Evidential Conditional Neural Processes

Deep Shankar Pandey, Qi Yu

Rochester Institute of Technology
{dp7972, qi.yu}@rit.edu

Abstract

The Conditional Neural Process (CNP) family of models offer a promising direction to tackle few-shot problems by achieving better scalability and competitive predictive performance. However, the current CNP models only capture the overall uncertainty for the prediction made on a target data point. They lack a systematic fine-grained quantification on the distinct sources of uncertainty that are essential for model training and decision-making under the few-shot setting. We propose Evidential Conditional Neural Processes (ECNP), which replace the standard Gaussian distribution used by CNP with a much richer hierarchical Bayesian structure through evidential learning to achieve epistemic-aleatoric uncertainty decomposition. The evidential hierarchical structure also leads to a theoretically justified robustness over noisy training tasks. Theoretical analysis on the proposed ECNP establishes the relationship with CNP while offering deeper insights on the roles of the evidential parameters. Extensive experiments conducted on both synthetic and real-world data demonstrate the effectiveness of our proposed model in various few-shot settings.

Introduction

Meta-learning (Finn, Abbeel, and Levine 2017) offers a powerful vehicle to tackle the challenges of learning from limited data. It formulates learning into two phases: meta-training that learns the global (meta) knowledge shared across tasks and meta-testing that adapts the global knowledge to the limited data from few-shot testing tasks. While meta-learning achieves improved generalization capability by leveraging the meta-knowledge obtained from the meta-training tasks, few-shot tasks arising in the testing phase may deviate significantly from the training tasks. Furthermore, data in many real-world applications may be highly noisy, incomplete, or corrupted. These, when coupled with the weakly supervised signal from limited training data, make few-shot learning inherently uncertain and challenging.

Among existing meta-learning models, metric-based approaches (Vinyals et al. 2016; Snell, Swersky, and Zemel 2017; Chen et al. 2021) have achieved high predictive accuracy for few-shot classification problems. However, most metric-based models are not designed to output uncertainty, limiting their applicability to many real-world problems. Meanwhile, gradient-based approaches, such as MAML (Finn, Abbeel, and Levine 2017), have been extended to achieve uncertainty-aware meta-learning through Bayesian modeling. MAML formulates meta-learning as a bi-level optimization problem that requires expensive Hessian-gradient products during meta-learning along with other challenges such as training stability. First order approximations and alternatives of MAML, such as Reptile (Nichol, Achiam, and Schulman 2018), require time consuming gradient based adaptation during inference limiting their applications. Extending such models for uncertainty quantification (Yoon et al. 2018) may further increase the computational costs.

Different from deep learning (DL) models, Gaussian Processes (GPs) (Williams and Rasmussen 2006) offer a principled way to quantify uncertainty. By combining Bayesian modeling and kernel methods, a GP outputs a distribution over functions, where the kernel serves as a fixed prior that determines the smoothness of the functions as a specific form of meta-knowledge.

However, GPs, in their original form, suffer from a high computational cost for inference. Their generalization capability may also be limited due to the restricted priors induced from the fixed kernel functions, lacking the flexibility to adapt to the training data. This also significantly hinders GPs from being used as an effective meta-learning model, which needs to encode the meta-knowledge learned from other tasks in support of few-shot learning from new tasks.

The recently developed conditional neural processes (CNPs) (Garnelo et al. 2018a), neural processes (NPs) (Garnelo et al. 2018b), and their extensions provide a suite of effective meta-learning models, which bring together the benefits of GP's uncertainty capabilities and the DL models' flexibility of adapting to the data. Besides offering better scalability, rapid inference, and competitive predictive performance (Kim et al. 2019; Gordon et al. 2020), these models also naturally quantify uncertainty by simulating a stochastic process like a GP. However, the current NP models are sensitive to the outliers in the training tasks and lack competitive performance. Alternatively, CNP family of models achieve strong performance but only capture the overall uncertainty for the prediction made on a target data point. They lack a systematic fine-grained quantification of the different sources of uncertainty. Simply, CNP based models approximate the predictive distribution on a target data point by predicting both the mean and variance of a Gaussian. However, the variance term itself does not offer deeper insight on the two distinct sources of uncertainty: (i) lack of knowledge by the model (epistemic) or (ii) noise inherent in the data (aleatoric). Identifying the source of uncertainty can offer effective means to improve the model training (*e.g.,* by collecting more training data or constructing more informative sets) and facilitate critical decision-making (*e.g.,* whether to include humans in the loop).

In this paper, we propose Evidential Conditional Neural Processes (ECNPs), which provide novel and nontrivial extensions to CNP family of models with principled uncertainty quantification and decomposition. Being an CNP, an ECNP inherits all attractive model behaviors from the CNP family, including competitive predictive performance and scalability. By integrating evidential learning, an ECNP replaces a simple Gaussian distribution of CNP models with a much richer hierarchical Bayesian structure that leads to a robust neural process model with accurate epistemic-aleatoric uncertainty decomposition capabilities without any additional computational overhead. Such decomposition allows us to separate uncertainty caused by the noise in the data and the model's lack of knowledge on the target data point when making a prediction. Our main contributions are:

- The integration of evidential learning with CNPs results in a novel family of evidential conditional neural processes that are robust to outliers in meta-training and provides fine-grained uncertainty decomposition, both of which are essential for few-shot learning.
- A thorough theoretical analysis on the proposed ECNPs establishes the relationship with CNPs while offering deeper insights on the roles of the evidential parameters and why ECNPs are more suitable for few-shot learning.

Related Works

We discuss existing works that are most relevant to the proposed evidential neural processes in this section. Some additional related works ((Jøsang 2016; Sensoy, Kaplan, and Kandemir 2018; Pandey and Yu 2022b; Kandemir et al. 2021; Charpentier et al. 2022; Kingma and Welling 2013)) are covered in the Appendix (Pandey and Yu 2022a).

Uncertainty-aware Meta-Learning. There have been increasing efforts (Yoon et al. 2018; Pandey and Yu 2022b; Gordon et al. 2018; Grant et al. 2018; Ravi and Beatson 2019) to develop meta-learning models that can quantify uncertainty. Uncertainty information can be achieved through an ensemble of a diverse set of meta-learning models as in Bayesian MAML (Yoon et al. 2018). Uncertainty can also be estimated by considering a hierarchical model for meta-learning and carrying out Bayesian inference. To this end, ABML (Ravi and Beatson 2019) considers a hierarchical Bayesian model and uses amortized variational inference across tasks to obtain the uncertainty information. LLAMA (Grant et al. 2018) shows MAML as inference in a hierarchical Bayesian model with empirical Bayes and uses Laplace approximation to obtain Gaussian distribution for the posterior distribution that effectively captures the uncertainty. PLATIPUS (Finn, Xu, and Levine 2018) extends MAML using amortized variational inference to learn a distribution over prior model parameters that captures the uncertainty. These meta-learning approaches are computationally expensive and may lack rapid inference capabilities.

Neural Process Family. Neural Process (NP)-based models (Garnelo et al. 2018a,b; Kim et al. 2019; Gordon et al. 2020) offer computationally efficient alternatives to existing uncertainty-aware meta-learning approaches as inference in NP is a computationally cheap forward pass through an encoder-decoder architecture. Generative Query Networks (GQN) (Eslami et al. 2018) can be seen as one of the earliest NP models that use a generation network and a query network to tackle scene representation and autonomous scene understanding problems. Conditional Neural Processes (CNP) (Garnelo et al. 2018a) generalize GQN using an encoder-aggregator-decoder architecture. Neural processes (Garnelo et al. 2018b) further generalize CNPs by introducing a latent variable in the encoder-decoder architecture. Attentive Neural processes (ANP) (Kim et al. 2019) replace the mean aggregation in CNP with multi-headed attention that learns to attend to the most relevant context points leading to significantly better target embedding and improved results at the cost of increased computational cost from the attention mechanism. Convolutional Conditional Neural Processes (ConvCNPs) (Gordon et al. 2020) achieve translation equivariance using a functional space representation for the context set. CNAPS (Requeima et al. 2019) and Simple CNAPS (Bateni et al. 2020) extend the NP models to handle few-shot classification tasks. Various evaluation metrics such as Inclusion@K and Uncertainty Increase (Grover et al. 2019) were introduced to better analyze the uncertainty capabilities of neural process models. Le *et al.* (Le et al. 2018) and Naderiparizi *et al.* (Naderiparizi et al. 2020) studied the impact of architecture choices and different optimization objective choices for NP and CNP models. GNP (Bruinsma et al. 2021) and FullConvGN (Markou et al. 2021) extended the CNP models to handle predictive correlations, i.e., the dependencies in output.

As discussed earlier, CNPs and their variants can only capture the overall predictive uncertainty on the target points. NPs recover the posterior distribution of the model after being exposed to the context points by introducing a global latent variable. However, NPs require approximation procedures and usually resort to computationally expensive sampling schemes for model training/inference. The proposed ECNPs address these critical gaps by integrating evidential learning with CNP models through an evidential hierarchical Bayesian prior with a much richer representation power to support fine-grained uncertainty decomposition while achieving robust predictions and being computationally efficient in few-shot settings.

Evidential Neural Processes

Problem Setup: Consider a meta-dataset $\mathcal{M} = \{\mathcal{D}^i\}_{i=1}^M$, which consists of a collection of datasets/tasks. Each task $\mathcal{D} = (\mathcal{C}, \mathcal{T}) = \{(x_n, y_n)\}_{n=1}^{N_c+N_t}$ consists of a context set (*a.k.a.,* support set) $\mathcal{C} = \{X_c, Y_c\} = \{(x_n, y_n)\}_{n=1}^{N_c}$, a col-

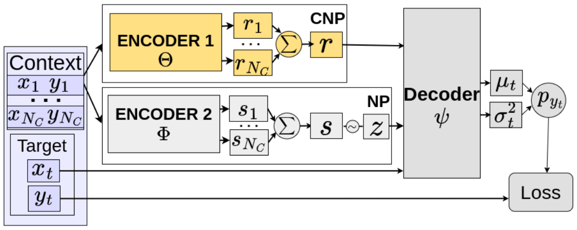


Figure 1: CNP and NP Models

lection of N_c input-output pairs, and the target set (*a.k.a.,* the query set) $\mathcal{T} = \{X_t, Y_t\} = \{(x_t, y_t)\}_{t=1}^{N_t}$ a collection of N_t input-output pairs. Meta-learning occurs in two phases: 1) meta-training where both the context set and target set information is available to the model, and 2) meta-testing where the model is provided with context set information and evaluated based on the performance over target set inputs.

Uncertainty Analysis via CNP and NP

A CNP, as shown by the top branch of Figure 1 , has a deterministic mapping from a context set $\mathcal{C}$ and a target input x_t to its prediction. Specifically, each context point is embedded by encoder Θ to representations $r_1, ..., r_{N_C}$, aggregated to a representation r, and passed through the decoder to obtain the parameters of the predictive distribution. The predictive distribution is assumed to be a Gaussian, where the decoder outputs the mean μ_t and the variance σ_t^2: $P_\theta(y_t|x_t; \mathcal{C}) = \mathcal{N}(y_t|\mu_t, \sigma_t^2)$. The variance term σ_t^2 captures the overall uncertainty. A NP model, as shown by the bottom branch of Figure 1, introduces a latent variable z to produce a distribution over functions given the same context set $\mathcal{C}$. Specifically, each context point is embedded by encoder Φ to representations $s_1, ..., s_{N_C}$, aggregated to a representation s, and passed through a NN to obtain the parameters for the latent distribution. The latent variable induced distribution over functions allows NPs to model both epistemic and aleatoric uncertainty for each target point prediction. Assume that the latent variable follows a distribution $q(z)$ and by sampling from this distribution, we obtain $z_{1,..,L} \sim q(z)$. For each sampled z_l and a target point x_t, the decoder outputs a predictive distribution $p(y_t|x_t, z_l) = \mathcal{N}(\mu_l, \sigma_l^2)$. As a result, we can obtain the epistemic uncertainty as the variance of the mean outputs $\text{Var}[\mu_l]$ and the aleatoric uncertainty as the expected variance $\mathbb{E}[\sigma_l^2]$.

However, there are two key limitations for NP-based uncertainty decomposition. First, it requires sampling of the latent variable, which may make the overall inference computationally expensive, especially when learning from a large number of tasks. Second, the fine-grained uncertainties are obtained indirectly (*e.g.,* via MC sampling) and thus it becomes challenging to guide the model to correct its inherent mistakes regarding fine-grained uncertainties during training. Moreover, both CNP and NP lack robustness to outliers in the training tasks. The proposed ECNP addresses the key limitations of both CNP and NPs. We present an evidential extension of CNP that outputs the aleatoric and epistemic uncertainty directly from the deterministic path while ensuring a robust prediction given a noisy context set. The introduced hierarchical structure explicitly captures fine-grained uncertainty enabling the model to correct its mistakes in the fine-grained uncertainties.

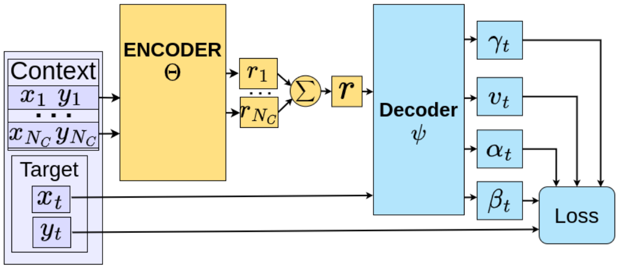


Figure 2: ECNP Model

Evidential Conditional Neural Process (ECNP)

We extend the CNP model to an evidential neural process. To this end, as in (Amini et al. 2020), we assume that the likelihood function is a Gaussian with an unknown mean and variance. We place an evidential prior over the mean and variance and train the neural process to output the hyperparameters of the evidential distribution using the limited information of the context set and target input. Moreover, we train the evidential model to be confident for the correct prediction and output low evidence (*i.e.,* confidence) when the model's predictions are incorrect. Our evidential conditional model introduces insignificant computational overhead and is deterministic while being expressive in uncertainty quantification. In particular, it can quantify both aleatoric and epistemic uncertainty with a single forward pass through the network without any sampling as in the NP.

Uncertainty and evidence quantification by ECNP. In the ECNP model, we consider a hierarchical Bayesian structure in which each target observation y_t is a sample from a Gaussian $\mathcal{N}(y_t|\mu, \sigma)$, whose mean and variance are governed by a higher-order Normal-Inverse-Gamma prior (Bishop and Nasrabadi 2006):

$$\text{NIG}(\mu, \sigma^2|\mathbf{p}_t) = \mathcal{N}(\mu|\gamma_t, \frac{\sigma^2}{v_t})\Gamma^{-1}(\sigma^2|\alpha_t, \beta_t) \qquad (1)$$

where $\mathbf{p}_t = (\gamma_t, v_t, \alpha_t, \beta_t)$ and Γ^{-1} is an inverse-gamma distribution. Intuitively, the context set $\mathcal{C}$ interacts with the meta knowledge in the meta-learning model to output the prior NIG parameters $\mathbf{p}_t = (\gamma_t, v_t, \alpha_t, \beta_t)$. When being exposed a new target point x_t, this prior will interact with the Gaussian likelihood $p(y_t|x_t) = \mathcal{N}(\mu, \sigma^2)$ to produce a Student-t predictive distribution given by

$$\begin{aligned} p(y_t|x_t, \mathbf{p}_t) &= \int_{\mu,\sigma^2} p(y_t|x_t, \mu, \sigma^2)\text{NIG}(\mu, \sigma^2|\mathbf{p}_t)\text{d}\mu\text{d}\sigma^2 \\ &= \frac{\Gamma(\alpha_t + \frac{1}{2})}{\Gamma(\alpha_t)}\sqrt{\frac{v_t}{2\pi\beta_t(1+v_t)}}\Big(1 + \frac{v_t(y_t - \gamma_t)^2}{2\beta_t(1+v_t)}\Big)^{-(\alpha_t + \frac{1}{2})} \\ &= \text{St}\Big(y_t; \gamma_t, \frac{\beta_t(1+v_t)}{v_t\alpha_t}, 2\alpha_t\Big) \qquad (2) \end{aligned}$$

As a result, the prediction for a target point x_t is

$$\hat{y}_t = \mathbb{E}_{p(y_t|x_t,\mathbf{p}_t)}[y_t] = \int y_t p(y_t|x_t,\mathbf{p}_t)\mathrm{d}y_t = \gamma_t \quad (3)$$

Given the predicted evidential parameters, the NIG distribution is fully characterized, which allows us to evaluate Var$[\mu]$ and $\mathbb{E}[\sigma^2]$ that can be used to quantify the aleatoric (AL) and epistemic (EP) uncertainty (Amini et al. 2020), respectively:

$$\mathtt{AL} = \mathbb{E}[\sigma^2] = \frac{\beta_t}{\alpha_t - 1}, \quad \mathtt{EP} = \mathrm{Var}[\mu] = \frac{\beta_t}{v_t(\alpha_t - 1)} \quad (4)$$

By leveraging the conjugacy between the NIG prior and the Gaussian likelihood, it can be shown that after interacting with N i.i.d. data samples, the posterior is still a NIG$(\mu, \sigma^2|\mathbf{p}_N)$, where $\mathbf{p}_N = (\gamma_N, v_N, \alpha_N, \beta_N)$ with

$$v_N = v + N, \quad \alpha_N = \alpha + \frac{N}{2} \quad (5)$$

Thus, both v and α can be naturally interpreted as the evidence (in the form of pseudo counts) to quantify the confidence on the prior mean and the prediction of a target data sample, respectively. Furthermore, β denotes the initial variance of the model and (4) shows that a large β leads to a low confidence in the model's prediction, which implies lack of evidence. By aggregating all evidence related parameters, ECNP is able to quantify the overall model confidence as

$$\mathcal{E}_t = v_t + \alpha_t + \frac{1}{\beta_t} \quad (6)$$

A more detailed posterior analysis on the hierarchical model for evidence quantification is provided in the Appendix.

Training ECNP. In this evidential framework, learning is formulated as an evidence acquisition process and the model is trained to maximize the likelihood of model evidence. Equivalently, we train the model to minimize the negative log-likelihood of the model given by

$$L_t^{\mathrm{NLL}} = \log \frac{\Gamma(\alpha_t)\sqrt{\frac{\pi}{v_t}}}{\Gamma(\alpha_t + \frac{1}{2})} - \alpha_t \log(2\beta_t(1 + v_t)) + \\ \big(\alpha_t + \frac{1}{2}\big) \log\big((y_t - \gamma_t)^2 v_t + 2\beta_t(1 + v_t)\big) \quad (7)$$

We further introduce an evidence regularization term to encourage the model to output low evidence/confidence when the predictions are incorrect:

$$\mathcal{L}_t^{\mathrm{R}} = |y_t - \gamma_t| \times \mathcal{E}_t \quad (8)$$

The regularization term L_t^{R} penalizes the evidence of highly confident wrong predictions. In other words, the model is trained to output a low value of v_t and α_t and high β_t values when the prediction is wrong leading to high uncertainty in the predictions. Finally, the model is expected to output high epistemic uncertainty at regions far from the observed context points as only the meta-knowledge is available for predictions at those points. To this end, a novel kernel-based regularization term is introduced as

$$\mathcal{L}_t^{\mathrm{KER}} = v_t \times D(x_t, \mathcal{C}) \quad (9)$$

where $D(x_t, \mathcal{C})$ is a distance function that measures the minimum Euclidean distance between the target point input x_t and the context set $\mathcal{C}$. When the target input is far away from the context set, this kernel loss dominates the overall loss leading to small v_t values and equivalently high epistemic uncertainty (EP).

The overall loss in the evidential model is the regularized sum of the model evidence loss, evidence regularization loss, and the kernel regularization loss:

$$\mathcal{L} = \sum_{t=1}^{N_t} \mathcal{L}_t^{\mathrm{NLL}} + \lambda_1 \mathcal{L}_t^{\mathrm{R}} + \lambda_2 L_t^{\mathrm{KER}} \quad (10)$$

where λ_1 and λ_2 are regularization terms.

Theoretical Analysis

In this section, we present our theoretical results that show the superiority of the ECNP and reveal the deeper connection between the proposed ECNPs and the NP models. These theoretical results help to justify why ECNPs provide a more principled way to conduct meta-learning over few-shot tasks than the CNP family of models.

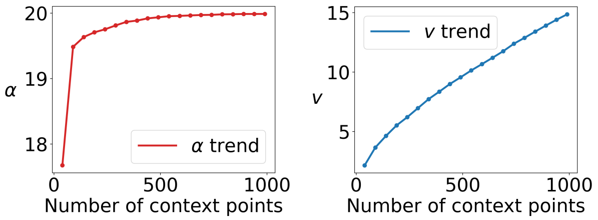
Theorem 1. *The ECNP model with a hierarchical Bayesian structure in the decoder is guaranteed to be more robust to outliers in the training tasks as compared to the CNP models that use a Gaussian structure.*

The detailed proof is provided in the Appendix. Intuitively, when the model evidence is finite (*i.e.,* $\alpha_t < \infty$), the outliers will be assigned a lower weight than normal data samples when evaluating the gradient for model update. When $\alpha_t \to \infty$, the model will behave similarly to the CNP model and be less robust to outliers. In our model, the hierarchical bayesian structure leads to the heavy tailed t predictive distribution enabling outlier robustness. Similar outlier robustness can, in theory, be introduced in the CNP models by modifying the CNP decoder to directly parameterize the heavy distributions (e.g. Student t distribution) and training to minimize the log likelihood under the new distribution. Empirical studies of such robust distributions for CNP models can be an interesting future work. However, such modeling would lack efficient and fine-grained uncertainty quantification capabilities, a major focus of our work.

Theorem 2. *The conditional neural process is one instance of an evidential neural process when two of the evidential hyperparameters meet the following conditions: (i) $\alpha_t \to \infty$; (ii) $\alpha_t v_t = const.$*

A detailed proof is provided in the Appendix.

Interpretation of evidential parameters. The theoretical results given above not only establish the important relationship between ECNPs and CNPs but also unveil some key insights on why ECNPs along with an evidential Bayesian hierarchical prior is fundamentally more suitable for meta-learning based few-shot learning. As discussed in Section above, both evidential parameters v_t and α_t can be interpreted as evidence of the model (in the form of pseudo counts). Meanwhile, the hierarchical structure of the NIG prior as defined in (1) indicates that v_t and α_t capture the



(a) Change of α

(b) Change of v

Figure 3: Evidence change on a complex task

evidence at different levels, where v_t corresponds to the evidence collected for the global knowledge in the form of the prior mean (*i.e.,* γ) whereas α_t provides evidence on the local knowledge in the form of the variance (*i.e.,* σ) on the per data sample level. Theorem 2 shows that CNP primarily focus on improving the local knowledge by allowing α_t to grow while keeping v_t very small (due to $\alpha_t v_t = \text{const}$). While this has the effect of using an uninformative prior (by assigning minimal evidence v_t to the prior mean), it misses the opportunity to incorporate useful global knowledge that can be obtained through meta-learning from other relevant tasks. While using an uninformative prior is encouraged in a regular learning setting with sufficient training data, it is inherently inadequate for the few-shot setting, where there is not enough labeled data to support model training.

By leveraging a more expressive Bayesian hierarchical structure, ECNPs effectively address the key limitations of the CNPs as outlined above. In particular, they allow the evidence v_t to grow with the global knowledge, which is particularly important for more complex few-shot tasks where the meta-knowledge could play a more critical role. Figure 3 shows the change in local (*i.e.,* α) and global (*i.e.,* v) evidence for different number of context points in a complex few-shot task (*i.e.,* image completion and details are provided in the experiment section). It is interesting to see that α grows fast and then shows a much slower increasing trend, which implies that the local knowledge may already reach the limit. On the other hand, v continues to grow, which indicates that adding new context points can help retrieve more relevant global knowledge acquired through meta-learning. Meanwhile, the prediction error also continues to decrease (see Figure 20 in the Appendix), which demonstrates effective knowledge transfer achieved by the ECNP model.

Experiments

Datasets. For function regression experiments, we consider two synthetic datasets i) sinusoidal function regression (Gondal et al. 2021), and ii) regression on sample functions from a Gaussian process (Garnelo et al. 2018a). The sinusoidal regression function is of the form $y = A\sin(x+\phi), A \in [0.1, 5.0], \phi \in [0, \pi]$ and $x \in [-5, 5]$ and the GP is defined by a squared-exponential kernel with length scale of 0.6, variance of 1.0 and $x \in [-2, 2]$. Each function regression task is defined by a K-shot context set with $K + u$ data points in the target set where $u \sim U(3, K)$, and $U(a, b)$ represents a uniform distribution in range (a, b). Moreover, the function regression models are trained for 30,000 meta-iterations using a batch of 8 tasks and evaluated on 2,000 test tasks. For Image completion experiments, we consider three benchmark datasets: MNIST (Deng 2012) CelebA (Liu et al. 2015), and Cifar10 (Krizhevsky, Hinton et al. 2009). The details of the benchmark datasets are summarized in Appendix Table 4. Image completion task is created by randomly selecting a subset of the set points (input-output pairs) from an image. Specifically, each position in the image grid is the input and the pixel value (*e.g.,* the RGB value) is the output. We randomly select 50 points to make the context set, use the remaining points in the image to make the target set, and train models for 50 epochs using a batch of 8 tasks, and evaluate the model on the test set.

Baselines. We consider three baseline models: Neural Processes (NP) (Garnelo et al. 2018b), Conditional Neural Processes (CNP) (Garnelo et al. 2018a), and the Attentive Neural Process (ANP) (Kim et al. 2019) For a fair comparison to the baselines, we consider the evidential equivalent of the baselines with the same encoder and decoder architectures. Specifically, for our evidential models, we consider two variants: i) ECNP: evidential model with deterministic path similar to CNP, and ii) ECNP-A: the evidential model with multi-head attention mechanism in encoder similar to ANP. Additional details of the model architecture and training are presented in the Appendix.

Performance Evaluation

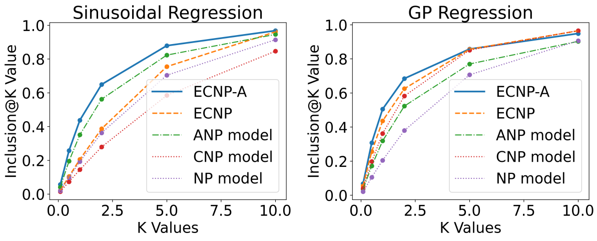
In this set of experiments, we report the generalization performance in terms of Mean Squared Error (MSE) along with three uncertainty based evaluation metrics: Log Likelihood (LL), Inclusion @K, and Uncertainty-Increase (Grover et al. 2019) for all the models on function regression and image completion tasks. We consider Inclusion@K with $K = 1$ in Table 1 and Table 2. Inclusion and Uncertainty-Increase have been developed to analyze and compare the uncertainty estimates of NP based models. Additional details along with comparisons are presented in the Appendix. We also empirically verify their robustness to outliers for function regression and image completion tasks. Limited by space, we present ablation studies in the Appendix.

Function regression. In the function regression problem, the model has to learn the underlying function based on the limited information of the context set and the meta-knowledge. Table 1 shows the results for 5-shot regression experiments. Our model improves the generalization performance compared to the the corresponding baseline model across almost all the datasets. Moreover, when considering the uncertainty metrics, as shown in Table 1 and Figure 4, our model considerably improves over the baselines.

2D image completion. We consider image completion experiments similar to (Eslami et al. 2018), where the model needs to infer the underlying function $f : [0, 1]^2 \rightarrow [0, 1]^{ch}$ ($ch-$ number of channels) to make prediction for each image pixel position in the target set given the context set. Table 2 compares our model with the baselines for 50-shot experiments. As can be seen, our model leads to comparable to improved performance than corresponding baselines

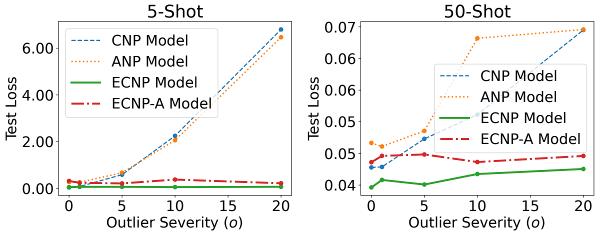
Dataset:	Sinusoidal Regression			GP Regression		
Model	MSE(↓)	Inclusion@K(↑)	Unc. Increase(↑)	MSE(↓)	Inclusion@K(↑)	Unc. Increase(↑)
NP	0.1050±0.0200	0.192 ± 0.040	0.563 ± 0.005	0.348±0.0116	0.205 ± 0.019	0.686 ± 0.007
CNP	0.0458±0.0074	0.144 ± 0.025	0.590 ± 0.006	0.3158±0.0038	0.362 ± 0.027	0.783 ± 0.002
ANP	0.3561±0.1084	0.351 ± 0.046	0.785 ± 0.048	0.3219±0.0124	0.318 ± 0.014	**0.875 ± 0.026**
ECNP	**0.0391±0.0078**	0.205 ± 0.018	0.608 ± 0.013	**0.3084±0.0014**	0.435 ± 0.02	0.798 ± 0.003
ECNP-A	0.2932±0.0956	**0.437 ± 0.044**	**0.814 ± 0.030**	0.3258±0.0162	**0.505 ± 0.038**	**0.875 ± 0.042**

Table 1: Comparison on 5-Shot Regression Problem



(a) Sinusoidal Regression (b) GP Regression

Figure 4: Impact of K to Inclusion@K



(a) Sinusoid (b) MNIST

Figure 5: Outlier robustness performance

in terms of MSE and log likelihood. We compare the uncertainty behavior (using Inclusion@K ($K = 1$) and uncertainty increase) of the representative CNP and the corresponding ECNP models in Table 3. Additional results are presented in the Appendix.

As can be seen in function regression and 2D image completion experiments, our model has better uncertainty characteristics than the baselines which is mainly due to the fine-grained and accurate uncertainty guidance capabilities in our hierarchical model. Our model explicitly captures the aleatoric and epistemic uncertainties through the evidence parameters (β, α, v). Furthermore, the model is guided during training to have accurate overall uncertainty via the evidence regularization in (8), and accurate epistemic uncertainty from the kernel regularization in (9). Such uncertainty guidance leads to more accurate uncertainty performance in our model. These results empirically validate our model's generalization performance and superiority over other comparison baselines.

Outlier robustness. Due to the hierarchical Bayesian structure leading to a heavy tailed predictive distribution, our model is theoretically guaranteed to be robust to outliers in the training tasks. Here, we empirically validate the claim by experimenting with 5-shot sinusoidal regression and 50-shot MNIST image completion (results on other datasets and settings are presented in the Appendix).

To make the noisy training task, we randomly select one target point in all the training tasks and apply an additive transformation $y_t = y_t + o$ to make it an outlier (here o determines the outlier severity). We train the models on the noisy tasks (*i.e.,* tasks with outlier), and after training, we evaluate on clean test set tasks. Figure 5 (a)-(b) shows the comparison results of the ECNP and ECNP-A models with their corresponding baselines of CNP and ANP models. Across both experiments, ECNP models remain robust to the outlier as their test set performance remains relatively unaffected even for severe outliers. Such outlier robustness in our model can be attributed to the heavy tailed predictive distribution that is inherently introduced by the hierarchical structure. In comparison, the baseline models lack the required robustness characteristics and their performance degrades severely as the outlier becomes extreme. Such baseline models may require additional mechanisms to handle the outliers, something our model can automatically do. These results empirically validate the robustness superiority for the proposed ECNP model.

Effectiveness of Uncertainty Decomposition

In this set of experiments, we show that the proposed ECNP models can capture fine-grained uncertainty to best support few-shot learning through epistemic-aleatoric (EP-AL) uncertainty decomposition that can enable active context set construction and effective meta-knowledge transfer.

EP-AL decomposition. Our proposed model can perform Epistemic-Aleatoric uncertainty decomposition for any test task. Here, we compare the predicted uncertainty for the proposed ECNP model with the respective CNP baseline in sinusoidal regression task. Both models are trained for 20,000 iterations using training tasks with data in range $[-5, 5]$. As shown in Figure 6 (c)-(d), outside the training range (*i.e.,* $x_t \in [5, 10]$), prediction from both CNP and ECNP is inaccurate as expected. The CNP model continues to remain confident in regions far from the data whereas our ECNP model correctly outputs high epistemic uncertainty in the regions far away from the observed data.

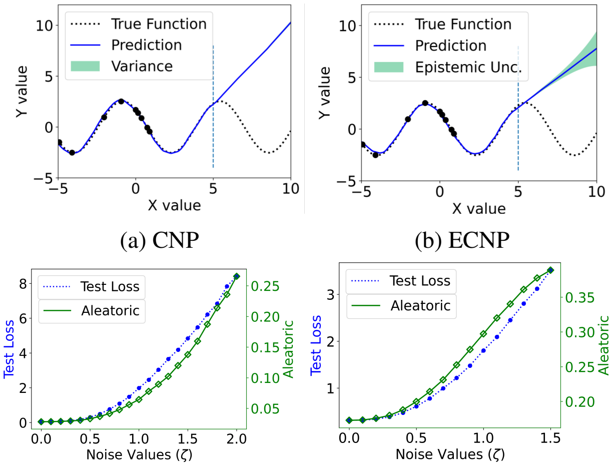
Next, we experiment with noisy test tasks to analyze the aleatoric uncertainty of our proposed model. We consider a model trained on clean 5-shot regression tasks and evaluate on 5-shot noisy test tasks. Specifically, we add random Gaussian noise to the context set of the test tasks ($y_c = y_c + \zeta\epsilon, \epsilon \sim \mathcal{N}(0, 1)$) and vary the level of noise (*i.e.,* ζ) to study the model behavior. Figure 6 visualizes the impact of the noise on the predicted performance (MSE) and the

Dataset	MNIST		Cifar10		CelebA	
Model	MSE(↓)	LL(↑)	MSE(↓)	LL(↑)	MSE(↓)	LL(↑)
NP	0.048±0.001	0.538±0.010	0.027±0.000	0.434±0.003	0.025±0.000	0.433±0.006
CNP	0.044±0.001	0.710±0.009	0.023±0.000	0.576±0.005	0.021±0.001	0.660±0.004
ANP	0.045±0.001	0.702±0.007	0.017±0.000	**0.765**±0.004	**0.014**±0.000	0.850±0.002
ECNP	**0.041**±0.002	**0.734**±0.014	0.022±0.001	0.601±0.004	0.020±0.000	0.694±0.004
ECNP-A	0.043±0.001	0.713±0.013	**0.016**±0.001	0.764±0.004	**0.014**±0.000	**0.852**±0.002

Table 2: Comparison on 50-Shot Image Completion Problems

Metric: **Inclusion@K** (↑)		
Dataset	CNP model	ECNP model
MNIST	0.622 ± 0.001	**0.828 ± 0.000**
Cifar10	0.129 ± 0.005	**0.144 ± 0.003**
CelebA	0.133 ± 0.004	**0.156 ± 0.003**
Metric: **Uncertainty Increase** (↑)		
Dataset	CNP model	ECNP model
MNIST	0.306 ± 0.000	**0.524 ± 0.000**
Cifar10	0.505 ± 0.005	**0.531 ± 0.004**
CelebA	0.519 ± 0.003	**0.541 ± 0.003**

Table 3: Comparison of CNP and ECNP models



(a) CNP

(b) ECNP

(c) Sinusoidal Regression

(d) GP Regression

Figure 6: (a)-(b) ECNP vs. CNP on a sinusoidal task; (c)-(d) ECNP performance for noisy test tasks

model's predicted aleatoric uncertainty for two datasets averaged across 2000 test tasks. As expected, the model's predictive accuracy decreases as tasks become more noisy. Our proposed model accurately identifies the noisy tasks and outputs more aleatoric uncertainty as tasks become more noisy showing the effectiveness of our model's predicted aleatoric uncertainty in identifying noisy tasks.

Active context set construction. The proposed ECNP model can capture both aleatoric and epistemic uncertainty in a single forward pass. We investigate the effectiveness of the captured epistemic uncertainty in a context point selection experiment. We randomly select a test image and both models start with random 10 context points indicated by the context (Figure 7), which represents the pixel positions that are included in the context set. CNP model and ECNP model randomly select the next 100 context points. ECNP-Active model iteratively queries the epistemic uncertainty for different target positions and includes the queried data with the greatest epistemic uncertainty in the context set for the next iteration. By including the most informative points in the context set using the epistemic uncertainty information, ECNP-Active performs significantly better than the other models (Figure 7 (b)) illustrating the effectiveness of our proposed model's uncertainty.

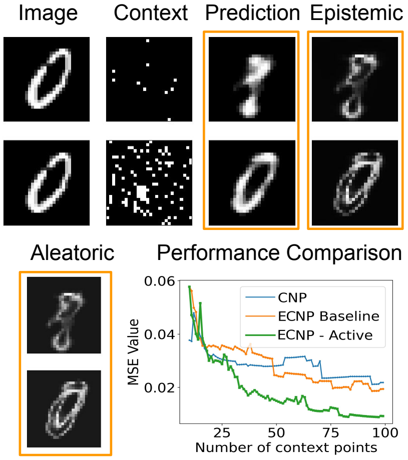


Figure 7: Active context set construction

Ablations Study and Additional Experiments

We carry out a detailed ablation study to investigate some key model parameters. The results along with some additional illustrative examples are presented in the Appendix.

Conclusion

We propose evidential conditional neural processes, that can conduct epistemic-aleatoric uncertainty decomposition in few-shot learning. ECNPs introduce a hierarchical Bayesian structure to replace the standard Gaussian distribution. The hierarchical bayesian structure enables the model to quantify fine-grained uncertainty in an efficient way. Moreover, our theoretical results reveal a deep connection with the CNP models and further justify why a richer hierarchical structure provides a more principled way to capture the meta-knowledge through higher-order priors, making it fundamentally more suitable for meta-learning over few-shot tasks. Experiments over various 1D regression and 2D image completion tasks demonstrate the superiority of our proposed model and its uncertainty capabilities.

Acknowledgements

This research was supported in part by an NSF IIS award IIS-1814450 and an ONR award N00014-18-1-2875. The views and conclusions contained in this paper are those of the authors and should not be interpreted as representing any funding agency.

References

Amini, A.; Schwarting, W.; Soleimany, A.; and Rus, D. 2020. Deep evidential regression. *Advances in Neural Information Processing Systems*, 33: 14927–14937.

Bateni, P.; Goyal, R.; Masrani, V.; Wood, F.; and Sigal, L. 2020. Improved few-shot visual classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 14493–14502.

Bishop, C. M.; and Nasrabadi, N. M. 2006. *Pattern recognition and machine learning*, volume 4. Springer.

Bruinsma, W. P.; Requeima, J.; Foong, A. Y.; Gordon, J.; and Turner, R. E. 2021. The Gaussian neural process. *arXiv preprint arXiv:2101.03606*.

Charpentier, B.; Borchert, O.; Zügner, D.; Geisler, S.; and Günnemann, S. 2022. Natural Posterior Network: Deep Bayesian Predictive Uncertainty for Exponential Family Distributions. In *International Conference on Learning Representations*.

Chen, Y.; Liu, Z.; Xu, H.; Darrell, T.; and Wang, X. 2021. Meta-Baseline: Exploring Simple Meta-Learning for Few-Shot Learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 9062–9071.

Deng, L. 2012. The mnist database of handwritten digit images for machine learning research [best of the web]. *IEEE signal processing magazine*, 29(6): 141–142.

Eslami, S. A.; Rezende, D. J.; Besse, F.; Viola, F.; Morcos, A. S.; Garnelo, M.; Ruderman, A.; Rusu, A. A.; Danihelka, I.; Gregor, K.; et al. 2018. Neural scene representation and rendering. *Science*, 360(6394): 1204–1210.

Finn, C.; Abbeel, P.; and Levine, S. 2017. Model-agnostic meta-learning for fast adaptation of deep networks. In *International Conference on Machine Learning*, 1126–1135. PMLR.

Finn, C.; Xu, K.; and Levine, S. 2018. Probabilistic model-agnostic meta-learning. *Advances in neural information processing systems*, 31.

Garnelo, M.; Rosenbaum, D.; Maddison, C.; Ramalho, T.; Saxton, D.; Shanahan, M.; Teh, Y. W.; Rezende, D.; and Eslami, S. A. 2018a. Conditional neural processes. In *International Conference on Machine Learning*, 1704–1713. PMLR.

Garnelo, M.; Schwarz, J.; Rosenbaum, D.; Viola, F.; Rezende, D. J.; Eslami, S.; and Teh, Y. W. 2018b. Neural processes. *arXiv preprint arXiv:1807.01622*.

Gondal, M. W.; Joshi, S.; Rahaman, N.; Bauer, S.; Wuthrich, M.; and Schölkopf, B. 2021. Function Contrastive Learning of Transferable Meta-Representations. In *International Conference on Machine Learning*, 3755–3765. PMLR.

Gordon, J.; Bronskill, J.; Bauer, M.; Nowozin, S.; and Turner, R. E. 2018. Meta-learning probabilistic inference for prediction. *arXiv preprint arXiv:1805.09921*.

Gordon, J.; Bruinsma, W. P.; Foong, A. Y. K.; Requeima, J.; Dubois, Y.; and Turner, R. E. 2020. Convolutional Conditional Neural Processes. In *International Conference on Learning Representations*.

Grant, E.; Finn, C.; Levine, S.; Darrell, T.; and Griffiths, T. 2018. Recasting Gradient-Based Meta-Learning as Hierarchical Bayes. In *International Conference on Learning Representations*.

Grover, A.; Tran, D.; Shu, R.; Poole, B.; and Murphy, K. 2019. Probing Uncertainty Estimates of Neural Processes. http://bayesiandeeplearning.org/2019/papers/125.pdf. Accessed: 2022-01-01.

Jøsang, A. 2016. *Subjective logic*. Springer.

Kandemir, M.; Akgül, A.; Haussmann, M.; and Unal, G. 2021. Evidential Turing Processes. *arXiv preprint arXiv:2106.01216*.

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

Kingma, D. P.; and Welling, M. 2013. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*.

Krizhevsky, A.; Hinton, G.; et al. 2009. Learning multiple layers of features from tiny images. https://www.cs.toronto.edu/~kriz/cifar.html. Accessed: 2022-01-01.

Le, T. A.; Kim, H.; Garnelo, M.; Rosenbaum, D.; Schwarz, J.; and Teh, Y. W. 2018. Empirical evaluation of neural process objectives. In *NeurIPS workshop on Bayesian Deep Learning*, volume 4.

Liu, Z.; Luo, P.; Wang, X.; and Tang, X. 2015. Deep Learning Face Attributes in the Wild. In *Proceedings of International Conference on Computer Vision (ICCV)*.

Markou, S.; Requeima, J.; Bruinsma, W. P.; and Turner, R. E. 2021. Efficient Gaussian Neural Processes for Regression. *CoRR*, abs/2108.09676.

Naderiparizi, S.; Chiu, K.; Bloem-Reddy, B.; and Wood, F. 2020. Uncertainty in Neural Processes. *arXiv preprint arXiv:2010.03753*.

Nichol, A.; Achiam, J.; and Schulman, J. 2018. On first-order meta-learning algorithms. *arXiv preprint arXiv:1803.02999*.

Pandey, D. S.; and Yu, Q. 2022a. Evidential Conditional Neural Processes. *arXiv preprint arXiv:2212.00131*.

Pandey, D. S.; and Yu, Q. 2022b. Multidimensional Belief Quantification for Label-Efficient Meta-Learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 14391–14400.

Ravi, S.; and Beatson, A. 2019. Amortized Bayesian Meta-Learning. In *International Conference on Learning Representations*.

Requeima, J.; Gordon, J.; Bronskill, J.; Nowozin, S.; and Turner, R. E. 2019. Fast and flexible multi-task classification using conditional neural adaptive processes. *Advances in Neural Information Processing Systems*, 32: 7959–7970.

Sensoy, M.; Kaplan, L.; and Kandemir, M. 2018. Evidential deep learning to quantify classification uncertainty. *Advances in neural information processing systems*, 31.

Snell, J.; Swersky, K.; and Zemel, R. 2017. Prototypical networks for few-shot learning. *Advances in neural information processing systems*, 30.

Vinyals, O.; Blundell, C.; Lillicrap, T.; Wierstra, D.; et al. 2016. Matching networks for one shot learning. *Advances in neural information processing systems*, 29: 3630–3638.

Williams, C. K.; and Rasmussen, C. E. 2006. *Gaussian processes for machine learning*, volume 2. MIT press Cambridge, MA.

Yoon, J.; Kim, T.; Dia, O.; Kim, S.; Bengio, Y.; and Ahn, S. 2018. Bayesian model-agnostic meta-learning. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, 7343–7353.