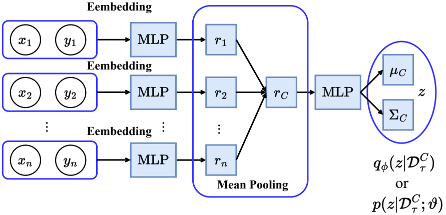
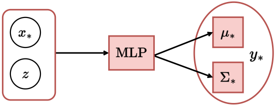
Overview

Introduction

Learning Modules in NPs



(a) Approximate or Learned Functional Priors



$$p(\mathcal{D}_\tau^T|z;\vartheta) = \prod_{i=1}^{n+m} p(y_i|[x_i, z];\vartheta)$$

(b) Generative Distributions

Figure: Encoder-Decoder Structures of NPs.

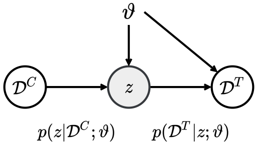


Figure: Deep Latent Variable Models for NPs. The model involves a functional prior distribution $p(z|\mathcal{D}^C;\vartheta)$ and a functional generative distribution $p(\mathcal{D}^T|z;\vartheta)$.

NPs as Exchangeable Stochastic Processes

Notations. Distribution of tasks $\tau \sim p(\mathcal{T})$; Context data points $\mathcal{D}_\tau^C = \{(x_i, y_i)\}_{i=1}^n$; Target data points $\mathcal{D}_\tau^T = \{(x_i, y_i)\}_{i=1}^{n+m}$.

The *generative process* in vanilla NPs (well defined exchangeable $\mathcal{SP}$s based on *de Finneti*'s theorem [1]):

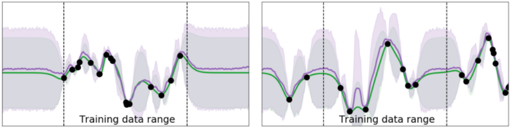
$$\rho_{x_{1:n+m}}(y_{1:n+m}) = \int p(z) \prod_{i=1}^{n+m} \mathcal{N}(y_i; \mu(x_i, z), \Sigma(x_i, z)) dz \tag{1}$$

The *probabilistic inference* of NPs in Meta Learning tasks:

$$\underbrace{p(\mathcal{D}_\mathcal{T}^T | \mathcal{D}_\mathcal{T}^C; \vartheta)}_{\text{Marginal Likelihood}} = \prod_{\tau \in \mathcal{T}} \left[\int \underbrace{p(\mathcal{D}_\tau^T | z; \vartheta)}_{\text{Generative Likelihood}} \underbrace{p(z | \mathcal{D}_\tau^C; \vartheta)}_{\text{Functional Prior}} dz \right] \tag{2}$$

Application of NPs

Stochastic processes can utilize correlations among data points for prediction. As a LVM, the NP [2] approximates $p(y_T | [x_c, y_c], x_T; \vartheta)$ with $\mathbb{E}_{p(z|\mathcal{D}^C;\vartheta)}[p(y_T | x_T, z; \vartheta)]$.



- Data efficient and safe control, *e.g.* fitting $p(\Delta s | s, a; \mathcal{M})$ in robotics systems $\mathcal{M}$ [3].
- Bayesian optimization or active learning to progressively query informative data points [4].
- Posterior sampling to encourage efficient exploration in meta RL [5].

Research Background

Meta Learning Functional Priors

Optimization Objective. In meta learning, the optimization objective of our interest is the marginal log-likelihood in Eq. (3).

$$\max_{\vartheta} \sum_{\tau \in \mathcal{T}} \ln \left[\int p(\mathcal{D}_{\tau}^{T} | z; \vartheta) \underbrace{p(z | \mathcal{D}_{\tau}^{C}; \vartheta)}_{\text{Functional Prior}} dz \right] \tag{3}$$

For simple notations, we consider one task τ to derive equations[1].

$$\mathcal{L}(\vartheta) = \ln \left[\int p(\mathcal{D}_{\tau}^{T} | z; \vartheta) p(z | \mathcal{D}_{\tau}^{C}; \vartheta) dz \right] \tag{4}$$

Further, the conditional independence $p(\mathcal{D}_{\tau}^{T} | z; \vartheta) = \prod_{i=1}^{n+m} p(y_i | [x_i, z]; \vartheta)$ is satisfied in the NP family [2, 6].

[1] Meta training and testing phases are implemented in a batch of tasks consistent with Eq. (2)/(3).

Advances in the NP Family

Performance Improvement via Inductive Biases. The vanilla NP suffers performance bottleneck [7]. Most of existing work study functional representations $q_\phi(z|\mathcal{D}^C)$ with NPs through a lens of structural inductive biases:

Table 4: Summary of Typical Neural Process Related Models (Meta-Testing Scenarios). The recognition model and the generative model respectively correspond to the encoder and the decoder in the family of neural processes. Here $[x_C, y_C]$ are context data points and we consider (x_*, y_*) a data point in target dataset.

Models	Recognition Model	Generative Model	Inductive Bias
CNP (Garnelo et al., 2018a)	$z = f_\phi(x_C, y_C)$	$p_\theta(y_*\|[x_*, z])$	conditional functional
NP (Garnelo et al., 2018b)	$q_\phi(z\|[x_C, y_C])$	$p_\theta(y_*\|[x_*, z])$	global functional
ANP (Kim et al., 2019; 2021)	$q_{\phi_1}(z\|[x_C, y_C])$ $f_{\phi_2}(z_*\|[x_C, y_C], x_*)$	$p_\theta(y_*\|[x_*, z, z_*])$	global functional local embedding
FCRL (Gondal et al., 2021)	$f_\phi(z\|[x_C, y_C])$	$p_\theta(y_*\|[x_*, z])$	contrastive functional
ConvNP (Foong et al., 2020)	$p_\phi(z\|[x_C, y_C])$	$p_\theta(y_*\|[x_*, z])$	convolutional functional
Conv-CNP (Gordon et al., 2019)	$f_\phi(z_*\|[x_C, y_C], x_*)$	$p_\theta(y_*\|[x_*, z_*])$	convolutional functional
FNP (Louizos et al., 2019)	$f_\phi(z_*\|[x_C, y_C], x_*)$	$p_\theta(y_*\|z_*)$	latent DAG

Existing Research Issues in this Domain

Apart from improving expressiveness of functional priors via structural inductive biases, we investigate the following topics to advance the research:

- **Source of Inference Suboptimality in NPs.**
  What about diagnosing the *inference suboptimality* through a lens of *optimization objectives*?
  [**Hints:** revisit the formulation of NPs from a scratch]

- **Randomness in Functional Representations (Posterior Predictive Dist. vs Prior Predictive Dist.).**
  What is the relationship between *uncertainty in priors*, *the extent of observability* and *complexity of function families*?
  [**Hints:** analyze the meta learned statistics]

Optimization Gaps & Statistical Traits

Given $p(z|\mathcal{D}_\tau^C;\vartheta)$ and $p(\mathcal{D}_\tau^T|z;\vartheta)$, the exact functional posterior can be obtained by the Bayes rule:

$$p(z|\mathcal{D}_\tau^T;\vartheta) = \frac{p(\mathcal{D}_\tau^T|z;\vartheta)p(z|\mathcal{D}_\tau^C;\vartheta)}{\int p(\mathcal{D}_\tau^T|z;\vartheta)p(z|\mathcal{D}_\tau^C;\vartheta)dz}.$$

Exact ELBO for NPs. Following the VI operation, we can connect the exact ELBO with the log-likelihood in Eq. (5).

$$\mathcal{L}(\vartheta) = \underbrace{\mathbb{E}_{q_\phi(z)}\left[\ln \frac{p(\mathcal{D}_\tau^T, z|\mathcal{D}_\tau^C;\vartheta)}{q_\phi(z)}\right]}_{\text{Exact ELBO}} + \underbrace{D_{KL}\left[q_\phi(z) \parallel p(z|\mathcal{D}_\tau^T;\vartheta)\right]}_{\text{Posterior Approximation Gap}} \tag{5}$$

$$\mathcal{L}_{\text{ELBO}}(\vartheta,\phi) = \mathbb{E}_{q_\phi(z)}\left[\ln p(\mathcal{D}_\tau^T|z;\vartheta)\right] - D_{KL}\left[q_\phi(z) \parallel p(z|\mathcal{D}_\tau^C;\vartheta)\right] \tag{6}$$

Approximate ELBO in vanilla NPs

Approximate ELBO for NPs: simply replacing the intractable prior $p(z|\mathcal{D}_\tau^C;\vartheta)$ with the approximate prior $q_\phi(z|\mathcal{D}_\tau^C)$ in $\mathcal{L}_{\text{ELBO}}(\vartheta,\phi)$.

$$\mathcal{L}_{\text{NP}}(\vartheta,\phi) = \mathbb{E}_{q_\phi(z)}\Bigg[\ln \underbrace{p(\mathcal{D}_\tau^T|z;\vartheta)}_{\text{Generative Likelihood}}\Bigg] - \underbrace{D_{KL}\left[q_\phi(z)\parallel q_\phi(z|\mathcal{D}_\tau^C)\right]}_{\text{Consistent Regularizer}} \tag{7}$$

Remark (1)

Optimizing Eq. (7) cannot guarantee finding optimal or locally optimal solutions for the maximization over $\sum_{\tau\in\mathcal{T}} \ln p(\mathcal{D}_\tau^T|\mathcal{D}_\tau^C;\vartheta)$.

There is not strict sign for $\mathcal{L}(\vartheta)$ and $\mathcal{L}_{\text{NP}}(\vartheta,\phi)$ to see which is greater:

$$\mathcal{L}(\vartheta) \geq \mathcal{L}_{\text{ELBO}}(\vartheta,\phi), \quad \mathcal{L}(\vartheta) \ngeq \mathcal{L}_{\text{NP}}(\vartheta,\phi) \quad \forall\vartheta\in\Theta \text{ and } \phi\in\Phi \tag{8}$$

Other Available Optimization Objectives

Now we turn to other tractable optimization objectives in NPs.

Conditional Neural Processes (CNPs) [6]: deterministic functional embedding.

$$\mathcal{L}_{\text{CNP}}(\vartheta) = \mathbb{E}_{p(z|\mathcal{D}_\tau^C;\vartheta)}\left[\ln p(\mathcal{D}_\tau^T|z;\vartheta)\right] \quad \text{with} \quad p(z|\mathcal{D}_\tau^C;\vartheta) = \delta(|z-\hat{z}|) \tag{9}$$

Monte Carlo Maximum Likelihood Neural Processes (ML-NPs) [8][2]: direct optimize the functional prior via MC estimates.

$$\mathcal{L}_{\text{ML-NP}}(\vartheta) = \ln\left[\frac{1}{B}\sum_{b=1}^{B}\exp\left(\ln p(\mathcal{D}_\tau^T|z^{(b)};\vartheta)\right)\right] \quad \text{with} \quad z^{(b)} \sim p(z|\mathcal{D}_\tau^C;\vartheta) \tag{10}$$

[2] We remove the convolutional modules from ConvNPs since the optimization objective is our focus.

Multi-sample Prediction & Functional Prior Collapse

Multi-sample Prediction: NP models need to run B stochastic forward pass by sampling $z^{(b)} \sim p(z|\mathcal{D}_\tau^C;\vartheta)$ and then compute the log-likelihoods as $\ln\left[\frac{1}{B}\sum_{b=1}^{B} p(\mathcal{D}_\tau^T|z^{(b)};\vartheta)\right]$.

Asymptotic Behavior: Given an average predictive error measure β, the number of context points n and $\mathcal{D}_\tau^T$, $\beta(\mathcal{D}_\tau^T;n)\downarrow$ are decreased when increasing $n\uparrow$ in prediction.

Definition (Prior Collapse)

The functional prior $p(z|\mathcal{D}_\tau^C;\vartheta) = \mathcal{N}(z;\mu_\vartheta(\mathcal{D}_\tau^C),\Sigma_\vartheta(\mathcal{D}_\tau^C))$ is said to collapse when $\text{Tr}[\Sigma_\vartheta(\mathcal{D}_\tau^C)] = \sum_{i=1}^{d}\sigma_i^2 \approx 0$ with $\Sigma_\vartheta(\mathcal{D}_\tau^C) = \text{diag}[\sigma_1^2,\dots,\sigma_d^2]$.

Tractable Optimization via Expectation Maximization

Variational EM-Steps

The basic idea is illustrated in Figure. In detail, we iteratively construct the lower bound $\mathcal{L}(\vartheta_K)$ and maximize the surrogate function $\mathcal{L}(\vartheta; \vartheta_K)$.

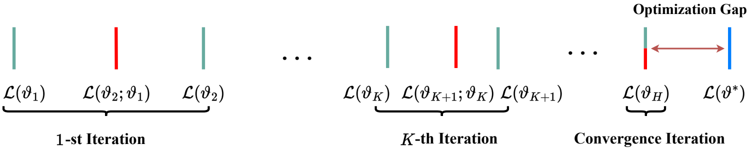


Figure: Expectation Maximization for NPs. Green lines indicate the E-steps while the red lines are M-steps. After convergence, the gap $\mathcal{L}(\vartheta_H) - \mathcal{L}(\vartheta_{H-1})$ is close to zero and the algorithm results in at least a local optimum.

Pseudo Algorithm

Algorithm 1: Variational Expectation Maximization for NPs.

Input : Task distribution $p(\mathcal{T})$; Task batch size, Number of particles, Initialized ϑ and η.

Output: Meta-trained parameters ϑ and η.

```
for k = 1 to K do
    E-step #1: k ← k + 1 and reset the variational posterior
        q_ϕ(z) = p(z|D_τ^T; ϑ_k) in Eq. (5);
    if Use the Functional Prior as the Proposal then
        Reset q_η(z|D_τ^T) = p(z|D_τ^C; ϑ_k);
    else
        E-step #2: update the proposal η_k = arg min_η L_KL(η; η_{k−1}, ϑ_k) in Eq.
            (29) according to operations in Appendix (E.3.1);
    end
    M-step: optimize surrogate functions ϑ_{k+1} = arg max_ϑ L_SI-NP(ϑ; η_k, ϑ_k) in
        Eq. (12);
end
```

Surrogate Function for Exact NPs

Valid Surrogate Function. Here ϑ_k denotes the parameter of NPs in the k-th iteration. In Eq. (6), we replace the approximate posterior with the last time updated $p(z|\mathcal{D}_\tau^T;\vartheta_k)$ in Algorithm (1).

$$\mathcal{L}(\vartheta;\vartheta_k) = \mathbb{E}_{p(z|\mathcal{D}_\tau^T;\vartheta_k)}\left[\ln p(\mathcal{D}_\tau^T, z|\mathcal{D}_\tau^C;\vartheta) - \ln \underbrace{p(z|\mathcal{D}_\tau^T;\vartheta_k)}_{\text{fixed functional posterior}}\right] \tag{11}$$

Proposition (1)

The proposed meta learning function $\mathcal{L}(\vartheta;\vartheta_k)$ in Eq. (11) is a surrogate function w.r.t. the log-likelihood of the meta learning dataset.

$$\max_{\vartheta}\mathcal{L}(\vartheta;\vartheta_k) \Leftrightarrow \max_{\vartheta}\mathcal{L}_{\text{EM}}(\vartheta;\vartheta_k) = \mathbb{E}_{p(z|\mathcal{D}_\tau^T;\vartheta_k)}\left[\ln p(\mathcal{D}_\tau^T, z|\mathcal{D}_\tau^C;\vartheta)\right] \tag{12}$$

SI-NPs. We sample l.v.s from a proposal distribution $z^{(b)} \sim q_{\eta_k}(z|\mathcal{D}_\tau^T)$ and optimize the objective via Self-normalized Importance sampling [9].

$$\mathcal{L}_{\text{EM}}(\vartheta;\vartheta_k) = \mathbb{E}_{q_\eta}\left[\frac{p(z|\mathcal{D}_\tau^T;\vartheta_k)}{q_\eta(z|\mathcal{D}_\tau^T)} \ln p(\mathcal{D}_\tau^T, z|\mathcal{D}_\tau^C;\vartheta)\right] \approx \sum_{b=1}^{B} \hat{\omega}^{(b)} \ln p(\mathcal{D}_\tau^T, z^{(b)}|\mathcal{D}_\tau^C;\vartheta)$$

$$= \sum_{b=1}^{B} \underbrace{\hat{\omega}^{(b)}}_{\text{Importance Weight}} \left[\ln \underbrace{p(\mathcal{D}_\tau^T|z^{(b)};\vartheta)}_{\text{Generative Likelihood}} + \ln \underbrace{p(z^{(b)}|\mathcal{D}_\tau^C;\vartheta)}_{\text{Functional Prior Likelihood}}\right] = \mathcal{L}_{\text{SI-NP}}(\vartheta;\eta_k,\vartheta_k) \tag{13}$$

where $\omega^{(b)} = p(\mathcal{D}_\tau^T|z^{(b)};\vartheta_k)p(z^{(b)}|\mathcal{D}_\tau^C;\vartheta_k)/q_{\eta_k}(z^{(b)}|\mathcal{D}_\tau^T)$ and $\hat{\omega}^{(b)} = \frac{\omega^{(b)}}{\sum_{b'=1}^{B}\omega^{(b')}}$.

Important Note: We set $q_{\eta_k}(z^{(b)}|\mathcal{D}_\tau^T) = p(z^{(b)}|\mathcal{D}_\tau^C;\vartheta_k)$ as the default.

Connections between Optimization Objectives

Proposition (2)

With one Monte Carlo sample used in Eq. (13), the presumed diagonal Gaussian prior $p(z|\mathcal{D}_\tau^C;\vartheta)$ will collapse. Hence, SI-NP in Eq. (14) is equivalent with CNP in Eq. (9).

$$\mathcal{L}_{SI\text{-}NP}(\vartheta;\eta_k,\vartheta_k) \approx \mathbb{E}_{p(z|\mathcal{D}_\tau^C;\vartheta_k)}\left[\ln \underbrace{p(\mathcal{D}_\tau^T|z;\vartheta)}_{\textit{Generative Likelihood}}\right] + \underbrace{\mathbb{E}_{p(z|\mathcal{D}_\tau^C;\vartheta_k)}\left[\ln p(z|\mathcal{D}_\tau^C;\vartheta)\right]}_{\textit{Prior Collapse Term}} \tag{14}$$

Hints: Perform the limit analysis *w.r.t.* the equation below.

$$\begin{aligned}\mathcal{L}_{\text{SI-NP}} &= \mathbb{E}_{p(z|\mathcal{D}_\tau^C;\vartheta)}\left[\ln p(\mathcal{D}_\tau^T|z;\vartheta)\right] + \mathbb{E}_{p(z|\mathcal{D}_\tau^C;\vartheta_k)}\left[\ln p(z|\mathcal{D}_\tau^C;\vartheta)\right] \\ &\approx \sum_{i=1}^{n+m} \ln p(y_i|[x_i,\mu_\vartheta+\hat{\epsilon}\Sigma_\vartheta^{\frac{1}{2}};\vartheta]) - \left(\frac{1}{2}\ln(2\pi) + \sum_{i=1}^{d}\left[\ln\sigma_i + \frac{(\mu_i-\hat{z}_i)^2}{2\sigma_i^2}\right]\right)\end{aligned} \tag{15}$$

Connections between Optimization Objectives

Connections. A table is given to explain the relationship between these mentioned objectives. Remember that the loglikelihood of meta datast is $\mathcal{L}(\vartheta) = \ln\left[\int p(\mathcal{D}_\tau^T|z;\vartheta)p(z|\mathcal{D}_\tau^C;\vartheta)dz\right]$.

Table 3: A Summary of Optimization Objectives in NPs Family. We list the available optimization objectives in Section (3)/(4). For Importance Weighted Estimates, multiple Monte Carlo samples are required in meta training.

Optimization Objective	Connection with $\mathcal{L}(\vartheta)$ in Eq. (4)	Importance Weighted Estimates
$\mathcal{L}_{\text{NP}}(\vartheta,\phi)$	Approximate ELBO	✗
$\mathcal{L}_{\text{CNP}}(\vartheta)$	Biased Estimate	✗
$\mathcal{L}_{\text{ML-NP}}(\vartheta)$	Biased Estimate	✓
$\mathcal{L}_{\text{SI-NP}}$ (Ours)	Biased Estimate	✓

Experiments

Research Questions & Learning Tasks

Research Questions. (i) Can variational EM based models SI-NPs achieve a better local optimum than vanilla NPs? (ii) What is the role of randomness in functional priors?

Baselines & Evaluations. Since our concentration is on optimization objectives in NPs family, we compare to NPs [2], and CNPs [6], ML-NPs [8] in experiments.

Benchmarks. We include curve fitting, image completion and other tasks (Sim2Real tasks in Lotka-Volterra/Predator-Prey systems).

Combination with Structural Inductive Biases. We take the attention augmentation the same as that in [7] as an example to examine the performance.

Gaussian Processes. Kernels, respectively Matern$-\frac{5}{2}$, RBF, and Periodic, are used to generate diverse function distributions.

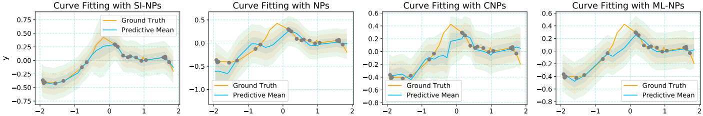


Figure: Examples of Curve Fitting in RBF Kernel Cases. The plots report predictive mean functions with ±3 standard deviations.

Table 1: Test average log-likelihoods of target data points for 1-dimensional Gaussian process dataset with various kernels (reported standard deviations in 5 runs). For each run, we randomly sample 1000 functions as tasks to evaluate.

#	Matern $-\frac{5}{2}$	RBF	Periodic
$\mathcal{L}_{\text{GP}}$ (Oracle)	0.821$\pm$0.03	1.18$\pm$0.013	0.833$\pm$0.017
$\mathcal{L}_{\text{NP}}$ (Garnelo et al., 2018b)	-0.225$\pm$0.03	-0.183$\pm$0.03	-0.611$\pm$0.034
$\mathcal{L}_{\text{CNP}}$ (Garnelo et al., 2018a)	0.295$\pm$0.017	0.463$\pm$0.023	-0.533$\pm$0.009
$\mathcal{L}_{\text{ML-NP}}$ (Foong et al., 2020)	0.303$\pm$0.013	0.439$\pm$0.009	-0.547$\pm$0.036
$\mathcal{L}_{\text{SI-NP}}$ (ours)	0.305$\pm$0.006	0.493$\pm$0.007	-0.532$\pm$0.036

CIFAR10/SVHN/MNIST/FMNIST dataset. A random number of pixels are masked to complete. Given the context pixel locations and values $[x_C, y_C]$, we need to learn a map from each 2-D pixel location $x \in [0,1]^2$ to pixel values $y \in \mathbb{R}$ or $y \in \mathbb{R}^3$.

Image completion results. We report average log-likelihoods with varying numbers of context points in random cases.

Table 2: Test average log-likelihoods with reported standard deviations for image completion in MNIST/FMNIST/SVHN/CIFAR10 (5 runs). We test the performance of different optimization objectives in both context data points and target data points. Except CNPs, we use 32 Monte Carlo samples from the functional prior to evaluate the average log-likelihoods.

#	MNIST context	MNIST target	FMNIST context	FMNIST target	SVHN context	SVHN target	CIFAR10 context	CIFAR10 target
$\mathcal{L}_{\text{NP}}$	$0.81_{\pm 0.006}$	$0.73_{\pm 0.007}$	$0.83_{\pm 0.007}$	$0.73_{\pm 0.009}$	$3.19_{\pm 0.02}$	$3.07_{\pm 0.02}$	$2.35_{\pm 0.04}$	$2.03_{\pm 0.02}$
$\mathcal{L}_{\text{CNP}}$	$1.05_{\pm 0.005}$	$0.99_{\pm 0.008}$	$0.95_{\pm 0.007}$	$0.90_{\pm 0.009}$	$\mathbf{3.57}_{\pm 0.003}$	$3.48_{\pm 0.004}$	$2.71_{\pm 0.004}$	$2.53_{\pm 0.006}$
$\mathcal{L}_{\text{ML-NP}}$	$1.06_{\pm 0.004}$	$0.99_{\pm 0.006}$	$0.94_{\pm 0.008}$	$0.89_{\pm 0.007}$	$3.51_{\pm 0.008}$	$3.43_{\pm 0.006}$	$2.60_{\pm 0.005}$	$2.41_{\pm 0.005}$
$\mathcal{L}_{\text{SI-NP}}$ (ours)	$\mathbf{1.09}_{\pm 0.006}$	$\mathbf{1.02}_{\pm 0.004}$	$\mathbf{0.98}_{\pm 0.004}$	$\mathbf{0.94}_{\pm 0.005}$	$\mathbf{3.57}_{\pm 0.003}$	$\mathbf{3.50}_{\pm 0.003}$	$\mathbf{2.75}_{\pm 0.004}$	$\mathbf{2.60}_{\pm 0.005}$

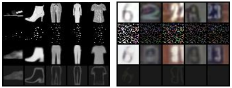


Figure: Completed Images with SI-NPs.

Observations. SI-NP achieves the best performance in all image datasets. The performance gaps between SI-NPs and NPs are remarkable. All baselines exhibit the asymptotic behaviors in Fig. (8).

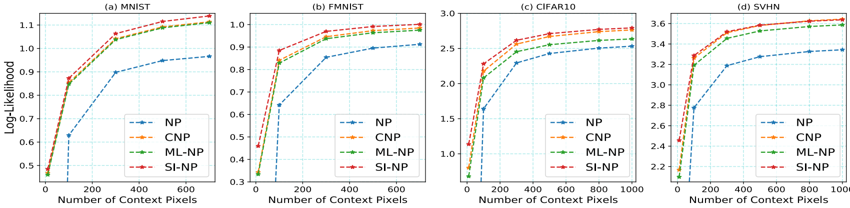


Figure: Asymptotic Performance in Image Completion. For MNIST/FMNIST datasets, the numbers of context pixels in testing are $\{10, 100, 300, 500, 700\}$. For CIFAR10/SVHN datasets, the numbers are $\{10, 100, 300, 500, 800, 1000\}$.

Observations. The scale of SI-NPs' trace values coincides with the semantics complexity of datasets: CIFAR10>SVHN>FMNIST>MNIST.

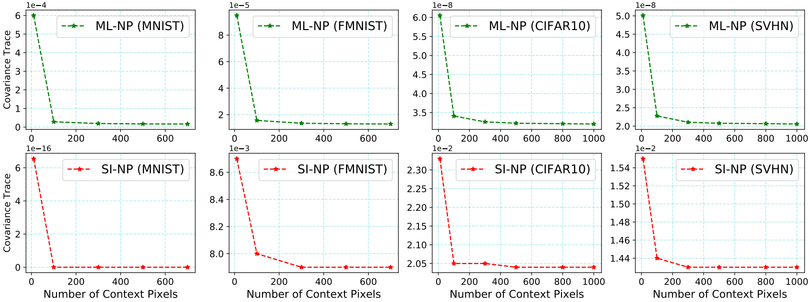


Figure: Statistics of Learned Functional Priors in ML-NPs/SI-NPs. The trace of learned functional priors' covariance matrices $\mathrm{Tr}[\Sigma_\vartheta(\mathcal{D}_\tau^C)]$ is computed based on $p(z|\mathcal{D}_\tau^C;\vartheta) = \mathcal{N}(z;\mu_\vartheta(\mathcal{D}_\tau^C),\Sigma_\vartheta(\mathcal{D}_\tau^C))$.

Additional Experiments

Augmenting SI-NPs with Structural Inductive Biases. We take the addition of attention networks [7] as an example to show the performance.

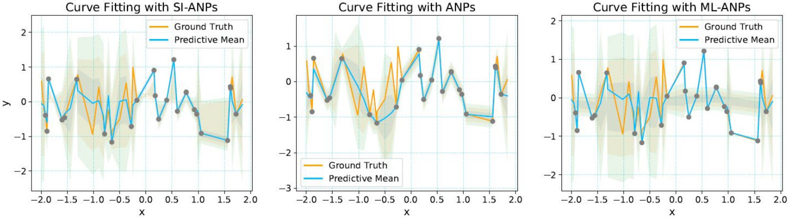


Table 8: Test average log-likelihoods of target data points with reported standard deviations for image completion in MNIST/FMNIST/SVHN/CIFAR10 (4 runs). All NP models are augmented with attention networks. Same as in the main paper, we test the performance of different optimization objectives in target data points. We use 32 Monte Carlo samples from the functional prior to evaluate the average log-likelihoods.

#	MNIST	FMNIST	SVHN	CIFAR10
$\mathcal{L}_{\text{ANP}}$ (Kim et al., 2019)	$1.173_{\pm 0.008}$	$1.101_{\pm 0.01}$	$4.011_{\pm 0.005}$	$3.605_{\pm 0.016}$
$\mathcal{L}_{\text{ML-ANP}}$ (Foong et al., 2020)	$1.216_{\pm 0.003}$	$1.172_{\pm 0.009}$	$4.017_{\pm 0.002}$	$3.545_{\pm 0.01}$
$\mathcal{L}_{\text{SI-ANP}}$ (ours)	$1.212_{\pm 0.004}$	$\mathbf{1.174}_{\pm 0.005}$	$\mathbf{4.040}_{\pm 0.002}$	$\mathbf{3.710}_{\pm 0.028}$

Conclusion & Outlook

Summary of MoE-NPs

Primary empirical findings. (1) Training NPs with EM algorithms results in better (local) optimum; (2) Randomness of SI-NPs' functional priors relates with complexity of function families.

Existing limitations in SI-NPs. (1) More inference particles required in meta training; (2) Additional structural inductive biases required for better performance.

Future work. (1) Functional uncertainty decomposition (global l.v. vs learned output variance param.) (2) Reasonable optimization objectives for independent proposal distributions.

[1] G Jay Kerns and Gábor J Székely. Definetti's theorem for abstract finite exchangeable sequences. *Journal of Theoretical Probability*, 19(3):589–608, 2006.

[2] Marta Garnelo, Jonathan Schwarz, Dan Rosenbaum, Fabio Viola, Danilo J Rezende, SM Eslami, and Yee Whye Teh. Neural processes. *arXiv preprint arXiv:1807.01622*, 2018.

[3] Alexandre Galashov, Jonathan Schwarz, Hyunjik Kim, Marta Garnelo, David Saxton, Pushmeet Kohli, SM Eslami, and Yee Whye Teh. Meta-learning surrogate models for sequential decision making. *arXiv preprint arXiv:1903.11907*, 2019.

[4] ChangYong Oh, Efstratios Gavves, and Max Welling. Bock: Bayesian optimization with cylindrical kernels. In *International Conference on Machine Learning*, pages 3868–3877. PMLR, 2018.

[5] Kate Rakelly, Aurick Zhou, Chelsea Finn, Sergey Levine, and Deirdre Quillen. Efficient off-policy meta-reinforcement learning via probabilistic context variables. In *International conference on machine learning*, pages 5331–5340. PMLR, 2019.

[6] Marta Garnelo, Dan Rosenbaum, Christopher Maddison, Tiago Ramalho, David Saxton, Murray Shanahan, Yee Whye Teh, Danilo

Rezende, and SM Ali Eslami. Conditional neural processes. In *International Conference on Machine Learning*, pages 1704–1713. PMLR, 2018.

[7] Hyunjik Kim, Andriy Mnih, Jonathan Schwarz, Marta Garnelo, Ali Eslami, Dan Rosenbaum, Oriol Vinyals, and Yee Whye Teh. Attentive neural processes. In *International Conference on Learning Representations*, 2019.

[8] Andrew Foong, Wessel Bruinsma, Jonathan Gordon, Yann Dubois, James Requeima, and Richard Turner. Meta-learning stationary stochastic process prediction with convolutional neural processes. *Advances in Neural Information Processing Systems*, 33:8284–8295, 2020.

[9] Surya T Tokdar and Robert E Kass. Importance sampling: a review. *Wiley Interdisciplinary Reviews: Computational Statistics*, 2(1):54–60, 2010.