Appendix

Organization of Appendix. We first present the proof of the theoretical results presented in the main paper. We then present the analysis of the hierarchical Bayesian structure introduced in the ECNP model that extends evidential learning to meta-learning models and discuss additional relevant works in evidential learning. Next, we provide the details of the datasets, experimental settings, model architectures, and discuss complexity. We then carry out an ablation study to demonstrate the effect of different parameters. We then provide additional experimental results across different datasets and settings that demonstrate the effectiveness of the proposed evidential model. Finally, we discuss some limitations of the work that we aim to address in our future works. The codes for all the experiments in this paper can be found at this link[1].

Proofs of Theoretical Results

In this section, we show the proofs of the theoretical results presented in the main paper.

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.

Definition 3. *Outlier:* Consider a task defined by an underlying generating function $f(\cdot)$. A data point (x_o, y_o) in the target set of the task is an outlier with severity os if the value of the output y_o deviates from the ground truth value y_{true} with a margin of os. i.e $|y_{true} - y_o| > os$, $y_{true} = f(x_o)$.

Proof. Consider we have a task with the context set $\mathcal{C}$, and N_t input output pairs of the form (x_t, y_t) in the target set. The meta-learning model has to be able to correctly predict for each target set input x_t after learning from the context set $\mathcal{C}$.

Consider a CNP model that outputs the mean γ_t and variance s_t^2 for a target input x_t given the context set $\mathcal{C}$. Consider the model parameters ψ_γ output the prediction for the target set i.e. $f_{\psi_\gamma}(x_t|\mathcal{C}) = \gamma_t$. Now, for a task with N_t points in the target set, the Maximum Likelihood estimate of the model parameters ψ_γ is given by

$$\frac{\partial \mathcal{L}}{\partial \psi_\gamma} = \frac{\partial\big(\sum_{t=1}^{N_t} L_t^{NLL}\big)}{\partial \psi_\gamma} \tag{11}$$

$$= \frac{\partial\big(\sum_{t=1}^{N_t} -\log \mathcal{N}(y_t|\gamma_t, s_t^2)\big)}{\partial \psi_\gamma} = 0 \tag{12}$$

$$or, \sum_{t=1}^{N_t} s_t^{-2}(y_t - \gamma_t)\frac{\partial \gamma_t}{\partial \psi_\gamma} = 0 \tag{13}$$

Consider the ECNP model that outputs the prediction γ_t along with the evidential parameters v_t, α_t and β_t leading to scale parameter s_t and $2\alpha_t$ degrees of freedom. Consider the model is trained without regularization and assume that the model parameters ψ_γ output the prediction γ_t are i.e. $f_{\psi_\gamma}(x_t|\mathcal{C}) = \gamma_t$. Now, for a task with N_t points in the target set, the Maximum Likelihood estimate of the model parameters ψ_γ is given by

$$\frac{\partial \mathcal{L}}{\partial \psi_\gamma} = \frac{\partial\big(\sum_{t=1}^{N_t} L_t^{NLL}\big)}{\partial \psi_\gamma} = 0 \tag{14}$$

$$or, \frac{\partial\big(\sum_{t=1}^{N_t} -\log St(y_t|\gamma_t, s_t, 2\alpha_t)\big)}{\partial \psi_\gamma} = 0 \tag{15}$$

$$or, \sum_{t=1}^{N_t} \frac{\partial\Big((\alpha_t + \frac{1}{2})\log(2\alpha_t + s_t^{-2}(y_t - \gamma_t)^2\Big)}{\partial \gamma_t}\frac{\partial \gamma_t}{\partial \psi_\gamma} = 0 \tag{16}$$

$$or, \sum_{t=1}^{N_t} \frac{(2\alpha_t + 1)}{(2\alpha_t + \delta_t^2)} s_t^{-2}(y_t - \gamma_t)\frac{\partial \gamma_t}{\partial \psi_\gamma} = 0 \tag{17}$$

$$or, \sum_{t=1}^{N_t} w_t s_t^{-2}(y_t - \gamma_t)\frac{\partial \gamma_t}{\partial \psi_\gamma} = 0 \tag{18}$$

where $\delta_t^2 = s_t^{-2}(y_t - \gamma_t)^2$ is the Mahalanobis distance between the prediction and the ground truth, and $w_t = \frac{(2\alpha_t+1)}{(2\alpha_t+\delta_t^2)}$ is the outlier dependent scaling factor. As the outliers in the target set of training tasks become more extreme, δ_t^2 increases, outlier scaling factor w_t decreases proportionally for ECNP model to down-weight the impact of the outliers in estimation of the model parameters, effectively enabling the ECNP model to be robust to outliers. □

Remark 1. The ECNP model is least robust to outliers when the ECNP model realizes the CNP model, *i.e.,* $2\alpha_t \rightarrow \infty$, & $\alpha_t v_t = const \implies w_t = 1$.

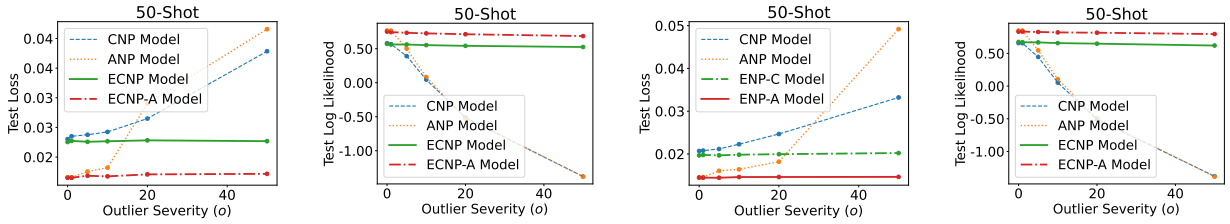
Remark 2. The robustness ECNP model and the maximum likelihood estimate of the parameters ψ_γ are unaffected by the proposed kernel based regularization.

Impact to generalization: As shown by Theorem 1, ECNP is robust to outliers in training tasks. The robustness ensures that the model learns from true signals, avoiding outliers, which is expected to improve generalization compared to a less robust model.

Empirical Validation

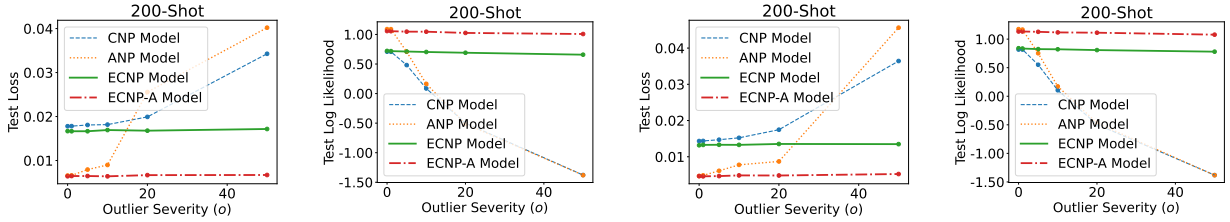
We carry out experiments with Cifar10 and CelebA datasets across 50-shot and 200-shot settings (Figure 8 and Figure 9) to empirically validate above theoretical claims. We consider the CNP and ANP models as the baselines and compare with their evidential extensions: ECNP and ECNP-A. For all the evidential models, we set both λ_1 and λ_2 to 0.1. In absence of any outliers in the training tasks, our evidential model shows comparable to marginally better performance. As the outlier in training tasks become more extreme, the baseline models start to break down and their performance degrades significantly, In comparison, our model continues to remain robust to outliers for different severity level across all datasets and settings. Experiments clearly demonstrate superiority of our proposed model for outlier robustness.

[1] Source codes:https://github.com/pandeydeep9/ECNP



(a) Cifar10-MSE (b) Cifar10-LL (c) CelebA-MSE (d) CelebA-LL

Figure 8: Impact of outlier to NP based models for different 50-shot image completion tasks.



(a) Cifar10-MSE (b) Cifar10-LL (c) CelebA-MSE (d) CelebA-LL

Figure 9: Impact of outlier to NP based models for different 200-shot image completion tasks.

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) $\alpha_t v_t = const$.

Proof. Let $\frac{v_t \alpha_t}{\beta_t(1+v_t)} = k$. Now, consider one instance of our model when $2\alpha_t \to \infty$ and $v_t \alpha_t$ is a constant. The condition $v_t \alpha_t = const$ can easily be satisfied *e.g.,* by setting $v_t = \frac{1}{\alpha_t}$. In this case, $k \to \frac{1}{\beta}$ and we get

$$p(y_t|x_t, \mathbf{p}_t) \propto \left(1 + \frac{k(y_t - \gamma_t)^2}{2\alpha_t}\right)^{-(\alpha_t+1/2)} \tag{19}$$

$$= e^{-\frac{k(y_t-\gamma_t)^2}{2} + O(\frac{1}{2\alpha_t})} \tag{20}$$

The predictive distribution is an exponential quadratic function w.r.t. y_t, which gives rise to a Gaussian $y_t \sim \mathcal{N}(y_t|\gamma_t, \beta_t^{-1})$. This matches the predictive distribution output by a CNP model. □

Posterior Analysis of the Hierarchical Bayesian Model for Evidence Quantification

In the ECNP, we assume a hierarchical Bayesian model in which each observation y_n is a sample from a Gaussian with unknown mean and unknown variance, with a higher-order Normal-Inverse-Gamma prior $\text{NIG}(\mu, \sigma^2|\mathbf{p})$ over the Gaussian likelihood function

$$y_n \sim \mathcal{N}(\mu, \sigma^2) \tag{21}$$

$$\mu \sim \mathcal{N}(\mu|\gamma_p, \sigma^2 v_p^{-1}), \qquad \sigma^2 \sim \Gamma^{-1}(\sigma^2|\alpha_p, \beta_p) \tag{22}$$

$$\text{NIG}(\mu, \sigma^2|\mathbf{p}) = \mathcal{N}(\mu|\gamma_p, \frac{\sigma^2}{v_p})\Gamma^{-1}(\sigma^2|\alpha_p, \beta_p) \tag{23}$$

where $\mathbf{p} = (\gamma_p, v_p, \alpha_p, \beta_p)$ represents the parameters of the NIG distribution and Γ^{-1} represents the inverse-gamma distribution. The evidential hyperparameters are governed by data observations where each new observation contributes to the model's predictive behavior.

To illustrate the model behavior, let us assume that we observe N i.i.d. data points $Y_N = \{y_1, y_2, ...y_N\}$ and study the impact of the observations on our hierarchical evidential model. Due to the conjugacy between the prior and the likelihood, the posterior is also an NIG distribution:

$$p(\mu, \sigma^2|Y_N) \propto \text{NIG}(\mu, \sigma^2|\mathbf{p}) \prod_{n=1}^{N} \mathcal{N}(y_n|\mu, \sigma^2) \tag{24}$$

This posterior factorizes as $p(\mu, \sigma^2|Y_N) = p(\mu|\sigma^2, Y_N)p(\sigma^2|Y_N)$. The conditional posterior of the mean $(p(\mu|\sigma^2, Y_N))$ is

$$p(\mu|\sigma^2, Y_N) = \mathcal{N}(\mu|\gamma_N, \sigma^2 v_N^{-1}) \tag{25}$$

where the parameters of the conditional distribution are given by

$$v_N = v_p + N \tag{26}$$

$$\gamma_N = \frac{v_p}{v_N}\gamma_p + \frac{N}{v_N}\bar{y}_N, \quad \bar{y}_N = \frac{1}{N}\sum_{n=1}^{N} y_n \tag{27}$$

Here, v_N effectively serves as the evidence for the observations: as we collect more data, v_N increases leading to reduced variance and more confident predictions. The posterior over the variance $(p(\sigma^2|Y_N))$ is an inverse-gamma dis-

tribution of the form

$$p(\sigma^2|Y_N) = \int p(\mu, \sigma^2|Y_N)\mathrm{d}\mu \tag{28}$$

$$\propto_{\sigma^2} \int \prod_{n=1}^{N} \mathcal{N}(y_n|\mu, \sigma^2)\mathcal{N}(\mu|\gamma_p, \frac{\sigma^2}{v_p})\Gamma^{-1}(\sigma^2|\alpha_p, \beta_p)\mathrm{d}\mu \tag{29}$$

$$= \Gamma^{-1}(\sigma^2|\alpha_N, \beta_N) \tag{30}$$

where the parameters of the posterior are

$$\alpha_N = \alpha_p + \frac{N}{2} \tag{31}$$

$$\beta_N = \beta_p + \frac{1}{2}\sum_{n=1}^{N}(y_n - \bar{y}_N)^2 + \frac{N v_p}{2(v_p + N)}(\bar{y}_N - \gamma_p)^2 \tag{32}$$

The parameters α_N and β_N contribute to the model's confidence (*i.e.,* the model evidence) indirectly through the higher-order Inverse gamma (IG) distribution. The expected value of σ^2 is $\frac{\beta_N}{\alpha_N - 1}$. When α_N is high and β_N is small, the IG samples $\sigma^2 \sim \Gamma^{-1}(\alpha_N, \beta_N)$ are close to zero indicating low variance and high confidence in prediction. Conversely, when α_N is small and β_N increases, the variance σ^2 increases indicating low confidence in the prediction. Based on the above analysis, we define the evidence for the prediction ($\mathcal{E}$) in the evidential hierarchical model as

$$\mathcal{E}_N = v_N + \alpha_N + \frac{1}{\beta_N} \tag{33}$$

In this hierarchical Bayesian model, the N training data observations interact with the prior distribution NIG(μ, σ^2) to output the hyperparameters for the NIG posterior. Equivalently, in the proposed evidential conditional neural processes as shown in Figure 2 right, the context set $\mathcal{C}$ interacts with the meta knowledge in the meta-learning model to output the posterior NIG parameters $\mathbf{p}_t = (\gamma_t, v_t, \alpha_t, \beta_t)$ for a target input x_t. The parameter γ_t corresponds to the prediction, and the remaining NIG parameters work together to quantify the aleatoric uncertainty, the epistemic uncertainty, and the evidence for the prediction.

Related works on Evidential Deep Learning

In Evidential Deep Learning models, ideas from Subjective Logic (Jøsang 2016) are used to equip Deep Learning models with accurate uncertainty quantification capabilities. Evidential Deep learning has been extended to both classification and regression problems. EDL(Sensoy, Kaplan, and Kandemir 2018) introduces higher-order evidential Dirichlet prior for the multinomial likelihood in classification problems that enables the deterministic neural network model to capture different uncertainty characteristics. Units-ML (Pandey and Yu 2022b) extends EDL for few-shot classification. ETP (Kandemir et al. 2021), an improvement on EDL, develops an uncertainty-aware classification model by integrating parametric Bayesian and evidential Bayesian model into a complete Bayesian model that addresses the issue of total calibration in classification. DER (Amini et al. 2020) extends evidential learning to regression problems by introducing a NIG prior for the Gaussian likelihood that leads to effective aleatoric-epistemic uncertainty quantification. NatPN (Charpentier et al. 2022) develops an unified evidential deep learning model for both classification and regression by introducing exponential family distributions as effective prior distributions. Compared to the above evidential works, our work can be seen as a novel extension of DER work (Amini et al. 2020) to the meta-learning setting that enables fine-grained uncertainty quantification in the few-shot regression tasks.

Details of Datasets and Experimental Settings

In this work, we consider two synthetic regression experiments (sinusoidal regression and GP) and three real-world benchmark datasets for image completion experiments: MNIST, Cifar10, and CelebA. For the synthetic regression experiments, we consider K-shot tasks with additional u samples (i.e. effectively $K + u$ samples where $u \sim U(3, K)$, represents sampling from a uniform distribution in range $(3, K)$) in the target set of training tasks and 400 samples in the target set of test tasks. For image completion experiments, we consider K random pixel positions ($K = 50/200$) in the context set and all the pixel positions in the target set. The details of the image datasets are presented in Table 4.

Table 4: Dataset Details

Characteristic	MNIST	CelebA	Cifar10
Image Size	28×28	32*32	32×32
Channels (ch)	1	3	3
Training Images	60,000	162,770	50,000
Test Images	10,000	19,962	10,000

Details of Uncertainty Metrics

In this work, we use Mean Squared Error (MSE) and Log Likelihood (LL) to compare the generalization performance; and consider Inclusion@K and Uncertainty-Increase (Grover et al. 2019) to evaluate the uncertainty estimates of the models.

Inclusion@K(I(k)) is defined as:

$$I(k) = \mathbb{E}_{x\sim\text{Uniform}(\mathcal{X})}[\mathbb{I}(|f(x) - m(x, \mathcal{S})| < ks(x, \mathcal{S})] \tag{34}$$

where the model outputs the prediction $m(x, \mathcal{S})$ and uncertainty $s(x, \mathcal{S})$ for the task $\mathcal{T}$ defined by the true function $f(.)$. The task has support set $\mathcal{S}$, query set $\mathcal{Q}$, and $\mathcal{X}$ represents the entire set of inputs in the task. To compute inclusion, we consider the variance of the predictive distribution as the uncertainty $s(x, \mathcal{S})$. Moreover, we consider query set inputs for $\mathcal{X}$.

Uncertainty-Increase(UI) is defined as:

$$UI = \frac{\sum_{x\sim\text{Uniform}(\mathcal{X})} \mathbb{I}(s(x, \mathcal{S}) - NN(x, \mathcal{S}))}{|\mathcal{X}|} \tag{35}$$

where $NN(x, \mathcal{S})$ represents the uncertainty of the datapoint in support set S that is closest to the datapoint x. In our experiments, we consider query set inputs for $\mathcal{X}$ and $|\mathcal{X}|$ represents the number of datapoints in the query set.

In our model, the proposed kernel based regularization encourages the model to correct its epistemic uncertainty that is expected to lead to accurate predictive uncertainty and improved performance on Uncertainty increase metric. Similarly, the proposed evidence regularization term encourages the model to have low confidence for wrong predictions. Such regularization are expected to lead improved uncertainty characteristics in our model.

Model Architectures and Setup

Our ECNP model can capture both the aleatoric and epistemic uncertainty in the few-shot tasks in a single forward pass without the need of sampling. In our experiments, for the context set encoder, we use a 4 layer neural network with 128 dimensional hidden layers that leads to 128 dimensional features. For the decoder, we use the 3 layer neural network with 130 dimensional input (129 dimensional input for function regression experiments), and 128 dimensional hidden layers across all experiments. In all the models with attention mechanism, we use multihead cross-attention with 8 heads in the encoder similar to (Kim et al. 2019). To obtain the evidential hyperparameters for ECNP, we transform the output representation using a 2 layer neural network with a 64 dimensional hidden layer. We apply ReLU activation function in the intermediate layers and apply the softplus activation on the final layer to obtain the evidential parameters. In the NP model with latent variable, we sample 5 instances from the latent variable to train (*i.e.,* ELBO estimation) and evaluate the model similar to (Garnelo et al. 2018b). In the NP model, the reparameterization trick of variational-auto-encoders (Kingma and Welling 2013) is used for the Gaussian distribution. Unless specified, for the CifarFS and CelebA evidential neural process experiments, we set $\lambda_1 = 0.01$, $\lambda_2 = 0.01$, and for all remaining experiments, we set $\lambda_1 = 0.1$, and $\lambda_2 = 0.1$. For the NIG hyperparameters, we set $\beta_t = f_\theta(.) + 0.2$, and upper bound the α_t and v_t to 20. Models are trained with the learning rate of 0.001 and Adam optimizer. For the quantitative results (Table 1, 2, 3), we average the results over 5 independent runs of the model and report the mean and standard deviation. The experiments use Pytorch, and are carried out on a 8GB GeForce RTX 2070 SUPER GPU-enabled PC and on a cluster with 8GB P4 Nvidia GPU using resources at (of Technology 2019). The codes for both the baselines and evidential models are available at https://anonymous.4open.science/r/ENP-DB67/README.md.

Complexity Discussion

Compared to other meta-learning works such as MAML (Finn, Abbeel, and Levine 2017), the proposed model has rapid inference capabilities, and computationally cheaper training. During training, MAML-based models formulate meta-learning as a bilevel optimization problem that introduces computationally expensive Hessian-gradient products for global model parameter update. Specifically, when training on one task, for each inner loop update over the support set, one additional hessian term needs to be computed for the global parameter update that leads to multiple forward-backward passes over the network. MAML's bayesian extensions for uncertainty quantification further increase computational cost. For instance, BayesianMAML introduces expensive ensembling of MAML models for uncertainty quantification. In contrast, CNP/ECNP training does not involve any bi-level optimization/hessian gradient products, and learning from a task only involves one forward pass and one backward pass, making it computationally cheap.

Also, during inference, optimization based models are relatively slower/computationally expensive as they need to update the model with the support set information over K gradient steps, whereas CNP/ECNP inference only requires a single forward pass through the network, a highly desirable characteristic in meta-learning algorithms. A similar model, VERSA (Gordon et al. 2018), is also computationally cheap to train, achieves rapid inference for prediction, and has uncertainty quantification capabilities. However, to quantify uncertainty, VERSA requires multiple rounds of sampling from the posterior distribution over the class weights of the linear classifier network. In contrast, ECNP leverages evidential learning for uncertainty quantification which avoids posterior sampling, making it even faster than VERSA.

Ablation Study

For the loss function of our ECNP model given by (10), we introduce two novel regularization terms: incorrect evidence regularizer L_t^{R} and epistemic uncertainty regularizer L_t^{KER} to guide the model to have accurate uncertainty estimation. The contribution of these terms to model training is controlled by the two parameters λ_1 and λ_2, respectively. Here, we study the impact of these hyperparameters in model training and performance. Figure 10 shows the impact of λ_1 on the test set accuracy, average test set epistemic uncertainty, and aleatoric uncertainty on 10-shot function regression tasks from the GP dataset. The model tends to underestimate the uncertainty values (*i.e.,* epistemic and aleatoric uncertainties) when λ_1 is low whereas a large λ_1 value causes the model's uncertainty to be large even for accurate model predictions. The model's training is optimal when there is a good balance between minimizing the NLL loss and minimizing the incorrect evidence (*e.g.,* $\lambda_1 = 0.1$).

The regularization parameter λ_2 controls the uncertainty estimation in regions far away from the context points. Figure 11 shows the impact of λ_2 to the model behavior as training progresses in a 5-shot sinusoidal regression problem. A large λ_2 leads to a sensitive model that outputs very a high epistemic uncertainty as the target data points become far from the context set observations and also hurts the model's generalization performance (*i.e.,* average test loss). Conversely, a very low λ_2 value does not train the model to consider the neighborhood information for uncertainty. The regularization term λ_2 provides the model with flexibility to consider the neighborhood information (*i.e.,* $L_t^{\text{KER}} = v_t \times D(x_t, \mathcal{C})$ in determining the uncertainty. Next, we carry out extrapolation experiments where the trained model is evaluated outside its training data range (*i.e.,* $x > 5.0$). Figure 12 visualizes the epistemic uncertainty for a random task. When the kernel based regularization term is introduced in training (*i.e.,* $\lambda_2 > 0$), the model accurately outputs a high epistemic uncertainty outside the training data range, which is desirable.

Impact of Regularization to Model's Uncertainty Characteristics

We also study the impact of regularization terms to uncertainty using MNIST dataset over 50-shot Image Completion experiments. Figure 13 shows the effect of different evidence regularization values λ_1 when $\lambda_2 = 0.1$. As can be seen, larger regularization leads to improved Inclusion performance without any impact the Uncertainty Increase metrics. Figure 14 shows the effect of kernel based regularization when evidence regularization term $\lambda_1 = 0.1$. Reasonable value of kernel based regularization helps improve the Uncertainty Increase metric without any impact to the Inclusion. Finally, very large values of the uncertainty regularization terms (both evidence regularization and kernel based regularizations) hurt the model's generalization capabilities as shown in Figure 13 (c) and Figure 14 (c).

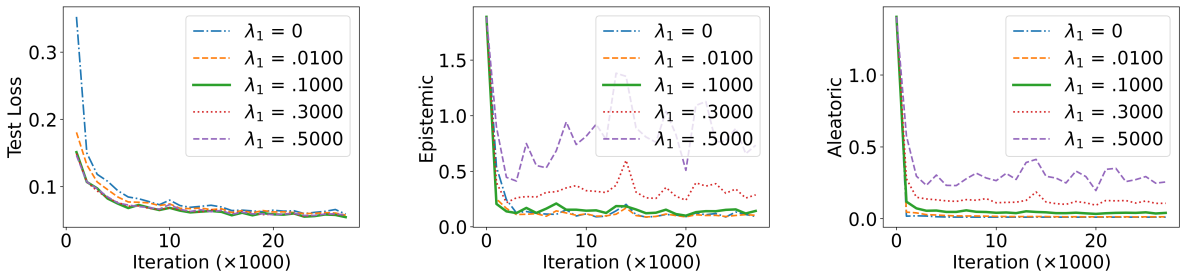
Additional Experimental Results

We present additional qualitative results with GP samples and CelebA that visualize the estimated uncertainty. We also conduct a deeper analysis on the behavior of the evidential parameters that reveal important insights on how the proposed evidential neural processes effectively combine the learning from few-shot samples (*i.e.,* context data) and the meta-knowledge from other tasks to achieve accurate prediction performance and fine-grained uncertainty quantification.

Regression experiments on GP tasks. We trained the ECNP model for 20,000 iterations on 10-shot GP tasks and evaluated the model performance on a random test task. Figure 15 shows the model's behavior on random GP tasks for ECNP as we increase the number of data points on the context set. The model outputs a high uncertainty at regions where the model has not observed the context point/s and it believes that the meta knowledge is not sufficient for accurate prediction. A low uncertainty will be predicted otherwise. For instance, in Figure 15 (b), the model outputs confident correct predictions in the region between the 3rd and 4th context points even though they are relatively far away from both context points. This may be due to that the meta-knowledge learned from other tasks is rich enough to lower the uncertainty. As we increase the number of observations in the context set, the model's confidence increases along with the predictive accuracy as indicated by Figures 15 (c) and 15 (d).

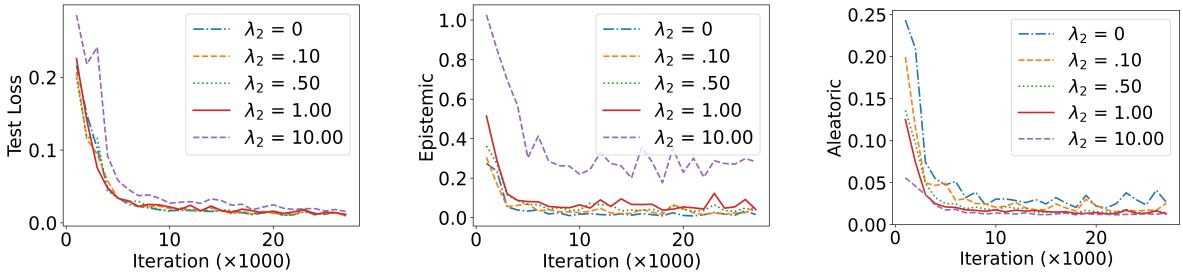
Image completion experiments on CelebA. Figure 16 shows the qualitative results of the evidential attentive neural process model on a random CelebA test task. The model was trained for 50 epochs using 200-shot CelebA tasks with $\lambda_1 = 0.1$ and $\lambda_2 = 1.0$. As we increase the number of context points in the task (indicated by Context Mask CM), the average epistemic uncertainty decreases rapidly whereas the aleatoric uncertainty decreases at a slower pace. This could be because there may be inherent noises associated with the few-shot tasks that may not be addressed by newly added context points. Furthermore, from (16) and (17), the epistemic and aleatoric uncertainties vary by a factor of v_t, which captures the meta-knowledge according to our previous discussions. More context points could allow the model to better relate to other meta-training tasks, leading to a larger v_t as shown by 16 (b). This also contributes to a faster decrease of the epistemic uncertainty. Moreover, with additional data points on the context set, the model has greater evidence for its prediction and the model's predictive accuracy increases as indicated by decrease in the MSE error.

Figure 17 visualizes the three higher-order hyperparameters α, β, and v along with the epistemic and aleatoric uncertainties for a CelebA test task with varying number of context points. We average the evidential hyperparameters and uncertainty across the 3 channels for illustration. When the number of context points in the task is low, the epistemic uncertainty and aleatoric uncertainty show similar trends. As we observe more data, the model evidence (*i.e.,* α, β, and v) gets updated to reflect the model's knowledge. When there are a few context points in the task, these hyperparameter



(a) Test Loss Trend (b) Epistemic Uncertainty Trend (c) Aleatoric Uncertainty Trend

Figure 10: Impact of Regularization parameter λ_1 in a 10-shot GP function regression task



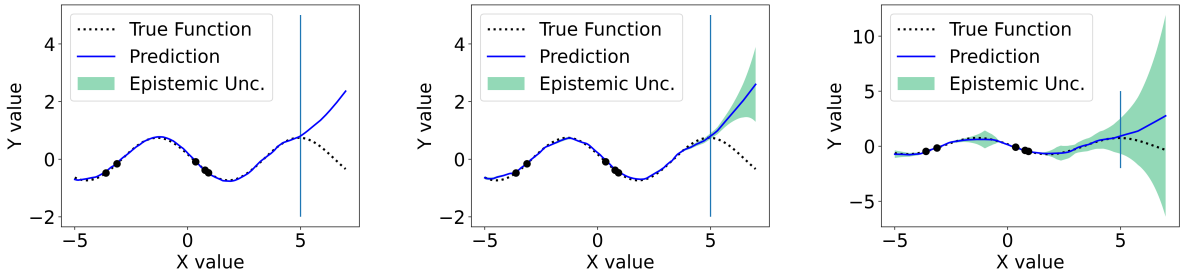
(a) Test Loss Trend (b) Epistemic Uncertainty Trend (c) Aleatoric Uncertainty Trend

Figure 11: Impact of regularization parameter (λ_2) in a 5-shot sinusoidal function regression task

values are guided by the meta-knowledge. As the model observes more context points, the model integrates the meta-knowledge with the information to update the evidential parameters. Specifically, the model considers the evidential hyperparameters α and β to estimate the aleatoric uncertainty that seems to be high around the edges and the boundaries. The model considers the evidential hyperparameter v along with α and β to estimate the epistemic uncertainty that decreases gradually with more data in the context set. Similar trends are observed across the experiments.

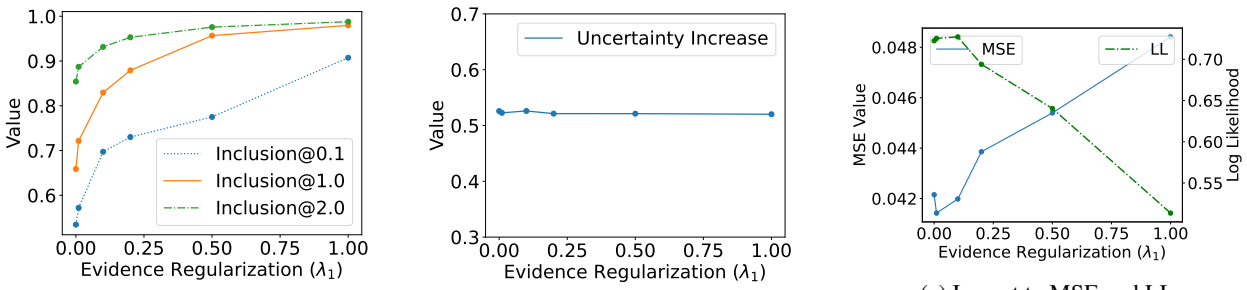
Evidential parameters guided local vs. meta-learning. We further inspect the evidential hyperparameters to better understand the model behavior. We first experiment with relatively simple function regression tasks over the sinusoid dataset. We train the evidential conditional neural process model with $\lambda_1 = 0/0.1$ and $\lambda_2 = 1.0$ for 30,000 meta iterations and evaluate on a random test task. The test task consists of 400 target points from which we randomly select K-shots in the context set. We track the average MSE error and the evidential hyperparameters α, β, and v on the target set across all the test tasks. Figures 18 and 19 show the results of the experiments where we observe an interesting trend. As can be seen from Figure 18 (d) and Figure 19 (d), the model performance converges after a few context points (*i.e.,* < 10) along with the (average) predicted evidence score. Meanwhile, both α and v predicted on the testing points also stop increasing. This implies that adding additional context points no longer helps to improve learning from local data (*i.e.,* context points), which is captured by α based on the hierarchical structure defined in (1)-(3). Similarly, it does not help to learn from the meta-knowledge, either, which is captured by v. This example clearly shows that how the proposed model effectively combines the learning from the local context data points while leveraging meta-knowledge from other similar few-shot task to achieve a fast convergence for relatively simple tasks.

For a more challenging CelebA task as shown in Figure 20, we observe a similar trend for α, which increases along with the addition of context points but starts to converge after a number of context points have been included. However, v shows a very different trend that continues to increase along with the addition of context points. The different behavior in these two evidential parameters precisely captures how the proposed model conducts effective learning for more challenging tasks. In particular, for such tasks, the local data (*i.e.,* context points) contribute relatively less since they are inherently limited in the few-shot setting for more complex tasks. Meanwhile, meta-knowledge is expected to play a more important role given a potential large number of training tasks available for the model to learn the meta-knowledge (and using the context points related to the learned meta-knowledge). This exactly matches the faster convergence of α and continuous growth of v. Furthermore, we also observe a more random trend in β due to a higher noise ratio in the image completion tasks. Finally, the MSE/Evidence trend matches the changes on α and v: evidence continues to grow due to the contribution from v and MSE converges at a slower pace than the simpler tasks and its decrease at the later stages of the leaning is mainly attribute to the increase of v (*i.e.,* the meta-knowledge).



(a) $\lambda_2 = 0$ (b) $\lambda_2 = 1.0$ (c) $\lambda_2 = 10.0$

Figure 12: Model behavior for different λ_2 values in a 10-shot sinusoidal function regression task



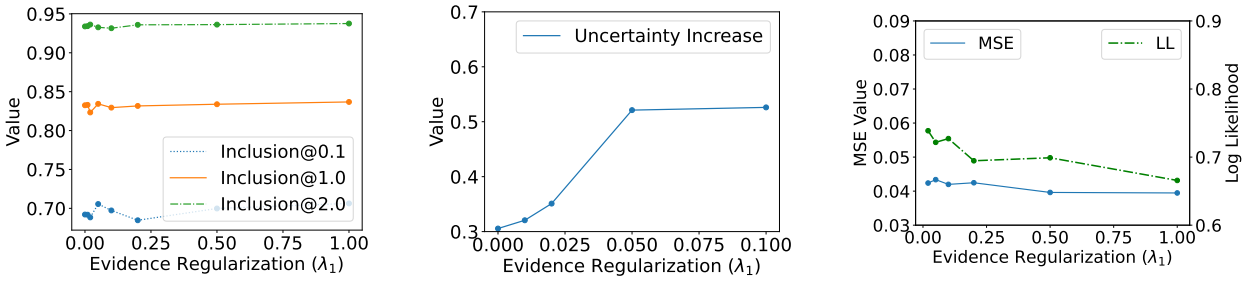
(a) Impact to Inclusion@K (b) Impact to Uncertainty Increase (c) Impact to MSE and LL

Figure 13: Impact of Evidence Regularization (λ_1) to different uncertainty metrics

Limitations and Future Work

In this work, we focus on the CNP family because of their rapid inference, scalability, and competitive predictive performance. More importantly, they naturally quantify uncertainty by simulating a stochastic process like a GP. We introduce a novel hierarchical Bayesian structure that can be viewed as a general augmentation to the CNP family of models to achieve fine-grained uncertainty decomposition and theoretically guaranteed robustness. Proposed novel structure can be combined with and enhance other models of the CNP family. To this end, we experiment with ConvCNP (Gordan et.al, 2020), a recent improvement of CNP, on 5-shot GP regression. The results are ConvCNP: MSE: 0.268, LL: -0.239, Evidential-ConvCNP: MSE: 0.228, LL:-0.012, which shows the potential of our method to augment recent CNP models. We leave additional exploration of the effectiveness of the proposed structure to other CNP works as a future work.

We developed evidential meta-learning model for fine-grained uncertainty quantification. The proposed ECNP model introduces two additional hyperparameters and requires hyperparameter tuning. Moreover, in this work, we experimented on few-shot regression tasks with 1D regression and 2D image completion. For the considered datasets, a relatively simple distance function such (the Euclidean distance) was effective for kernel based regularization. However, epistemic uncertainty guidance using kernel-based regularization in more challenging datasets such as image and videos may require better and efficient design of the distance function $D(.)$ (e.g. distance/similarity in the embedding space). Also, it can be an interesting future work to extend this work to other meta-learning approaches to equip them with fine-grained uncertainty quantification capabilities in a computationally efficient manner. We now plan to address the issues and experiment on larger datasets such as ImageNet with deeper neural networks as our future work.

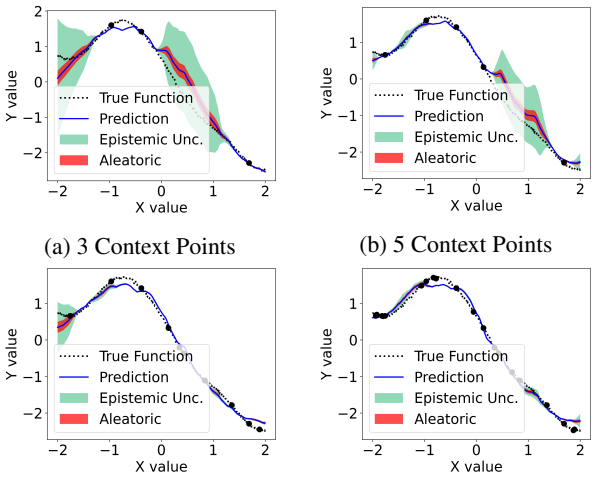


(a) Impact to Inclusion@K

(b) Impact to Uncertainty Increase

(c) Impact to MSE and LL

Figure 14: Impact of Kernel Based Regularization (λ_2) to different uncertainty metrics



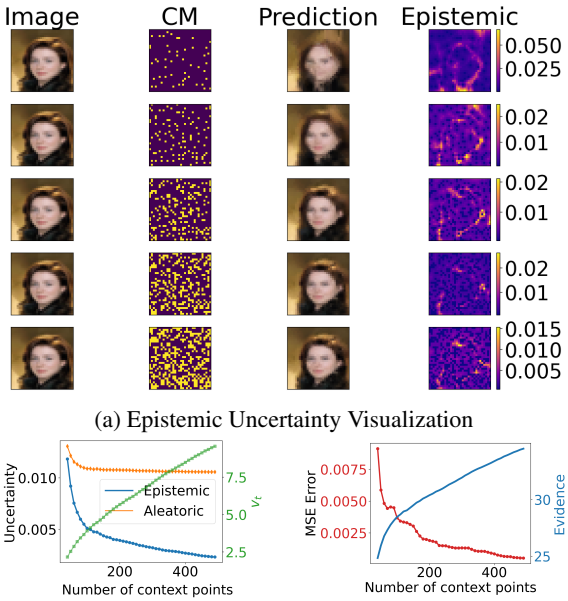
(a) 3 Context Points

(b) 5 Context Points

(c) 10 Context Points

(d) 20 Context Points

Figure 15: ECNP model performance on a GP function regression task

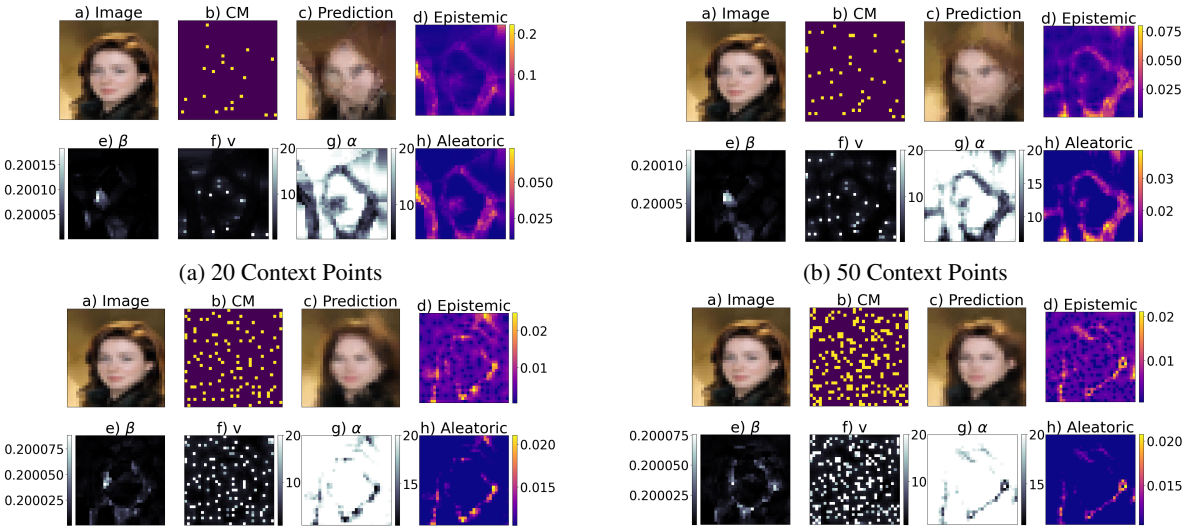


(a) Epistemic Uncertainty Visualization

(b) Uncertainty Trends

(c) MSE/Evidence Trend

Figure 16: Evidential ANP model performance on a CelebA task



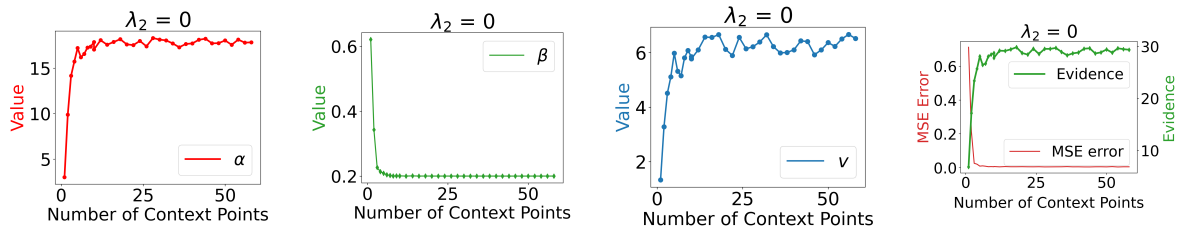
(a) 20 Context Points

(b) 50 Context Points

(c) 100 Context Points

(d) 200 Context Points

Figure 17: Visualization of the evidential hyperparameters and predicted uncertainty on a CelebA task



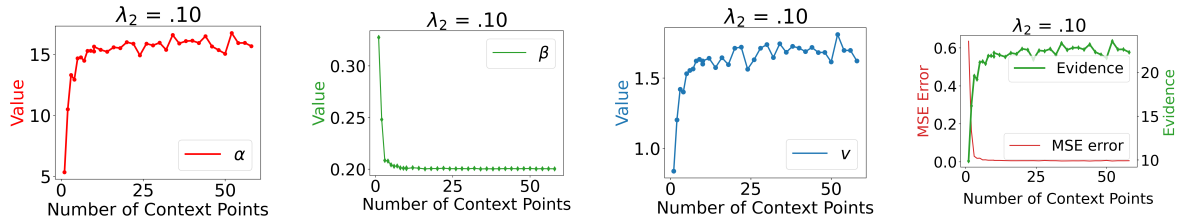
(a) α trend

(b) β trend

(c) v trend

(d) MSE/Evidence trend

Figure 18: Evidential CNP model performance on a Sinusoid Regression task for $\lambda_2 = 0$



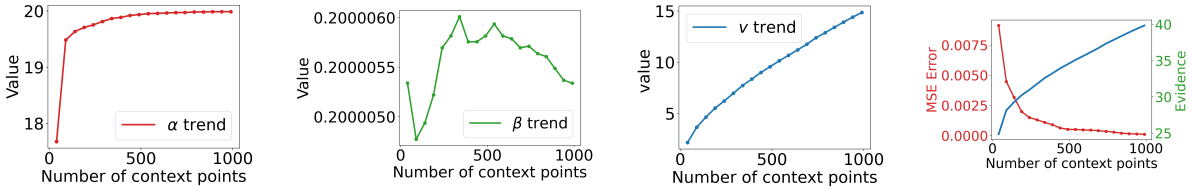
(a) α trend

(b) β trend

(c) v trend

(d) MSE/Evidence trend

Figure 19: Evidential CNP model performance on a Sinusoid Regression task for $\lambda_2 = 0.1$



(a) α trend

(b) β trend

(c) v trend

(d) MSE/Evidence trend

Figure 20: Evidential ANP model performance on a CelebA task for $\lambda_2 = 1.0$