Exploiting Inferential Structure in Neural Processes

Dharmesh Tailor[1] **Mohammad Emtiyaz Khan**[2] **Eric Nalisnick**[1]

[1]University of Amsterdam, Amsterdam, Netherlands
[2]RIKEN Center for AI Project, Tokyo, Japan

Abstract

Neural Processes (NPs) are appealing due to their ability to perform fast adaptation based on a context set. This set is encoded by a latent variable, which is often assumed to follow a simple distribution. However, in real-word settings, the context set may be drawn from richer distributions having multiple modes, heavy tails, *etc.* In this work, we provide a framework that allows NPs' latent variable to be given a rich prior defined by a graphical model. These distributional assumptions directly translate into an appropriate aggregation strategy for the context set. Moreover, we describe a message-passing procedure that still allows for end-to-end optimization with stochastic gradients. We demonstrate the generality of our framework by using mixture and Student-*t* assumptions that yield improvements in function modelling and test-time robustness.

1 INTRODUCTION

Many real-world tasks require models to make predictions in new scenarios on short notice. For example, climate models are often asked to make predictions at novel locations [Vaughan et al., 2022]. Data is collected in well-populated regions but predictions for remote regions (*e.g.* mountain ranges, forests *etc.*) are desirable as well. Neural processes (NPs) Garnelo et al. [2018b] are models designed for situations such as this. At test time, the model is seeded with a context data set from the target setting that (hopefully) allows the NP to make accurate predictions despite the possibly novel conditions. This behavior is implemented using an efficient encoder architecture that scales linearly with respect to the size of the context set and is permutation-invariant to its order (see Fig. 1a). Unfortunately, NPs still have shortcomings that make them brittle for this ambitious use case. For example, NPs commonly underfit [Kim et al., 2019] and suffer from limited representation power [Wagstaff et al., 2019]. Previous work has attempted to fix these problems by enriching the encoder's architecture, *e.g.* attention [Kim et al., 2019], transformers [Nguyen and Grover, 2022], and convolutions [Gordon et al., 2020, Foong et al., 2020].

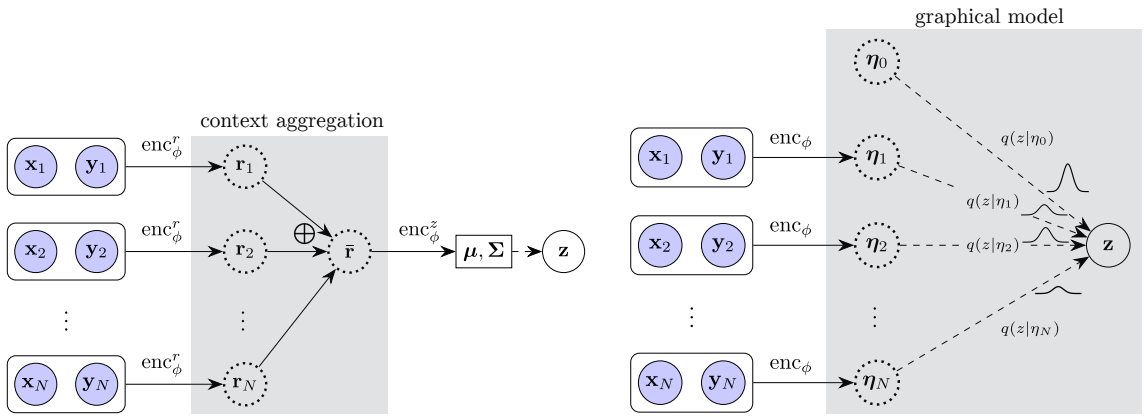
We consider an alternative approach that incorporates the structure and assumptions of the data into NPs' latent variable. To accomplish this, we consider priors defined by a probabilistic graphical model (PGM). While incorporating rich PGMs may seem like it would hinder the scalability that makes NPs an attractive model, we show it does not. Using *structured inference networks* [Lin et al., 2018] (see Fig. 1b), we can still train NPs end-to-end, using variational message passing for the PGM [Winn et al., 2005] and stochastic gradients for the neural networks. Within our framework, encoding the context set becomes analogous to inference in the PGM. This means that the aggregation operation over the context set is completely and automatically determined by the choice of PGM prior. We show that—under a simple Gaussian PGM—our approach recovers *Bayesian Aggregation* (BA) [Volpp et al., 2020].

In this paper, we describe a general framework for placing PGM priors on NPs with latent variables. We primarily focus on cases that exhibit conditional conjugacy but provide some discussion of the fully non-conjugate case as well. We show the explicit updates for mixture priors and Student-*t* assumptions. In the experiments, we show NPs with these mixture and heavy-tail assumptions—which we term *mixture* and *robust* Bayesian Aggregation, respectively—demonstrate improved performance in regression and image completion.

2 BACKGROUND

Problem setup NPs assume a partition of the data into a *context* set and a *target* set. The former is used by the model to seed adaptation. The latter is a set of points from

Accepted for the 39th Conference on Uncertainty in Artificial Intelligence (UAI 2023).



(a) Sum-Decomposition Network (b) Structured Inference Network

Figure 1: A set of context points $(\mathbf{x}_i, \mathbf{y}_i)$ can be aggregated using a sum-decomposition architecture, shown on the left. A shared network enc^r_ϕ is used for all followed by a pre-specified aggregation operation that pools all embeddings. The output is passed to a further network enc^z_ϕ. We propose an alternative, shown on the right, that does not require an additional network and rather uses inference on a probabilistic graphical model (PGM) to automatically aggregate the embeddings. Different types of exponential-family distributions and mixtures, denoted here by $q(z|\eta_i)$, can be used resulting in different aggregation mechanisms.

the same domain as the context set and for which we will make predictions. Specifically, we denote the context set for the lth task as $\mathcal{D}_c^{(l)} = \{\mathbf{x}_{c,i}^{(l)}, \mathbf{y}_{c,i}^{(l)}\}_{i=1}^{N_{l,c}}$, where $\mathbf{x}$ denotes a feature vector and $\mathbf{y}$ the corresponding response vector. The target set for the lth task is denoted similarly as $\mathcal{D}_t^{(l)} = \{\mathbf{x}_{t,i}^{(l)}, \mathbf{y}_{t,i}^{(l)}\}_{i=1}^{N_{l,t}}$. At test time, for a new task l_*, we observe $\mathcal{D}_c^{(l_*)}$ and $\{\mathbf{x}_{t,i}^{(l_*)}\}_{i=1}^{N_{l_*,t}}$. The target responses $\{\mathbf{y}_{t,i}^{(l_*)}\}_{i=1}^{N_{l_*,t}}$ are unobserved, and our goal is to predict them.

Neural Processes *Neural processes* [Garnelo et al., 2018b] frame few-shot learning as a multi-task learning problem [Heskes, 2000], employing a conditional latent variable model with context/target splits on task-specific datasets as shown in Fig. 2. Training amounts to maximising the following *conditional* marginal likelihood across L tasks:

$$\begin{aligned}&\sum_{l=1}^{L} \log p_\theta(\mathcal{D}_t^{(l)} \mid \mathcal{D}_c^{(l)}) \\ &\gtrapprox \sum_{l=1}^{L} \mathbb{E}_{q_\phi(\mathbf{z}|\mathcal{D}_c^{(l)} \cup \mathcal{D}_t^{(l)})} \left[\log p_\theta(\mathcal{D}_t^{(l)} \mid \mathbf{z})\right] \\ &\quad - \mathrm{D}_{\mathrm{KL}} \left[q_\phi(\mathbf{z}|\mathcal{D}_c^{(l)} \cup \mathcal{D}_t^{(l)}) \,\|\, q_\phi(\mathbf{z}|\mathcal{D}_c^{(l)})\right]\end{aligned} \tag{1}$$

where $\mathbf{z}_l$ is the task-specific latent variable and θ is the global parameter that is shared across tasks. The marginalization over task-specific latent variables is typically intractable hence approximate inference is used,

$$\begin{aligned}p_\theta(\mathbf{z} \mid \mathcal{D}_c) &= \tfrac{1}{p_\theta(\mathcal{D}_c)} \textstyle\prod_{i=1}^{N_c} p_\theta(\mathbf{y}_{c,i} \mid \mathbf{x}_{c,i}, \mathbf{z})\, p_\theta(\mathbf{z}) \\ &\approx q_\phi(\mathbf{z} \mid \mathcal{D}_c).\end{aligned} \tag{2}$$

where we have dropped task indices for notational simplicity. The variational approximation is amortised, meaning a recognition network is used. For a Gaussian approximation, the mean and variance are parameterized by neural networks (NNs) that take as input sets of datapoints: $q_\phi(\mathbf{z} \mid \mathcal{D}_c) = \mathcal{N}(\mathbf{z} \mid \widetilde{\boldsymbol{\mu}}, \widetilde{\boldsymbol{\Sigma}})$ with $\left(\widetilde{\boldsymbol{\mu}}, \widetilde{\boldsymbol{\Sigma}}\right) = \mathrm{enc}_\phi(\mathcal{D}_c)$. Throughout we assume covariance matrices have diagonal structure, resulting in fully-factorized Gaussian distributions.

Neural Process Extensions NPs come in two variants, the aforementioned (latent) NP [Garnelo et al., 2018b] and the conditional NP (CNP) [Garnelo et al., 2018a]. Instead of a latent variable, CNPs directly learn the predictive conditional distribution via a maximum likelihood meta-training procedure. Consequently they lack the ability to produce coherent function samples since each point is generated independently. Most subsequent work has improved upon NPs (and CNPs) through architectural modifications such as attention [Kim et al., 2019, 2022], convolutions [Gordon et al., 2020, Foong et al., 2020], mixtures [Wang and Van Hoof, 2022], equivariance [Kawano et al., 2021], and adaptation [Requeima et al., 2019]. Comparatively fewer works have attempted to improve the distributional assumptions of the latent variable. Two works have attempted to employ hierarchical [Wang and Van Hoof, 2020] and non-parametric [Flam-Shepherd et al., 2018] formulations to solve this problem.

Sum-Decomposition Networks The inference networks for NPs must have at least two properties. The first is that they make no assumptions about the size of the context set. The second is that the encoder be invariant to the ordering of context points. A common way to satisfy these criteria is by having the encoder take the form of a sum-decomposition network [Edwards and Storkey, 2017, Zaheer et al., 2017]:

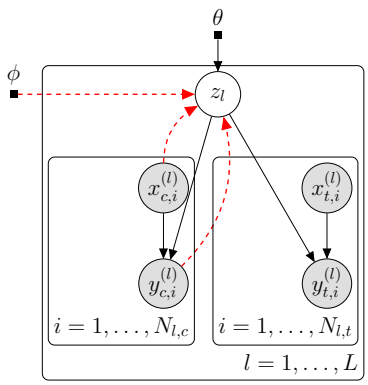


Figure 2: Generative process for multi-task learning with context/target splits along with the variational approximation (- - →).

$$\bar{\mathbf{r}} = \frac{1}{N_c} \sum_{i=1}^{N_c} \mathbf{r}_{c,i} \quad \text{with } \mathbf{r}_{c,i} = \text{enc}_\phi^r(\mathbf{x}_{c,i}, \mathbf{y}_{c,i}), \tag{3}$$

where $\mathbf{r}_i$ are datapoint-wise encodings given by a NN which are then aggregated. The aggregation operation is typically taken to be a simple average in NPs but other operators are valid as long as they are permutation-invariant. Thus the amortisation goes one level further with parameter sharing across context points. Finally, the aggregated representation $\bar{\mathbf{r}}$ is passed to a further NN to give the variational parameters. In the case of Gaussian posterior we have $\left(\widetilde{\boldsymbol{\mu}}, \widetilde{\boldsymbol{\Sigma}}\right) = \text{enc}_\phi^z(\bar{\mathbf{r}})$.

Bayesian Context Aggregation Volpp et al. [2020] propose a novel aggregation mechanism, derived by constructing a surrogate conditional latent variable (CLV) model in which the datapoint-wise encodings $\mathbf{r}_i$ are interpreted as noisy observations of the underlying Gaussian latent variable. Another encoder network then evaluates the observation noise for each datapoint $\boldsymbol{\sigma}_{c,i}^2 = \text{enc}_\phi^{\text{sigma}}(\mathbf{x}_{c,i}, \mathbf{y}_{c,i})$. By choosing a Gaussian prior with mean $\boldsymbol{\mu}_0$ and variance $\boldsymbol{\sigma}_0^2$, Bayesian inference in this surrogate CLV results in the following aggregation mechanism:

$$\begin{aligned} \widetilde{\boldsymbol{\sigma}}^2 &= \left[\boldsymbol{\sigma}_0^{-2} + \sum_{i=1}^{N_c} \boldsymbol{\sigma}_{c,i}^{-2}\right]^{-1} \\ \widetilde{\boldsymbol{\mu}} &= \boldsymbol{\mu}_0 + \widetilde{\boldsymbol{\sigma}}^2 \circ \sum_{i=1}^{N_c} \left(\mathbf{r}_{c,i} - \boldsymbol{\mu}_0\right) / \boldsymbol{\sigma}_{c,i}^2 \end{aligned} \tag{4}$$

where $\mathbf{a} \circ \mathbf{b}$ and $\mathbf{a}/\mathbf{b}$ denote element-wise product and division respectively between vector $\mathbf{a}$ and $\mathbf{b}$. Aggregation operates directly in the latent-space which forgoes the need for a further NN. Bayesian aggregation is a strict generalization of mean aggregation since the latter is recovered when a non-informative prior is used along with uniform observation variances. Eq. 4 can be seen as re-weighting context points, with the weights given by $\boldsymbol{\sigma}_{c,i}^{-2}$. This has some resemblance to self-attention mechanisms that have been adapted for neural processes [Kim et al., 2019] however the key difference is that the weights are computed without consideration of other context points.

3 STRUCTURED INFERENCE NETWORKS

We describe our approach that reframes individual context embeddings of the sum-decomposition architecture as factors in a probabilistic graphical model. To motivate this, let us consider the case where the likelihood and prior in Eq. 2 are given by *conjugate* exponential-family distributions. The posterior can then be obtained analytically [Wainwright and Jordan, 2008]. The prior can be expressed as $p(\mathbf{z}) = h(\mathbf{z}) \exp\left[\langle \mathbf{T}(\mathbf{z}), \boldsymbol{\eta}_0 \rangle - A(\boldsymbol{\eta}_0)\right]$, where $\boldsymbol{\eta}_0$ is the natural parameters, $\mathbf{T}(\mathbf{z})$ is the sufficient statistics, $h(\mathbf{z})$ is the base measure and $A(\boldsymbol{\eta}_0)$ is the log-partition function. Due to conjugacy, the likelihood can be expressed in the same form as the prior for some natural parameters $\boldsymbol{\eta}_y(\mathbf{z})$ and sufficient statistics $\mathbf{T}(\mathbf{y}_{c,i})$,

$$\begin{aligned} p(\mathbf{y}_{c,i}|\mathbf{z}) &= h(\mathbf{y}_{c,i}) \exp\left[\langle \mathbf{T}(\mathbf{y}_{c,i}), \boldsymbol{\eta}_y(\mathbf{z}) \rangle - A\left(\boldsymbol{\eta}_y(\mathbf{z})\right)\right] \\ &\propto \exp\left(\underbrace{\begin{bmatrix} \boldsymbol{\eta}_y(\mathbf{z}) \\ -A\left(\boldsymbol{\eta}_y(\mathbf{z})\right) \end{bmatrix}}_{=\mathbf{T}(\mathbf{z})}^\top \underbrace{\begin{bmatrix} \mathbf{T}(\mathbf{y}_{c,i}) \\ 1 \end{bmatrix}}_{=\boldsymbol{\eta}_{c,i}(\mathbf{y}_{c,i})} \right) \end{aligned}$$

where we have excluded the feature vector for simplicity. Then the posterior distribution is,

$$p(\mathbf{z}|\mathcal{D}) \propto h(\mathbf{z}) \exp\left[\left\langle \mathbf{T}(\mathbf{z}), \boldsymbol{\eta}_0 + \sum_{i=1}^{N} \boldsymbol{\eta}_{c,i}(\mathbf{y}_{c,i}) \right\rangle\right] \tag{5}$$

where it is clear that the posterior natural parameters are given by simply adding the *sufficient statistics* of $\{\mathbf{y}_{c,i}\}_{i=1}^{N_c}$ to the prior natural parameters. The computation in Eq. 5 is strikingly similar to the sum-decomposition architecture in Eq. 3. There, the encodings $\text{enc}_\phi^r(\mathbf{x}_{c,i}, \mathbf{y}_{c,i})$ play a similar role to the sufficient statistics $\boldsymbol{\eta}_{c,i}(\mathbf{y}_{c,i})$ in that they are aggregated to obtain the contextual representation; but they are non-linear embeddings which could be much more expressive.

3.1 NEURAL SUFFICIENT STATISTICS

Using the recipe for conjugacy in the exponential-family, we aim to construct a variational distribution of the same form but replacing the sufficient statistics by *neural sufficient statistics*. While training the end-to-end NP still requires variational inference (due to the NNs in the encoder and decoder), having this form will allow for efficient, conjugate updates to the distribution over the latent variables. This is given as,

$$\begin{aligned} &q_\phi(\mathbf{z}|\mathcal{D}_c) \\ &\quad \propto \exp\left[\left\langle \mathbf{T}(\mathbf{z}), \boldsymbol{\eta}_{\phi_{\text{PGM}}} + \sum_{i=1}^{N_c} \mathbf{f}_{\phi_{\text{NN}}}(\mathbf{x}_{c,i}, \mathbf{y}_{c,i}) \right\rangle\right] \end{aligned} \tag{6}$$

where $\mathbf{f}_{\phi_{\text{NN}}}(\cdot)$ is a neural network for amortized and gradient-based construction of neural sufficient statistics [Wu et al., 2020] and $\boldsymbol{\eta}_{\phi_{\text{PGM}}}$ is the prior natural parameters. This is

an instance of a structured inference network (SIN) [Lin et al., 2018] in which the variational distribution takes a factorized form consisting of the prior and N *factors* which are sometimes referred to as deep observational likelihoods [Johnson et al., 2016]. This can be seen by rewriting Eq. 6 as,

$$q_\phi(\mathbf{z} \,|\, \mathcal{D}_c) = \frac{1}{Z_c(\phi)} \underbrace{\left[\prod_{i=1}^{N_c} \exp\left(\left\langle \mathbf{T}(\mathbf{z}), \mathbf{f}_{\phi_{\text{NN}}}(\mathbf{x}_{c,i}, \mathbf{y}_{c,i}) \right\rangle \right) \right]}_{\text{NN factors}} \times \underbrace{\left[q(\mathbf{z}; \phi_{\text{PGM}}) \right]}_{\text{prior}} \quad (7)$$

where $Z_c(\phi)$ is the normalization constant. The framework is also flexible to allow for the prior parameters to be fitted, $\phi := \{\phi_{\text{NN}}, \phi_{\text{PGM}}\}$. The neural sufficient statistics give rise to factors that can be conjugate to the prior by construction. Whilst this may seem like an arbitrary construction, in the specific case of Gaussianity, the optimal variational approximation decomposes into the prior and local Gaussian factors that approximate the likelihood [Nickisch and Rasmussen, 2008, Opper and Archambeau, 2009].

Furthermore, the variational lower bound resulting from using the structured inference network (Eq. 7) in Eq. 1 contains a term that resembles the entropy of the individual factors (see App. A.1 for further details). This is further evidence that the factors are approximating the sufficient statistics of the likelihood; this is due to the link between statistical sufficiency and information-maximizing representations of the data [Chen et al., 2021].

The structured inference network improves several aspects of the sum-decomposition network. The first improvement is the variational parameters are now computed directly from the context embeddings. This eliminates the need for an additional neural network reducing the total number of encoder parameters. The second improvement is the introduction of an explicit prior distribution whose parameters are aggregated along with the context embeddings. The third improvement is the aggregation mechanism is determined directly from the exponential-family parameterization. Natural parameterization, as shown in Eq. 6, leads to sum-pooling which is equivalent to mean-pooling in Eq. 3 within a constant of proportionality. However, expectation parameterization leads to weighted aggregation mechanisms as we demonstrate below for two cases with Gaussian prior and Mixture of Gaussian prior.

3.2 BAYESIAN CONTEXT AGGREGATION AS SIN WITH GAUSSIAN ASSUMPTIONS

For a Gaussian prior, the resulting conjugate exponential-family distribution is also Gaussian. By substituting this in Eq. 7 with $\mathbf{T}(\mathbf{z}) := \{\mathbf{z}, \mathbf{z}\mathbf{z}^\top\}$ for the factors we obtain,

$$q_\phi(\mathbf{z} \,|\, \mathcal{D}_c) = \tfrac{1}{Z_c(\phi)} \left[\prod_{i=1}^{N_c} \mathcal{N}(\mathbf{z} \,|\, \mathbf{m}_{c,i}, \mathbf{V}_{c,i}) \right] \times \left[\mathcal{N}(\mathbf{z} \,|\, \boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0) \right]$$

where $\left(\mathbf{m}_{c,i}, \mathbf{V}_{c,i}\right) \leftarrow \mathbf{f}_{\phi_{\text{NN}}}(\mathbf{x}_{c,i}, \mathbf{y}_{c,i})$ are the moment parameters of the factors (mean and variance) evaluated by the recognition network and prior moments $\phi_{\text{PGM}} := \{\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0\}$. Then the variational distribution is also Gaussian $q_\phi(\mathbf{z} \,|\, \mathcal{D}_c) = \mathcal{N}(\mathbf{z}; \widetilde{\boldsymbol{\mu}}, \widetilde{\boldsymbol{\Sigma}})$ with posterior moments given by,

$$\begin{aligned} \widetilde{\boldsymbol{\Sigma}}^{-1} &= \textstyle\sum_{i=1}^{N_c} \mathbf{V}_{c,i}^{-1} + \boldsymbol{\Sigma}_0^{-1} \\ \widetilde{\boldsymbol{\mu}} &= \widetilde{\boldsymbol{\Sigma}} \left(\textstyle\sum_{i=1}^{N_c} \mathbf{V}_{c,i}^{-1} \mathbf{m}_{c,i} + \boldsymbol{\Sigma}_0^{-1} \boldsymbol{\mu}_0 \right). \end{aligned} \quad (8)$$

The normalization constant $Z_c(\phi)$ is also available in closed-form. We also assume diagonal covariance matrices throughout thereby ensuring no expensive matrix operations are performed for aggregation. This Gaussian-based procedure is equivalent to the previously proposed Bayesian aggregation mechanism [Volpp et al., 2020]. This can be seen by a straightforward manipulation of Eq. 8 to give the incremental update form of Eq. 4 (see App. A.2 for proof).

3.3 MIXTURE BAYESIAN AGGREGATION

We now consider a more expressive prior distribution, the mixture of Gaussian (MoG) prior, which despite being a conditionally-conjugate exponential-family distribution, results in closed-form updates when the factors are chosen to be Gaussian,

$$q_\phi(\mathbf{z} \,|\, \mathcal{D}_c) = \tfrac{1}{Z_c(\phi)} \left[\prod_{i=1}^{N_c} \mathcal{N}(\mathbf{z} \,|\, \mathbf{m}_{c,i}, \mathbf{V}_{c,i}) \right] \times \left[\textstyle\sum_{k=1}^{K} \pi_k \mathcal{N}(\mathbf{z} \,|\, \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \right].$$

The factors' moment parameters are evaluated in the same way as Sec. 3.2 and the prior parameters are given by $\phi_{\text{PGM}} := \{\pi_k, \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k\}_{k=1}^K$ where $\sum_{k=1}^K \pi_k = 1$. Then the variational distribution also takes a MoG form $q_\phi(\mathbf{z} \,|\, \mathcal{D}_c) = \sum_{k=1}^K \widetilde{\pi}_k \mathcal{N}(\mathbf{z} \,|\, \widetilde{\boldsymbol{\mu}}_k, \widetilde{\boldsymbol{\Sigma}}_k)$ with posterior parameters given by,

$$\begin{aligned} \widetilde{\boldsymbol{\Sigma}}_k^{-1} &= \textstyle\sum_{i=1}^{N_c} \mathbf{V}_{c,i}^{-1} + \boldsymbol{\Sigma}_k^{-1} \\ \widetilde{\boldsymbol{\mu}}_k &= \widetilde{\boldsymbol{\Sigma}}_k \left(\textstyle\sum_{i=1}^{N_c} \mathbf{V}_{c,i}^{-1} \mathbf{m}_{c,i} + \boldsymbol{\Sigma}_k^{-1} \boldsymbol{\mu}_k \right) \\ \widetilde{\pi}_k &= \pi_k \; C_k \;/\; Z_c \end{aligned} \quad (9)$$

for $k = 1, \ldots, K$ where $Z_c = \sum_{k=1}^K \widetilde{\pi}_k$. The updates for the mean and variance of each component Gaussian takes an identical form to Eq. 8. The update for the mixing proportions requires evaluation of the normalization constant C_k for each Gaussian component which is stated in App. A.3. A MoG prior may be a beneficial modelling assumption if we expect the data to arise from multiple sources. We refer to this approach as *mixture* Bayesian Aggregation (mBA), a generalization of BA that is recovered when the number of mixture components K is set to 1.

4 BEYOND CONJUGACY

The conjugate case is attractive due to its analytic properties, but it may be a poor modelling assumption for real-world data. By relaxing the need for conjugacy between the factors and the prior, we can allow for a wider range of distributional assumptions. Yet non-conjugacy implies the posterior can no longer be expressed analytically since the evidence $Z_c(\phi)$ is intractable. We can instead form a lower bound on the evidence by introducing an approximating distribution $\tilde{q}$,

$$\log Z_c(\phi) \geq \mathbb{E}_{\tilde{q}(\mathbf{z})}\left[\log q_\phi(\mathbf{z}, \mathcal{D}_c)\right] + \mathcal{H}\left(\tilde{q}(\mathbf{z})\right) \tag{10}$$

with joint distribution $q_\phi(\mathbf{z}, \mathcal{D}_c)$ corresponding to Eq. 7 and $\mathcal{H}(q) = \mathbb{E}_q\left[-\log q(\mathbf{z})\right]$ is the entropy. The variational posterior now depends implicitly on the neural sufficient statistics and prior parameters. Whilst this approach would be generally applicable, without further assumptions every forward pass through the NP would require solving the stochastic optimization problem that is maximizing Eq. 10. This closely resembles a recently proposed aggregation mechanism for set embedding termed equilibrium aggregation [Bartunov et al., 2022] which also considers an optimization-based formulation. They show that under certain conditions (*e.g.* choice of initialization, regularization strength *etc.*), convergent dynamics are observed with a small number of gradient-descent steps.

These techniques could also be adapted here, but instead we consider certain restrictions that lead to a more tractable approach while still allowing for expressive PGMs. We start by restricting $\tilde{q}$ to be a mean-field distribution [Jordan et al., 1999]. Considering a certain partition of $\mathbf{z}$ into M disjoint groups over which $\tilde{q}$ factorises, we have $\tilde{q}(\mathbf{z}) = \prod_{j=1}^{M} \tilde{q}_j(\mathbf{z}_j)$. Now the stationary point $\tilde{q}^*(\mathbf{z})$ of Eq. 10 satisfies (Bishop [2006], Eq. 10.9),

$$\log \tilde{q}_j^*(\mathbf{z}_j) = \mathbb{E}_{\tilde{q}^*_{/j}(\mathbf{z}_{/j})}\left[\log q_\phi\left(\mathbf{z}, \mathcal{D}_c\right)\right] + \text{cnst.} \tag{11}$$

where $\mathbf{z}_{/j}$ indicates all variables excluding the jth group using which the expectation of the log joint is taken with respect to. We further suppose that the factors and prior specify a conditionally-conjugate exponential-family system and that each factor in $\tilde{q}(\mathbf{z})$ belongs to the same exponential-family as the complete conditional distribution in $q_\phi(\mathbf{z}, \mathcal{D}_c)$. Then the expectation in Eq. 11 takes a closed-form expression resulting in tractable updates for each factor of the mean-field posterior. We can iteratively optimize each $\mathbf{z}_j$ whilst holding the others fixed using Eq. 11. This is often referred to as coordinate ascent variational inference (CAVI) [Blei et al., 2017] which is a special case of variational message passing [Winn et al., 2005]. A closely related approach is differentiable EM (DIEM) [Kim, 2021] that frames set embedding as maximum-a-posteriori estimation. This performs expectation-maximization (EM) updates and can be viewed as a special case of our method.

4.1 ROBUST BAYESIAN AGGREGATION

We now consider a specific instance of a conditionally-conjugate exponential-family system in which the coordinate-ascent updates arising from Eq. 11 give rise to a novel, weighted aggregation mechanism. This extends the all-Gaussian assumptions of Sec. 3.2 by introducing a Gamma prior over the precision of each Gaussian factor, whose marginal form is a heavy-tailed Student-*t* distribution, as well as a hierarchical prior. This is adapted from previous work on robust Bayesian interpolation [Tipping and Lawrence, 2005] which demonstrated robustness to outliers and corruptions in the targets with this model specification. The probabilistic model is,

$$q_\phi(\mathbf{z}, \alpha, \boldsymbol{\beta}) = \left[\prod_{i=1}^{N_c} \mathcal{N}(\mathbf{z}|\mathbf{m}_{c,i}, \beta_i^{-1}\mathbf{V}_{c,i})\right] \times \left[q(\mathbf{z}|\alpha)q(\alpha)\prod_{i=1}^{N_c} q(\beta_i)\right]$$

with $q(\mathbf{z}|\alpha) = \mathcal{N}(\mathbf{z}|\mathbf{0}, \alpha^{-1}\mathbf{I})$, $q(\alpha) = \mathcal{G}(\alpha|a_0, b_0)$ and $q(\beta_i) = \mathcal{G}(\beta_i|c_0, c_0)$. The factors' moment parameters are evaluated as before and $\phi_{\text{PGM}} = \{a_0, b_0, c_0\}$. Each factor and $q(\beta_i)$ together can be viewed as the hierarchical form of the Student-*t* distribution,

$$\mathcal{T}(\mathbf{z}|\mathbf{m}_{c,i}, \mathbf{V}_{c,i}, c_0) = \int \mathcal{N}(\mathbf{z} \,|\, \mathbf{m}_{c,i}, \beta_i^{-1}\mathbf{V}_{c,i}) \cdot \mathcal{G}(\beta_i|c_0, c_0)\, d\beta_i.$$

The marginal form would render the updates arising from Eq. 11 intractable hence we restrict ourselves to the joint specification. Next we introduce a mean-field distribution with a corresponding factorization and functional form to the prior, $\tilde{q}(\mathbf{z}, \alpha, \boldsymbol{\beta}) = \tilde{q}(\mathbf{z})\tilde{q}(\alpha)\tilde{q}(\boldsymbol{\beta})$. where $\tilde{q}(\mathbf{z}) = \mathcal{N}(\mathbf{z}|\widetilde{\boldsymbol{\mu}}, \widetilde{\boldsymbol{\Sigma}})$, $\tilde{q}(\alpha) = \mathcal{G}(\alpha|\widetilde{a}, \widetilde{b})$ and $\tilde{q}(\boldsymbol{\beta}) = \prod_{i=1}^{N_c} \mathcal{G}(\beta_i|\widetilde{c}, \widetilde{d}_i)$. Plugging this into Eq. 11, we can derive the following updates for the parameters of the mean-field posterior,

$$\widetilde{\boldsymbol{\Sigma}}^{-1} = \textstyle\sum_{i=1}^{N_c} \mathbb{E}[\beta_i]\mathbf{V}_{c,i}^{-1} + \mathbb{E}[\alpha]\mathbf{I} \tag{12}$$

$$\widetilde{\boldsymbol{\mu}} = \widetilde{\boldsymbol{\Sigma}} \textstyle\sum_{i=1}^{N_c} \mathbb{E}[\beta_i]\mathbf{V}_{c,i}^{-1}\mathbf{m}_{c,i}, \tag{13}$$

$$\widetilde{a} = a_0 + \tfrac{D}{2} \tag{14}$$

$$\widetilde{b} = b_0 + \tfrac{1}{2}\mathbb{E}[\mathbf{z}^\top\mathbf{z}] \tag{15}$$

$$\widetilde{c} = c_0 + \tfrac{D}{2} \tag{16}$$

$$\widetilde{d}_i = c_0 + \tfrac{1}{2}\Big(\mathbf{m}_{c,i}^\top\mathbf{V}_{c,i}^{-1}\mathbf{m}_{c,i} - 2\mathbf{m}_{c,i}^\top\mathbf{V}_{c,i}^{-1}\mathbb{E}[\mathbf{z}] + \operatorname{tr}(\mathbf{V}_{c,i}^{-1}\mathbb{E}[\mathbf{z}\mathbf{z}^\top])\Big) \tag{17}$$

with expectations given by $\mathbb{E}[\mathbf{z}] = \widetilde{\boldsymbol{\mu}}$, $\mathbb{E}[\mathbf{z}\mathbf{z}^\top] = \widetilde{\boldsymbol{\mu}}\widetilde{\boldsymbol{\mu}}^\top + \widetilde{\boldsymbol{\Sigma}}$, $\mathbb{E}[\alpha] = \widetilde{a}/\widetilde{b}$ and $\mathbb{E}[\beta_i] = \widetilde{c}/\widetilde{d}_i$. We iterate through the updates in the order presented above. At convergence or after a fixed number of iterations, we only keep the posterior over $\mathbf{z}$ which is passed to the NP decoder.

The new procedure is an extension of BA. This can be seen by initializing the Gamma parameters to $\widetilde{a} = \widetilde{b} = \widetilde{c} = \widetilde{d}_i =$

Table 1: *mixture Bayesian Aggregation (mBA)*. We observe a boost in performance over BA with increasing number of components (K). We also exceed the performance of NP with Self-Attention (NP+SA) on OOD data. This is demonstrated on the task of 2D image completion (EMNIST).

	Predictive Log-Likelihood ↑				**RMSE ↓**			
	Seen classes (0-9)		Unseen classes (10-46)		Seen classes (0-9)		Unseen classes (10-46)	
	context	target	context	target	context	target	context	target
NP	$0.701_{\pm 0.065}$	$0.589_{\pm 0.061}$	$0.567_{\pm 0.059}$	$0.405_{\pm 0.039}$	$0.201_{\pm 0.018}$	$0.218_{\pm 0.014}$	$0.244_{\pm 0.014}$	$0.265_{\pm 0.009}$
NP+SA	$\mathbf{0.977}_{\pm 0.006}$	$\mathbf{0.840}_{\pm 0.005}$	$0.823_{\pm 0.007}$	$0.609_{\pm 0.008}$	$0.127_{\pm 0.002}$	$\mathbf{0.165}_{\pm 0.001}$	$0.177_{\pm 0.002}$	$0.224_{\pm 0.002}$
NP-BA	$0.866_{\pm 0.097}$	$0.708_{\pm 0.076}$	$0.749_{\pm 0.118}$	$0.537_{\pm 0.093}$	$0.154_{\pm 0.027}$	$0.193_{\pm 0.014}$	$0.193_{\pm 0.033}$	$0.238_{\pm 0.018}$
NP-mBA (K=2)	$0.952_{\pm 0.121}$	$0.778_{\pm 0.083}$	$0.855_{\pm 0.143}$	$0.623_{\pm 0.097}$	$0.128_{\pm 0.033}$	$0.181_{\pm 0.017}$	$0.162_{\pm 0.038}$	$0.221_{\pm 0.020}$
NP-mBA (K=3)	$0.953_{\pm 0.108}$	$0.777_{\pm 0.079}$	$0.857_{\pm 0.132}$	$0.623_{\pm 0.103}$	$0.128_{\pm 0.031}$	$0.180_{\pm 0.015}$	$0.162_{\pm 0.038}$	$0.221_{\pm 0.020}$
NP-mBA (K=5)	$0.975_{\pm 0.083}$	$0.792_{\pm 0.064}$	$\mathbf{0.883}_{\pm 0.105}$	$\mathbf{0.642}_{\pm 0.084}$	$\mathbf{0.122}_{\pm 0.025}$	$0.177_{\pm 0.011}$	$\mathbf{0.155}_{\pm 0.031}$	$\mathbf{0.217}_{\pm 0.016}$

1 and running a single step of Eqs. 12 and 13, then BA is recovered (when standard normal prior is used). However, compared with BA, Eq. 12 presents an alternate approach to updating the prior precision which now depends on the aggregated contextual representation through Eq. 15. This adaptive mechanism resembles that of a data-dependent prior [Tipping, 2001].

Another crucial difference in Eqs. 12 and 13 is that the neural sufficient statistics (in natural parameterization) are now weighted by the 1st moment of the variational noise distribution. The context-dependent variability in this term is given by Eq. 17 which is evaluated by using the individual context embedding as well as the aggregated contextual representation given by the moments of $\tilde{q}(\mathbf{z})$. Viewed as a weighted aggregation scheme, this has close resemblance to self-attention which evaluates a similarity function (*e.g.* Euclidean norm, dot-product *etc.*) between all pairs of context points. In the case of corruptions to the context set, such outliers can be downweighted, improving the robustness of neural processes. We refer to our approach as *robust* Bayesian aggregation (rBA).

We run the coordinate-wise updates for a fixed number of steps and observe that in practice convergent dynamics appear within a few steps as shown in Fig. 5b. We note that all operations are fully-differentiable and we backprop through the *unrolled* steps for gradient-based learning. It may be possible to incorporate implicit differentiation techniques, for instance as used in Deep Equilibrium Models [Bai et al., 2019], to improve the efficiency and reduce the memory of the proposed method.

5 EXPERIMENTS

We conduct experiments to assess the performance of our mixture and robust variants[1] of Bayesian Aggregation (BA) against vanilla BA [Volpp et al., 2020]. We also benchmark against NPs with regular encoder architecture (NP) [Garnelo et al., 2018b] and an NP with self-attention module (NP+SA) [Kim et al., 2019]. Similar to Volpp et al. [2020], we only evaluate on NP-based models with a single latent path (i.e. no deterministic path). We do not consider encoder architectures that evaluate task-specific contextual representations such as Attentive NP [Kim et al., 2019] and leave the combination of BA variants with cross-attention style mechanisms to future work. The decoder network architecture is fixed across all models such that any performance differences can be attributed solely to the encoder architecture. To reduce the effects of differences in network architecture between BA and non-BA encoders, we keep the BA encoder size close to but less that of the non-BA architectures. Each model is trained with 5 different seeds and we report the mean and standard deviations from these runs. Please refer to App. B for further details on the model specification and training configuration.

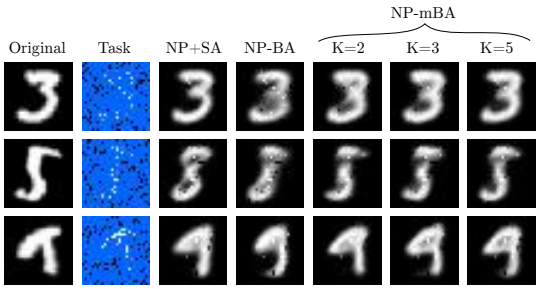
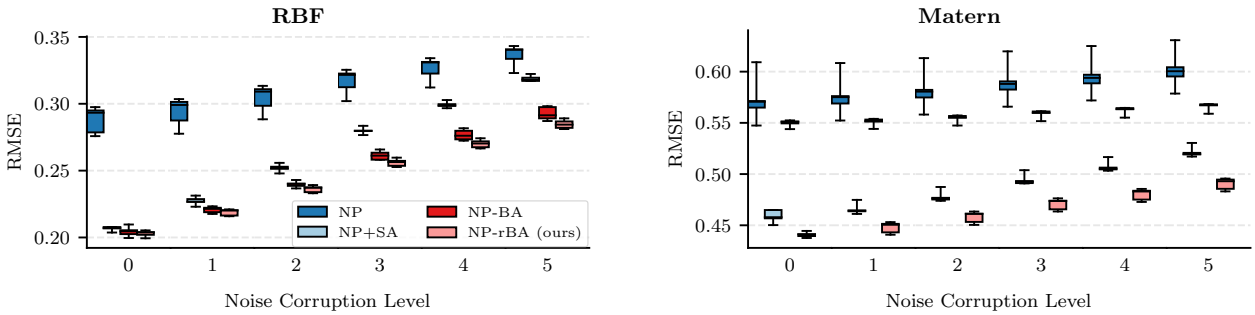


Figure 3: Completed images on EMNIST (0-9) from 100 context points ("task").

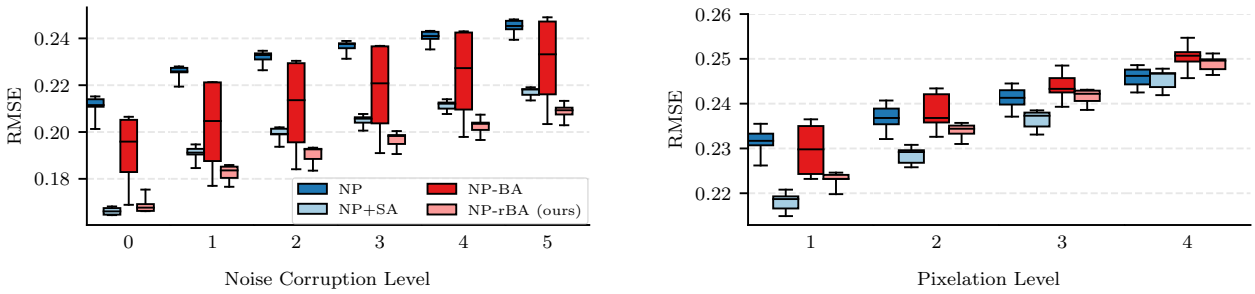
We consider the following tasks,

- **1-D Regression** We train on functional samples drawn from Gaussian Process (GP) priors with RBF and Matern-5/2 kernel. The specification of the RBF kernel is taken from Lee et al. [2020] where both the lengthscale and output variance is varied (GP hyperparameters). The matern-5/2 kernel is taken from Gordon et al. [2020] where the hyperparameters are fixed.
- **Image completion** A subset of pixels for a given image is presented with the task to predict the remaining pixels.

[1]For a reference implementation see: `https://github.com/dvtailor/np-structured-inference`.



(a) 1D regression with functional samples from GP kernel.



(b) 2D image completion on EMNIST dataset.

Figure 4: *robust Bayesian Aggregation (rBA).* We demonstrate gains to test-time robustness in the presence of corruptions to the context set as compared with BA. In (a), context set function values are perturbed by heavy-tail noise of varying magnitude. In (b), we consider common image noise perturbations (left) and image pixelation (right), of varying intensity.

Therefore, we can understand the inputs to be coordinates of each pixel and the pixel intensities representing the targets (i.e. 2-D regression). We consider the EMNIST dataset [Cohen et al., 2017] – a dataset of handwritten digits and letters comprising of grayscale 28x28 images. During training, we restrict images to the first 10 classes. We evaluate on two settings: in-distribution data (held-out images from the first 10 classes) and out-of-distribution (OOD) data (images from the remaining 37 classes).

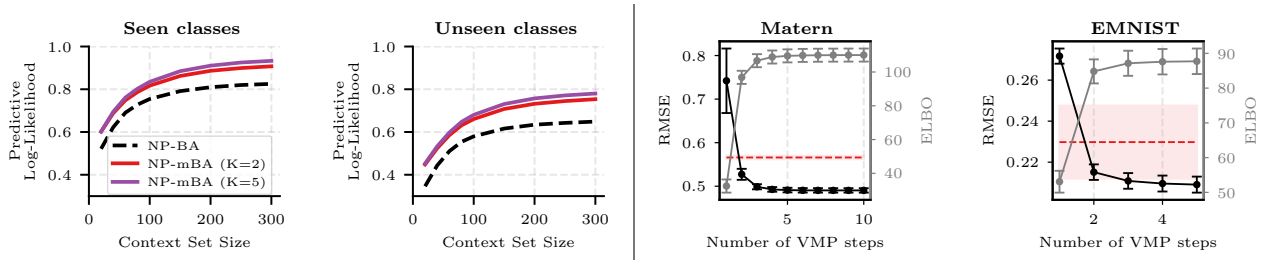
5.1 MIXTURE BAYESIAN AGGREGATION

We initialize the mixture prior in the PGM as follows: the prior mixing proportions are set to uniform, the prior variance is set to 1 and the prior means are sampled from a zero-mean Gaussian with standard deviation 0.1. The latter setting is to encourage diversity in the posterior mixture components and prevent collapse to a unimodal Gaussian. The task of image completion on the EMNIST dataset is considered. Table 1 clearly demonstrates the utility of additional mixing components with improved function modelling as the number of components is increased from 2 to 5 (see Fig. 3). This is also shown in dependence of the context set size in Fig. 5a. The reconstruction error is also indicated ("context"). On OOD data, NP-mBA even exceeds the performance of NP with Self-Attention (NP+SA) for all numbers of components considered. This is despite NP+SA having a considerably larger number of parameters (see Table 2). We also observe little change in the run-time when the number of components is increased.

5.2 ROBUST BAYESIAN AGGREGATION

We demonstrate the improved robustness of our robust Bayesian Aggregation (NP-rBA) to corruptions in the context sets at test-time (see Fig. 6). For the 1D regression task, we extend the model-data mismatch setting from Lee et al. [2020] where the function values are corrupted by Student-*t* noise of increasing magnitude, $\varepsilon \sim \gamma \cdot \mathcal{T}(2.1)$ with $\gamma \in [0.05, 0.08, 0.11, 0.13, 0.15]$. For the image completion setting, we only consider the in-distribution data and perform noise corruption and pixelation as specified in [Hendrycks and Dietterich, 2019]. For the noise corruption, we average over four different types of noise, namely Gaussian, Shot, Poisson and Impulse noise, of increasing severity. Pixelation involves downsampling the image to different reduced resolutions and then upsampling back to the original resolution. For our NP-rBA, we run the message passing algorithm for 5 steps (at both train and test time) except for Matern where we run for 10 steps.

In Fig. 4, we observe a consistent gain over BA across the different tasks and corruptions. With the exception of



(a) Boost with increasing mixture components

(b) Boost with increasing message-passing steps

Figure 5: (a) Across a range of context set sizes, there is a gain in performance with increasing mixture components over BA. This is demonstrated on the task of 2D image completion (error bars are omitted for visual clarity). (b) By increasing the number of message-passing (VMP) steps, we observe a monotonic improvement in both the accuracy of the target function and on the lower bound to the evidence of the PGM. With just 2 steps, there is a gain over vanilla BA (- -). This is shown for test-time corruption to the context set of the highest intensity. During training, a fixed number of steps is used (10 and 5 for Matern and EMNIST respectively).

Table 2: Number of model parameters and training time (s). Run-time is measured on the image completion experiment on a GTX 1080, for a single epoch (batch size 100). NP-mBA is run with 5 components and NP-rBA with 5 message-passing steps.

	# Parameters	Training (s)
NP	248,962	8.1
NP+SA	282114	11.2
NP-BA		10.0
NP-mBA	166,914	16.3 (*)
NP-rBA		16.1

(*) 10 latent samples

corruption by pixelation, NP-rBA also shows improvement over NP with Self-Attention (NP+SA). Fig. 5b demonstrates speed-accuracy trade-off where increasing the number of steps at test-time leads to monotonic improvement in performance.

6 CONCLUSION

We propose structured inference networks for context aggregation in Neural Processes (NPs). This change is attractive for several reasons: (i) the local encodings now have a clear interpretation as neural sufficient statistics, (ii) the aggregation step is predetermined by and follows from the probabilistic assumptions and inference routine, and (iii) structured priors are straightforward to incorporate. We show that an existing context aggregation mechanism, Bayesian Aggregation (BA), is recovered by imposing Gaussianity assumptions in the PGM.

By imposing different modelling assumptions, we demonstrate alternative context aggregation mechanisms can be derived. In particular, we consider, (1) Mixture of Gaussian

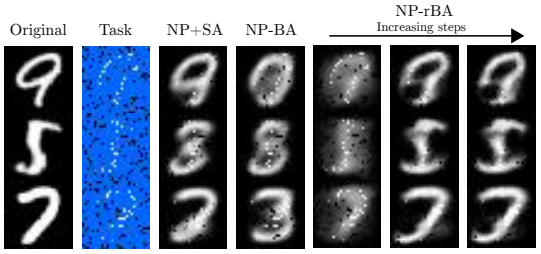


Figure 6: Completed images on EMNIST (0-9) from 100 context points ("task") that have been corrupted by Gaussian noise.

prior and (2) Student-t assumptions, which give rise to two novel variants of BA, namely *mixture* and *robust* Bayesian Aggregation. We demonstrate improvements to the functional modelling and test-time robustness of NPs without any increase in the parameterization of the encoder.

For future work, we look to consider more general PGM structures, *e.g.* temporal transition structure for modelling time-series data or even causal structure via directed acyclic graphs (DAGs). It is worth reiterating that there is effectively no systematic limitation to our framework since, if the user deems there to be too much computational overhead in using an expressive PGM, the assumptions can be simplified (*e.g.* Gaussianity) thereby recovering existing strategies.

Author Contributions

List of Authors: Dharmesh Tailor (D.T.), Mohammad Emtiyaz Khan (M.E.K.), Eric Nalisnick (E.N.).

D.T. and E.N. conceived the original idea. This was then discussed with M.E.K. D.T. derived and implemented the new aggregation strategies and ran the experiments, with regular feedback from E.N. D.T. wrote the first draft of the paper, after which all authors contributed to writing.

Acknowledgements

We would like to thank Qi Wang (University of Amsterdam) for his helpful feedback and discussions.

References

Shaojie Bai, J. Zico Kolter, and Vladlen Koltun. Deep Equilibrium Models. *Advances in Neural Information Processing Systems*, 2019.

Sergey Bartunov, Fabian B. Fuchs, and Timothy P. Lillicrap. Equilibrium aggregation: encoding sets via optimization. In *Proceedings of the 38th Conference on Uncertainty in Artificial Intelligence*, 2022.

Christopher M. Bishop. *Pattern Recognition and Machine Learning*. Springer, 2006.

David M. Blei, Alp Kucukelbir, and Jon D. McAuliffe. Variational Inference: A Review for Statisticians. *Journal of the American Statistical Association*, 2017.

Yanzhi Chen, Dinghuai Zhang, Michael U. Gutmann, Aaron Courville, and Zhanxing Zhu. Neural Approximate Sufficient Statistics for Implicit Models. In *Proceedings of the International Conference on Learning Representations*, 2021.

Gregory Cohen, Saeed Afshar, Jonathan Tapson, and Andre van Schaik. EMNIST: Extending MNIST to handwritten letters. In *International Joint Conference on Neural Networks*, 2017.

Harrison Edwards and Amos Storkey. Towards a Neural Statistician. In *Proceedings of the International Conference on Learning Representations*, 2017.

Daniel Flam-Shepherd, Yuxiang Gao, and Zhaoyu Guo. Stick-Breaking Neural Latent Variable Models. *NeurIPS Workshop on Bayesian Deep Learning*, 2018.

Andrew Y. K. Foong, Wessel P. Bruinsma, Jonathan Gordon, Yann Dubois, James Requeima, and Richard E. Turner. Meta-Learning Stationary Stochastic Process Prediction with Convolutional Neural Processes. In *Advances in Neural Information Processing Systems*, 2020.

Marta Garnelo, Dan Rosenbaum, Christopher Maddison, Tiago Ramalho, David Saxton, Murray Shanahan, Yee Whye Teh, Danilo Rezende, and S. M. Ali Eslami. Conditional Neural Processes. In *Proceedings of the 35th International Conference on Machine Learning*, 2018a.

Marta Garnelo, Jonathan Schwarz, Dan Rosenbaum, Fabio Viola, Danilo J. Rezende, S. M. Eslami, and Yee Whye Teh. Neural Processes. *ICML Workshop on Theoretical Foundations and Applications of Deep Generative Models*, 2018b.

Jonathan Gordon, Wessel P. Bruinsma, Andrew Y. K. Foong, James Requeima, Yann Dubois, and Richard E. Turner. Convolutional Conditional Neural Processes. In *Proceedings of the International Conference on Learning Representations*, 2020.

Dan Hendrycks and Thomas Dietterich. Benchmarking Neural Network Robustness to Common Corruptions and Perturbations. In *Proceedings of the International Conference on Learning Representations*, 2019.

T. M. Heskes. Empirical Bayes for Learning to Learn. In *Proceedings of the 17th International Conference on Machine Learning*, 2000.

Matthew J. Johnson, David K. Duvenaud, Alex Wiltschko, Ryan P. Adams, and Sandeep R. Datta. Composing graphical models with neural networks for structured representations and fast inference. In *Advances in Neural Information Processing Systems*, 2016.

Michael I. Jordan, Zoubin Ghahramani, Tommi S. Jaakkola, and Lawrence K. Saul. An Introduction to Variational Methods for Graphical Models. *Machine learning*, 1999.

Makoto Kawano, Wataru Kumagai, Akiyoshi Sannai, Yusuke Iwasawa, and Yutaka Matsuo. Group Equivariant Conditional Neural Processes. In *Proceedings of the International Conference on Learning Representations*, 2021.

Hyunjik Kim, Andriy Mnih, Jonathan Schwarz, Marta Garnelo, Ali Eslami, Dan Rosenbaum, Oriol Vinyals, and Yee Whye Teh. Attentive Neural Processes. In *Proceedings of the International Conference on Learning Representations*, 2019.

Mingyu Kim, Kyeong Ryeol Go, and Se-Young Yun. Neural Processes with Stochastic Attention: Paying more attention to the context dataset. In *Proceedings of the International Conference on Learning Representations*, 2022.

Minyoung Kim. Differentiable Expectation-Maximization for Set Representation Learning. In *Proceedings of the International Conference on Learning Representations*, 2021.

Juho Lee, Yoonho Lee, Jungtaek Kim, Eunho Yang, Sung Ju Hwang, and Yee Whye Teh. Bootstrapping Neural Processes. In *Advances in Neural Information Processing Systems*, 2020.

Wu Lin, Nicolas Hubacher, and Mohammad Emtiyaz Khan. Variational Message Passing with Structured Inference Networks. In *Proceedings of the International Conference on Learning Representations*, 2018.

Tung Nguyen and Aditya Grover. Transformer Neural Processes: Uncertainty-Aware Meta Learning Via Sequence Modeling. In *Proceedings of the 39th International Conference on Machine Learning*, 2022.

Hannes Nickisch and Carl Edward Rasmussen. Approximations for Binary Gaussian Process Classification. *Journal of Machine Learning Research*, 2008.

Manfred Opper and Cédric Archambeau. The Variational Gaussian Approximation Revisited. *Neural Computation*, 2009.

James Requeima, Jonathan Gordon, John Bronskill, Sebastian Nowozin, and Richard E. Turner. Fast and Flexible Multi-Task Classification Using Conditional Neural Adaptive Processes. In *Advances in Neural Information Processing Systems*, 2019.

Michael E. Tipping. Sparse Bayesian Learning and the Relevance Vector Machine. *Journal of Machine Learning Research*, 2001.

Michael E. Tipping and Neil D. Lawrence. Variational inference for Student-t models: Robust Bayesian interpolation and generalised component analysis. *Neurocomputing*, 2005.

Anna Vaughan, Will Tebbutt, J. Scott Hosking, and Richard E. Turner. Convolutional conditional neural processes for local climate downscaling. *Geoscientific Model Development*, 2022.

Michael Volpp, Fabian Flürenbrock, Lukas Grossberger, Christian Daniel, and Gerhard Neumann. Bayesian Context Aggregation for Neural Processes. In *Proceedings of the International Conference on Learning Representations*, 2020.

Edward Wagstaff, Fabian Fuchs, Martin Engelcke, Ingmar Posner, and Michael A. Osborne. On the Limitations of Representing Functions on Sets. In *Proceedings of the 36th International Conference on Machine Learning*, 2019.

Martin J. Wainwright and Michael I. Jordan. Graphical Models, Exponential Families, and Variational Inference. *Foundations and Trends in Machine Learning*, 2008.

Qi Wang and Herke Van Hoof. Doubly Stochastic Variational Inference for Neural Processes with Hierarchical Latent Variables. In *Proceedings of the 37th International Conference on Machine Learning*, 2020.

Qi Wang and Herke Van Hoof. Learning Expressive Meta-Representations with Mixture of Expert Neural Processes. In *Advances in Neural Information Processing Systems*, 2022.

John Winn, Christopher M. Bishop, and Tommi Jaakkola. Variational Message Passing. *Journal of Machine Learning Research*, 2005.

Hao Wu, Heiko Zimmermann, Eli Sennesh, Tuan Anh Le, and Jan-Willem van de Meent. Amortized Population Gibbs Samplers with Neural Sufficient Statistics. In *Proceedings of the 37th International Conference on Machine Learning*, 2020.

Manzil Zaheer, Satwik Kottur, Siamak Ravanbakhsh, Barnabas Poczos, Russ R. Salakhutdinov, and Alexander J. Smola. Deep Sets. In *Advances in Neural Information Processing Systems*, 2017.