empirical performance using attention mechanisms, yet frequently produce jittery samples due to shifting attention patterns (Kim et al., 2019). Furthermore, NPs fail to ensure consistency in the context sets as their conditional distributions do not relate through Bayes' rule (Kim et al., 2019). We refer to Appendix D for a primer on these methods.

Several models have sought to address these limitations. ConvCNPs enforce stationary into the stochastic process by decoding functional embeddings using convolutional NNs (Gordon et al., 2019). Gaussian NPs model predictive correlations (i.e. covariances), which provides universal approximation guarantees but struggle with computational scalability (Bruinsma et al., 2021). Recently, Autoregressive NPs were introduced which exhibit promising performance on a range of problems but are hampered by simple distributions, typically Gaussian, for variables early in the auto-regressive generation (Bruinsma et al., 2023).

2.3. Probabilistic Denoising Diffusion Models

Diffusion models depend on two procedures: a *forward* and a *reverse* process. The forward process consists of a Markov chain, which incrementally adds random noise to the data. The reverse process is tasked with inverting this chain. The *forward process* starts from the data distribution $q(\boldsymbol{s}_0)$ and iteratively corrupts samples by adding small amounts of noise, which leads to a fixed Markov chain $q(\boldsymbol{s}_{0:T}) = q(\boldsymbol{s}_0) \prod_{t=1}^T q(\boldsymbol{s}_t \mid \boldsymbol{s}_{t-1})$ for a total of T steps. The corrupting distribution is Gaussian $q(\boldsymbol{s}_t \mid \boldsymbol{s}_{t-1}) = \mathcal{N}(\boldsymbol{s}_t; \sqrt{1-\beta_t}\boldsymbol{s}_{t-1}, \beta_t \mathbf{I})$. The magnitude of the noise in each step is controlled by a pre-specified variance schedule $\{\beta_t \in (0,1)\}_{t=1}^T$. Note that there is no learning involved in the forward process, it simply creates a sequence of random variables $\{\boldsymbol{s}_t\}_{t=0}^T$ which progressively look more like white noise $q(\boldsymbol{s}_T) \approx \mathcal{N}(\mathbf{0}, \mathbf{I})$.

While the forward process is Markovian, the true reverse probability $q(\boldsymbol{s}_{t-1} \mid \boldsymbol{s}_t)$ requires the entire sequence. Therefore, the *reverse process* learns to approximate these conditional probabilities in order to carry out the reverse diffusion process. The approximation relies on the key observation that the reverse conditional probability is tractable when conditioned on the initial state $\boldsymbol{s}_0$: $q(\boldsymbol{s}_{t-1} \mid \boldsymbol{s}_0, \boldsymbol{s}_t) = \mathcal{N}(\boldsymbol{s}_{t-1}; \tilde{\boldsymbol{\mu}}(\boldsymbol{s}_0, \boldsymbol{s}_t), \tilde{\beta}_t \mathbf{I})$. As a result, the reverse process can be traversed by estimating the initial state $\boldsymbol{s}_0$ from $\boldsymbol{s}_t$ and t using a NN, which can then be passed to $\tilde{\boldsymbol{\mu}}$. In this work, we follow Ho et al. (2020) who directly parameterise $\tilde{\boldsymbol{\mu}}$ as $\boldsymbol{\mu}_\theta(\boldsymbol{s}_t, t) = \frac{1}{\sqrt{\alpha_t}}\left(\boldsymbol{s}_t - \frac{\beta_t}{\sqrt{1-\bar{\alpha}_t}}\boldsymbol{\epsilon}_\theta(\boldsymbol{s}_t, t)\right)$ with $\alpha_t = 1 - \beta_t$ and $\bar{\alpha}_t = \prod_{j=1}^t \alpha_j$. In what follows, we refer to $\boldsymbol{\epsilon}_\theta$ as the noise model. The parameters θ of the noise model are optimised by minimising the objective $\mathbb{E}_{t, \boldsymbol{s}_0, \boldsymbol{\varepsilon}}[\boldsymbol{\varepsilon} - \boldsymbol{\epsilon}_\theta(\boldsymbol{s}_t, t)]$, where the expectation is taken over time $t \sim \mathcal{U}(\{1, 2, \ldots, T\})$, data $\boldsymbol{s}_0 \sim q(\boldsymbol{s}_0)$, and $\boldsymbol{s}_t = \sqrt{\bar{\alpha}_t}\boldsymbol{s}_0 + \sqrt{1-\bar{\alpha}_t}\boldsymbol{\varepsilon}$ with $\boldsymbol{\varepsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$. For a trained network, generating samples from $q(\boldsymbol{s}_0)$ is done by running the reverse process starting from $\boldsymbol{s}_T \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$.



(a) GP Regression (b) Attentive Latent NP (c) Neural Diffusion Process (ours)

Figure 1: Posterior samples conditioned on a context dataset (black dots) for different probabilistic models.

3. Neural Diffusion Processes

In this section, we introduce Neural Diffusion Processes (NDPs), which define a probabilistic model over functions via their finite marginals. We focus our attention on the difference between NDPs and traditional diffusion models used for gridded data, such as images or audio.

3.1. Data, Forward and Reverse Process

NDPs generalise diffusion models to stochastic processes by allowing indexing of the random variables which are being diffused. This set of random variables corresponds to the function's marginals, which represent a more flexible type of random variable compared to, say, an image. In an image, pixel values are organised on a predefined grid (height $\times$ width) with an implicit order. Contrarily, function values lack ordering and do not reside on a predetermined grid. Consequently, samples extracted from a NDP should be evaluable throughout their input domain.

Data NDPs require many datasets that consist of both, function inputs $\boldsymbol{x}$ and corresponding function values $\boldsymbol{y} = f(\boldsymbol{x})$. The functions $f : \mathbb{R}^D \to \mathbb{R}$ may be synthetically created (e.g., step functions, Sec. 6.1.2), drawn from a Gaussian process (Sec. 6.2.1), or correspond to the pixel maps in images (Sec. 6.1.3). Crucially, in this meta-learning approach, the NDP needs numerous examples drawn from a distribution over f to learn an empirical covariance from the data, $\{(\boldsymbol{x}^i \in \mathbb{R}^{N\times D}, \boldsymbol{y}^i \in \mathbb{R}^N)\}_{i=1}^M$ where $\boldsymbol{y}^i = f^i(\boldsymbol{x}^i)$ assuming we evaluate the function at N random location in its domain. This setup is akin to NPs, the key difference being that NDPs do not necessitate the separation of the dataset into a context and target set during training.

Forward process Let $\boldsymbol{x}_0 = \boldsymbol{x}$ and $\boldsymbol{y}_0 = \boldsymbol{y}$ then NDPs gradually add noise following

$$q\Big(\begin{bmatrix}\boldsymbol{x}_t\\ \boldsymbol{y}_t\end{bmatrix} \mid \begin{bmatrix}\boldsymbol{x}_{t-1}\\ \boldsymbol{y}_{t-1}\end{bmatrix}\Big) = \mathcal{N}\big(\boldsymbol{y}_t; \sqrt{1-\beta_t}\boldsymbol{y}_{t-1}, \beta_t\big).$$

This corresponds to adding Gaussian noise to the function values $\boldsymbol{y}_t$ while keeping the input locations $\boldsymbol{x}_t$ fixed across time. At $t = T$ the function values $\boldsymbol{y}_t$ should be indistinguishable to samples from $\mathcal{N}(\mathbf{0}, \mathbf{I})$, while $\boldsymbol{x}_T = \boldsymbol{x}_{T-1} = \ldots = \boldsymbol{x}_0$. In Sec. 6.3, we discuss a generalisation of this scheme where we perturb both the function inputs and outputs to obtain a diffusion model that can sample the joint $p(\boldsymbol{x}_0, \boldsymbol{y}_0)$. For now, we focus on the conditional $p(\boldsymbol{y}_0 \mid \boldsymbol{x}_0)$ as this is the distribution of interest in supervised learning.

Backward kernel NDPs parameterise the backward Markov kernel using a NN that learns to de-noise the corrupted function values $\boldsymbol{y}_t$. In contrast, the input locations $\boldsymbol{x}_t$ have not been corrupted in the forward process which makes reversing their chain trivial. This leads to a parameterised backward kernel p_θ of the form

$$p_\theta\Big(\begin{bmatrix}\boldsymbol{x}_{t-1}\\ \boldsymbol{y}_{t-1}\end{bmatrix} \mid \begin{bmatrix}\boldsymbol{x}_t\\ \boldsymbol{y}_t\end{bmatrix}\Big) = \mathcal{N}\Big(\boldsymbol{y}_{t-1}; \boldsymbol{\mu}_\theta(\boldsymbol{x}_t, \boldsymbol{y}_t, t), \tilde{\beta}_t\Big), \quad (3)$$

where the mean is parameterised through the noise $\boldsymbol{\epsilon}_\theta$ as

$$\boldsymbol{\mu}_\theta(\boldsymbol{x}_t, \boldsymbol{y}_t, t) = \frac{1}{\sqrt{\alpha_t}}\big(\boldsymbol{y}_t - \frac{\beta_t}{\sqrt{1-\bar{\alpha}_t}}\boldsymbol{\epsilon}_\theta(\boldsymbol{x}_t, \boldsymbol{y}_t, t)\big). \quad (4)$$

The noise model NN $\boldsymbol{\epsilon}_\theta : \mathbb{R}^{N\times D} \times \mathbb{R}^N \times \mathbb{R} \to \mathbb{R}^N$ has as inputs the function inputs $\boldsymbol{x}_t$, the corrupted function values $\boldsymbol{y}_t$, and time t. The network is tasked with predicting the noise that was added to $\boldsymbol{y}_0$ to obtain $\boldsymbol{y}_t$. The design of the network is specific to the task of modelling functions and as such differs from current approaches. In the next section we discuss its particular architecture, but for now we want to stress that it is critical for the noise model to have access to the input locations $\boldsymbol{x}_t$ to make such predictions.

Objective Following Ho et al. (2020), but substituting the NDP's forward and backward transition densities leads to the objective

$$\mathcal{L}_\theta = \mathbb{E}_{t, \boldsymbol{x}_0, \boldsymbol{y}_0, \boldsymbol{\varepsilon}}\Big[\|\boldsymbol{\varepsilon} - \boldsymbol{\epsilon}_\theta(\boldsymbol{x}_0, \boldsymbol{y}_t, t)\|^2\Big], \quad (5)$$

with $\boldsymbol{y}_t = \sqrt{\bar{\alpha}_t}\boldsymbol{y}_0 + \sqrt{1-\bar{\alpha}_t}\boldsymbol{\varepsilon}$ and where we made use of the fact that $\boldsymbol{x}_t$ equals $\boldsymbol{x}_0$ for all timesteps t. Full derivation of the objective is given in Appendix A.

3.2. Prior and Conditional Sampling

Prior Using a trained noise model $\boldsymbol{\epsilon}_\theta$, one can obtain prior function draws from the NDP at a specific set of input locations $\boldsymbol{x}_0$ by simulating the reverse process. That is, starting

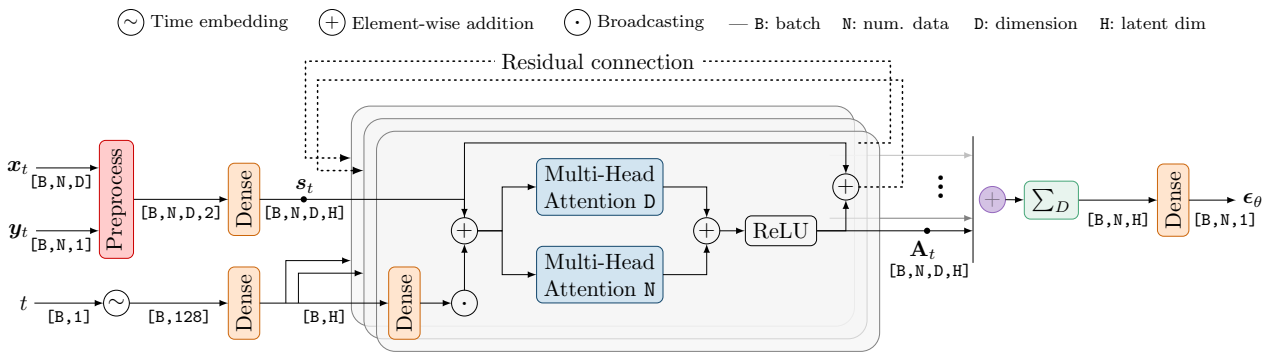


Figure 2: Architecture of the noise prediction model, utilised at each step within the Neural Diffusion Process. The greyed box represents the bi-dimensional attention block, as discussed in Section 4.2.

from $\boldsymbol{y}_T \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ and iteratively using the backward kernel p_θ from Eq. (3) for time $t = T, \ldots, 1$. This procedure leads to the samples from the prior illustrated in the left panes of Figs. 3 and 4.

Conditional NDPs draw samples from the conditional distribution $p(\boldsymbol{y}_0^*, | \, \boldsymbol{x}_0^*, \mathcal{D})$, where $\mathcal{D} = (\boldsymbol{x}_0^c \in \mathbb{R}^{M \times D}, \boldsymbol{y}_0^c \in \mathbb{R}^M)$ is the *context* dataset, using a slight adaptation of REPAINT algorithm (Lugmayr et al., 2022).

The conditional sampling scheme, given as pseudocode in Appendix B.2, works as follows. Start by sampling $\boldsymbol{y}_T^* \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$. Then, for each time t, sample a noisy version of the context $\boldsymbol{y}_t^c$ using the forward process

$$\boldsymbol{y}_t^c \sim \mathcal{N}\Big(\sqrt{\bar{\alpha}_t}\boldsymbol{y}_0^c, (1 - \bar{\alpha}_t)\, \mathbf{I}\Big). \tag{6}$$

Continue by collecting the union of the noisy context and target set in $\boldsymbol{y}_t = \{\boldsymbol{y}_t^*, \boldsymbol{y}_t^c\}$. Similarly, collect the inputs as $\boldsymbol{x}_0 = \{\boldsymbol{x}_0^*, \boldsymbol{x}_0^c\}$. Finish the step by sampling from the backward kernel p_θ using the collected inputs $\boldsymbol{x}_0$ and function values $\boldsymbol{y}_t$

$$\boldsymbol{y}_{t-1} \sim \mathcal{N}\Big(\frac{1}{\sqrt{\alpha_t}}\big(\boldsymbol{y}_t - \frac{\beta_t}{\sqrt{1-\bar{\alpha}_t}}\epsilon_\theta(\boldsymbol{x}_0, \boldsymbol{y}_t, t)\big), \tilde{\beta}_t \mathbf{I}\Big). \tag{7}$$

Simulating this scheme from $t = T, \ldots, 1$ ensures that for each backward step we leverage the context dataset. Consistent with the findings of Lugmayr et al. (2022), we found that in practice the sample quality improves by repeating Eqs. (6) and (7) multiple times (e.g., 5) per time step. We show conditional samples in Figs. 1c, 3 and 4 using our proposed algorithm.

4. Noise Model Architecture

NDPs implement the noise model $\boldsymbol{\epsilon}_\theta$ as a NN. Here, we review its architecture and key components. In principle, *any* NN could be used. However, if we wish for NDPs to mimic stochastic processes (SPs), the noise model must learn to generate a prior distribution over functions. We expect such a prior to possess several key symmetries and properties, which will heavily influence our choice of architecture.

4.1. Input Size Agnosticity

Before addressing the NN invariances and equivariances, we focus our attention to a key property of the network: the NDP network is agnostic to dataset size N and dimension D. That is, the weights of the network do *not* depend on the size of the inputs (i.e. N nor D). This has as important practical consequences that it is not required to train different NDPs for datasets with different size or dimensionality. This makes it possible to train only a single model that handles downstream tasks with different N or D values. To achieve this functionality, NDPs start by reshaping the inputs $(\boldsymbol{x}_t \in \mathbb{R}^{N \times D}, \boldsymbol{y}_t \in \mathbb{R}^N)$ to $\mathbb{R}^{N \times D \times 2}$ by replicating the $\boldsymbol{y}_t$ outputs D times before concatenating them with $\boldsymbol{x}_t$.

4.2. Bi-Dimensional Attention Block

A key property of stochastic processes is an *equivariance* to the ordering of inputs. As such, shuffling the order of data points in the context dataset $\mathcal{D}$ or the order at which we make predictions should not affect the probability of the data (i.e., the data is exchangeable). Secondly, we also expect an *invariance* in the ordering of the input dimensions. Consider, for example, a dataset consisting of two features, the weight and height of people. We would not expect the posterior function to be different if we would swap the order of the columns in the training data. This is an important invariance encoded in many GP kernels (e.g., Matérn, RBF) but often overlooked in the neural net literature.

We accommodate for both desiderata using our proposed *bi-dimensional attention block*. We denote this block by $\mathbf{A}_t : \mathbb{R}^{N \times D \times H} \to \mathbb{R}^{N \times D \times H}$ as it acts on the preprocessed inputs $(\boldsymbol{x}_t, \boldsymbol{y}_t)$. At its core, the block consists of two multi-

head self-attention (MHSA) layers (Vaswani et al., 2017) (see Fig. 2). The MHSA layers act on different axes of the input: one attends to the input dimension axis d, while the other attends across the dataset sequence axis n. The outputs of the two are subsequently summed and passed through a non-linearity. This process is repeated multiple times by feeding the output back into the next bi-dimensional attention block using residual connections. Concretely, the ℓ^{th} block is defined as

$$\mathbf{A}_t^{\ell}(\boldsymbol{s}_t^{\ell-1}) = \mathbf{A}_t^{\ell-1} + \sigma\left(\text{MHSA}_d(\boldsymbol{s}_t^{\ell-1}) + \text{MHSA}_n(\boldsymbol{s}_t^{\ell-1})\right),$$

for $\ell = \{1, \dots, L\}$ with σ the ReLU function, $\boldsymbol{s}_t^0 = \boldsymbol{s}_t$ (i.e., the output of the preprocessing step) and $\mathbf{A}_t^0 = \mathbf{0}$. Kossen et al. (2021) introduced a similar block which operates sequentially across datapoints and attributes, whereas ours acts in parallel.

To summarise, the bi-dimensional attention block is *equivariant* to the order of the dimensions and dataset sequence, as formalised by

Proposition 4.1. *Let Π_N and Π_D be the set of all permutations of indices $\{1, \dots, N\}$ and $\{1, \dots, D\}$, respectively. Let $\boldsymbol{s} \in \mathbb{R}^{N \times D \times H}$ and $(\pi_n \circ \boldsymbol{s}) \in \mathbb{R}^{N \times D \times H}$ denote a tensor where the ordering of indices in the* FIRST *dimension are given by $\pi_n \in \Pi_N$. Similarly, let $(\pi_d \circ \boldsymbol{s})$ denote a tensor where the ordering of indices in the* SECOND *dimension are given by $\pi_d \in \Pi_D$. Then, $\forall \pi_n, \pi_d \in \Pi_N \times \Pi_D$:*

$$\pi_d \circ \pi_n \circ \mathbf{A}_t(\boldsymbol{s}) = \mathbf{A}_t\left(\pi_d \circ \pi_n \circ \boldsymbol{s}\right). \tag{8}$$

Proof. Follows from the invariance of MHSA and the commutativity of π_d and π_n as detailed in Appendix C.2. □

The final noise model output $\boldsymbol{\epsilon}_\theta$ is obtained by summing the output of the different bi-dimensional attention layers (see purple '+' in Fig. 2). This is followed by a sum over the input dimension axis (green block). These final operations introduce an invariance over input dimensionality, while preserving the equivariance over the dataset ordering. We summarise the noise model's invariance and equivariance properties in the following proposition

Proposition 4.2. *Let π_n and π_d be defined as in Proposition 4.1, then $\boldsymbol{\epsilon}_\theta$ satisfies*

$$\pi_n \circ \boldsymbol{\epsilon}_\theta(\boldsymbol{x}_t, \boldsymbol{y}_t, t) = \boldsymbol{\epsilon}_\theta(\pi_n \circ \pi_d \circ \boldsymbol{x}_t, \pi_n \circ \boldsymbol{y}_t, t). \tag{9}$$

Proof. The claim follows from Proposition 4.1 and Zaheer et al. (2017) as shown in Appendix C.3. □

Crucially, by directly encoding these properties into the noise model, the NDP produces a set of random variables $\{y_t^1, \dots, y_t^n\}$ at each time step t that are exchangeable as defined in Eq. (1).

5. On Meta-Learning Consistency

Neural Diffusion Processes (NDPs) are generative models that define a probabilistic model over functions via their finite marginals. As prescribed by the Kolmogorov extension theorem, these finite marginals originate from a stochastic process if they satisfy the conditions of exchangeability and consistency. Proposition 4.2 has established that the exchangeability condition can be achieved by parametrising the noise model using the proposed bi-dimensional attention model, rendering the network permutation-equivariant. Consequently, we direct our focus to the more intricate topic of marginal consistency.

We first note that when the true noise $\boldsymbol{\varepsilon}$ is accessible, the generative model described by the reverse process exhibits marginal consistency across all finite marginals, thereby corresponding to samples from a stochastic process. However, in NDPs the noise model is approximated by a neural network $\boldsymbol{\epsilon}_\theta \approx \boldsymbol{\varepsilon}$. This approximation leads to NDPs forfeiting consistency within the generative process p_θ.

Even though NDPs cannot guarantee consistency as per Eq. (2), they do embed a particular form of it. Consider $\boldsymbol{y} = \boldsymbol{y}^* \cup \boldsymbol{y}^c = \boldsymbol{y}^{*\prime} \cup \boldsymbol{y}^{c\prime}$ where, $\boldsymbol{y}^c$ and $\boldsymbol{y}^{c\prime}$ denote two sets of contexts and $\boldsymbol{y}^*$ and $\boldsymbol{y}^{*\prime}$ represent the targets. When the union of these sets are identical, NDPs guarantee that the conditionals $p(\boldsymbol{y}^*|\boldsymbol{y}^c)$ and $p(\boldsymbol{y}^{*\prime}|\boldsymbol{y}^{c\prime})$ are consistent among each other and the joint $p(\boldsymbol{y})$. Consequently, NDPs can generate $2^{|\boldsymbol{y}|}$ consistent marginals from this joint.

While this is arguably a restricted form of consistency, it is a property that (A)NPs do not possess (Kim et al., 2019). As detailed in Appendix D, when it comes to conditioning, NDPs are unlike (A)NPs. In (A)NPs the predictive distribution over the targets is learnt in an amortised fashion by a neural network, using the context as inputs. NDPs, on the contrary, draw conditional samples using the joint distribution over contexts and targets, as described in Sec. 3.2.

Continuing the example, in cases where the unions of the target and context sets in NDPs are not equal, i.e. $\boldsymbol{y}^* \cup \boldsymbol{y}^c \neq \boldsymbol{y}^{*\prime} \cup \boldsymbol{y}^{c\prime}$, NDPs cannot ensure consistency and must resort to meta-learning for an approximation. It has been posited by Foong et al. (2020), that it is in practice often more beneficial to refrain from enforcing consistency into the network as this limits the choices of network architectures, and can lead to constraints on the learning performance. This view is confirmed by our experiments, which show that NDP's predictive performance closely matches that of optimal models, despite their lack of KET consistency.

6. Experimental Evaluation

In this section we aim to address the following two questions: Firstly, what additional advantages do NDPs offer

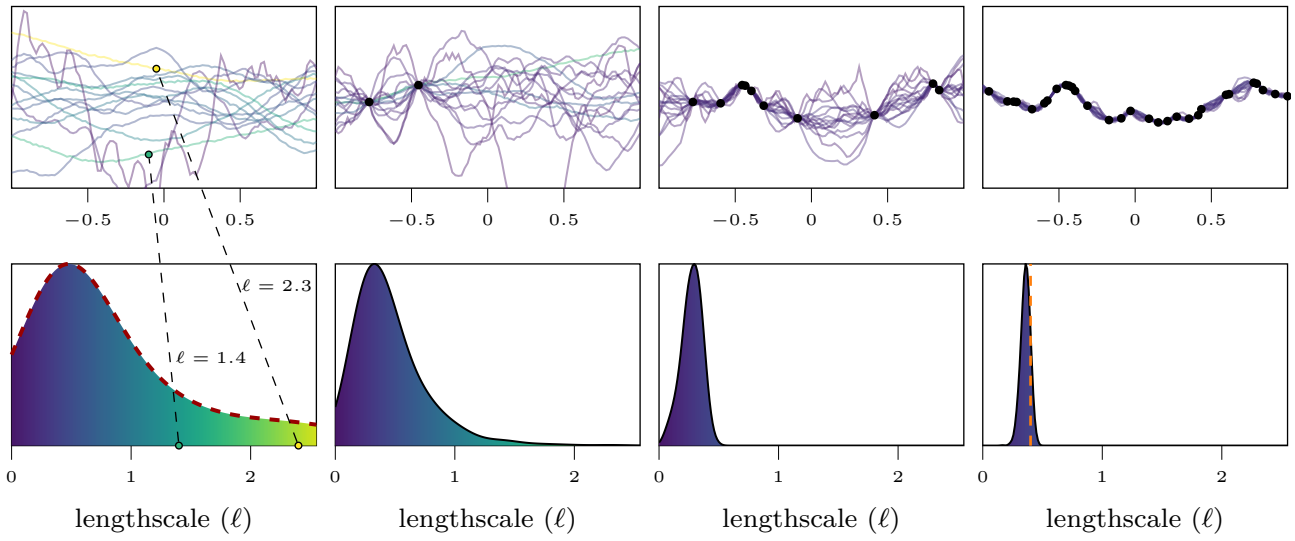


Figure 3: Hyperparameter marginalisation: Samples from the NDP, conditioned on an increasing number of data points (black dots), are illustrated in the top row. A sample is coloured according to its most likely lengthscale. The bottom row shows a histogram of likely lengthscales from the produced samples. As more data points are provided, the distribution of likely lengthscales converges from the prior over lengthscales to the lengthscale that was used to produce the data ($\ell = 0.3$).

beyond GPs? Secondly, how do NDPs stand in relation to NPs concerning performance and applicability? Furthermore, we introduce an innovative method for global optimisation of black-box functions in Sec. 6.3.1, which is predicated on modelling the joint distribution $p(x, y)$. All experiments (except Sec. 6.3.1) share the same model architecture illustrated in Fig. 2. Comprehensive details regarding the specific model configurations and training times can be found in Appendix F.1.

6.1. Emulating Gaussian Processes

6.1.1. Hyperparameter marginalisation

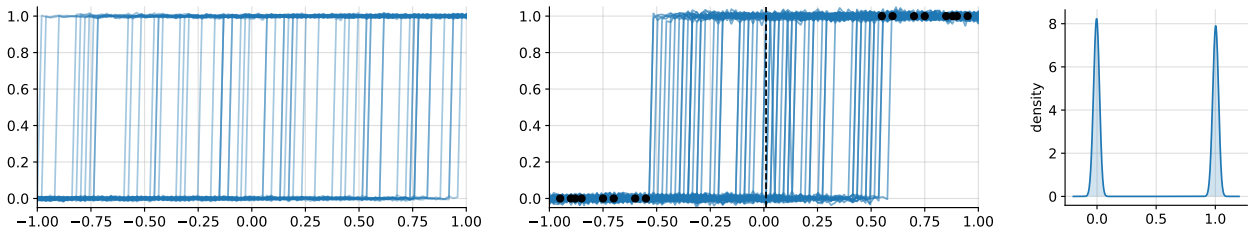
Conventionally, GP models optimise a point estimate of the hyperparameters. However, it is well known that marginalising over the hyperparameters can lead to significant performance improvements, albeit at an extra computational cost (Lalchand & Rasmussen, 2020; Simpson et al., 2021b). A key capability of the NDP is its ability to produce realistic conditional samples across the full gamut of data on which it was trained. It therefore in effect marginalises over the different hyperparameters it was exposed to, or even different kernels. To demonstrate this, we train a NDP on a synthetic dataset consisting of samples from a GP with a Matérn-$\frac{3}{2}$ kernel, with a prior on the lengthscale of $\log \mathcal{N}(\log(0.5), \sqrt{0.5})$. In the top row of Fig. 3 we show samples from this NDP for an increasing number of data points. A sample is coloured according to its most likely lengthscale. As the NDP has no notion of a lengthscale, we infer the most likely lengthscale by retrospectively fitting a set of GPs and matching the lengthscale of the GP corresponding to the highest marginal likelihood to the sample. The bottom row shows the histogram of lengthscales from the produced samples. We observe that the NDP covers the prior at first and then narrows down on the true data-generating lengthscale as more data is observed. Effectively marginalising over the lengthscale posterior.

6.1.2. Non-Gaussian Posteriors

The predictions of a GP for a finite set of data points follow a multivariate normal distribution. While this allows for convenient analytic manipulations, it also imposes a restrictive assumption. For example, it is impossible for a GP to represent a 1D step function when the step occurs at a random location within its domain (Neal, 1998). In Fig. 4, we train a NDP on a prior that consists of functions that take a jump at a random location within the interval $[-1, 1]$. Unlike a GP, we observe that the NDP is able to correctly sample from the prior in (a), as well as from the conditional in (b). In (c) we show the a marginal of the NDP's posterior, which correctly captures the bimodal behaviour. This experiment highlights that the NDP can infer a data-driven distribution that need not be Gaussian, which is impossible for GPs.

6.1.3. Image regression

In the next experiment, we apply NDPs to the task of image regression. The specific goal is to predict pixel values based on their coordinates within the normalized range of $[-2, 2]$. For the MNIST dataset, our task simplifies to predicting a



(a) Prior (b) Conditional (c) $p(y \mid x = 0.0)$

Figure 4: Representing a step function using NDPs. Figure (a) and (b) show samples from the model's prior and conditional distribution, respectively. Figure (c) illustrates the non-Gaussian posterior a NDP can capture, which a GP, by definition, can not do.

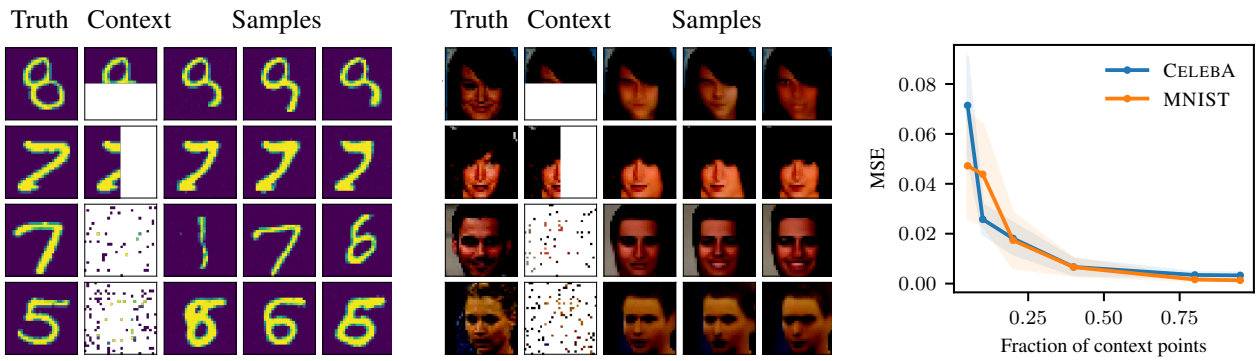


Figure 5: NDPs for image regression on MNIST and CelebA (32×32). Figures (a) and (b) show conditional samples where the context datasets are from top to bottom: the upper and left half of the pixels and a random selection of 5% and 10% of the pixels. Figure (c) plots the MSE of the NDP's predictions for an increasing number of context points.

single output value that corresponds to grayscale intensity. However, when tackling the CelebA 32×32 dataset, we deal with the added complexity of predicting three output values for each pixel to represent the RGB colour channels. Crucially, for these image-based tasks, the architecture of our bi-dimensional attention block has been modified to accommodate for the distinct nature of image data in which the order of the input coordinates do matter. This is done by removing the MHSA layer over D in the architecture of Fig. 2. Notably, we keep the attention over the input ordering N as this is key when treating images as datasets.

Figure 5 shows samples from the NDP conditioned on a variety of contexts (top, left, random 5% and 10% of pixels). We observe that the NDP is able to effectively learn the covariance over digits and human faces from the data — a very challenging task for any non-parametric kernel method. In (c) we measure the MSE of the NDP's predicted output (computed by drawing 5 samples from the conditional and taking the mean) and the target image. We observe an almost perfect match (low MSE) by increasing the context size. The MSE is computed on the normalised pixel values $[0, 1]$.

6.2. Comparison to Neural Processes

We now compare NDPs to a range of NP models.

6.2.1. Regression on Synthetic Data

We evaluate NDPs and NPs on two synthetic datasets, following the experimental setup from Bruinsma et al. (2021) but extending it to multiple input dimensions D. We use a Squared Exponential and a Matérn-$5/2$ kernel, where the lengthscale is set to $\ell = \sqrt{D}/4$. We corrupt the samples by white noise $\mathcal{N}(0, 0.05^2)$ and train the models on 2^{14} examples. At test time, the context dataset contains between 1 and $10 \times D$ points, whereas the target set has a fixed size of 50. We report the test log-likelihood of 128 examples.

For the NDP, the model's likelihood is computed using samples from the conditional distribution of the model. Utilising these samples, the empirical mean and covariance are computed, which are then used to fit a multivariate Gaussian. Given our understanding that the true posterior is Gaussian, this strategy allows for a valid and meaningful comparison with the other methods. Per test example, we used 128 samples to estimate the mean and covariance.

Table 1: Mean test log-likelihood (↑) ± 1 standard error estimated over 128 test samples. Statistically significant best non-GP model is in **bold**. '–' stands for computationally infeasible models.

	Squared Exponential			Matérn-$\frac{5}{2}$		
	$D=1$	$D=2$	$D=3$	$D=1$	$D=2$	$D=3$
GP (truth)	0.67 ± 0.03	-0.45 ± 0.03	-0.94 ± 0.03	0.19 ± 0.03	-0.85 ± 0.03	-1.14 ± 0.02
NDP (ours)	$\mathbf{0.48} \pm 0.04$	$\mathbf{-0.67} \pm 0.05$	$\mathbf{-1.16} \pm 0.04$	-0.00 ± 0.04	$\mathbf{-1.05} \pm 0.03$	$\mathbf{-1.33} \pm 0.03$
GNP	$\mathbf{0.51} \pm 0.02$	-0.98 ± 0.02	-1.36 ± 0.02	$\mathbf{0.14} \pm 0.02$	-1.12 ± 0.02	-1.37 ± 0.02
ConvNP	-0.41 ± 0.06	-1.16 ± 0.03	–	-0.63 ± 0.05	-1.22 ± 0.02	–
ANP	-0.55 ± 0.05	-1.19 ± 0.03	-1.36 ± 0.02	-0.70 ± 0.04	-1.23 ± 0.02	-1.37 ± 0.02
GP (diagonal)	-0.88 ± 0.07	-1.04 ± 0.04	-1.22 ± 0.04	-0.98 ± 0.06	-1.21 ± 0.04	-1.31 ± 0.03
ConvCNP	-0.74 ± 0.07	-1.20 ± 0.03	-1.36 ± 0.02	-0.87 ± 0.06	-1.25 ± 0.03	-1.37 ± 0.02

Table 1 presents the performance of different models: the ground-truth GP, the ground truth GP with diagonal covariance, Gaussian NPs (Bruinsma et al., 2021, GNP), Convolutional CNPs (Gordon et al., 2019, ConvCNP), Convolutional NPs (Foong et al., 2020, ConvNP) and Attentive NPs (Kim et al., 2019, ANP). The table demonstrates that for $D = 1$, the performance of NDPs is on par with that of GNPs, a model designed specifically to capture Gaussianity. Moreover, both NDPs and GNPs significantly surpass the performance of other methods and nearly reach the level of the ground truth. However, when we scale up to $D = 2$ and 3, NDPs uniquely stand out as the only method that maintains competitiveness while scaling with relative ease.

6.2.2. Bayesian Optimisation

We now tackle four black-box optimisation problems in dimensions three to six. The probabilistic models are part of a Bayesian optimisation (BO) loop in which they are used as a surrogate of the expensive black-box objectives. At each iteration, the surrogate is evaluated at 128 random locations in the input domain, and the input corresponding to the minimum value is selected as the next query point. The objective is evaluated at this location and added to the context dataset (akin to Thompson sampling BO (Shahriari et al., 2015)). Figure 6 shows the regret (distance from the true minimum) for the different models. We observe that the NDP almost matches the performance of GPR, which is the gold standard model for this type of task. NDPs also outperform the NPs and random search strategies. The important difference between GPR and the NDP is that the NDP requires *no training* during the BO loop, whereas the GPR is retrained at every step.

6.3. Modelling Function Inputs and Outputs Jointly

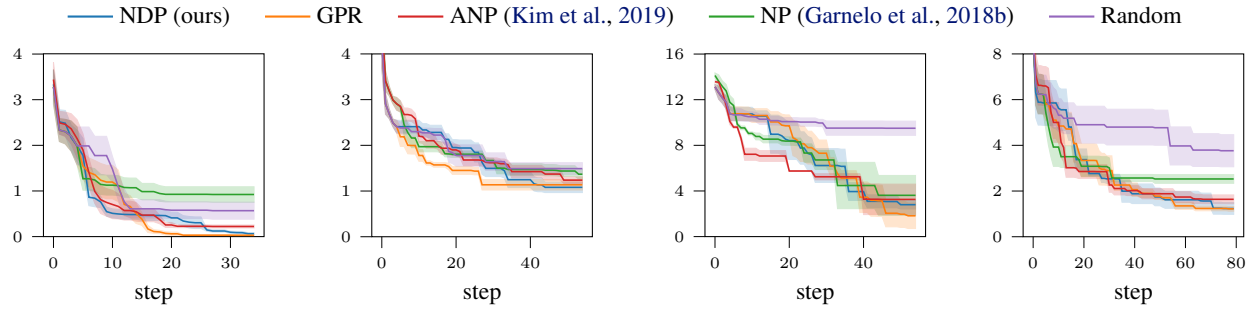
So far, we have used NDPs to model $p(\boldsymbol{y} \mid \boldsymbol{x}, \mathcal{D})$. This is a natural choice as in regression it is typically the only quantity of interest to make predictions. However, we can extend NDPs to model the joint $p(\boldsymbol{x}, \boldsymbol{y} \mid \mathcal{D})$. For this, during the forward process we corrupt both the function inputs and outputs with additive Gaussian noise. The task of the reverse process now consists of denoising both the corrupted inputs $\boldsymbol{x}_t$ and outputs $\boldsymbol{y}_t$, which leads to the objective

$$\mathcal{L}_\theta = \mathbb{E}_{t,\boldsymbol{x}_0,\boldsymbol{y}_0,\boldsymbol{\varepsilon}_x,\boldsymbol{\varepsilon}_y}\Big[\|\boldsymbol{\varepsilon}_x - \boldsymbol{\epsilon}_\theta^x(\boldsymbol{x}_t, \boldsymbol{y}_t, t)\|^2 + \|\boldsymbol{\varepsilon}_y - \boldsymbol{\epsilon}_\theta^y(\boldsymbol{x}_t, \boldsymbol{y}_t, t)\|^2\Big], \quad (10)$$

where we highlighted the differences with Eq. (5) in orange. Importantly, in this case we require a noise model for both the inputs and outputs: $\boldsymbol{\epsilon}_\theta^x$ and $\boldsymbol{\epsilon}_\theta^y$, resp. We detail the architecture of this NN in the supplementary (Fig. 8). We design it such that $\boldsymbol{\epsilon}_\theta^x$ and $\boldsymbol{\epsilon}_\theta^y$ share most of the weights apart from the final invariance layers as explained in Sec. 4. The inputs to the NN are now the corrupted inputs $\boldsymbol{x}_t = \sqrt{\bar{\alpha}_t}\boldsymbol{x}_0 + \sqrt{1-\bar{\alpha}_t}\boldsymbol{\varepsilon}_x$ and outputs $\boldsymbol{y}_t = \sqrt{\bar{\alpha}_t}\boldsymbol{y}_0 + \sqrt{1-\bar{\alpha}_t}\boldsymbol{\varepsilon}_y$. This is in contrast to the previous NDP in which $\boldsymbol{x}_t = \boldsymbol{x}_0$ for all t.

6.3.1. Global optimisation using NDPs

By taking advantage of the NDP's ability to model the full joint distribution $p(\boldsymbol{x}, \boldsymbol{y} \mid \mathcal{D})$ we conceive of a new global optimisation strategy. Consider conditioning the NDP on the current belief about the minima y^*. A NDP allows us to obtain samples from $p(x^* \mid y^*)$ which provides information about where the minima lie in the input domain. In Fig. 7 we illustrate a global optimisation routine using this idea. In each step, we sample a target from $p(y^*)$, which describes our belief about the minima. For the experiment, we sample y^* from a truncated-normal, where the mean corresponds to the minimum of the observed function values and the variance corresponds to the variance of the observations. The next query point is selected by sampling $p(x^* \mid y^*, \mathcal{D})$, thereby systematically seeking out the global minimum. This experiment showcases NDP's ability to model the complex interaction between inputs and outputs of a function —a task on which it was not trained.



(a) Hartmann 3D (b) Rastrigin 4D (c) Ackley 5D (d) Hartmann 6D

Figure 6: The regret of several probabilistic models used in Thompson sampling-based Bayesian Optimisation.

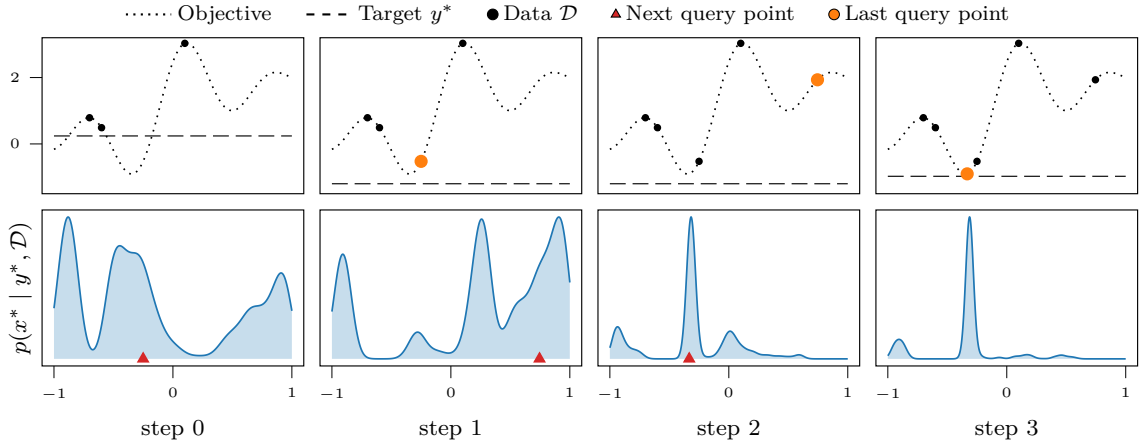


Figure 7: Global optimisation by diffusing the input locations. We condition the model at each step on a target y^* value (dashed line), and use the NDP to sample from $p(x^* \mid y^*, \mathcal{D})$. The panels in the upper row illustrate the progression of the optimisation. Bottom row shows the distribution $p(x^* \mid y^*, \mathcal{D})$ and the red triangles mark the next selected query point.

7. Conclusion

We proposed Neural Diffusion Processes (NDPs), a denoising diffusion generative model for learning distribution over functions, and generating prior and conditional samples. NDPs generalise diffusion models to infinite-dimensional functions by allowing the indexing of the function's marginals. We introduced the bi-dimensional attention block, which wires dimension and sequence equivariance into the architecture such that it satisfies the basic properties of a stochastic process. We empirically show that NDPs are able to capture functional distributions that are richer than Gaussian processes, and more accurate than neural processes. We concluded the paper by proposing a novel optimisation strategy based on diffusing the inputs. The code is available at `https://github.com/vdutor/neural-diffusion-processes`.

Limitations As with other diffusion models, we found that the sample quality improves with the number of diffusion steps. This does however lead to slower inference times. Techniques for accelerating the inference process could be incorporated to ameliorate this issue. Secondly, as is common with NNs we found that it is important for test input points to lie within the training range, as going beyond leads to poor performance.

Acknowledgements

We would like to recognise the contributions of Carl Henrik Ek, Erik Bodin, Sebastian Ober, and anonymous reviewers for their insightful feedback and constructive suggestions. We're also thankful to Yeh Whye Teh, Emile Mathieu, and Michael Hutchinson for their valuable insights, particularly concerning the topic of consistency. Their inputs have substantially shaped the presentation of this work.

References

Berkeley, J., Moss, H. B., Artemev, A., Pascual-Diaz, S., Granta, U., Stojic, H., Couckuyt, I., Qing, J., Loka, N., Paleyes, A., Ober, S. W., and Picheny, V. Trieste, 2022. URL https://github.com/secondmind-labs/trieste.

Bond-Taylor, S. and Willcocks, C. G. infinite-diff: Infinite resolution diffusion with subsampled mollified states. *arXiv preprint arXiv:2303.18242*, 2023.

Bruinsma, W., Markou, S., Requeima, J., Foong, A. Y. K., Vaughan, A., Andersson, T., Buonomo, A., Hosking, S., and Turner, R. E. Autoregressive Conditional Neural Processes. In *International Conference on Learning Representations*, February 2023. URL https://openreview.net/forum?id=OAsXFPBfTBh.

Bruinsma, W. P., Requeima, J., Foong, A. Y., Gordon, J., and Turner, R. E. The Gaussian neural process. In *Advances in Approximate Bayesian Inference*, 2021.

Dubois, Y., Gordon, J., and Foong, A. Y. Neural process family. http://yanndubs.github.io/Neural-Process-Family/, September 2020.

Foong, A. Y. K., Bruinsma, W. P., Gordon, J., Dubois, Y., Requeima, J., and Turner, R. E. Meta-Learning Stationary Stochastic Process Prediction with Convolutional Neural Processes. *arXiv:2007.01332 [cs, stat]*, November 2020.

Fortuin, V., Baranchuk, D., Rätsch, G., and Mandt, S. GP-VAE: Deep probabilistic time series imputation. In *International Conference on Artificial Intelligence and Statistics*, 2020.

Franzese, G., Rossi, S., Rossi, D., Heinonen, M., Filippone, M., and Michiardi, P. Continuous-time functional diffusion processes. *arXiv preprint arXiv:2303.00800*, 2023.

Garnelo, M., Rosenbaum, D., Maddison, C., Ramalho, T., Saxton, D., Shanahan, M., Teh, Y. W., Rezende, D. J., and Eslami, S. M. A. Conditional neural processes. In *International Conference on Machine Learning*, 2018a.

Garnelo, M., Schwarz, J., Rosenbaum, D., Viola, F., Rezende, D. J., Eslami, S. M. A., and Teh, Y. W. Neural processes. 2018b.

Gordon, J., Bruinsma, W. P., Foong, A. Y. K., Requeima, J., Dubois, Y., and Turner, R. E. Convolutional conditional neural processes. In *International Conference on Learning Representations*, 2019.

Ho, J., Jain, A., and Abbeel, P. Denoising diffusion probabilistic models. In *Neural Information Processing Systems*, 2020.

Hoogeboom, E., Satorras, V. G., Vignac, C., and Welling, M. Equivariant diffusion for molecule generation in 3D. In *International Conference on Machine Learning*, pp. 8867–8887, 2022.

Kerrigan, G., Ley, J., and Smyth, P. Diffusion Generative Models in Infinite Dimensions, 2022.

Kim, H., Mnih, A., Schwarz, J., Garnelo, M., Eslami, S. M. A., Rosenbaum, D., Vinyals, O., and Teh, Y. W. Attentive neural processes. In *International Conference on Learning Representations*, 2019.

Kong, Z., Ping, W., Huang, J., Zhao, K., and Catanzaro, B. Diffwave: A versatile diffusion model for audio synthesis. In *International Conference on Learning Representations*, 2020.

Kossen, J., Band, N., Gomez, A. N., Lyle, C., Rainforth, T., and Gal, Y. Self-attention between datapoints: Going beyond individual input-output pairs in deep learning. *arXiv:2106.02584*, 2021.

Lalchand, V. and Rasmussen, C. E. Approximate inference for fully Bayesian Gaussian process regression. In *Advances in Approximate Bayesian Inference*, 2020.

Lim, J. H., Kovachki, N. B., Baptista, R., Beckham, C., Azizzadenesheli, K., Kossaifi, J., Voleti, V., Song, J., Kreis, K., Kautz, J., Pal, C., Vahdat, A., and Anandkumar, A. Score-based diffusion models in function space, 2023.

Liu, S., Sun, X., Ramadge, P. J., and Adams, R. P. Task-agnostic amortized inference of Gaussian process hyperparameters. In *Neural Information Processing Systems*, 2020.

Lugmayr, A., Danelljan, M., Romero, A., Yu, F., Timofte, R., and Van Gool, L. Repaint: Inpainting using denoising diffusion probabilistic models. 2022.

Luo, S. and Hu, W. Diffusion probabilistic models for 3d point cloud generation. In *Computer Vision and Pattern Recognition*, 2021.

Mishra, S., Flaxman, S., Berah, T., Pakkanen, M., Zhu, H., and Bhatt, S. pivae: Encoding stochastic process priors with variational autoencoders. *arXiv preprint arXiv:2002.06873*, 2020.

Neal, R. M. Regression and classification using gaussian process priors. In Bernardo, J. M., Berger, J. O., Dawid, J. W., and Smith, A. F. M. (eds.), *International Conference on Learning Representations*. Oxford University Press, 1998.

Nichol, A., Dhariwal, P., Ramesh, A., Shyam, P., Mishkin, P., McGrew, B., Sutskever, I., and Chen, M. Glide:

Towards photorealistic image generation and editing with text-guided diffusion models. *arXiv preprint arXiv:2112.10741*, 2021.

Phillips, A., Seror, T., Hutchinson, M., De Bortoli, V., Doucet, A., and Mathieu, E. Spectral Diffusion Processes, November 2022. URL http://arxiv.org/abs/2209.14125.

Pidstrigach, J., Marzouk, Y., Reich, S., and Wang, S. Infinite-Dimensional Diffusion Models for Function Spaces, February 2023. URL http://arxiv.org/abs/2302.10130.

Ramesh, A., Dhariwal, P., Nichol, A., Chu, C., and Chen, M. Hierarchical text-conditional image generation with CLIP latents. *arXiv preprint arXiv:2204.06125*, 2022.

Rasmussen, C. E. and Williams, C. K. *Gaussian processes for machine learning*. MIT Press, 2006.

Shahriari, B., Swersky, K., Wang, Z., Adams, R. P., and De Freitas, N. Taking the human out of the loop: A review of bayesian optimization. *Proceedings of the IEEE*, 2015.

Simpson, F., Davies, I., Lalchand, V., Vullo, A., Durrande, N., and Rasmussen, C. E. Kernel identification through transformers. In *Neural Information Processing Systems*, 2021a.

Simpson, F., Lalchand, V., and Rasmussen, C. E. Marginalised Gaussian processes with nested sampling. In *Neural Information Processing Systems*, 2021b.

Sohl-Dickstein, J., Weiss, E. A., Maheswaranathan, N., and Ganguli, S. Deep unsupervised learning using nonequilibrium thermodynamics. *arXiv preprint arXiv:1503.03585*, 2015.

Song, Y. and Ermon, S. Improved techniques for training score-based generative models. In *Neural Information Processing Systems*, volume 33, 2020.

van der Wilk, M., Dutordoir, V., John, S., Artemev, A., Adam, V., and Hensman, J. A framework for interdomain and multioutput Gaussian processes. *arxiv preprint arXiv:2003.01115*, 2020.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. Attention is all you need. In *Neural Information Processing Systems*, 2017.

Wang, C. and Neal, R. M. Gaussian process regression with heteroscedastic or non-gaussian residuals. *arXiv preprint arXiv:1212.6246*, 2012.

Wilson, A. G., Hu, Z., Salakhutdinov, R., and Xing, E. P. Deep kernel learning. In Gretton, A. and Robert, C. C. (eds.), *Proceedings of the 19th International Conference on Artificial Intelligence and Statistics*, volume 51 of *Proceedings of Machine Learning Research*, pp. 370–378, Cadiz, Spain, 09–11 May 2016. PMLR.

Xu, M., Yu, L., Song, Y., Shi, C., Ermon, S., and Tang, J. Geodiff: A geometric diffusion model for molecular conformation generation. *arXiv preprint arXiv:2203.02923*, 2022.

Zaheer, M., Kottur, S., Ravanbakhsh, S., Poczos, B., Salakhutdinov, R. R., and Smola, A. J. Deep sets. In *Neural Information Processing Systems*, 2017.

A. Derivation of the Loss

In this section we derive the objectives given in Eqs. (5) and (10), following a path similar to that presented in Ho et al. (2020). Let $\boldsymbol{s}_t = (\boldsymbol{x}_t, \boldsymbol{y}_t)$ be the state that combines both the inputs and observations. We wish to find the set of parameters θ which maximise the likelihood given the initial state, $p_\theta(\boldsymbol{s}_0)$. While a direct evaluation of the likelihood appears intractable,

$$p_\theta(\boldsymbol{s}_0) = \int p_\theta(\boldsymbol{s}_{0\ldots T}) d\boldsymbol{s}_{1\ldots T}\,,$$

this may be recast in a form which allows for a comparison to be drawn between the forward and reverse trajectories (Sohl-Dickstein et al., 2015)

$$p_\theta(\boldsymbol{s}_0) = \int p_\theta(\boldsymbol{s}_T) q(\boldsymbol{s}_{1\ldots T}|\boldsymbol{s}_0) \prod_{t=1}^{T} \frac{p_\theta(\boldsymbol{s}_{t-1}|t)}{q(\boldsymbol{s}_t|\boldsymbol{s}_{t-1})} d\boldsymbol{s}_{1\ldots T}\,.$$

The appeal of introducing the reverse process is that it is tractable when conditioned on the first state $\boldsymbol{s}_0$, taking a Gaussian form

$$q(\boldsymbol{s}_{t-1} \mid \boldsymbol{s}_t, \boldsymbol{s}_0) = \mathcal{N}(\boldsymbol{s}_{t-1} \mid \tilde{\boldsymbol{\mu}}(\boldsymbol{s}_t, \boldsymbol{s}_0), \tilde{\beta}_t \mathbf{I}).$$

Our loss function reflects a lower bound on the negative log likelihood

$$\mathbb{E}[-\log p_\theta(\boldsymbol{s}_0)] \leq \mathbb{E}_{q(\boldsymbol{s}_{0:T})}\left[\log \frac{q(\boldsymbol{s}_{1:T} \mid \boldsymbol{s}_0)}{p_\theta(\boldsymbol{s}_{0:T})}\right] := \mathcal{L}_\theta\,,$$

which may be decomposed into two edge terms and a sum over the intermediate steps, as follows

$$\mathcal{L}_\theta = \mathbb{E}_q\left[L_T + \sum_{t=2}^{T} L_{t-1} + L_0\right]\,,$$

where

$$\begin{aligned} L_0 &= -\log p_\theta(\boldsymbol{s}_0 \mid \boldsymbol{s}_1)\,, \\ L_{t-1} &= \mathrm{KL}(q(\boldsymbol{s}_{t-1} \mid \boldsymbol{s}_t, \boldsymbol{s}_0) \| p_\theta(\boldsymbol{s}_{t-1} \mid \boldsymbol{s}_t))\,, \\ L_T &= \mathrm{KL}(q(\boldsymbol{s}_T \mid \boldsymbol{s}_0) \| p_\theta(\boldsymbol{s}_T))\,. \end{aligned} \tag{11}$$

We can write

$$L_{t-1} = \mathbb{E}_{\boldsymbol{s}_0, \boldsymbol{\varepsilon}}\left[\frac{1}{2\sigma^2} \|\tilde{\boldsymbol{\mu}}(\boldsymbol{s}_t, \boldsymbol{s}_0) - \boldsymbol{\mu}_\theta(\boldsymbol{s}_t, t)\|^2\right]\,, \tag{12}$$

where the mean is a function of the previous and first state

$$\tilde{\boldsymbol{\mu}}(\boldsymbol{s}_t, \boldsymbol{s}_0) = \frac{\sqrt{\alpha_t}(1-\bar{\alpha}_{t-1})}{1-\bar{\alpha}_t}\boldsymbol{s}_t + \frac{\sqrt{\bar{\alpha}_{t-1}}\beta_t}{1-\bar{\alpha}_t}\boldsymbol{s}_0$$

and the variance

$$\tilde{\beta}_t = \frac{1-\bar{\alpha}_{t-1}}{1-\bar{\alpha}_t}\beta_t\,.$$

It is helpful to consider the relationship between the initial and final states

$$\boldsymbol{s}_0 = \frac{1}{\sqrt{\bar{\alpha}_t}}(\boldsymbol{s}_t - \sqrt{1-\bar{\alpha}_t}\boldsymbol{\varepsilon}_s),$$

where $\boldsymbol{\varepsilon}_s = (\boldsymbol{\varepsilon}_x, \boldsymbol{\varepsilon}_y)$, so that we can rewrite the mean as

$$\tilde{\boldsymbol{\mu}}(\boldsymbol{s}_t, \boldsymbol{\varepsilon}_s) = \frac{1}{\sqrt{\alpha_t}}(\boldsymbol{s}_t - \frac{\beta_t}{\sqrt{1-\bar{\alpha}_t}}\boldsymbol{\varepsilon}_s)\,. \tag{13}$$

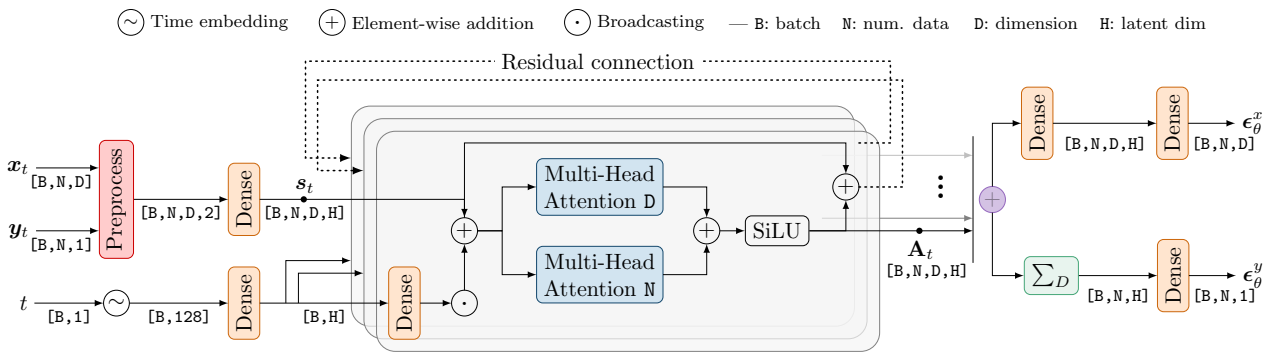


Figure 8: Architecture of the NDP's NN noise models $\boldsymbol{\epsilon}_\theta^x$ and $\boldsymbol{\epsilon}_\theta^y$. Compared to Fig. 2 this architecture has two outputs, one to predict the corruption on the inputs $\boldsymbol{x}_t$ and one to predict the corruption on the function outputs $\boldsymbol{y}_t$. Both output share a lot of weights in the network and only bifurcate in the last few layers.

Equation 12 can now be expressed as

$$L_{t-1} = \mathbb{E}_{\boldsymbol{s}_0, \boldsymbol{\varepsilon}_s}\left[\frac{\beta_t^2}{2\sigma^2\alpha_t(1-\bar{\alpha}_t)}\|\boldsymbol{\varepsilon}_s - \boldsymbol{\epsilon}_\theta^s\|^2\right]. \tag{14}$$

Finally, since the variance schedule is fixed, and the edge terms are not found to improve empirical performance, our simplified training objective is given by

$$\mathcal{L}_\theta = \mathbb{E}_{\boldsymbol{x}_0, \boldsymbol{y}_0, \boldsymbol{\varepsilon}_x, \boldsymbol{\varepsilon}_y, t}\left[\|\boldsymbol{\varepsilon}_x - \boldsymbol{\epsilon}_\theta^x(\boldsymbol{x}_t, \boldsymbol{y}_t, t)\|^2 + \|\boldsymbol{\varepsilon}_y - \boldsymbol{\epsilon}_\theta^y(\boldsymbol{x}_t, \boldsymbol{y}_t, t)\|^2\right]. \tag{15}$$

So far we have derived the objective of the 'full' NDP model (i.e., the model which diffuses both $\boldsymbol{x}$ and $\boldsymbol{y}$) given in Eq. (10) of the main paper. In Fig. 8 we detail the architecture for the noise models $\boldsymbol{\epsilon}_\theta^x(\cdot)$ and $\boldsymbol{\epsilon}_\theta^y(\cdot)$.

B. Algorithms

In this section we list pseudo-code for training and sampling NDPs.

B.1. Training

Algorithm 1 Training

input Input distribution $q(\boldsymbol{x}_0)$ and covariance function k_ψ, with prior over hyperparameters p_ψ. Noise schedule β_t for $t \in \{1, 2, \ldots, T\}$. A loss function L (e.g., MSE or MAE).

begin
  Precompute $\gamma_t = \sqrt{1-\bar{\alpha}_t}$ and $\bar{\alpha}_t = \prod_{j=1}^{t}(1-\beta_j)$.
  for $i = 1, 2, \cdots, N_{\text{iter}}$ **do**
    Sample $\boldsymbol{x}_0 \sim q(\boldsymbol{x}_0)$, $\psi \sim p_\psi$, $\boldsymbol{y}_0 \sim \mathcal{N}(\mathbf{0}, k_\psi(\boldsymbol{x}_0, \boldsymbol{x}_0) + \sigma^2\mathbf{I})$.
    Sample $\boldsymbol{\varepsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$, and $t \sim \mathcal{U}(\{1, \ldots, T\})$.
    Compute $\boldsymbol{y}_t = \sqrt{\bar{\alpha}_t}\boldsymbol{y}_0 + \gamma_t\boldsymbol{\varepsilon}$.
    Update θ using gradient $\nabla_\theta L(\boldsymbol{\varepsilon}, \boldsymbol{\epsilon}_\theta(\boldsymbol{x}_0, \boldsymbol{y}_t, t))$.
  end for

B.2. Prior and Conditional Sampling

In practice, the code to sample the prior is a special case of the conditional code with an empty context dataset and $U = 1$. However, here we list them both for the clarity of the exposition.

Algorithm 2 Prior Sampling

input A set of input locations $\boldsymbol{x}_0$, a NDP noise model $\boldsymbol{\epsilon}_\theta(\cdot)$. γ_t and $\bar{\alpha}_t$ for a given noise schedule β_t.

begin
Sample a random initial state and $\boldsymbol{y}_T$ from $\mathcal{N}(\mathbf{0}, \mathbf{I})$.
for $t = T, T-1, \ldots, 1$ **do**
Sample using backward kernel:

$$\boldsymbol{y}_{t-1} \sim \mathcal{N}\big(\frac{1}{\sqrt{1-\beta_t}}(\boldsymbol{y}_t - \frac{\beta_t}{\gamma_t}\boldsymbol{\epsilon}_\theta(\boldsymbol{x}_0, \boldsymbol{y}_t, t)), \frac{\gamma_t^2}{\gamma_{t-1}^2}\beta_t\mathbf{I}\big).$$

end for
output $(\boldsymbol{x}_0, \boldsymbol{y}_0)$

Algorithm 3 Conditional Sampling

input A context dataset $\mathcal{D} = \{([\boldsymbol{x}_0^c]_i \in \mathcal{X}, [\boldsymbol{y}_0^c]_i \in \mathbb{R})\}_{i=1}^N$. A NDP noise model $\boldsymbol{\epsilon}_\theta(\cdot)$. γ_t and $\bar{\alpha}_t$ for a given noise schedule β_t.

begin
Sample $\boldsymbol{y}_T^* \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$.
Let $\boldsymbol{x}_0 = [\boldsymbol{x}_0^c, \boldsymbol{x}_0^*]$
for $t = T, T-1, \ldots, 1$ **do**
for $u = 1, \ldots, U$ **do**
Sample context points t steps forward

$$\boldsymbol{y}_t^c \sim \mathcal{N}\Big(\sqrt{\bar{\alpha}_t}\boldsymbol{y}_0^c, (1-\bar{\alpha}_t)\,\mathbf{I}\Big). \tag{16}$$

Let $\boldsymbol{y}_t = [\boldsymbol{y}_t^c, \boldsymbol{y}_t^*]$.
Sample backward Eq. (3)

$$\boldsymbol{y}_{t-1} \sim \mathcal{N}\big(\frac{1}{\sqrt{1-\beta_t}}(\boldsymbol{y}_t - \frac{\beta_t}{\gamma_t}\boldsymbol{\epsilon}_\theta(\boldsymbol{x}_0, \boldsymbol{y}_t, t)), \frac{\gamma_t^2}{\gamma_{t-1}^2}\beta_t\mathbf{I}\big).$$

Diffuse forward by one step

$$\boldsymbol{y}_t \sim \mathcal{N}\big(\sqrt{1-\beta_t}\boldsymbol{y}_{t-1}, \beta_t\mathbf{I}\big)$$

end for
end for
output $(\boldsymbol{x}_0, \boldsymbol{y}_0)$

C. Proofs

In this section, we shall formally demonstrate that NDP's noise model adheres to the symmetries associated with permutations of the datapoint orderings and the permutations of the input dimensions. We focus our attention to the full noise model of Fig. 8 because Fig. 2 is a simplification of it in which we only keep a single output. We start by proving properties of its main building block: the bi-dimensional attention block. We can then straightforwardly prove the equivariance and invariance properties that hold for the NDP's noise model. Before that, we start by setting the notation.

C.1. Notation, Definitions and Preliminary lemmas

Notation. Let $\boldsymbol{s} \in \mathbb{R}^{N \times D \times H}$ be a tensor of rank (or dimension) three, where we refer to each dimension according to the following convention:

$$\text{shape}(\boldsymbol{s}) = [N, D, H],$$

where N stands for the sequence length, D the input dimensionality and H the embedding.

In the next definitions we will use NumPy-based indexing and slicing notation. We assume the reader is familiar with this convention. Most notably, we use a colon (:) to reference every element in a dimension.

Definition C.1. Let Π_N be the set of all permutations of indices $\{1, \dots, N\}$. Let $\pi_n \in \Pi_N$ and $\boldsymbol{s} \in \mathbb{R}^{N \times D \times H}$. Then $(\pi_n \circ \boldsymbol{s}) \in \mathbb{R}^{N \times D \times H}$ denotes a tensor where the ordering of indices in the first dimension are reshuffled (i.e., permuted) according to π_n. We write

$$\pi_n \circ \boldsymbol{s} = \boldsymbol{s}_{\pi_n(1), \pi_n(2), \dots, \pi_n(N); :; :}$$

Definition C.2. Let Π_D be the set of all permutations of indices $\{1, \dots, D\}$. Let $\pi_d \in \Pi_D$ and $\boldsymbol{s} \in \mathbb{R}^{N \times D \times H}$. Then $(\pi_d \circ \boldsymbol{s}) \in \mathbb{R}^{N \times D \times H}$ denotes a tensor where the ordering of indices in the second dimension are reshuffled (i.e., permuted) according to π_d. We write

$$\pi_d \circ \boldsymbol{s} = \boldsymbol{s}_{:, \pi_d(1), \pi_d(2), \dots, \pi_d(D); :}.$$

Definition C.3. A function $f : \mathbb{R}^{N \times D \times H} \to \mathbb{R}^{N \times D \times H}$ is *equivariant to* Π if for any permutation $\pi \in \Pi$, we have

$$f(\pi \circ \boldsymbol{s}) = \pi \circ f(\boldsymbol{s}).$$

Definition C.4. A function $f : \mathbb{R}^{N \times D \times H} \to \mathbb{R}^{N \times D \times H}$ is *invariant to* Π if for any permutation $\pi \in \Pi$, we have

$$f(\pi \circ \boldsymbol{s}) = f(\boldsymbol{s}).$$

Invariance in layman's terms means that the output is not affected by a permutation of the inputs.

Lemma C.5. *The composition of equivariant functions is equivariant.*

Proof. Let f and g be equivariant to Π, then for all $\pi \in \Pi$:

$$(f \circ g)(\pi \circ \boldsymbol{s}) = f(g(\pi \circ \boldsymbol{s})) = f(\pi \circ g(\boldsymbol{s})) = \pi \circ f(g(\boldsymbol{s})),$$

which shows that the composition, $f \circ g$, is equivariant as well. □

Lemma C.6. *An element-wise operation between equivariant functions remains equivariant.*

Proof. Let f and g be equivariant to Π, then for all $\pi \in \Pi$ and an element-wise operation $\oplus$ (e.g., addition), we have

$$(f \oplus g)(\pi \circ \boldsymbol{s}) = f(\pi \circ \boldsymbol{s}) \oplus g(\pi \circ \boldsymbol{s}) = (\pi \circ f(\boldsymbol{s})) \oplus (\pi \circ g(\boldsymbol{s})) = \pi \circ (f \oplus g).$$

which shows that the composition, $f \oplus g$, is equivariant as well. □

Lemma C.7. *A function $f : \mathbb{R}^{N \times D \times H} \to \mathbb{R}^{N \times D \times H}$ which applies the same function g to all its rows, i.e. $f : \boldsymbol{s} \mapsto [g(\boldsymbol{s}_{1;:;:}), g(\boldsymbol{s}_{2;:;:}), \dots, g(\boldsymbol{s}_{N;:;:})]$ with $g : \mathbb{R}^{D \times H} \to \mathbb{R}^{D \times H}$ is equivariant in the first dimension.*

Proof. Follows immediately from the structure of f. □

C.2. Bi-Dimensional Attention Block

With the notation, definitions and lemmas in place, we now prove that the bi-dimensional attention block is equivariant in its first and second dimension, respectively to permutations in the set Π_N and Π_D. We start this section by formally defining the bi-dimensional attention block and its main components attention components Attn_N and Attn_D. Finally, in Appendix C.3 we prove the properties of the noise models $\boldsymbol{\epsilon}_\theta^x$ and $\boldsymbol{\epsilon}_\theta^y$ making use of the results in this section.

Definition C.8. Ignoring the batch dimension B, let $\mathbf{A}_t : \mathbb{R}^{N \times D \times H} \to \mathbb{R}^{N \times D \times H}; \boldsymbol{s} \mapsto \mathbf{A}_t(\boldsymbol{s})$ be the bi-dimensional attention block. As illustrated in Fig. 8, it operates on three dimensional tensors in $\mathbb{R}^{N \times D \times H}$ and applies attention across the first and second dimension using Attn_N and Attn_D, respectively. The final output $\mathbf{A}_t(\boldsymbol{s})$ is obtained by summing the two attention outputs, before applying an element-wise non-linearity.

We now proceed by defining the component Attn_D and its properties. Subsequently, in Definition C.11 we define Attn_N and its properties. Finally, we combine both to prove Proposition 4.1 in the main paper about the bi-dimensional attention block.

Definition C.9. Let $\text{Attn}_D : \mathbb{R}^{N\times D\times H} \to \mathbb{R}^{N\times D\times H}$ be a self-attention block (Vaswani et al., 2017) acting across D. Let σ be a softmax activation function operating on the last dimension of a tensor. Then, Attn_D is defined as

$$\text{Attn}_D(\boldsymbol{s})[n,d,h] = \sum_{d'=1}^{D} \boldsymbol{\Sigma}_{n,d,d'}(\boldsymbol{s})\boldsymbol{s}^v_{n,d',h}, \quad \text{where} \quad \boldsymbol{\Sigma}_{n,d,d'}(\boldsymbol{s}) = \sigma\big(\frac{1}{\sqrt{H}}\sum_l \boldsymbol{s}^k_{n,d,\ell}\boldsymbol{s}^q_{n,d',\ell}\big)$$

given a linear projection of the inputs $\boldsymbol{s}$ which maps them into keys (k), queries (q) and values (v)

$$\boldsymbol{s}^k_{n,d,\ell} = \sum_j \boldsymbol{s}_{n,d,j}\mathbf{W}^k_{j,\ell}, \quad \boldsymbol{s}^q_{n,d',\ell} = \sum_j \boldsymbol{s}_{n,d,j}\mathbf{W}^q_{j,\ell}, \quad \boldsymbol{s}^v_{n,d',h} = \sum_j \boldsymbol{s}_{n,d,j}\mathbf{W}^v_{j,\ell}.$$

Proposition C.10. *Attn_D is equivariant to Π_N and Π_D (i.e., across sequence length and input dimensionality).*

Proof. We prove the equivariance to Π_N and Π_D separately. First, from the definition we can see that Attn_D is a function that acts on each element row of $\boldsymbol{s}$ separately. Thus, by Lemma C.7, Attn_D is equivariant to Π_N.

Next, we want to prove equivariance to Π_D. We want to show that for all $\pi_d \in \Pi_D$:

$$\text{Attn}_D(\pi_d \circ \boldsymbol{s}) = \pi_d \circ \text{Attn}_D(\boldsymbol{s}).$$

The self-attention mechanism consists of a matrix multiplication of the attention matrix Σ and the projected inputs. We start by showing that the attention matrix is equivariant to permutations in Π_D

$$\begin{aligned}
\boldsymbol{\Sigma}_{n,d,d'}(\pi_d \circ \boldsymbol{s}) &= \sigma\Big(\frac{1}{\sqrt{H}}\sum_\ell \big(\pi_d \circ \boldsymbol{s}\big)^k_{n,d,\ell}\big(\pi_d \circ \boldsymbol{s}\big)^q_{n,d,\ell}\Big) \\
&= \sigma\Big(\frac{1}{\sqrt{H}}\sum_{\ell,j} \big(\boldsymbol{s}_{n,\pi_d(d),j}\mathbf{W}^k_{j,\ell}\big)\big(\boldsymbol{s}_{n,\pi_d(d'),j}\mathbf{W}^q_{j,\ell}\big)\Big) \\
&= \boldsymbol{\Sigma}_{n,\pi_d(d),\pi_d(d')}(\boldsymbol{s})
\end{aligned}$$

It remains to show that the final matrix multiplication step restores the row-equivariance

$$\begin{aligned}
\text{Attn}_D(\pi_d \circ \boldsymbol{s}) &= \sum_{d'} \boldsymbol{\Sigma}_{n,d,d'}(\pi_d \circ \boldsymbol{s})\big(\pi_d \circ \boldsymbol{s}\big)^v_{n,d',h} \\
&= \sum_{d'} \boldsymbol{\Sigma}_{n,\pi_d(d),\pi_d(d')}(\boldsymbol{s})\boldsymbol{s}^v_{n,\pi_d(d'),h} \\
&= \sum_{d'} \boldsymbol{\Sigma}_{n,\pi_d(d),d'}(\boldsymbol{s})\boldsymbol{s}^v_{n,d',h} = \pi_d \circ \text{Attn}_D(\boldsymbol{s}).
\end{aligned}$$

This concludes the proof. □

We continue by defining and proving the properties of the second main component of the bi-dimensional attention block: Attn_N.

Definition C.11. Let $\text{Attn}_N : \mathbb{R}^{N\times D\times H} \to \mathbb{R}^{N\times D\times H}$ be a self-attention block (Vaswani et al., 2017) acting across N (i.e. the sequence length). Let σ be a softmax activation function operating on the last dimension of a tensor. Then Attn_N is defined as

$$\text{Attn}_N(\boldsymbol{s})[n,d,h] = \sum_{n'=1}^{N} \boldsymbol{\Sigma}_{n,d,n'}(\boldsymbol{s})\boldsymbol{s}^v_{n',d,h}, \text{ where } \quad \boldsymbol{\Sigma}_{n,d,n'}(\boldsymbol{s}) = \sigma\big(\frac{1}{\sqrt{H}}\sum_l \boldsymbol{s}^k_{n,d,\ell}\boldsymbol{s}^q_{n',d,\ell}\big)$$

given a linear projection of the inputs $\boldsymbol{s}$ which maps them into keys (k), queries (q) and values (v)

$$\boldsymbol{s}^k_{n,d,\ell} = \sum_j \boldsymbol{s}_{n,d,j}\mathbf{W}^k_{j,\ell}, \quad \boldsymbol{s}^q_{n',d,\ell} = \sum_j \boldsymbol{s}_{n',d,j}\mathbf{W}^q_{j,\ell}, \quad \boldsymbol{s}^v_{n',d,h} = \sum_j \boldsymbol{s}_{n',d,j}\mathbf{W}^v_{j,\ell}.$$

Proposition C.12. *Attn_N is equivariant to Π_N and Π_D (i.e., across sequence length and input dimensionality).*

Proof. Follows directly from Proposition C.10 after transposing the first and second dimension of the input. □

Finally, we have the necessary ingredients to prove Proposition 4.1 from the main paper.

Proposition C.13. *The bi-dimensional attention block* $\mathbf{A}_t$ *is equivariant to* Π_D *and* Π_N.

Proof. The bi-dimensional attention block simply adds the output of Attn_D and Attn_N, followed by an element-wise non-linearity. Therefore, as a direct application of Lemma C.6 the complete bi-dimensional attention block remains equivariant to Π_N and Π_D. □

C.3. NDP Noise Model

By building on the equivariant properties of the bi-dimensional attention block, we prove the equivariance and invariance of the NDP's noise models, denoted by $\boldsymbol{\epsilon}_\theta^x$ and $\boldsymbol{\epsilon}_\theta^y$, respectively. We refer to Fig. 8 for their definition. In short, $\boldsymbol{\epsilon}_\theta^x$ consists of adding the output of several bi-dimensional attention blocks followed by dense layers operating on the last dimension. Similarly, $\boldsymbol{\epsilon}_\theta^y$ is constructed by summing the bi-dimensional attention blocks, but is followed by a summation over D before applying a final dense layer.

The following propositions hold:

Proposition C.14. *The function* $\boldsymbol{\epsilon}_\theta^x$ *is equivariant to* Π_D *and* Π_N.

Proof. The output $\boldsymbol{\epsilon}_\theta^x$ is formed by element-wise summing the output of bi-dimensional attention layers. Directly applying Lemma C.5 and Proposition C.13 completes the proof. □

Proposition C.15. *The function* $\boldsymbol{\epsilon}_\theta^y$ *is equivariant to* Π_N.

Proof. The summation over D does not affect the equivariance over Π_N from the bi-dimensional attention blocks as it can be cast as a row-wise operation. □

Proposition C.16. *The function* $\boldsymbol{\epsilon}_\theta^y$ *is invariant to* Π_D.

Proof. Follows from the equivariance of the bi-dimensional blocks and (Zaheer et al., 2017, Thm. 7). □

D. A Primer on Neural Processes

While Neural Diffusion Processes (NDPs) and Neural Processes (NPs) share a similar goal, they differ significantly in their approach.

NPs learn a function that maps inputs $x \in \mathbb{R}^d$ to outputs $y \in \mathbb{R}$. NPs specifically define a family of conditional distributions that allows one to model an arbitrary number of targets $\boldsymbol{x}^*$ by conditioning on an arbitrary number of observed contexts, referred to as the context dataset $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^n$.

NPs use an encoder-decoder architecture to define the conditional distributions. The encoder is a NN which operates on the context dataset to output a dataset representation $r = enc(\mathcal{D})$. Using this representation, a decoder predicts the function output at a test location. Conditional NPs (Garnelo et al., 2018a) define the target distribution as a Gaussian which factorises over the different target points

$$p(\boldsymbol{y}^* \mid \boldsymbol{x}^*, \mathcal{D}) = \prod_{i=1}^{n} \mathcal{N}(y_i^* \mid dec_\mu(x_i^*, r), dec_{\sigma^2}(x_i^*, r)), \tag{17}$$

where the mean and variance of the Gaussians are given by decoding the context dataset representation r and targets x_i^*. When the context dataset is empty $\mathcal{D} = \varnothing$, NPs typically set the representation r to a fixed vector.

A latent NP (Garnelo et al., 2018b) relies on a similar encoder-decoder architecture but now the encoder is used to parameterise a global latent variable z. Conditioned on z the likelihood factorises over the target points

$$p(\boldsymbol{y}^* \mid \boldsymbol{x}^*, \mathcal{D}) = \int_z \prod_{i=1}^{n} \mathcal{N}\left(y_i^* \mid dec_\mu(x_i^*, r, z), dec_{\sigma^2}(x_i^*, r, z)\right) p(z \mid r)\, \mathrm{d}z. \tag{18}$$

Table 2: Variability in the violin plots of Fig. 10 measured by the standard deviation on the means.

model	μ	σ^2	q_{50}	$q_{90} - q_{10}$	$q_{75} - q_{25}$
GP	0.005687	0.001491	0.005245	0.010979	0.011858
NDP	0.019796	0.002985	0.018961	0.018718	0.010045

The idea behind having a global latent variable is to model various instances of the stochastic process.

The encoder and decoder networks are trained by maximising a lower bound to the log likelihood over different function realisations, where a random subset of points is placed in the context and target sets. In all experiments we use an existing opensource package `https://github.com/wesselb/neuralprocesses` for the NP baselines as it provides well-tested and finely-tuned model configurations for a variety of tasks.

E. On Marginal Consistency of Neural Processes and Neural Diffusion Processes

When discussing marginal consistency, there are two different settings that one needs to consider.

1. For a given context dataset, does marginal consistency, $p(y_0 \mid \mathcal{D}) = \int p(y_0, y_1, \ldots, y_n \mid \mathcal{D}) \, \mathrm{d}y_{1:n}$, hold?
2. Are the conditional distributions related through Bayes' rule: $p(y_0 \mid \mathcal{D}) = \int p(y_0 \mid \mathcal{D}) p(y_1 \mid y_0, \mathcal{D}) \, \mathrm{d}y_1$?

It is straightforward to prove that the family of NP models satisfy the first condition. This follows directly from the factorisation of the target distribution in Eqs. (17) and (18). We refer to Dubois et al. (2020) for a concise overview of the proofs. As a result, NPs will for a given context dataset satisfy the KET conditions. However, it is important to note that consistency of NPs is not satisfied within contexts (Kim et al., 2019). To see this, consider the case where we marginalise over y_1 and apply Bayes' rule

$$p(y_0 \mid \mathcal{D}) = \int_{y_1} p(y_1 \mid \mathcal{D}) \, p(y_0 \mid y_1, \mathcal{D}) \, \mathrm{d}y_1. \tag{19}$$

The distribution produced by the NP on the LHS of the equation need not match the distribution one would obtain if you append the additional point (x_1, y_1) to the context dataset on the RHS. This is a consequence of the NNs, which are responsible for encoding and decoding the context dataset, and can not guarantee consistency when their input changes. This holds true for latent and conditional NPs, which means neither of them are marginally consistent when their context dataset changes.

NDPs can not mathematically guarantee marginal consistency as defined by Eq. (2) for an arbitrary number of points n and sequences $\{x_i\}_{i=1}^n$. Hence, they do not satisfy the criteria for Kolmogorov Extension Theorem. However, it is important to note that NDPs differ from NPs in their approach to constructing a predictive distribution. In NPs, the predictive distribution is generated directly by the NN. In NDPs, on the contrary, the predictive distribution is constructed through conditioning the joint.

E.1. Empirical Evaluation

While NDPs do not mathematically satisfy marginal consistency for arbitrary sequences, the following experiment demonstrates that they approximate consistency through meta-learning. Similar to NPs, we argue that this behaviour is induced by the maximum log-likelihood objective, which can be seen as minimising the KL between the consistent samples from the data-generating process and the NDP's output.

To illustrate this behaviour, let us sample from the joint $p(y_0 \cup \{y_i\}_{i=1}^n) \mid \mathcal{D})$, and investigate the dependence of the marginal $p(y_0)$ as we vary the other test inputs $\{x_i\}_{i=1}^n$. We set y_0 to correspond to the prediction at $x = 0$ and study 5 different random configurations for the other input locations $\{x_i\}_{i=1}^n$, numbered #1 to #5. In Fig. 9, we show the posterior samples $p(y_0 \cup \{y_i\}_{i=1}^n \mid \mathcal{D})$, the difference between the 5 plots is that we evaluate the posterior at different locations $\{x_i\}_{i=1}^n$ in $[-1, 1]$. For each configuration we make sure that $x = 0.0$ is included such that we can investigate the marginal $p(y_0)$.

Figure 10 shows the empirical distribution across samples of the mean μ, variance σ^2, median q_{50} and quantile widths of $p(y_1)$ for each of the different input configurations. As a reference, we also add the statistics of GP at $x = 0$ samples, which

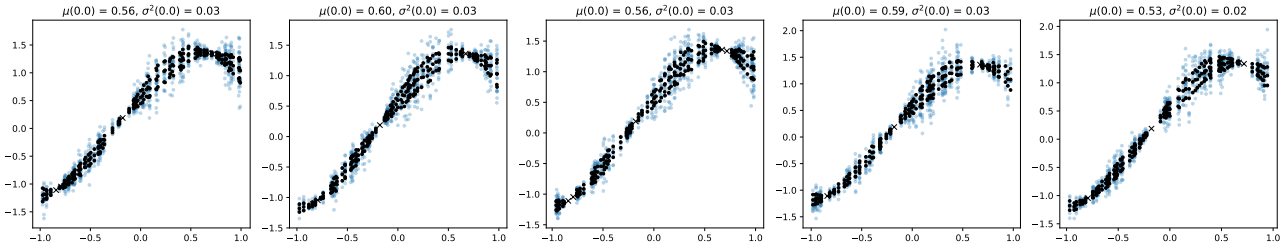


Figure 9: Samples (blue) from the NDP with empirical 25 and 75th quantiles (black). In each of the 5 plots we vary the locations at which the posterior is evaluated. We make sure that $x = 0$ is included in the test points, which allows us to compute the empirical mean, variance and quantiles of y_0, which corresponds to the sample evaluated at $x = 0$.

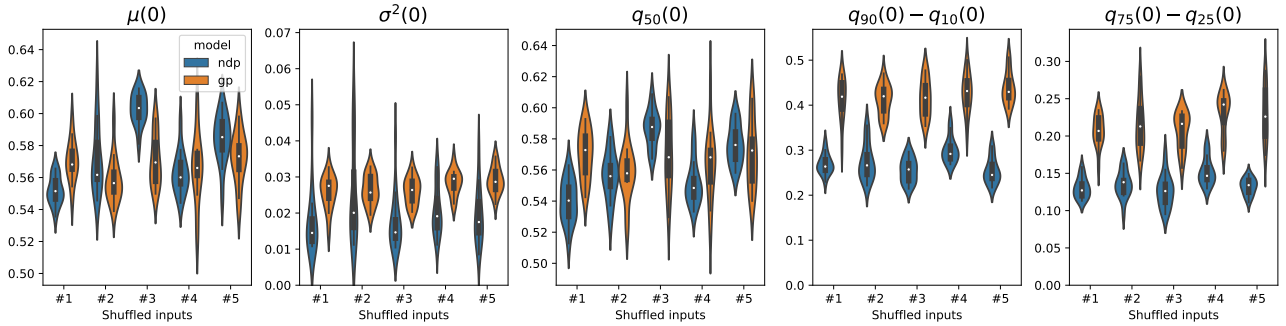


Figure 10: Boxplot of the empirical mean, variance, median and quantile widths of Fig. 9 by repeating the experiment 20 times for each random input configuration. We compare the NDP across different input configurations and agains a GP.

are drawn from a consistent distribution, to the figure. Consistency is observed when the violin plots of the statistics do not change as of a result of varying $\{x\}_{i=1}^{n}$ (i.e., #1, #2, ..., #5 for each model appear the same).

We can quantify this variability by computing the standard deviation on the means of the violin plots across the varying inputs. The numerical values of this are given in Table 2. We observe that variation in the quantiles are found to be lower than 10%. By comparing with the equivalent performance of a GP, we find that the bulk of the variations are due to stochastic fluctuations as a result of finite sample size as opposed to the inconsistencies associated with the NDP.

F. Additional Information on the Experiments

F.1. Experimental Setup and Overview

Here we provide more detailed information describing how the numerical experiments were conducted.

All experiments share the same model architecture illustrated in Figure 2, there are however a number of model parameters that must be chosen. An L1 (i.e., Mean Absolute Error, MAE) loss function was used throughout. We use four or five bi-dimensional attention blocks, each consisting of multi-head self-attention blocks (Vaswani et al., 2017) containing a representation dimensionality of $H = 64$ and 8 heads. Each experiment used either 500 or 1000 diffusion steps, where we find larger values produce more accurate samples at the expense of computation time. Following Nichol et al. (2021) we use a cosine-based scheduling of β_t during training. The Adam optimiser is used throughout. Our learning rate follows a cosine-decay function, with a 20 epochs linear learning rate warm-up to a maximum learning rate of $\eta = 0.001$ before decaying. All NDP models were trained for 250 epochs with the exception of the lengthscale marginalisation experiment, which was trained for 500 epochs. Each epoch contained 4096 example training $(\boldsymbol{y}_0, \boldsymbol{x}_0)$ pairs. Training data was provided in batches of 32, with each batch containing data with the same kernel hyperparameters but different realisations of prior GP samples. The complete configuration for each experiment is given in Table 3

Experiments were conducted on a 32-core machine and utilised a single Tesla V100-PCIE-32GB GPU. Training of each

model used in the experiments takes no longer than 30 minutes, except for the image regression experiments.

Table 3: Experiment configuration and training time.

Experiment regression	Symthetic data	Hyperparameter marginalisation	Step	High dim BO	1D opt	Image regression
Epochs	250	500	250	250	250	100
Total samples seen	1024k	2048k	1024k	1024k	1024k	-
Batch size	32	32	32	32	32	32
Loss	L1	L1	L1	L1	L1	L1
LR decay	cosine	cosine	cosine	cosine	cosine	cosine
LR init	$2e^{-5}$	0.001	0.001	0.001	0.001	$2e^{-5}$
LR warmup epochs	20	20	20	20	20	20
Num blocks	4	5	5	5	5	5
Representation dim (H)	64	64	64	64	64	64
Num heads	8	8	8	8	8	8
Num timesteps (T)	500	1000	1000	500	2000	500
β schedule	cosine	cosine	cosine	cosine	cosine	cosine
Num points (N)	60/70/80	100	100	256	100	$H \times W$
Deterministic inputs	True	True	True	True	False	True
Training time	17m	33m	16m	21m	16m	10h

Time embedding The diffusion step t is a crucial input of the NN noise estimator as the model needs to be able to differentiate between noise added at the start or the end of the process. Following Vaswani et al. (2017) we use a cyclic 128-dimensional encoding vector for each step

$$t \mapsto [\sin(10^{\frac{0\times4}{63}}t), \sin(10^{\frac{1\times4}{63}}t), \ldots, \sin(10^{\frac{64\times4}{63}}t), \cos(10^{\frac{0\times4}{63}}t), \ldots, \cos(10^{\frac{64\times4}{63}}t)] \in \mathbb{R}^{128}$$

F.2. Illustrative Figure

For the creation of Figs. 1 and 11, we use GPflow (van der Wilk et al., 2020) for the GP regression model using a kernel that matches the training data: a squared exponential with lengthscale set to 0.2. The other baseline, the Attentive Latent NP, is a pre-trained model which was trained on a dataset with the same configuration. The ALNP model originates from the reference implementation of Dubois et al. (2020).

F.3. Hyperparameter Marginalisation

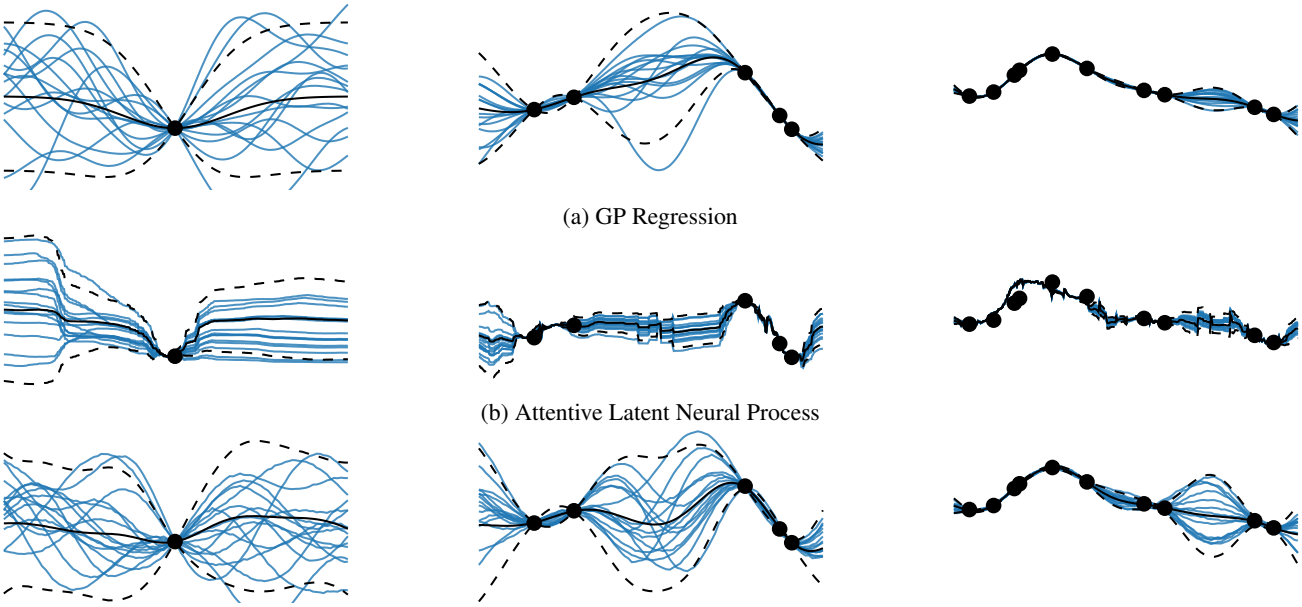
In this experiment, we provide training data to the NDP in the form of unique prior samples from a ground truth GP model. The input for each prior sample, $\boldsymbol{x}_0$, is deterministically spaced across $[-1, 1]$ with $N = 100$. The corresponding output $\boldsymbol{y}_0$ is sampled from a GP prior $\boldsymbol{y}_0|\boldsymbol{x}_0 \sim \mathcal{N}\big(\mathbf{0}, k(\boldsymbol{x}_0, \boldsymbol{x}_0) + \sigma^2\big)$ where the kernel is a stationary Matern-$\frac{3}{2}$ kernel. The noise variance is set to $\sigma^2 = 10^{-6}$ and the kernel variance is fixed to $\sigma_k^2 = 1.0$ throughout. We place a log-normal prior on the lengthscale $\log \mathcal{N}(\log 0.5, \sqrt{0.5})$.

F.4. Non-Gaussian Posteriors

The step function training data is created synthetically by drawing $u \sim \mathcal{U}[-1, 1]$ and letting

$$f(x) = 0 \text{ for } x \leq u, \qquad \text{and} \qquad f(x) = 1 \text{ for } x > u. \tag{20}$$

For completeness, we show the performance of two GP models on this data in Fig. 12. As per the definition, each marginal of a GP is Gaussian which makes modelling this data impossible.



(a) GP Regression

(b) Attentive Latent Neural Process

(c) Neural Diffusion Process (ours)

Figure 11: **1D regression:** The blue curves are posterior samples from different probabilistic models. We also plot the empirical mean and two standard deviations of the samples in black. From left to right we increase the number of data points (black dots) and notice how the process gets closer to the true underlying function.

F.5. Image Regression

In this experiment we train a NDP on two image datasets: MNIST and CelebA. We treat an image as a function with 2D inputs in $[1, H] \times [1, W]$ and 1D grey-scale outputs (MNIST) or 3D RGB output (CelebA). For both datasets we re-scale the 2D inputs to the unit box $[-2, 2] \times [-2, 2]$ and the outputs to normalised.

F.6. Regression on Synthetic Data

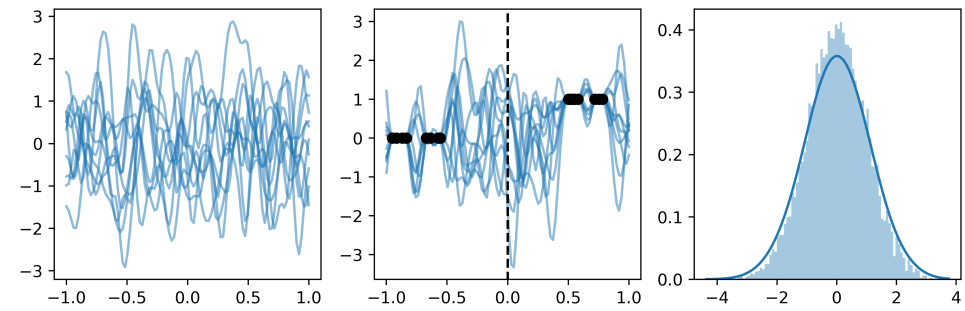
We adopt the same experimental setup as Wang & Neal (2012) to generate synthetic data, which includes Gaussian process (Squared Exponential (SE), MAT'ERN-$\frac{5}{2}$) sample paths. Figure 13 displays samples from each of these datasets, which are corrupted with observation noise having a variance of $\sigma^2 = 0.05^2$. We use a single lengthscale across all dimensions set to $\ell = 0.25\sqrt{D}$.

The training data is composed of 2^{14} sample paths, whereas the test dataset comprises 128 paths. For each test path, we sample the number of context points within a range of 1 and $10D$. We fix the number of target points to 50. The input range for both the training and test datasets, covering both context and target sets, is set to $[-2, 2]$.

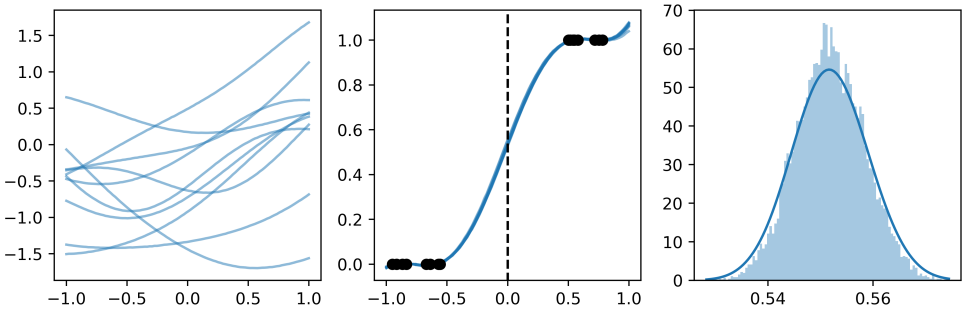
F.7. Bayesian Optimisation

In this experiment, we perform Bayesian optimisation on the Hartmann 3D & 6D, Rastrigin 4D and Ackley 5D objectives. We re-scale, without loss of generality, the inputs of the objectives such that the search space is $[-1, 1]^D$.

The baseline models, GPR and Random, originate from Trieste (Berkeley et al., 2022) —a TensorFlow/GPflow-based Bayesian Optimisation Python package. We benchmark two NDP models: Fixed and Marginalised. The Fixed NDP model is trained on Matérn-$\frac{5}{2}$ samples with a fixed lengthscale set to 0.5 along all dimensions. The Marginalised NDP model is trained on Matérn-$\frac{5}{2}$ samples originating from different lengthscales, drawn from a log-Normal prior.

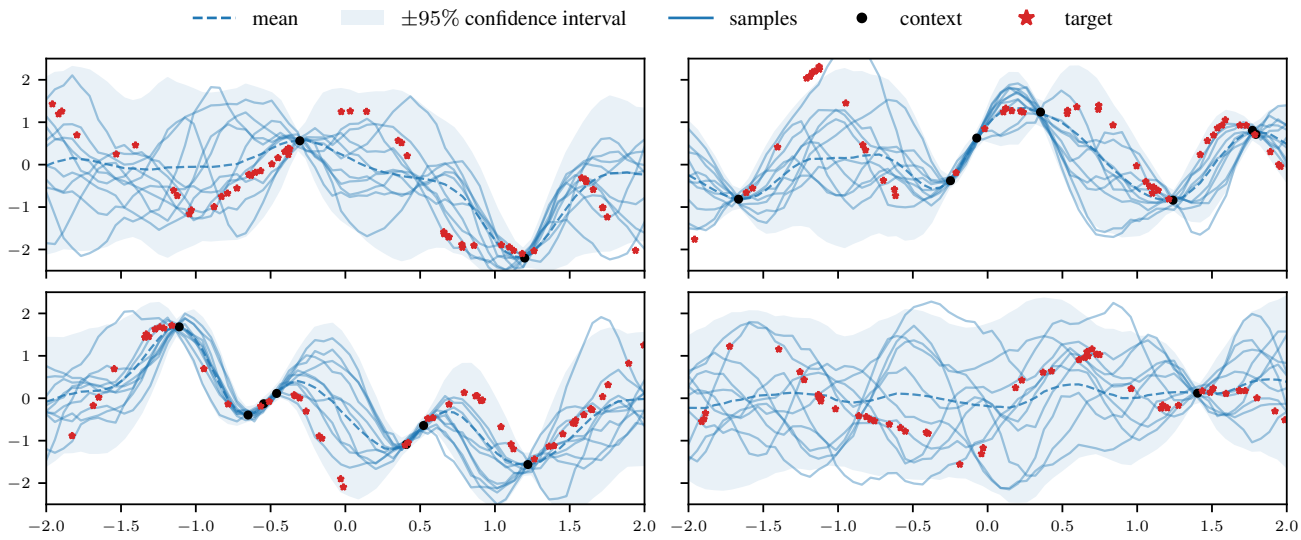


(a) Squared Exponential kernel

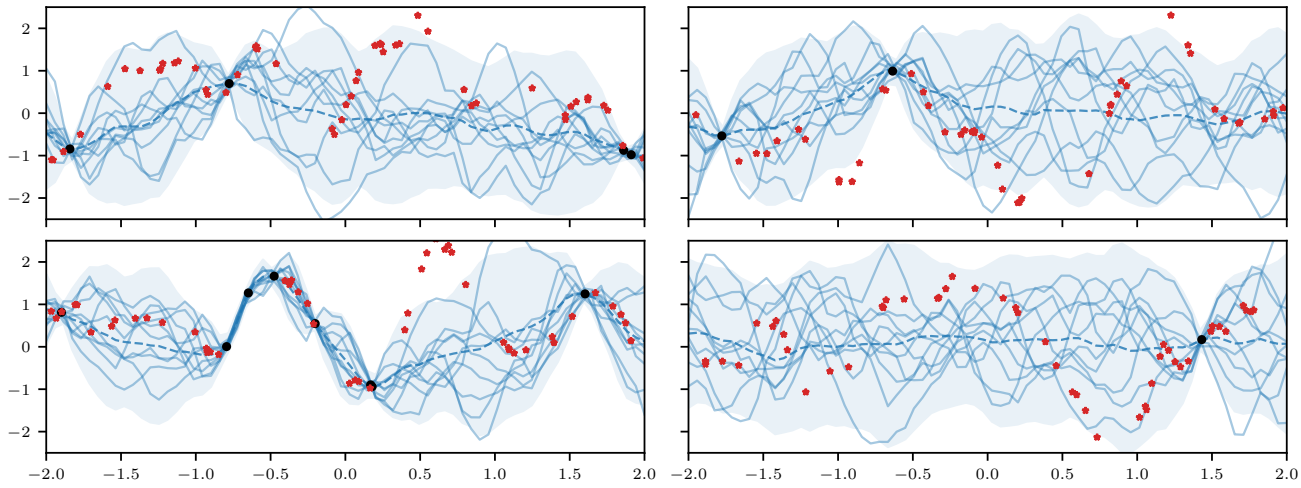


(b) Spectral Mixture kernel

Figure 12: Performance of Gaussian process models on step function data. The left panel shows samples from the prior. The middle panel shows conditional samples where the data is given by the black dots. The right panel shows the marginal at $x = 0$.



(a) Squared Exponential



(b) Matérn–$\frac{5}{2}$

Figure 13: Model predictions on 1D synthetic datasets. Black dots: random number of context points. Red crosses: 50 target points. Blue lines: samples of the model, mean and 95% confidence intervals using the samples.