Journal of Machine Learning for Biomedical Imaging. 2022:009. pp 1-? Submitted 09/2021; Published 04/2022
Special Issue: Information Processing in Medical Imaging (IPMI) 2021
Guest Editors: Aasa Feragen, Stefan Sommer, Julia Schnabel, Mads Nielsen

Distributional Gaussian Processes Layers for Out-of-Distribution Detection

Sebastian G. Popescu s.popescu16@imperial.ac.uk
Biomedical Image Analysis Group, Imperial College London, London, U.K.

David J. Sharp david.sharp@imperial.ac.uk
Computational, Cognitive & Clinical Neuroimaging Laboratory, Imperial College London, London, U.K.

James H. Cole james.cole@ucl.ac.uk
Centre for Medical Image Computing, University College London, London, U.K.

Konstantinos Kamnitsas konstantinos.kamnitsas@eng.ox.ac.uk
Department of Engineering Science, University of Oxford, Oxford, U.K.
Biomedical Image Analysis Group, Imperial College London, London, U.K.

Ben Glocker b.glocker@imperial.ac.uk
Biomedical Image Analysis Group, Imperial College London, London, U.K.

Abstract

Machine learning models deployed on medical imaging tasks must be equipped with out-of-distribution detection capabilities in order to avoid erroneous predictions. It is unsure whether out-of-distribution detection models reliant on deep neural networks are suitable for detecting domain shifts in medical imaging. Gaussian Processes can reliably separate in-distribution data points from out-of-distribution data points via their mathematical construction. Hence, we propose a parameter efficient Bayesian layer for hierarchical convolutional Gaussian Processes that incorporates Gaussian Processes operating in Wasserstein-2 space to reliably propagate uncertainty. This directly replaces convolving Gaussian Processes with a distance-preserving affine operator on distributions. Our experiments on brain tissue-segmentation show that the resulting architecture approaches the performance of well-established deterministic segmentation algorithms (U-Net), which has not been achieved with previous hierarchical Gaussian Processes. Moreover, by applying the same segmentation model to out-of-distribution data (i.e., images with pathology such as brain tumors), we show that our uncertainty estimates result in out-of-distribution detection that outperforms the capabilities of previous Bayesian networks and reconstruction-based approaches that learn normative distributions. To facilitate future work our code is publicly available[1].

Keywords: Gaussian Processes, Image Segmentation, Out-of-distribution Detection

1. Introduction

Deep learning methods have achieved state-of-the-art results on a plethora of medical image segmentation tasks to clinical risk assessment (Zhou et al., 2021; Tang, 2019; Imai et al., 2020). However, their application in clinical settings remains challenging due to issues pertaining to lack of reliability and miscalibration of confidence estimates. Reliably estimating uncertainty in predictions is also of vital interest in adjacent machine learning fields such as reinforcement

1. https://github.com/SebastianPopescu/DistGP_Layers

https://www.melba-journal.org/papers/2022:009.html.

learning, to guide exploration, or in active learning, to guide the selection of data points for the next iteration of labelling. Most research into incorporating uncertainty into medical image segmentation has gravitated around modelling inter-rater variability and the inherent aleatoric uncertainty associated to the dataset, which can be caused by noise or inter-class ambiguities, alongside modelling the uncertainty in parameters (Czolbe et al., 2021). However, less focus has been placed on how models behave when processing unexpected inputs which differ from the characteristics of the training data. Such inputs, often called anomalies, outliers or out-of-distribution samples, could possibly lead to deleterious effects in healthcare applications where predictive models may encounter data that is corrupted or from patients with diseases that the model is not trained for (Curth et al., 2019; Mårtensson et al., 2020).

Out-of-distribution (OOD) detection in medical imaging has been mostly approached through the lens of reconstruction-based techniques involving some form of encoder-decoder network trained on normative datasets (Chen et al., 2019, 2021). Conversely, we focus on enhancing task-specific models (e.g., a segmentation model) with reliable uncertainty quantification that enables outlier detection. Standard deep neural networks (DNNs), despite their high predictive performance, often exhibit unreasonably high confidence in predictions estimates when processing unseen samples that are not from the data manifold of the training set (e.g., in the presence of pathology under the hypothesis of training data being composed of normal subjects or in a more general setting the presence of motion artifacts never seen in training images). To alleviate this, Bayesian approaches that assign posteriors over weights (MC Dropout (Gal and Ghahramani, 2016b) included) or in function space (Repulsive Deep Ensembles (D'Angelo and Fortuin, 2021) included) have been proposed (Wilson and Izmailov, 2020). However, either assigning priors on weights or in function space does not necessarily lend itself to reliable OOD detection capabilities by virtue of inspecting the predictive variance of the model as was shown in Henning et al. (2021). The authors argue that both infinite-width networks, trained via the Neural Network Gaussian Process (NNGP) kernel (Lee et al., 2017), or finite-width networks trained via Hamiltonian Monte Carlo (Neal et al., 2011) are not reliable for OOD detection since they show that the associated NNGP kernel is not correlated with distances between objects in input space. This loss of *distance-awareness* after encoding data has catastrophic effects on OOD detection, as we will soon see. Similarly, Foong et al. (2019) describe a limitation in the expressiveness of the predictive uncertainty estimate given by mean-field variational inference (MFVI) when applied as the inference technique for Bayesian Neural Networks (BNNs). Concretely, MFVI fails in offering quality uncertainty estimates in regions between well-separated clusters of data, which the authors coin as *in-between* uncertainty, with potentially catastrophic consequences for active learning, Bayesian optimisation or robustness to out-of-distribution data. In this paper we follow an alternative approach, using Gaussian Processes (GP) as the building block for deep Bayesian networks.

The use of GPs for image classification has garnered interest in the past years. Hybrid approaches, whereby a DNN's embedding mechanism is trained end-to-end with a GP as the classification layer, were the first attempts to unify the two approaches (Bradshaw et al., 2017). The first convolutional kernel was proposed in Van der Wilk et al. (2017), constructed by aggregating patch response functions. This approach was stacked on feed forward GP layers applied in a convolutional manner, with promising improvements in accuracy compared to their shallow counterpart (Blomqvist et al., 2018).

We expand on the aforementioned work, by introducing a simpler convolutional mechanism, which does not require convolving GPs at each layer and hence alleviates the computational cost of optimizing over inducing points' locations residing in high dimensional spaces alongside the issues brought upon by multi-output GPs. We propose a plug-in Bayesian layer more amenable to CNN architectures. More concretely, we seek to replace each individual component of a standard convolutional layer in convolutional neural networks (CNNs), respectively the convolved filters and the non-linear activation function. Firstly, we impose constraints on the filter such that we have an upper bound on distances after the convolution with regards to distances between the same objects beforehand. This will ensure that objects which were close in previous layers will remain close going forward, which as we shall see later on is a fundamental property for reliable OOD detection. Moreover, directly using convolved filters as opposed to convolved GPs (Blomqvist et al., 2018) solves the issue with optimizing high-dimensional inducing points' locations alongside introducing a simpler mechanism by which we can introduce correlations between channels (Nguyen et al., 2014). Secondly, we replace the element-wise non-linear activation functions with Distributional Gaussian Processes (DistGP) (Bachoc et al., 2017) used in one-to-one mapping manner, essentially acting as a non-parametric activation function. A variant of DistGP used in a hierarchical setting akin to Deep Gaussian Processes (DGP) (Damianou and Lawrence, 2013) was shown to be better at detecting OOD due to both kernel and architecture design (Popescu et al., 2020). In this paper we will show that our proposed module is also suited for OOD detection on both toy/image data and biomedical scans.

In the remainder of this section we provide a deeper exploration of uncertainties used in literature for biomedical image segmentation, subsequently introducing the concept of *distance-awareness* and imposing smoothness constraints on learned representations in a deep network as prerequisites for reliable OOD detection. These two properties will be key to motivate the imposed constraints and architecture choice of our proposed probabilistic module later on.

1.1 Uncertainty quantification for biomedical imaging segmentation

While prediction uncertainty can be computed for standard neural networks by using the softmax probability, these uncertainty estimates are often overconfident (Guo et al., 2017; McClure et al., 2019). Research into Bayesian models has focused on a separation of uncertainty into two different types, aleatoric (data intrinsic) and epistemic (model parameter uncertainty). To formalize this difference, we consider a multi-class classification problem, with classes denoted as $\{y_1, \cdots, y_C\}$ and model parameters denoted by θ. We have the following predictive equation at testing time:

$$p(y_c|x^*, \mathbf{D}) = \int \underbrace{p(y_c \mid x^*, \Theta)}_{\text{Aleatoric Uncertainty}} \underbrace{p(\theta|\mathbf{D})}_{\text{Epistemic Uncertainty}} d\theta \tag{1}$$

Aleatoric uncertainty is irreducible, given by noise in the data acquisition process and has been considered in medical image segmentation (Monteiro et al., 2020), whereas epistemic uncertainty can be reduced by providing more data during model training. This has also been studied in segmentation tasks (Nair et al., 2020). Previous work proposed to account for the uncertainty in the learned model parameters using an approximate Bayesian inference

over the network weights (Kendall et al., 2015). However, it was shown that this method may produce samples that vary pixel by pixel and thus may not capture complex spatially correlated structures in the distribution of segmentations maps. The probabilistic U-Net (Kohl et al., 2018) produces samples with limited diversity due to the fact that stochasticity is introduced in the highest resolution level of the U-Net. To solve this issue, Baumgartner et al. (2019) introduce a hierachical structure between the different levels of the U-Net, hence introducing stochasticity at each level. Another improvement on the Probabilistic U-Net comes by adding variational dropout (Kingma et al., 2015) to the last layer to gain epistemic uncertainty quantification properties (Hu et al., 2019). All the models previously introduced relied on multiple annotations of the images with the intended goal of capturing this uncertainty in annotations with the aid of sampling from some form of latent variables which encode information about the whole image at varying scales of the U-Net. However, none of these previous works test how their models behave in the presence of outliers.

1.2 Distributional Uncertainty as a proxy for OOD detection

Besides the dichotomy consisting of aleatoric and epistemic uncertainty, reliably highlighting certain inputs which have undergone a domain shift (Lakshminarayanan et al., 2017) or out-of-distribution samples (Hendrycks and Gimpel, 2016) has garnered a lot of interest in the past years. Succinctly, the aim is to measure the degree to which a model knows when it does not know, or more precisely if a network trained on a specific dataset is evaluated at testing time on a completely different dataset (potentially from a different modality or another application domain), then the expectation is that the network should output high predictive uncertainty on this set of data points that are very far from the training data manifold.

A problem with introducing this new type of uncertainty is how to disentangle it from epistemic uncertainty. For example, in the Deep Ensembles paper (Lakshminarayanan et al., 2017), the authors propose to measure the disagreement between different sub-models of the deep ensemble $\sum_{m=1}^{M} KL\left[p(y \mid x; \theta_m) \| \mathbb{E}\left[p(y \mid X)\right]\right]$ for M sub-models with associated sub-model parameters θ_m and $\mathbb{E}\left[p(y \mid x)\right] = \frac{1}{M} \sum_{m=1}^{M} p(y \mid x; \theta_m)$ is the prediction of the ensemble. We remind ourselves that epistemic uncertainty can be reduced by adding more data. By this logic, epistemic uncertainty cannot be reduced outside the data manifold of our dataset since we don't add data points which do not stem from the same data generative pipeline (this is not true in the case of OOD detection models which explicitly use OOD samples during training/testing (Liang et al., 2017; Hafner et al., 2020)). Hence, epistemic uncertainty can only be reduced inside the data manifold and should be zero outside the data manifold (assuming model is *distance-aware*, which we will subsequently define). Conversely, our chosen measure for OOD detection should grow outside the data manifold and be close to or 0 inside the data manifold. With this in mind, the disagreement metric introduced in Lakshminarayanan et al. (2017) cannot achieve this separation, confounding the two types of uncertainty.

Malinin and Gales (2018) introduced for the first time the separation of total uncertainty into three components: epistemic, aleatoric and distributional uncertainty. To make the

distinction clearer, the authors argue that aleatoric uncertainty is a "known-unknown", whereby the model confidently states that an input data point is hard to classify (class overlap). Contrary, distributional uncertainty is an "unknown-unknown" due to the fact that the model is unfamiliar with the input space region that the test data comes from, thereby not being able to make confident predictions.

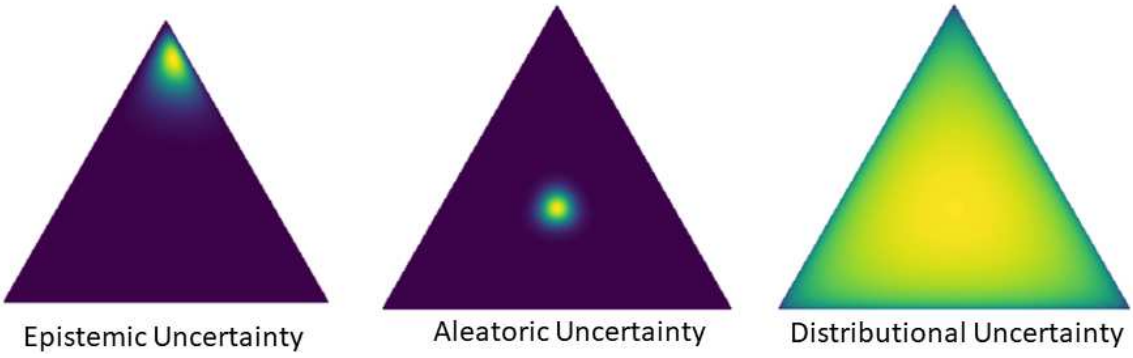


Figure 1: Probability simplex for a 3 class classification problem, where corners corresponds to a class; Each point represents a categorical distribution, with brighter colors indicating higher density of the underlying ensemble. **Epistemic Uncertainty** captures uncertainty in model parameters caused by lack of data or model non-identifiability, with the ensemble of the predictions being concentrated in a corner of the probability simplex albeit with an increased diversity; **Aleatoric Uncertainty** captures class overlap, with the ensemble of predictions being confidently mapped to the highest predictive entropy; **Distributional Uncertainty** captures domain shift, with the ensemble of predictions being centred in the middle with highest possible degree of diversity;

We briefly introduce the uncertainty decomposition mechanism introduced in Malinin and Gales (2018). Considering equation (1), by using Monte Carlo integration of above equation and computing the predictive entropy, we would not be able to discern between high predictive entropy due to aleatoric uncertainty (class overlap) or distributional uncertainty (dataset/domain shift). Hence, Malinin and Gales (2018) propose to introduce a latent variable μ over the categorical variables corresponding to each class, parametrized as a distribution over distributions on a simplex, $p(\mu|x^*, \theta)$. The intuition behind this Dirichlet distribution over the probability simplex is that OOD points should be scattered, whereas in-distribution points should concentrate. We can now re-write our predictive equation as:

$$p(y_c|x^*, \mathbf{D}) = \int\int \underbrace{p(y_c|x^*, \mu)}_{\text{Aleatoric Uncertainty}} \underbrace{p(\mu|x^*, \theta)}_{\text{Distributional Uncertainty}} \underbrace{p(\theta|\mathbf{D})}_{\text{Epistemic Uncertainty}} d\mu \, d\theta \quad (2)$$

The authors argue that using a measure of spread of the ensemble (after sampling from $p(\theta \mid \mathbf{D})$) will be more informative. We remind ourselves that Mutual Information between variable X and Y can be expressed in terms of the difference between entropy and conditional entropy: $I(X;Y) = H(P(X)) - H(P(X|Y))$. Hence we can use the Mutual Information measure between model predictions and Dirichlet parameters to obtain a better measure of

uncertainty. We integrate out over θ in the main equation and we get:

$$\underbrace{I[y, \mu | x^*, \mathbf{D}]}_{\text{Distributional Uncertainty}} = \underbrace{H[\mathbb{E}_{p(\mu|\mathbf{D})} p(y|x^*, \mu)]}_{\text{Total Uncertainty}} - \underbrace{\mathbb{E}_{p(\mu|\mathbf{D})}[H[P(y|x^*, \mu)]]}_{\text{Aleatoric Uncertainty}} \tag{3}$$

Connections between this uncertainty disentanglement framework specialized for DNN-parametrized Dirichlet distributions and uncertainty disentanglement in GP will be subsequently made clearer in subsection 2.3. Distributional uncertainty will be the key uncertainty score used throughout this paper to assess whether input data points are inside or outside the data manifold.

1.3 Distance awareness and smoothness for out-of-distribution detection

Perhaps the greatest inspiration and motivation for this paper resides in the theoretical framework introduced in Liu et al. (2020), by which the authors outline what are some key mathematical conditions for provably reliable OOD detection in DNNs. We commence by briefly outlining the ideas introduced in aforementioned paper.

For an abstract data generating distribution $p\left(y \mid x\right)$, where y is scalar, respectively $x \in \mathbb{X} \subset \mathbb{R}^D$ with the input data manifold being equipped with a suitable metric $\|\cdot\|_X$. We consider our training data $D = \{(x_i, y_i)_{1:n}\}$ to be sampled from a subset of the full input space $x_{in-d} \in \mathbb{X}$, where the *in-d* abbreviation stems from in-distribution. With this in mind, we can consider the in-distribution data generating distribution $p_{in-d}\left(y \mid x\right) = p\left(y \mid x,\ x \in \mathbb{X}_{in-d}\right)$, respectively the out-of-distribution data generating distribution $p_{ood}\left(y \mid x\right) = p\left(y \mid x,\ x \notin \mathbb{X}_{in-d}\right)$. Hence, it is safe to assume the full data generating distribution $p\left(y \mid x\right)$ is composed as a mixture of the in-distribution and OOD generating distributions:

$$\begin{aligned} p\left(y \mid x\right) &= p\left(y, x \in \mathbb{X}_{in-d} \mid x\right) + p\left(y, x \notin \mathbb{X}_{in-d} \mid x\right) && (4) \\ &= p\left(y \mid x, x \in \mathbb{X}_{in-d}\right) p\left(x \in \mathbb{X}_{in-d}\right) + p\left(y \mid x, x \notin \mathbb{X}_{in-d}\right) p\left(x \notin \mathbb{X}_{in-d}\right) && (5) \\ &= p_{in-d}\left(y \mid x\right) p\left(x \in \mathbb{X}_{in-d}\right) + p_{ood}\left(y \mid x\right) p\left(x \notin \mathbb{X}_{in-d}\right) && (6) \end{aligned}$$

Evidently, during training we are only learning $p_{in-d}\left(y \mid x\right)$ since we only have access to $D \subset \mathbb{X}_{in-d}$. Therefore, our model is completely in the dark with regards to $p_{ood}\left(y \mid x\right)$. These two data generating distributions more often than not are fundamentally different (e.g., having trained a model on T1w MRI scans, subsequently feeding it with T2w MRI scans, an imaging modality which has an almost inverse scaling to represent varying brain tissue). With this in mind, Liu et al. (2020) argue that the optimal strategy is for $p_{ood}\left(y \mid x\right)$ to be predicted as a uniform distribution, thus signalling the lack of knowledge of the model on this different input domain. We can now recall the distinction made by Malinin and Gales (2018), between "known-unknowns" (aleatoric uncertainty, e.g., class overlap) and "unknown-unknowns" (distributional uncertainty, e.g., domain shift), both of which have an uninformative predictive distribution (maximum predictive entropy). However, to disentangle these two types of uncertainty, we need a second-order type of uncertainty that basically scatters logit samples when distributional uncertainty is high, respectively accurately samples logits to maximum predictive entropy in the case of high aleatoric uncertainty. We now formalize this desiderata by a condition called "distance awareness" in Liu et al. (2020).

Definition 1 (Definition 1 in Liu et al. (2020)) *We consider the predictive distribution for unseen point $p(y^* \mid x^*)$ at testing time, for model trained on $\mathbb{X}_{in-d} \in \mathbb{X}$, with the data manifold being equipped with metric $\|\cdot\|_X$. Then, we can affirm that $p(y^* \mid x^*)$ is* distance-aware *if there exists summary statistic $u(x^*)$ of $p(y^* \mid x^*)$ that embeds the distance between $\mathbb{X}_{in-d}$ and x^*:*

$$u(x^*) = v\left[\mathbb{E}_{x \sim \mathbb{X}_{in-d}}\left[\|x^* - x\|_X^2\right]\right] \tag{7}$$

, where v is a monotonic function that increases with distance.

Definition 1 does not make any assumptions related to the architecture of the model from which the predictive distribution stems. In practice we would have the following composition to arrive at the logits $logit(x^*) = f \circ enc(x^*)$, where $enc(\cdot)$ represents a network that outputs the representation learning layer and $f(\cdot)$ is the output layer. In Liu et al. (2020) the authors propose the following two conditions to ensure that the composition is *distance-aware*:

- $f(\cdot)$ is *distance-aware*
- $\mathbb{E}_{x \sim \mathbb{X}_{in-d}}\left[\|x^* - X\|_X^2\right] \approx \mathbb{E}_{x \sim \mathbb{X}_{in-d}}\left[\|enc(x^*) - enc(X)\|_{enc(X)}^2\right]$

The last condition means that distances between data points in input space should be correlated with distances in learned representation, which is equipped with a $\|\cdot\|_{enc(X)}$ metric. In our work, we will use GPs as f, because as we will see in section 2.1, GPs are *distance-aware* functions. This enables us to build a *distance-aware* model that is more appropriate for OOD detection. Whereas GPs satisfy the *distance-aware* condition for the last layer predictor, we are still left with the question on how to maintain distances in the learned representation correlated to distances in the input layer. This will be subsequently dealt with.

Network smoothness constraints Throughout this paper we will consider the general term of "smoothness" of a model to mean the degree to which changes in the input have an effect on the output at a particular layer. The question now shifts into how can we quantify the smoothness of a network/function? In mathematical analysis a function $f : \mathbb{X} \to \mathbb{Y}$ is said to be k-smooth if the first k derivatives exist $\{f_\prime, f^{''}, \cdots, f^{(k)}\}$ and are continuous. We denote functions which have these properties as being of class C^k. For example, Gaussian Processes using squared exponential kernels are C^∞ since the squared exponential kernel is infinitely differentiable. However, such a definition and quantification of *smoothness* wouldn't aid us in ensuring the second condition of *distance-awareness*. For this, we shall use Lipschitz continuity, which is defined as follows: considering two metric spaces $\mathbb{X}$ and $\mathbb{Y}$ equipped with metrics $\|\cdot\|_X$ and $\|\cdot\|_Y$ and $f : \mathbb{X} \to \mathbb{Y}$ is Lipschitz continuous if $\exists$ real $K \geq 0$ such that $\forall x, y \in \mathbb{X}$ we have $\|f(x), f(y)\|_Y \leq K\|x, y\|_X$. Intuitively, for Lipschitz functions there is an upper limit in how much outputs can change with respect to distances in input space. It is perhaps better to highlight now that Lipschitz functions represent a global property. There are also locally Lipschitz continuous functions which respect the aforementioned condition just in a neighbourhood of x, respectively $B_r(x) = \{y \in \mathbb{X} : \|x, y\|_X \leq r\}$. Lastly, bi-Lipschitz continuity is defined as $\frac{1}{K}\|x, y\|_X \geq \|f(x), f(y)\|_Y \leq K\|x, y\|_X$, which is a property that avoids learning trivially smooth functions and maintains useful information (Rosca et al., 2020). With this in mind, recent work (Liu et al., 2020; van Amersfoort et al., 2021) have enforced the bi-Lipschitz property on feature extractors, thereby ensuring strong correlation

between distances between data points in input space, respectively in the representation learning layer.

1.4 Contributions

This work makes the following main contributions:

- We introduce a Bayesian layer that can act as a drop-in replacement for standard convolutional layers. Operating on stochastic layers with Gaussian distributions, we upper bound the convolved affine operators in Wasserstein-2 space, thus ensuring Lipschitz continuity. To introduce non-linearities, we apply DistGP element-wise on the output of the constrained affine operator, thereby obtaining non-parametric "activation functions" which ensure adequate quantification of distributional uncertainty at each layer.
- We derive theoretical requirements for the model to not suffer from *feature collapse*, with additional empirical results to support the theory.
- We demonstrate that a GP-based convolutional architecture can achieve competitive results in segmentation tasks in comparison to a U-Net.
- We show improved OOD detection results on both general OOD tasks and on medical images compared to previous OOD approaches such as reconstruction-based models.

2. Background

In this section we provide a brief review of the theoretical toolkit required for the remainder of the paper. We commence by laying out foundational material on GPs, followed by an introduction to attempts to sparsify GPs. Subsequently, we introduce an uncertainty disentanglement framework for sparse Gaussian Processes. We briefly define Wasserstein-2 distances and show how they can be used to define kernels operating on Gaussian distributions. Lastly, we introduce recent re-formulations of deep GPs through the lens of OOD detection.

2.1 Primer on Gaussian Processes

A Gaussian Process can be seen as a generalization of multivariate Gaussian random variables to infinite sets. We define this statement in more detail now. We consider $f(x)$ to be a stochastic field, with $x \in \mathbb{R}^d$ and we define $m(x) = \mathbb{E}[f(x)]$ and $C(x_i, x_j) = Cov[f(x_i), f(x_j)]$. We denote a Gaussian Process (GP) $f(x)$ as:

$$f(x) \sim GP(m(x), C(x_i, x_j)) \tag{8}$$

The covariance function have the condition to generate non-negative-definite covariance matrices, more specifically they have to satisfy: $\sum_{i,j} a_i a_j C(x_i, x_j) \geq 0$ for any finite set $\{x_1, \cdots, x_n\}$ and any real valued coefficients $\{a_1, \cdots, a_n\}$. Throughout this paper we will only consider second-order stationary processes which have constant means and $Cov[f(x_i), f(x_j)] = C(\|x_i - x_j\|)$. We can see that such covariance functions are invariant to translations.

Squared exponential/radial basis function kernel defines a general class of stationary covariance functions:

$$k^{SE}(x_i, x_j) = \sigma^2 \exp\left[\sum_{d=1}^{D} -\frac{(x_{i,d} - x_{j,d})^2}{l_d^2}\right] \tag{9}$$

, where we have written its definition in the anisotropic case. The emphasis on the domain will make more sense in subsequent subsections where we will introduce kernels operating on Gaussian measures. Intuitively, the lengthscale values $\{l_1^2, \cdots, l_D^2\}$ represent the strength along a particular dimension of input space by which successive values are strongly correlated with correlation invariably decreasing as the distance between points increases. Such a covariance function has the property of Automatic Relevance Determination (ARD) (Neal, 2012). Lastly, the kernel variance σ^2 controls the variance of the process, more specifically the amplitude of function samples.

A GP has the following joint distribution over finite subsets $\mathbb{X}_1 \in \mathbb{X}$ with function values $f(X_1) \in \mathbb{Y}$. Analogously for $\mathbb{X}_2$, with their union being denoted as $x = \{x_1, \cdots, x_n\}$.

$$\begin{pmatrix} f(x_1) \\ f(x_2) \end{pmatrix} = \mathcal{N}\left[\begin{pmatrix} m(x_1) \\ m(x_2) \end{pmatrix}, \begin{pmatrix} k(x_1, x_1), k(x_1, x_2) \\ k(x_2, x_1), k(x_2, x_2) \end{pmatrix}\right] \tag{10}$$

The following observation model is used:

$$p(y|f, x) = \prod_{i=1}^{N} p(y_n|f(x_n)) \tag{11}$$

, where given a supervised learning scenario, the dataset $D = \{x_i, y_i\}_{i=1,\cdots,n}$ can be shorthand denoted as $D = \{x, y\}$. In the case of probabilistic regression, we make the assumption that the noise is additive, independent and Gaussian, such that the latent function $f(x)$ and the observed noisy outputs y are defined by the following equation:

$$y_i = f(x_i) + \epsilon_i, \text{ where } \epsilon_i \sim \mathcal{N}\left(0, \sigma_{noise}^2\right) \tag{12}$$

To train a GP for regression tasks, one performs Marginal Likelihood Maximization of Type 2 over the following equation:

$$p(y) = \mathcal{N}\left(y \mid m, K_{ff} + \sigma_{noise}^2 \mathbb{I}_n\right) \tag{13}$$

$$\propto -(y-m)^\top \left(K_{ff} + \sigma_{noise}^2 \mathbb{I}_n\right)^{-1} (y-m) - \log|\, K_{ff} + \sigma_{noise}^2 \mathbb{I}_n \,| \tag{14}$$

by treating the kernel hyperparameters as point-mass.

We are interested in finding the posterior $p\left(f(x^*) \mid y\right)$ since the goal is to predict for unseen data points x^* which are different than the training set. We know that the joint prior over training set observations and testing set latent functions is given by:

$$\begin{pmatrix} y \\ f(x^*) \end{pmatrix} = \mathcal{N}\left[\begin{pmatrix} m(x) \\ m(x^*) \end{pmatrix}, \begin{pmatrix} k(x,x) + \sigma_{noise}^2 \mathbb{I}_n & k(x, x^*) \\ k(x^*, x) & k(x^*, x^*) \end{pmatrix}\right] \tag{15}$$

Now we can simply apply the conditional rule for multivariate Gaussians to obtain:

$$p\left(f(x^*) \mid y\right) = \mathcal{N}(f(x^* \mid m(x^*) + K_{f^*f}\left[K_{ff} + \sigma^2_{noise}\mathbb{I}_n\right]^{-1}\left[y - m(x)\right], \\ K_{f^*f^*} - K_{f^*f}\left[K_{ff} + \sigma^2_{noise}\mathbb{I}_n\right]^{-1} K_{ff^*}) \tag{16}$$

An illustration of this predictive distribution is given in Figure 2.

GP predictive variance as distributional uncertainty. GPs are clearly *distance-aware* provided we use a translation-invariant kernel. The summary statistics (see definition 1) for an unseen point is given by $u(x^*) = K_{f^*f^*} - K_{f^*f}K_{ff}^{-1}K_{ff^*}$, which is monotonically increasing as a function of distance (see Figure 2). Throughout this paper, we will use the non-parametric variance of sparse variants of GPs as a proxy for distributional uncertainty, which will be used to assess if inputs are in or outside the distribution.

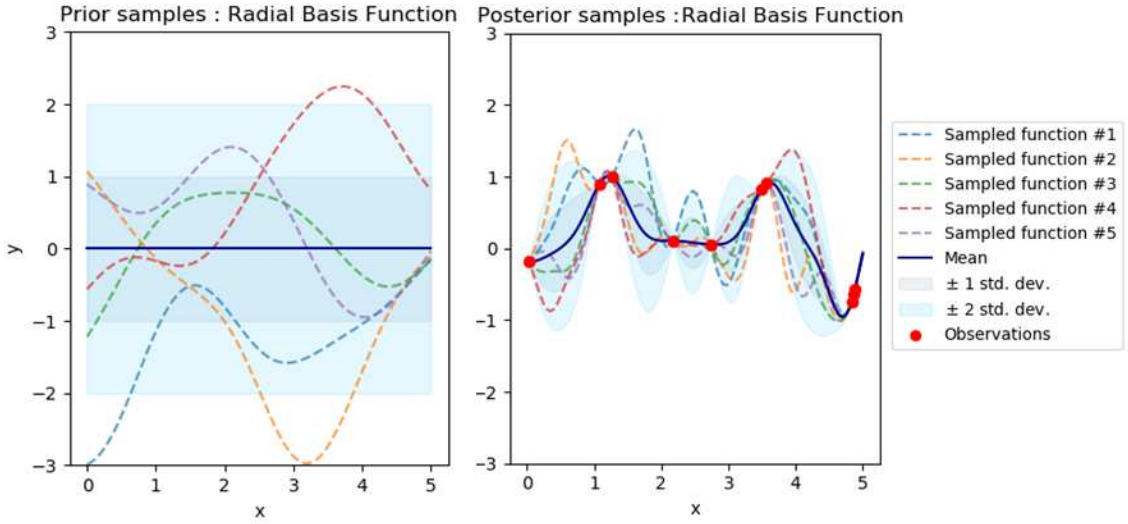


Figure 2: Left : GP prior samples using radial basis function kernel, which ensures a smooth function space hypothesis space; **Right :** GP samples conditioned on observations using radial basis function kernel. Predictive variance increases as input is further away from observations.

The usage of GP in real-world datasets is hindered by matrix inversion operations which have $\mathbb{O}(n^3)$ time, $\mathbb{O}(n^2)$ memory for training, where n is the number of data points in the training set. In the next subsection we will see how to avert having to incur these expensive computational budgets.

2.2 Sparse Variational Gaussian Processes

In this subsection we succinctly review commonly used probabilistic sparse approximations for Gaussian process regression. Quinonero-Candela and Rasmussen (2005) provides a unifying view of sparse approximations by placing each method into a common framework of analyzing their posterior and their *effective prior*, which will be shortly defined.

One way to avert the computationally expensive operators associated to the K_{ff} matrix is to modify the joint prior over $p(f, f^*)$ so that the respective terms depend on a matrix of lower rank, where the joint prior is defined as:

$$\begin{pmatrix} f(x) \\ f(x^*) \end{pmatrix} = \mathcal{N}\left[\begin{pmatrix} m(x) \\ m(x^*) \end{pmatrix}, \begin{pmatrix} k(x,x), k(x,x^*) \\ k(x^*,x), k(x^*,x^*) \end{pmatrix}\right] \tag{17}$$

We introduce an additional set of M latent variables $U = [U_1, \cdots, U_m] \in \mathbb{Y}$ with associated input locations $Z = [Z_1, \cdots, Z_m] \in \mathbb{X}$. Throughout this paper, the former will be entitled inducing point values, respectively the latter inducing point locations.

Due to the consistency property of Gaussian Processes (i.e., for probabilistic model as defined in equation (10) we have $p\left(f(x_1)\right) = \int p\left(f(x)\right)\ df(x_2)$, which ensures that if we marginalize a subset of elements, the remainder will remain unchanged.), one can marginalize out U to recover the initial joint prior over $p(f, f^*)$:

$$p(f, f^*) = \int p(f, f^*, U) dU = \int p(f, f^* \mid U) p(U)\ dU \tag{18}$$

, where $p(U) = \mathcal{N}\left(U \mid 0, K_{uu}\right)$ and K_{uu} is the kernel covariance matrix evaluated at Z.

All sparse approximations to GPs originate from the following approximation:

$$p(f, f^*) \approx q(f, f^*) = \int q(f^* \mid U) q(f \mid U) p(U) dU \tag{19}$$

which translates into a conditional independency between training and testing latent variables given U. Intuitively, the name "inducing points" for $\{Z, U\}$ was given for this precise property, that U induces the values for the training and testing set.

Titsias (2009) introduced the first variational lower bound comprising a probabilistic regressiom model over inducing points. More specifically, the authors applied variational inference in an augmented probability space that comprises training set latent function values F alongside inducing point latent function values U, more specifically using the following generative process in the case of a regression task:

$$p\left(U\right) = \mathcal{N}\left(U \mid 0, K_{uu}; Z\right) \tag{20}$$

$$p\left(F \mid U\right) = \mathcal{N}\left(F \mid K_{fu}K_{uu}^{-1}U, K_{ff} - Q_{ff}; Z, X\right) \tag{21}$$

$$p\left(y \mid F\right) = \mathcal{N}\left(y \mid F, \sigma^2_{noise}\right) \tag{22}$$

, where $Q_{ff} = K_{fu}K_{uu}^{-1}K_{uf}$. We explicitly denoted the dependence of either Z or X, however for decluttering reasons these notations will be dropped unless its not evident on what certain distributions depend on.

In terms of doing exact inference in this new model, respectively computing the posterior $p(f|y)$ and the marginal likelihood $p(y)$, it remains unchanged even with the augmentation of the probability space by U as we can marginalize $p(F) = \int p(F, U) dU$ due to the marginalization properties of Gaussian processes. Succintely, $p(F)$ is not changed by modifying the values of U, even though $p(F|U)$ and $p(U)$ do indeed change. This translates into the fundamental difference between variational parameters U and hyperparameters of the model

$\{\sigma^2_{noise}, \sigma^2, l_1^2, \cdots, l_D^2\}$, whereby the introduction of more variational parameters does not change the fundamental definition of the model before probability space augmentation.

Stochastic Variational Inference (SVI) (Hoffman et al., 2013) enables the application of VI for extremely large datasets, by virtue of performing inference over a set of global variables, which induce a factorisation in the observations and latent variables, such as in the Bayesian formulation of Neural Networks with distributions (implicit or explicit) over matrix weights. GP do no exhibit these properties, but by virtue of the approximate prior over testing and training latent functions for SGP approximations with inducing points U, which we remind here:

$$p(f, f^*) \approx q(f, f^*) = \int p(f \mid U)p(f^* \mid U)p(U)\ dU \tag{23}$$

this translates into a fully factorized model with respect to observations at training and testing time, conditioned on the global variables U.

Our goal is to approximate the true posterior distribution $p(F, U \mid y) = p(F \mid U, Y)p(U \mid Y)$ by introducing a variational distribution $q(F, U)$ and minimizing the Kullback-Leibler divergence:

$$KL\left[q(F,U) \| p(F,U \mid y)\right] = \int q(F,U) \log \frac{q(F,U)}{p(F,U \mid y)}\ dF\ dU \tag{24}$$

, where the approximate posterior factorized as $q(F, U) = p(F \mid U)q(U)$ and $q(U)$ is an unconstrained variational distribution over U. Following the standard VI framework we need to maximize the following variational lower bound on the log marginal likelihood:

$$\log p(y) \geq \int p(F \mid U)q(U) \log \frac{p(y \mid F)p(F \mid U)p(U)}{p(F \mid U)p(U)}\ dF\ dU \tag{25}$$

$$\geq \int q(U) \left[\int \log p(Y \mid F)p(F \mid U)\ dF + \log \frac{p(U)}{q(U)}\right]\ dU \tag{26}$$

We can now solve for the integral over F:

$$\int \log p(y|F)p(F|U)\ dF = \mathbb{E}_{p(F|U)}\left[-\frac{n}{2}\log(2\pi\sigma^2_{noise}) - \frac{1}{2\sigma^2_{noise}} Tr\left[yy^\top - 2yF^\top + FF^\top\right]\right] \tag{27}$$

$$= -\frac{n}{2}\log(2\pi\sigma^2_{noise}) - \frac{1}{2\sigma^2_{noise}} Tr[yy^\top - 2y\left(K_{fu}K_{uu}^{-1}U\right)^\top + \left(K_{fu}K_{uu}^{-1}U\right)\left(K_{fu}K_{uu}^{-1}U\right)^\top + K_{ff} - Q_{ff}] \tag{28}$$

$$= \log \mathcal{N}\left(y|K_{fu}K_{uu}^{-1}U, \sigma^2_{noise}\mathbb{I}_n\right) - \frac{1}{2\sigma^2_{noise}} Tr\left[K_{ff} - Q_{ff}\right] \tag{29}$$

We can now rewrite our variational lower bound as follows:

$$\log p(y) \geq \int q(U) \log \frac{\mathcal{N}\left(y \mid K_{fu}K_{uu}^{-1}U, \sigma^2_{noise}\mathbb{I}_n\right)p(U)}{q(U)}\ dU - \frac{1}{2\sigma^2_{noise}} Tr\left[K_{ff} - Q_{ff}\right] \tag{30}$$

The variational posterior is explicit in this case, respectively $q(F, U) = p(F \mid U; X, Z)q(U)$, where $q(U) = \mathcal{N}(U \mid m_U, S_U)$. Here, m_U and S_U are free variational parameters. Due to the

Gaussian nature of both terms we can marginalize U to arrive at $q(F) = \int p(F \mid U)q(U) = \mathcal{N}(F \mid \tilde{U}(x), \tilde{\Sigma}(x))$, where:

$$\tilde{U}(x) = K_{fu}K_{uu}^{-1}m_U \tag{31}$$

$$\tilde{\Sigma}(x) = K_{ff} - K_{fu}K_{uu}^{-1}\left[K_{uu} - S_U\right]K_{uu}^{-1}K_{uf} \tag{32}$$

The lower bound can be re-expressed as follows:

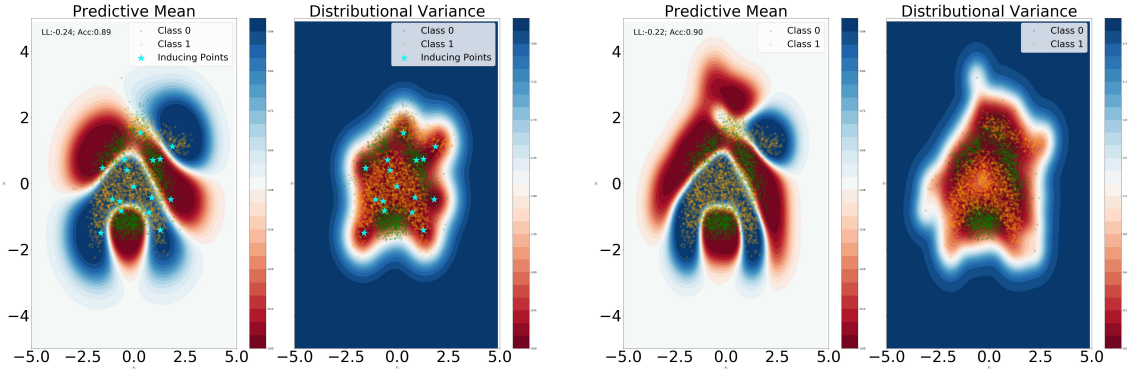
$$\log p(y) \geq \int q(U) \log \mathcal{N}_y\left(K_{fu}K_{uu}^{-1}U, \sigma^2_{noise}\mathbb{I}_n\right) dU - KL\left[q(U)\|p(U)\right] - \frac{1}{2\sigma^2_{noise}}Tr\left[K_{ff} - Q_{ff}\right] \tag{33}$$

We proceed to integrate out U, arriving at the following lower bound:

$$\mathcal{L}_{SVGP} = \mathcal{N}\left(y \mid K_{fu}K_{uu}^{-1}m_U, \sigma^2_{noise}\mathbb{I}_n\right) - \frac{1}{2\sigma^2_{noise}}Tr\left[K_{fu}K_{uu}^{-1}S_UK_{uu}^{-1}K_{uf}\right] \tag{34}$$

$$- \frac{1}{2\sigma^2_{noise}}Tr\left[K_{ff} - Q_{ff}\right] - KL\left[q(U)\|p(U)\right]$$

, where we can easily see that the last equation is factorized with respect to individual observations. This lower variational bound will be denoted as the sparse variational GP (SVGP). This bound is maximized with respect to variational parameters U and hyperparameters of the model $\{Z, \sigma^2_{noise}, \sigma^2, l_1^2, \cdots, l_D^2\}$. An illustration of SVGP trained on the "banana" dataset is given in Figure 3, showing similar behaviour to a GP only using a fraction of training set to obtain similar predictive distribution at testing time.



(a) SVGP (b) GP

Figure 3: Left: Predictive mean and variance of SVGP. Inducing points (teal stars) are tasked to compress the information present in the entire training set such that predictive equations conditioned on them are similar to ones conditioned on entire training set; **Right**: Predictive mean and variance of GP. Not all training points are crucial in devising the decision boundary.

2.3 Uncertainty decomposition in SVGP through evidential learning lens

In subsection 1.2 we have introduced the rationale behind the uncertainty decomposition framework introduced in Malinin and Gales (2018). We now expand on this topic on how to

separate uncertainties in deep evidential learning models (Amini et al., 2019) and make an analogy to how uncertainties are decomposed in SVGP.

For multi-class classification tasks, evidential learning directly parametrizes predictive distributions over the probability simplex. Hence, in comparison to Bayesian Deep Learning or Deep Ensembles it averts parametrizing the logit space, subsequently feeding it through a softmax function. Dirichlet distributions provide an obvious choice for defining a distribution over the $K-1$ dimensional probability simplex, having the following p.d.f.: $Dir\,(\mu,\alpha) = \frac{1}{\beta(\alpha)}\prod_{c=1}^{K}\mu_c^{\alpha_c-1}$, where $\beta(\alpha) = \frac{\prod_{c=1}^{K}\Gamma(\alpha_c)}{\Gamma(\alpha_0)}$ and $\alpha_0 = \sum_{c=1}^{K}\alpha_c$ with $\alpha_c \geq 0$. α_0 is called the precision, being similar to the precision of a Gaussian distribution, where larger α_0 will indicate a sharper distribution.

Dirichlet networks involve having a NN predict the concentration parameters of the Dirichlet distribution $\alpha = f_\theta(x)$, where predictions are made as follows: $\tilde{y} = arg\,\underset{c}{max}\{\frac{\alpha_c}{\alpha_0}\}_{c=1}^{K}$.

We remind ourselves the uncertainty decomposition framework laid out in subsection 1.2:

$$\underbrace{I[y,\mu|x^*,\mathbf{D}]}_{\text{Distributional Uncertainty}} = \underbrace{H[\mathbb{E}_{p(\mu|\mathbf{D})}p(y|x^*,\mu)]}_{\text{Total Uncertainty}} - \underbrace{\mathbb{E}_{p(\mu|\mathbf{D})}[H[P(y|x^*,\mu)]]}_{\text{Aleatoric Uncertainty}} \tag{35}$$

In the case of Dirichlet networks these uncertainty measures have analytic formulas:

$$\mathbb{E}_{p(\mu|\mathbf{D})}[H[P(y|x^*,\mu)]] = -\sum_{c=1}^{K}\frac{\alpha_c}{\alpha_0}\left[\psi(\alpha_c+1) - \psi(\alpha_0+1)\right] \tag{36}$$

$$I[y,\mu|x^*,\mathbf{D}] = -\sum_{c=1}^{K}\frac{\alpha_c}{\alpha_0}\left[\log\frac{\alpha_c}{\alpha_0} - \psi(\alpha_c+1) + \psi(\alpha_0+1)\right] \tag{37}$$

, where ψ is the digamma function. Epistemic uncertainty quantifies the spread in the Dirichlet distribution, hence α_0 can be used to measure it (Charpentier et al., 2021).

Exact inference is not tractable in GP on classification tasks due to the non-conjugacy between the GP prior and the non-Gaussian likelihood (Categorical or Bernoulli). Therefore, approximation are required such as the Laplace approximation (Williams and Barber, 1998), Expectation Propagation (Minka, 2013) or VI (Hensman et al., 2015). Milios et al. (2018) have proposed a method that circumvents these approximate inference techniques by re-branding the classification problem into a regression one, for which exact inference is possible. We commence to briefly lay out the Dirichlet-based GP Classification algorithm.

We consider the probability simplex $\pi = [\pi_1, \cdots, \pi_K] \sim Dir(\alpha)$. We can transform a multi-class classification task into a multi regression scenario where if $y_c = 1$ in a one-hot-encoding, then we can assign $\alpha_c = 1 + \alpha_\epsilon$, respectively $\alpha_c = \alpha_\epsilon$ for $0 \leq \alpha_\epsilon << 1$. The model has the following generative process:

$$\pi \sim Dir(\alpha) \tag{38}$$

$$p(y \mid \alpha) = Cat(\pi) \tag{39}$$

To sample from the Dirichlet distribution we use the following routine: $\pi_c = \frac{x_c}{\sum_{k=1}^{K} x_k}$ with $x_c \sim \Gamma(\alpha_c, 1)$ following the Gamma distribution. From this sampling procedure, we can

see that the generative process translates to independent Gamma likelihoods for each class. Intuitively, at this point in the derivation we need a GP to produce $x_c \geq 0$, since Gamma distributions are only defined on $\mathbb{R}^+$. Since the marginal GP over a subset of data is governed by a multivariate normal it will not satisfy this constraint. To obtain GP sampled functions that respect this constraints, we can use an exp function to transform it. With this in mind, we know that $x \sim \text{log-normal}\left(x \mid \mu, \sigma^2\right) \overset{d}{=} exp(\mathcal{N}\left(x \mid \mu, \sigma^2\right))$, with $\mathbb{E}\left[x\right] = \exp\left(\mu + \frac{\sigma^2}{2}\right)$ and $V\left[x\right] = \left[\exp\sigma^2 - 1\right]\exp\left(2\mu + \sigma^2\right)$. Hence, we can approximate $x_c \sim \gamma(\alpha_c, 1)$ with $\tilde{x_c} \sim \text{log-normal}\left(\tilde{\mu_c}, \tilde{\sigma_c^2}\right)$. To ensure a good approximation, the authors in Milios et al. (2018) propose using moment matching:

$$\mathbb{E}\left[x_c\right] = \alpha_c = \exp\left(\tilde{\mu_c}, \frac{\tilde{\sigma_c^2}}{2}\right) = \mathbb{E}\left[\tilde{x_c}\right] \tag{40}$$

$$V\left[x_c\right] = \alpha_c = \left[\exp\tilde{\sigma_c^2} - 1\right]\exp\left[2\tilde{\mu_c} + \tilde{\sigma_c^2}\right] = V\left[\tilde{x_c}\right] \tag{41}$$

with equality if $\tilde{\mu_c} = \log\alpha_c - \frac{\tilde{\sigma_c^2}}{2}$ and $\tilde{\sigma_c^2} = \log\left(\frac{1}{\alpha_c} + 1\right)$. We can re-express this approximation by taking a natural logarithm, obtaining $\log\tilde{x_c} \sim \mathcal{N}\left(\tilde{\mu_c}, \tilde{\sigma_c^2}\right)$. This translates into a heteroskedastic regression model $\tilde{\mu_c} = f_c + \mathcal{N}\left(0, \tilde{\sigma_c^2}\right)$, where $f_c \sim GP\left(0, K_{ff}\right)$. Hence, one can now apply the standard inference scheme for full GP or we can sparsify the model and apply the SVGP framework. At testing time, the expectation of class probabilities will be:

$$\mathbb{E}\left[\pi_{i,c}\right] = \int \frac{\exp f_{i,c}}{\sum_{k=1}^{C} \exp f_{i,k}} q(f_{i,c})\ df_{i,c} \tag{42}$$

which can be approximated via Monte Carlo integration. In the sparse scenario, $q(f_{i,c}) \sim \mathcal{N}\left(\tilde{U}(x_i), \tilde{\Sigma}(x_i)\right)$ similar to the predictive equations introduced in subsection 2.2. In conclusion, if using Dirichlet-based GP for Classification one can obtain similar estimates of aleatoric and distributional uncertainty in the space of the probability simplex as in equations (36) and (37) specific to Dirichlet Networks. However, for the purposes of this paper we intend to measure distributional uncertainty in the space of logits, as the formulas are simpler to compute and more intuitive from the viewpoint of *distance-awareness*.

As we have previously stated, GPs are *distance-aware*. Thus, they can reliably notice departures from the training set manifold. For SVGP we decompose the model uncertainty into two components:

$$h(\cdot) = \mathcal{N}(h \mid 0, K_{ff} - K_{fu}K_{uu}^{-1}K_{uf}) \tag{43}$$

$$g(\cdot) = \mathcal{N}(g \mid K_{fu}K_{uu}^{-1}m_U, K_{fu}K_{uu}^{-1}S_U K_{uu}^{-1}K_{uf}) \tag{44}$$

The $h(\cdot)$ variance captures the shift from within to outside the data manifold and will be denoted as *distributional uncertainty*. The variance $g(\cdot)$ is termed here as *within-data uncertainty* and encapsulates uncertainty present inside the data manifold. A visual depiction of the two is provided in Figure 14 (bottom). To capture the overall uncertainty in $h(\cdot)$,

thereby also capturing the spread of samples from it, we can calculate it's differential entropy as:

$$h(h) = \frac{n}{2}\log 2\pi + \frac{1}{2}\log \mid K_{ff} - K_{fu}K_{uu}^{-1}K_{uf} \mid + \frac{1}{2} \tag{45}$$

In practice we only use the diagonal terms of the Schur complement, hence the log determinant term will considerably simplify. Intuitively, if terms on the diagonal of the Schur complement have higher values, so will the distributional differential entropy. This OOD measure in logit space will be used throughout the rest of the paper.

2.4 Deep Gaussian Processes fail in propagating distributional uncertainty

Deep Gaussian Processes (DGP) were first introduced in Damianou and Lawrence (2013), as a multi-layered hierarchical formulation of GPs. Composition of processes has retains theoretical properties of underlying stochastic process (such as Kolmogorov extension theorem) while also ensuring a more diverse hypothesis space of process priors, or at least in theory as we shall later see.

We can view the DGP as a composition of functions, keeping in mind that this is only one way of defining this class of probabilistic models (Dunlop et al., 2018):

$$f_L(x) = f_L \circ ... \circ f_1(x) \tag{46}$$

with $f_l = \mathcal{GP}\left(m_l, k_l\left(\cdot,\cdot\right)\right)$. Assuming a likelihood function we can write the joint prior as:

$$p\left(y, \{f_l\}_{l=1}^{L}; X\right) = \underbrace{p(y \mid f_L)}_{\text{likelihood}} \underbrace{\prod_{l=1}^{L} p(f_l \mid f_{l-1})}_{\text{prior}} \tag{47}$$

with $p\left(f_l \mid f_{l-1}\right) \sim \mathcal{GP}\left(m_l(f_{l-1}), k_l\left(f_{l-1}, f_{l-1}\right)\right)$, where in the case we choose squared exponential kernels we have the following formula for the l-th layer:

$$k^{SE}(f_{l,i}, f_{l,j})_l = \sigma_l^2 \exp\left[\sum_{d=1}^{D_l} -\frac{\left(f_{l,i,d} - f_{l,j,d}\right)^2}{l_{l,d}^2}\right]$$

where D_l represents the number of dimensions of F_l and we introduce layer specific kernel hyperparameters $\{\sigma_l^2, l_{l,1}^2, \cdots, l_{l,D_l}^2\}$.

Analytically integrating this Bayesian hierarchical model is intractable as it requires integrating Gaussians which are present in a non-linear way. Moreover, to enable faster inference over our model we can augment each layer l with M_l inducing points' locations Z_{l-1}, respectively inducing points' values U_l resulting in the following augmented joint prior:

$$p\left(y, \{f_l\}_{l=1}^{L}, \{U_l\}_{l=1}^{L}; X, \{Z_l\}_{l=0}^{L-1}\right) = \underbrace{p(y|f_L)}_{\text{likelihood}} \underbrace{\prod_{l=1}^{L} p(f_l|f_{l-1}, U_l; Z_{l-1})p(U_l)}_{\text{prior}} \tag{48}$$

, where $p(f_l \mid f_{l-1}, U_l; Z_{l-1}) = \mathcal{N}\left(f_l \mid m_l(f_{l-1}) + K_{fu}K_{uu}^{-1}\left(U_l - m_l(Z_{l-1}), K_{ff} - K_{fu}K_{uu}^{-1}K_{uf}\right)\right)$. To perform SVI we introduce a factorised variational approximate posterior $q\left(\{U_l\}_{l=1}^{L}\right) =$

$\prod_{l=1}^{L} \mathcal{N}\left(U_l \mid m_{U_l}, S_{U_l}\right)$. Using a similar derivation as in the uncollapsed evidence lower bound for SVGPs, we can arrive at our ELBO for DGPs:

$$\mathcal{L}_{DGP} = \mathbb{E}_{q(\{f_l\}_{l=1}^{L})}\left[\log p\left(y \mid f_L\right)\right] - \sum_{l=1}^{L} KL\left[q(U_l) \| p(U_l)\right] \tag{49}$$

where $q\left(\{f_l\}_{l=1}^{L}\right) = \prod_{l=1}^{L} q\left(f_l \mid f_{l-1}\right)$ and $q\left(f_l \mid f_{l-1}\right) = \mathcal{N}\left(f_l \mid \tilde{U}_l(f_{l-1}), \tilde{\Sigma}_l(f_{l-1})\right)$, respectively:

$$\tilde{U}_l(f_{l-1}) = m_l(f_{l-1}) + K_{fu}K_{uu}^{-1}\left[m_{U_l} - m_l(Z_{l-1})\right] \tag{50}$$

$$\tilde{\Sigma}_l(f_{l-1}) = K_{ff} - K_{fu}K_{uu}^{-1}\left[K_{uu} - S_{U_l}\right]K_{uu}^{-1}K_{uf} \tag{51}$$

This composition of functions is approximated via Monte Carlo integration as introduced in the doubly stochastic variational inference framework for training DGPs (Salimbeni and Deisenroth, 2017).

In Popescu et al. (2020) the authors argued that total uncertainty in the hidden layers of a DGP will be higher for OOD data points in comparison to in-distribution data points only under a set of conditions. We briefly lay out the details here.

Without loss of generality for deeper architectures, we can consider the case of a DGP with two layers and zero mean functions which has the following posterior predictive equation:

$$q(F_2)(x) = \int p(F_2|F_1)q(F_1(x))dF_1 \tag{52}$$

, where $q(F_1(x)) = \mathcal{N}_{f_1}\left(\tilde{U}_1(x), \tilde{\Sigma}_1(x)\right)$. This is similar to the case of approximating GPs with uncertain inputs, in this case Multivariate Normals. In Girard (2004) they lay out a framework for obtaining Gaussian approxiations of GPs with uncertain inputs (in our case the uncertainty stems from the previous layer of the DGP), which when adapted to our case we obtain the following approximate moments for $q(F_2)(x)$:

$$m(F_2) = \tilde{U}_2(\tilde{U}_1(x)) \tag{53}$$

$$v(F_2) = \tilde{\Sigma}_2(\tilde{U}_1(x)) + \tilde{\Sigma}_1(x)\left[\frac{1}{2}\frac{\partial^2\tilde{\Sigma}_2(F_1)}{\partial^2 F_1}\Big|_{F_1=\tilde{U}_1(x)} + \left(\frac{\partial\tilde{U}_2(F_1)}{\partial F_1}\right)^2\Big|_{F_1=\tilde{U}_1(x)}\right] \tag{54}$$

In Popescu et al. (2020) they propose a realistic scenario which occurs frequently in practice, whereby the inducing points Z_l of particular layer are spread out such as to cover the entire spectrum of possible samples from the previous layer F_{l-1}. More precisely, we can consider an OOD data point x_{ood} in input space such that $\tilde{\Sigma}_1(x_{ood}) = \sigma^2$ and $\tilde{U}_1(x_{ood}) = 0$, respectively an in-distribution point x_{in-d} such that $\tilde{\Sigma}_1(x_{in-d}) = V_{in} \leq \sigma^2$ and $\tilde{U}_1(x_{in-d}) = M_{in}$. We also assume that Z_2 are equidistantly placed between $[-3\sigma, 3\sigma]$. The authors go on to show that the total variance in the second layer of x_{ood} will be higher than x_{in-d} if the following holds $\frac{\left(\frac{\partial\tilde{U}_2(F_1)}{\partial F_1}\right)^2\Big|_{F_1=M_{in}}}{\left(\frac{\partial\tilde{U}_2(F_1)}{\partial F_1}\right)^2\Big|_{F_1=0}} \leq \frac{\sigma^2}{V_{in}}$. One can rapidly infer that this

inequality holds if the absolute first order derivative of the parametric component of the SVGP around 0 is higher compared to any other value which might be evaluated at. This observation is to be made in conjunction with the fact that $\frac{\sigma^2}{V_{in}} \geq 1.0$, since the total variance of in-distribution points will be reduced compared to the prior variance.

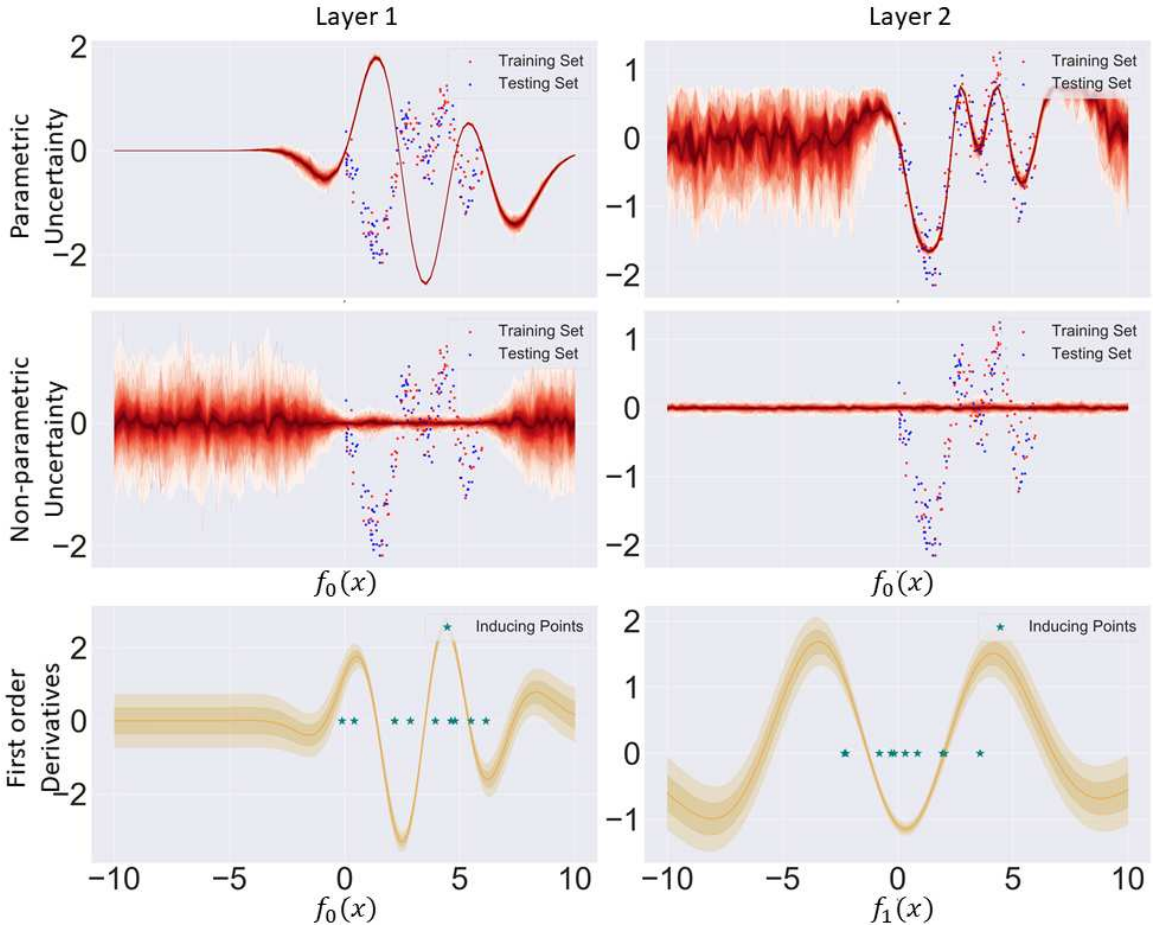


Figure 4: Layer-wise decomposition of uncertainty into parametric/epistemic and non-parametric/distributional for a zero mean function DGP, alongside first order derivatives. OOD points in input space x_{out} get mapped on average to 0 in $f_1(x_{\text{out}})$, which has a high absolute first order derivative causing the parametric uncertainty in f_2 to be high for x_{out}.

To gain some intuition as to what occurs in practice, we can consider a 2 layer DGP trained on a toy regression task, where we decompose the resulting posterior SVGP predictive equation into its parametric and non-parametric components for each layer with respect to input space (first two rows of Figure 4). To investigate whether our trained DGP respects the above inequality for propagating higher total uncertainty for OOD data points in comparison to in-distribution data points, we also need to predict what are the first-order derivatives with respect to the input stemming from the previous layer (last row of Figure 4). We encourage the reader to inspect McHutchon (2013) for an in-depth introduction to first order derivative of GPs. We can notice that the total variance is indeed higher for OOD data

points in the final layer, as this was brought upon by the high absolute value of the first order derivative around 0 in the second layer (OOD data points in the first hidden layer will have an expected value of 0). Intuitively, OOD data points in input space will have higher total uncertainty in output space due to the higher diversity of function values in the second layer. The diversity is caused by the high non-parametric uncertainty in the first hidden layer. Conversely, we can see that for in-distribution points the total variance in the first hidden layer is relatively small, hence the sampling will be close to deterministic, implicitly meaning that it will access only a very restricted set of function values in the second layer thus causing a relatively small total variance. Lastly, we remind ourselves that for GPs we can consider the non-parametric/distributional uncertainty as a proxy for OOD detection. From Figure 4 we can see that distributional uncertainty collapses in the second layer for any value in input space. This implies that DGPs are not *distance-aware*.

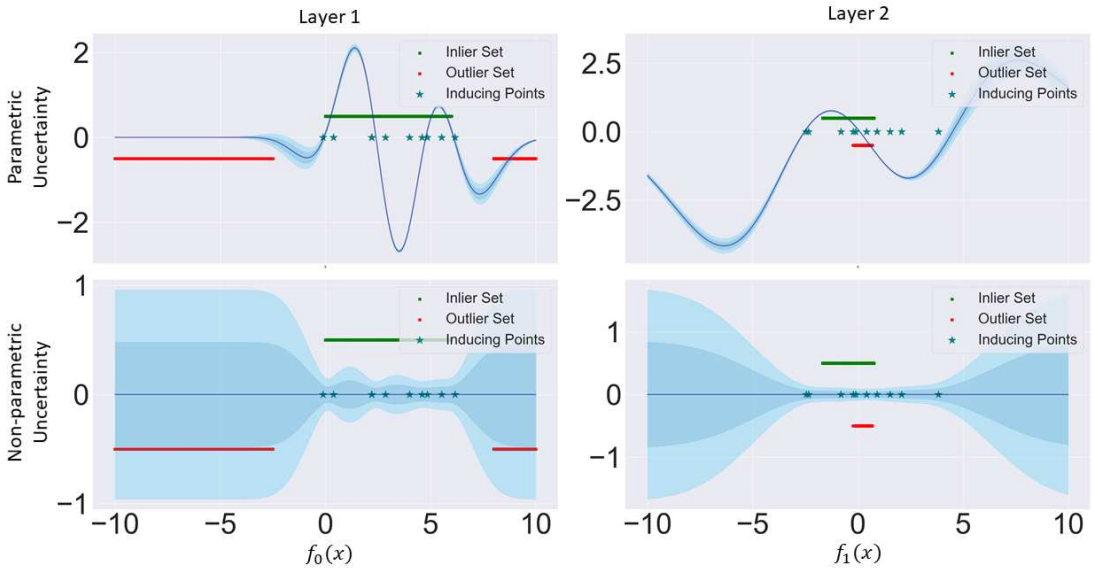


Figure 5: Layer-wise decomposition of uncertainty into parametric/epistemic and non-parametric/distributional for a zero mean function DGP. Outlier points are sampled close to inliers points in f_1, thereby causing their collapse of non-parametric variance since inducing points in f_1 are close to both outlier and inlier samples.

To understand what is causing this pathology, we take a simple case study of a DGP (zero mean function) with two hidden layers trained on a toy regression dataset (Figure 5). Taking a clear outlier in input space, say the data point situated at -7.5, it gets correctly identified as an outlier in the mapping from input space to hidden layer space as given by its distributional variance. However, its outlier property dissipates in the next layer after sampling, as it gets mapped to regions where the next GP assigns inducing point locations. This is due to points inside the data manifold getting confidently mapped between -2.0 and 1.0 in hidden layer space. Consequently, what was initially correctly identified as an outlier will now have its final distributional uncertainty close to zero. Adding further layers, will only compound this pathology.

2.5 Wasserstein-2 kernels for probability measures

As we have seen in the previous subsection, analytically integrating out the prior of a DGP is intractable in the case of using kernels operating in Euclidean space. However, the hidden layers of a DGP are intrinsically defined over probability measures (Gaussian in this case). This leads us to ponder whether we can obtain an analytically tractable formulation of DGPs by using kernels operating on probability measures, thereby we need a metric on probability measures which we subsequently introduce.

The Wasserstein space on $\mathbb{R}$ can be defined as the set $W_2(\mathbb{R})$ of probability measures on $\mathbb{R}$ with a finite moment of order two. We denote by $\Pi(\mu,\nu)$ the set of all probability measures Π over the product set $\mathbb{R}\times\mathbb{R}$ with marginals μ and ν, which are probability measures in $W_2(\mathbb{R})$. The transportation cost between two measures μ and ν is defined as:

$$T_2(\mu,\nu) = \inf_{\pi\in\Pi(\mu,\nu)} \int [x-y]^2 d\pi(x,y) \tag{55}$$

This transportation cost allows us to endow the set $W_2(\mathbb{R})$ with a metric by defining the quadratic Wasserstein distance between μ and ν as:

$$W_2(\mu,\nu) = T_2(\mu,\nu)^{1/2} \tag{56}$$

Theorem 2 (Theorem IV.1. in Bachoc et al. (2017)) *Let $k_W : W_2(\mathbb{R})\times W_2(\mathbb{R}) \to \mathbb{R}$ be the Wasserstein-2 RBF kernel defined as following:*

$$k^{W_2}(\mu,\nu) = \sigma^2 \exp \frac{-W_2^2(\mu,\nu)}{l^2} \tag{57}$$

then $k^{W_2}(\mu,\nu)$ is a positive definite kernel for any $\mu,\nu\in W_2(\mathbb{R})$, respectively σ^2 is the kernel variance, l^2 being the lengthscale.

A detailed proof of this theorem can be found in Bachoc et al. (2017).

Multiplication of positive definite kernels results again in a positive definite kernel, hence we arrive at the automatic relevance determination kernel based on Wasserstein-2 distances:

$$k^{W_2}([\mu_d]_{d=1}^D, [\nu_d]_{d=1}^D) = \sigma^2 \exp \sum_{d=1}^{D} \frac{-W_2^2(\mu_d,\nu_d)}{l_d^2} \tag{58}$$

Wasserstein-2 Distance between Gaussian distributions. Gaussian measures fulfill the condition of finite second order moment, thereby being a clear example of probability measures for which we can compute Wasserstein metrics. The Wasserstein-2 distance between two multivariate Gaussian distributions $\mathcal{N}(m_1,\Sigma_1)$ and $\mathcal{N}(m_2,\Sigma_2)$, which have associated Gaussian measures and implicitly the Wasserstein metric is well defined for them, has been shown to have the following form $\| m_1 - m_2 \|_2^2 + Tr\Big[\Sigma_1 + \Sigma_2 - 2\Big(\Sigma_1^{1/2}\Sigma_2\Sigma_1^{1/2}\Big)^{1/2}\Big]$ (Dowson and Landau, 1982), which in the case of univariate Gaussians simplifies to $|m_1 - m_2|^2 + |\sqrt{\Sigma_1} - \sqrt{\Sigma_2}|^2$. This last formulation will be used throughout this paper.

2.6 Distributional Deep Gaussian Processes & OOD detection

In the previous subsection we have seen that DGPs can easily fail in propagating distributional uncertainty forward. We now focus on the variant of DGPs introduced in Popescu et al. (2020) that was proven both theoretically and empirically to propagate distributional uncertainty throughout the hierarchy, thus ensuring *distance-awareness* properties. The insights gained from this subsection will constitute the departure point for our proposed model in the next section.

Distributional Gaussian Processes (DistGP) were first introduced in (Bachoc et al., 2017) to describe a shallow GP that operates on probability measures using a Wasserstein-2 based kernel as defined in equation (58).

We introduce the generative process of Distributional Deep Gaussian Processes (DDGP) for 2 layers:

$$p(F_1) \sim \mathcal{N}(0, K_{ff}) \tag{59}$$

$$F_1^{sth} = m(F_1) + \sqrt{v(F_1)}\epsilon,\ \epsilon \sim \mathcal{N}(0, \mathbb{I}_n) \tag{60}$$

$$F_1^{det} = \mathcal{N}(m(F_1), diag[v(F_1)]) \tag{61}$$

$$p(F_2) \sim \mathcal{N}\left(0, k_{hybrid}\left(\{F_1^{sth}, F_1^{det}\}, \{F_1^{sth}, F_1^{det}\}\right)\right) \tag{62}$$

, where the hybrid kernel is defined as follows:

$$k^{hybrid}(\mu_i, \mu_j) = k^E(x_i, x_j) \exp \sum_{d=1}^{D} \frac{-W_2^2(\mu_{i,d}, \mu_{j,d})}{l_d^2} \tag{63}$$

, where we denoted $\mu_i = \mathcal{N}(m(F_1(x_i)), \sigma_1^2)$; $\mu_j = \mathcal{N}(m(F_1(x_j)), \sigma_1^2)$ as the first two moments which are obtained through the F_1^{det} operation in the generative process. Intuitively this generative process implies keeping track of a *stochastic*, respectively *deterministic* component of the same SVGP at any given hidden layer, while the first layered is governed by a standard SVGP operating on Euclidean data. It is worthy to point out that for this probabilistic construction, the inducing points $\{Z_l\}_{l=1}^L$ have to reside in the space of multivariate Gaussians, hence $Z_l \sim \mathcal{N}(Z_l \mid \mu_{Z_l}, \Sigma_{Z_l})$. The first two moments are treated as hyperparameters that are optimized during training.

We can consider an OOD data point x_{ood} in input space such that $\tilde{\Sigma_1}(x_{ood}) = \sigma^2$ and $\tilde{U_1}(x_{ood}) = 0$, respectively an in-distribution point x_{in-d} such that $\tilde{\Sigma}_1(x_{in-d}) = V_{in} \leq \sigma^2$ and $\tilde{U}_1(x_{in-d}) = M_{in}$. We also assume that Z_2 are equidistantly placed between $[-3\sigma, 3\sigma]$. The authors go on to show that the total variance in the second layer of x_{ood} will be higher than x_{in-d} if the following holds $\sigma^2 >> Z_{2,var}$ and $V_{in} \approx Z_{2,var}$. To better understand this behaviour, we can consider a two-layered DGPs and DDGPs, we assume an in-distribution point to have low total variance in hidden layer F_1, respectively an OOD point to have high total variance. In the DGP case, upon sampling from $q(F_1(x_{in-d}))$ and $q(F_1(x_{ood}))$ we can end up with samples which are equally distance with respect to inducing points' location Z_1. If this occurs, then non-parametric variance (proxy for distributional variance) will be equal for x_{in-d} and x_{ood} in F_2. Hence, what was initially flagged as OOD in the first hidden layer will be considered as in-distribution by the second hidden layer. onversely, in the DDGP case and under the assumption that the variance of distributional inducing points' locations

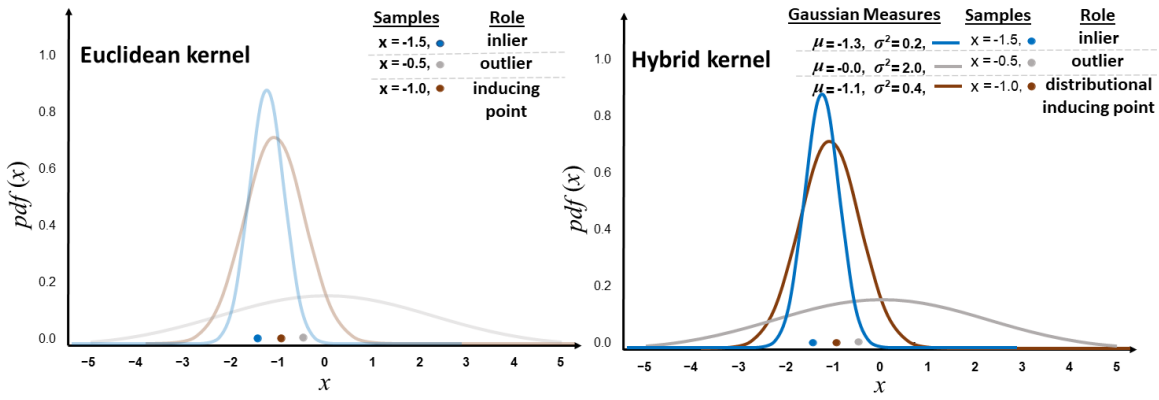


Figure 6: Conceptual difference between euclidean and hybrid kernel.

Z_2 is almost equal in distribution to the total variance of in-distribution points in F_2, the Wasserstein-2 component of the hybrid kernel will notice that there is a higher distance between the now distributional inducing point location and x_{ood}, as opposed of former with x_{in-d}. Then, the non-parametric variance of x_{ood} will be higher than that of x_{in-d}. A visual depiction of this case study is illustrated in Figure 6.

3. Distributional GP Layers

Shift towards single-pass uncertainty quantification Early methods for uncertainty quantification in Bayesian deep learning (BDL) have focused on estimating the variance of sample from difference sub-models, such as in using dropout (Gal and Ghahramani, 2016b), deep ensembles (Lakshminarayanan et al., 2017) or in sampling posterior network weights from a hypernetwork (Pawlowski et al., 2017). This results in slow uncertainty estimation at testing time, which can be critical in high-risk domains where speed is of essence (e.g., self-driving cars). Recent work in OOD detection has focused on estimating proxies for distributional uncertainty in a single-pass, such as in bi-Lipschitz regularized feature extractors for GP (van Amersfoort et al., 2021; Liu et al., 2020) or in parametrizing second-order uncertainty via neural networks within the framework of evidential learning (Charpentier et al., 2020; Amini et al., 2019). With this shift towards single-pass uncertainty quantification, DDGPs and the hybrid kernel introduced in subsection 2.6 are no longer appropriate since they involved sampling the features at each hidden layer. In next subsection we detail a deterministic variant which still preserves correlations between data points in the hidden layers.

Integrating GPs in convolutional architectures GP for image classification has garnered interest in the past years, with hybrid approaches, whereby a deep neural network embedding mechanism is trained end-to-end with GPs as the classification layer, being the first attempt to unify the two approaches (Wilson et al., 2016; Bradshaw et al., 2017). Garriga-Alonso et al. (2018) provided a conceptual framework by which classic CNN architectures are translated into the kernel of a shallow GP by exploiting the mathematical properties of the variance of weights matrices. Van der Wilk et al. (2017) proposed the first convolutional

kernel, constructed by aggregating patch response functions. Dutordoir et al. (2019) have attempted to solve the issue with complete spatial invariance of the convolutional kernel by adding an additional squared exponential kernel between the locations of two patches to account for spatial location, obtaining improvements in accuracy. To extend this shallow GP model to accommodate deeper architectures, Blomqvist et al. (2018) have proposed to use the convolutional GP on top of a succession of feed-forward GP layers which process data in a convolutional manner akin to standard convolutional layers. However, scaling this framework to modern convolutional architectures with large number of channels in each hidden layer is problematic for two reasons. Firstly, this would imply training high-dimensional multi-output GPs which still represents a research avenue on how to make it more efficient (Bruinsma et al., 2020). Ignoring correlations between channels would severely diminish the expressivity of the model. Secondly, the hidden layer GP which process data in a convolutional manner implies taking inducing points, with a dimensionality which scales linearly with the number of channels. This would imply optimization over high-dimensional spaces for each hidden layer, potentially leading to local minima. We will see later on an alternative to this framework for integrating GP in a convolutional architecture, one that is more amenable to modern convolutional architectures.

3.1 Deep Wasserstein Kernel Learning

3.1.1 Generative Process

We now write the generative process of this new probabilistic framework coined Deep Wasserstein Kernel Learning (DWKL) for 2 layers:

$$p(F_1) = \mathcal{N}\left(PCA(F_0), diag\left[K_{ff}\right]\right) \tag{64}$$

$$p(F_2) = \mathcal{N}\left[0, k^{W_2}\left(p(F_1), p(F_1)\right)\right] \tag{65}$$

Due to the introduction of PCA mean functions, data points in the hidden layer are now correlated. To make this clear, we can explicitly calculate it:

$$p\begin{pmatrix} F_{2,i} \\ F_{2,j} \end{pmatrix} \sim \mathcal{N}\left[\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma_2^2 \exp -\frac{-W_2^2(\mu_i,\mu_i)}{l^2} & \sigma_2^2 \exp -\frac{-W_2^2(\mu_i,\mu_j)}{l^2} \\ \sigma_2^2 \exp -\frac{-W_2^2(\mu_j,\mu_i)}{l^2} & \sigma_2^2 \exp -\frac{-W_2^2(\mu_j,\mu_j)}{l^2} \end{pmatrix}\right] \tag{66}$$

$$\sim \mathcal{N}\left[\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma_2^2 & K_{i,j}^{W_2} \\ K_{j,i}^{W_2} & \sigma_2^2 \end{pmatrix}\right] \tag{67}$$

where $\mu_i = \mathcal{N}\left(PCA(F_{0,i}), \sigma^2\right)$ and $\mu_j = \mathcal{N}\left(PCA(F_{0,j}), \sigma^2\right)$. If the PCA embeddings of x_i and x_j are different, then the Wasserstein-2 distance will be different than zero, hence introducing correlations.

3.1.2 Evidence lower bound

Deep Kernel Learning (DKL) (Wilson et al., 2016) is defined as a shallow GP with the input encoded by a neural network:

$$p(Y, F_L, U_L) = \underbrace{p(Y \mid F_L)}_{\text{likelihood}} \underbrace{p(F_L \mid U_L; Z_{L-1}, Enc(X))p(U_L)}_{\text{prior}} \tag{68}$$

,where $Enc(X)$ represents the input passed through a neural network encoder, providing a deterministic transformation of the data which is then fed into a SVGP operating on Euclidean data (using equation (9)).

We diverge from this approach by utilising stacked DistGP with Wasserstein-2 kernels as the encoder network, hence our transformed input given by a Gaussian distribution $q(F_{L-1})$. Using the first two moments of the penultimate layer, we introduce a DistGP so as to obtain the final predictions. The conditional equation for DistGP at arbitrary layer $l \geq 2$ is written as:

$$p(F_l \mid U_l; Z_{l-1}, F_{l-1}) = \mathcal{N}(F_l \mid K_{fu}^{W_2} K_{uu}^{W_2\,-1} U, K_{ff}^{W_2} - Q_{ff}^{W_2}) \tag{69}$$

, where we have inducing points $Z_l \sim \mathcal{N}(Z_l \mid \mu_{Z_l}, \Sigma_{Z_l})$ and uncertain input $F_{l-1} \sim \mathcal{N}\left(F_{l-1} \mid \tilde{U_{l-1}}(F_{l-2}), \tilde{\Sigma_{l-1}}(F_{l-2})\right)$. For computational reasons we take both $\tilde{\Sigma_{l-1}}(F_{l-2})$ and Σ_{Z_l} to be diagonal matrices. For $l = 1$ we have $q(F_1) \sim \mathcal{N}\left(F_1 \mid \tilde{U}_1(x), \tilde{\Sigma}_1(x)\right)$ which are the standard predictive equations for SVGP as given in equations (31) and (32) since the first layer is governed by a standard SVGP operating on Euclidean data.

The joint density prior of Deep Wasserstein Kernel Learning (DWKL) is given as:

$$\underbrace{p(Y|F)}_{\text{likelihood}} \underbrace{p(F_L \mid U_L; Z_{L-1}, Enc(X)) \prod_{l=1}^{L} p(U_l)}_{\text{prior}} \tag{70}$$

, where in our case $Enc(X) \sim \mathcal{N}\left[\tilde{U}_{L-1}(F_{L-2}), \tilde{\Sigma}_{L-1}(F_{L-2})\right]$ that acts as the uncertain input for the final distributional GP. We introduce a factorized posterior between layers and dimensions $q(F_L, \{U_l\}_{l=1}^L) = p(F_L|U_L; Z_{L-1}) \prod_{l=1}^L q(U_l)$, where $q(U_l)$ is taken to be a multivariate Gaussian with mean m_{U_l} and variance S_{U_l}. This gives the DWKL variational lower bound:

$$\mathcal{L}_{DKWL} = \mathbf{E}_{q(F_L, \{U_l\}_{l=1}^L)} p(Y \mid F_L) - \sum_{l=1}^{L} KL\left[q(U_l) \| p(U_l)\right] \tag{71}$$

, where $q(F_L) = \mathcal{N}(\tilde{U}_L(Enc(X)), \tilde{\Sigma}_L(Enc(X)))$. For $1 \leq l \leq L-1$, F_l act as features for the next kernel as opposed to random variables that need to be integrated out. We provide pseudo-code of the previously mentioned operations (see Algorithm 1).

3.2 Module Architecture

For ease of notation and graphical representation we describe the case of the input being a 2D image, with no loss of generality. We denote the image's representation $F_l \in \mathbb{R}^{H_l, W_l, C_l}$ with width W_l, height H_l and C_l channels at the l-th layer of a multi-layer model. F_0 is the image. Consider a square kernel of size $k_l \times k_l$. We denote with $F_l^{[p,k_l]} \in \mathbb{R}^{k_l, k_l, C_l}$ the p-th patch of F_l, which is the area of F_l that the kernel covers when overlaid at position p during convolution (e.g., orange square for a 3×3 kernel in Figure 7). We introduce the convolved $GP_0 : F_0^{[p,k_0]} \to \mathcal{N}(m, k)$ with $Z_0 \in \mathbb{R}^{k_0, k_0, C_0}$ to be the SGP operating on the Euclidean space of patches of the input image in a similar fashion to the layers introduced in Blomqvist

Algorithm 1: Deep Wasserstein Kernel Learning

Input: Euclidean data $X = F_0$
First layer is standard sparse variational GP
Variational Parameters: $U_1 \sim \mathcal{N}(m_{U_1},\ \Sigma_{U_1})$
Inducing Points: Euclidean space Z_0
$q(F_1) = \mathcal{N}(F_1 \mid K_{fu}^{SE} K_{uu}^{SE^{-1}} m_{U_1}, K_{ff}^{SE} - K_{fu}^{SE} K_{uu}^{SE^{-1}} (K_{uu}^{SE} - S_{U_1}) K_{uu}^{SE^{-1}} K_{uf}^{SE}$
for $l = 2$ **to** L **do**
  Hidden layers are distributional sparse variational GP
  Variational Parameters: $U_l \sim \mathcal{N}(m_{U_l},\ S_{U_l})$
  Inducing Points: $Z_{l-1} \sim \mathcal{N}(\mu_{Z_{l-1}},\ \Sigma_{Z_{l-1}})$
  Compute $K_{fu}^{W_2}$**:** $\sigma_l^2 \exp \sum_{d=1}^{D_l} \frac{-W_2^2(q(F_{l-1}[:,d]), Z_{l-1}[:,d])}{l_{l,d}^2}$
  Compute $K_{uu}^{W_2}$**:** $\sigma_l^2 \exp \sum_{d=1}^{D_l} \frac{-W_2^2(Z_{l-1}[:,d], Z_{l-1}[:,d])}{l_{l,d}^2}$
  $q(F_l) = \mathcal{N}(F_l \mid K_{fu}^{W_2} K_{uu}^{W_2{}^{-1}} m_{U_l}, K_{ff}^{W_2} - K_{fu}^{W_2} K_{uu}^{W_2{}^{-1}} \left[K_{uu}^{W_2} - S_{U_l}\right] K_{uu}^{W_2{}^{-1}} K_{uf}^{W_2}$
end for
Maximize ELBO: $\mathbb{E}_{q(F_L), \{q(U_l\}_{l=1}^L)} p(Y \mid F_L) - \sum_{l=1}^L KL\left[q(U_l) \| p(U_l)\right]$

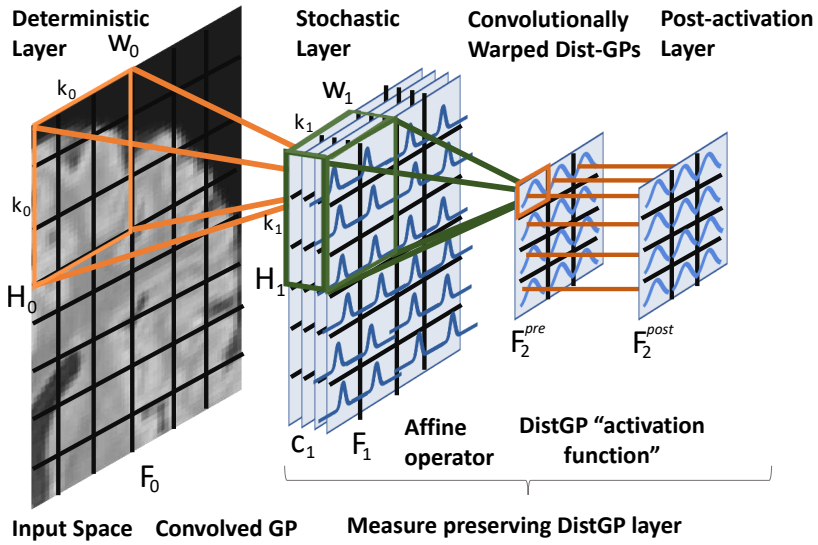


Figure 7: Schematic of measure-preserving DistGP layer. Sparse variational GP is convolved on input data to obtain first hidden layer. Affine operator is convolved on stochastic layer, propagating both mean and variance to obtain the pre-activation of the second hidden layer. Distributional GP is applied element-wise to introduce non-linearities and to propagate distributional uncertainty in the post-activation of the second hidden layer.

et al. (2018). For $1 \leq l \leq L$ we introduce affine operators $A_l \in \mathbb{R}^{k_l, k_l, C_{l-1}, C_{l,pre}}$ which are convolved on the previous stochastic layer in the following manner:

$$m(F_l^{pre}) = \text{Conv}_{2\text{D}}(m(F_{l-1}), A_l) \tag{72}$$
$$var(F_l^{pre}) = \text{Conv}_{2\text{D}}(var(F_{l-1}), A_l \odot A_l) \tag{73}$$

, where $\odot$ represents the Hadamard product. The affine operator is sequentially applied on the mean, respectively variance components of the previous layer F_{l-1} so as to propagate the Gaussian distribution to the next pre-activation layer F_l^{pre}. To obtain the post-activation layer, we apply a $DistGP_l : F_l^{pre,[p,1]} \rightarrow \mathcal{N}(m, k)$ in a many-to-one manner on the pre-activation patches to arrive at F_l^{post}. Figure 7 depicts this new module, entitled "Measure preserving DistGP" layer with pseudo-code offered in Algorithm 2. In Blomqvist et al. (2018) the convolved GP is used across the entire hierarchy, thereby inducing points are in high-dimensional space ($k_l^2 * C_l$). In our case, the convolutional process is replaced by an inducing points free affine operator, with inducing points in low-dimensional space ($C_{l,pre}$) for the DistGP activation functions. The affine operator outputs $C_{l,pre}$, which is taken to be higher than the associated output space of DistGP activation functions C_l. Hence, the affine operator can cheaply expand the channels, in constrast to the layers in Blomqvist et al. (2018) which would require high-dimensional multi-output GP. We motivate the preservation of distance in Wasserstein-2 space in the following section. Previous research has highlighted the importance of having an upper bound on $||h(x_1) - h(x_2)||_h \leq L_{upper}||x_1 - x_2||_x$, as it ensures a certain degree of robustness towards adversarial examples, since it prevents the hidden forward mappings from being overly sensitive to the conceptually meaningless perturbations in input space (Jacobsen et al., 2018; Sokolić et al., 2017; Weng et al., 2018). Conversely, the lower bound $||h(x_1) - h(x_2)||_h \geq L_{lower}||x_1 - x_2||_x$ ensures that the forward mappings do not become invariant to semantically meaningful changes in the input van Amersfoort et al. (2020).

3.3 Imposing Lipschitz Conditions in Convolutionally Warped DistGP

If a sample is identified as an outlier at certain layer, respectively being flagged with high variance, in an ideal scenario we would like to preserve that status throughout the remainder of the network. As the kernels operate in Wasserstein-2 space, the distance of a data point's first two moments with respect to inducing points is vital. Hence, we would like our network to vary smoothly between layers, so that similar objects in previous layers get mapped into similar spaces in the Wasserstein-2 domain. In this section, we accomplish this by quantifying the *"Lipschitzness"* of our "Measure preserving DistGP" layer and by imposing constraints on the affine operators so that they preserve distances in Wasserstein-2 space.

Proposition 3 *For a given DistGP F and a Gaussian distribution $\mu \sim \mathcal{N}(m_1, \Sigma_1)$ to be the centre of an annulus $B(x) = \{\nu \sim \mathcal{N}(m_2, \Sigma_2) \mid 0.125 \leq \frac{W_2(\mu,\nu)}{l^2} \leq 1.0$ and choosing any ν inside the ball we have the following Lipschitz bounds: $W_2(F(\mu), F(\nu)) \leq LW_2(\mu, \nu)$, where $L = (\frac{4\sigma^2}{l})^2 \left[\|K_{uu}^{-1}m\|_2^2 + \|K_{uu}^{-1}(K_{uu} - S)K_{uu}^{-1}\|_2\right]$ and l, σ^2 are the lengthscales and variance of the kernel.*

Proof is given in Appendix A.

Algorithm 2: Distributional Gaussian Processes Layers

Input: Euclidean data $X = F_0 \in \mathbb{R}^{H_0,W_0,C_0}$
First layer is convolved sparse variational $GP_0 : F_0^{[p,k_0]} \to F_1^{[p]}$
Variational Parameters: $U_1 \sim \mathcal{N}(m_{U_1},\ S_{U_1})$
Inducing Points: Euclidean space $Z_0 \in \mathbb{R}^{k_0,k_0,C_0}$
$q(F_1) = \mathcal{N}(F_1 \mid K_{fu}^{SE} K_{uu}^{SE^{-1}} m_{U_1}, K_{ff}^{SE} - K_{fu}^{SE} K_{uu}^{SE^{-1}} (K_{uu}^{SE} - S_{U_1}) K_{uu}^{SE^{-1}} K_{uf}^{SE}$
for $l = 2$ **to** L **do**
  affine operators: $A_l \in \mathbb{R}^{k_l,k_l,C_{l-1},C_{l,pre}}$
  $m(F_l^{pre}) = Conv_{2D}(m(F_{l-1}), A_l)$
  $v(F_l^{pre}) = Conv_{2D}(var(F_{l-1}), A_l^2)$
  Hidden layer activation functions are sparse variational GP
  $DistGP_l : F_l^{pre,[p,1]} \to F_l^{post,[p,1]}$
  Variational Parameters: $U_l \sim \mathcal{N}(m_{U_l},\ S_{U_l})$
  Inducing Points: $Z_{l-1} \sim \mathcal{N}(\mu_{Z_{l-1}},\ \Sigma_{Z_{l-1}})$
  Compute $K_{fu}^{W_2}$**:** $\sigma_l^2 \exp \sum_{d=1}^{D_l} \frac{-W_2^2(q(F_{l-1}[:,d]),Z_{l-1}[:,d])}{l_{l,d}^2}$
  Compute $K_{uu}^{W_2}$**:** $\sigma_l^2 \exp \sum_{d=1}^{D_l} \frac{-W_2^2(Z_{l-1}[:,d],Z_{l-1}[:,d])}{l_{l,d}^2}$
  $q(F_l^{post}) = \mathcal{N}(F_l \mid K_{fu}^{W_2} K_{uu}^{W_2{}^{-1}} m_{U_l}, K_{ff}^{W_2} - K_{fu}^{W_2} K_{uu}^{W_2{}^{-1}} \left[K_{uu}^{W_2} - S_{U_l}\right] K_{uu}^{W_2{}^{-1}} K_{uf}^{W_2}$
end for
Maximize ELBO: $\mathbb{E}_{q(F_L),\{q(U_l\}_{l=1}^L)} p(Y \mid F_L) - \sum_{l=1}^L KL\left[q(U_l) \| p(U_l)\right]$

Remark 4 *This theoretical result shows that DistGP* "activation functions" *have Lipschitz constants with respect to the Wasserstein-2 metric in both output and input domain. This will ensure that the distance between previously identified outliers and inliers will stay constant. However, it is worthy to highlight that we can only obtain locally Lipschitz continuous functions, given that we can only obtain Lipschitz constants for any Gaussian distribution ν inside the annulus $B(x) = \{\nu \sim \mathcal{N}(m_2, \Sigma_2) \mid 0.125 \leq \frac{W_2(\mu,\nu)}{l^2} \leq 1$. with respect to the centre of the annulus, μ.*

We are now interested in finding Lipschitz constants for the affine operator A that gets convolved to arrive at the pre-activation stochastic layer.

Proposition 5 *We consider the affine operator $A \in \mathbb{R}^{C,1}$ operating in the space of multivariate Gaussian distributions of size C. Consider two distributions $\mu \sim \mathcal{N}(m_1, \sigma_1^2)$ and $\nu \sim \mathcal{N}(m_2, \sigma_2^2)$, which can be thought of as elements of a hidden layer patch, then for the affine operator function $f(\mu) = \mathbb{N}(m_1 A, \sigma^2 A^2)$ we have the following Lipschitz bound: $W_2(f(\mu), f(\nu)) \leq L W_2(\mu, \nu)$, where $L = \sqrt{C}\|W\|_2^2$.*

Proof is given in Appendix A.

Remark 6 *We denote the l-th layer weight matrix, computing the c-th channel by column matrix $A_{l,c}$. We can impose the Lipschitz condition to Eq. 72, 73 by having constrained weight matrices with elements of the form $A_{l,c} = \frac{A_{l,1}}{C^{\frac{1}{2}} \sqrt{\sum_{c=1}^C W_{l,c}^2}}$.*

3.4 *Feature-collapse* in DistGP layers

In this subsection we delve deeper into the properties of DistGP layers from a function-space view. In light of recent interest into *feature collapse* van Amersfoort et al. (2020), which is the pathological phenomenon of having the representation layer collapse to a small finite set of values, with catastrophic consequences for OOD detection, we investigate what are the necessary conditions for our proposed network to collapse in feature space. Subsequently, we investigate if feature collapse is inherently encouraged by our loss function.

We commence by introducing notation conventions. We consider $\{u_l \in \mathbb{R}^{D_l}\}_{l=0:L}$ where D_l is the number of dimensions in the l-th layer of the hierarchy. We consider the following two functions $\Psi_l : u_{l-1} \to \mathcal{R}^{m_l}$ and $f_l : \mathcal{R}^{m_l} \to u_l$. To relate this notation to our construction of a DistGP layer introduced in section 3.2, m_l represents the number of dimensions of the warped GP (warping performed by affine deterministic layer; see dark green arrows in Figure 7). We denote by f_l to be the DistGP (mean function included) taking values in the space of continuous functions $C(u_l; \mathcal{R}^{m_l})$, which relates to the "activation function" construction from Figure 7. Then we have the following composition for a given DistGP layer:

$$u_l(x) = f_l\left(\Psi_l(u_{l-1})(x)\right) \tag{74}$$

One can easily see that DWKL can be recovered by taking $\Psi_l = id$, instead of the affine embedding. The first layer prior $p\begin{pmatrix} u_1(x) \\ u_1(x^*) \end{pmatrix}$ is defined as follows:

$$\mathcal{N}\left[\begin{pmatrix} m_1(x) \\ m_1(x^*) \end{pmatrix}, \begin{pmatrix} \sigma_1^2 & k^E\left(x, x^*\right) \\ k^E\left(x^*, x\right) & \sigma_1^2 \end{pmatrix}\right] \tag{75}$$

We now define the prior post-activation layers $p\begin{pmatrix} u_l(x) \\ u_l(x^*) \end{pmatrix}$ for $l \geq 2$ in the following recursive manner:

$$\mathcal{N}\left[\begin{pmatrix} m_l(x) \\ m_l(x^*) \end{pmatrix}, \begin{pmatrix} \sigma_l^2 & k^{W_2}\left(\mu_{l-1}(x), \mu_{l-1}(x^*)\right) \\ k^{W_2}\left(\mu_{l-1}(x^*), \mu_{l-1}(x)\right) & \sigma_l^2 \end{pmatrix}\right] \tag{76}$$

, where $\mu_l\left(x\right) = \mathcal{N}\left(m_{l-1}(x)W_l, \sigma_{l-1}^2 W_l^2\right)$, where $m_{l+1}(\cdot) = \overline{\overline{m_1(\cdot)W_1} \cdots W_l}$ and $\overline{m_1(x)W_1}$ signifies having the Principal Component Analysis (PCA) mean function of the first layer multiplied by W_1 and averaged across its dimensions.

Proposition 7 *We assume μ_0 to be bounded on bounded sets almost-surely. If at each layer we have satisfied the following inequality $D_l^2\langle\tilde{W}_l, \tilde{W}_l\rangle \leq 1$, respectively $\left[D_L * \langle\tilde{W}_L, \tilde{W}_L\rangle + \frac{\sigma_L^2}{2l_L^2}\right] \leq 1$, where D_l is the size of the l-th layer and $\tilde{W}_l$ represents a normalized version of the affine embedding W_l, we have the following result:*

$$P\left(\|u_n(x) - u_n(x^*)\|_2 \to 0\right) = 1 \tag{77}$$

The proof of Proposition 3 can be found in Appendix B.

Remark 8 *As we have previously outlined in the above derivation, if at each layer we have satisfied the following inequality* $D_l m_{l-1} D_{l-1} < \tilde{W}_l, \tilde{W}_l > \leq 1$*, respectively* $\left[m_l D_{L-1} * \langle \tilde{W}_L, \tilde{W}_L \rangle + \frac{\sigma_L^2}{2l_L^2} \right] \leq 1$ *then the network collapses to constant values. Intuitively, if the norm of* W_l *is not large enough, then it won't change the Gaussian random field too much. Furthermore, if* σ_l^2 *is larger, which translates in increased amplitude of the samples from the Gaussian random field, then the values will not collapse. As opposed to the hypothetical requirements for DGP Dunlop et al. (2018), we can immediately notice that for DistGP layers there is no requirement for the kernel variance and lengthscales from intermediate layers, relying solely on the last layer hyperparameters. Lastly, we can notice that as the width of the stochastic layers is increased, alongside warped layers through affine embedding, the conditions are less likely to be satisfied.*

3.5 Over-correlation in latent space

Ober et al. (2021) has highlighted a certain pathology in DKL applied to regression problems in the non-sparse scenario. The authors provide empirical examples of this pathology, whereby features in the representation learning layer are almost perfectly correlated, which would correspond to the feature collapse phenomenon as coined in van Amersfoort et al. (2020). We commence by briefly introducing the main results from that paper and then adapt them to the sparse scenario, which bears more resemblance to what occurs in practice.

Full GPs are trained via type-II maximum likelihood:

$$\log p(y) = \log \mathcal{N}\left(y \mid 0, K_{ff} + \sigma^2_{noise}\mathbb{I}_n\right) \tag{78}$$

$$\propto \underbrace{-\frac{1}{2}\log \mid K_{ff} + \sigma^2_{noise}\mathbb{I}_n \mid}_{\textit{complexity penalty}} \underbrace{-\frac{1}{2} y^\top \left(K_{ff} + \sigma^2_{noise}\mathbb{I}_n\right)^{-1} y}_{\textit{data fit}} \tag{79}$$

, where we define the squared exponential kernel $k^{SE}\left(x_i, x_j\right) = \sigma^2 \exp\left[\sum_{d=1}^{D} -\frac{\left(x_{i,d} - x_{j,d}\right)^2}{2l_d^2}\right]$ for $x_i, x_j \in \mathbb{R}^D$.

The authors in Ober et al. (2021) go on to show that at optimal values, the data fit term will converge towards $\frac{N}{2}$, where N is the number of training points. Hence, once the model has reached convergence, it can only increase its log-likelihood score by modifications to the *complexity penalty* term, which can be broken up as follows:

$$\frac{1}{2}\log \mid K_{ff} + \sigma^2_{noise}\mathbb{I}_n \mid = \frac{N}{2}\log \sigma_f^2 + \frac{1}{2}\log \mid \tilde{K_{ff}} + \tilde{\sigma^2_{noise}}\mathbb{I}_n \mid \tag{80}$$

, where we introduced the reparametrizations $K_{ff} = \sigma^2 \tilde{K_{ff}}$ and $\sigma^2_{noise} = \sigma^2 \tilde{\sigma^2_{noise}}$. We can easily see that if this term is to be minimized, one could decrease σ_f with the caveat that this would decrease model fit. Hence, the only solution is to have high correlations values in K_{ff} so as to get a determinant close to 0.

In the remainder of this subsection, we derive similar results to Ober et al. (2021) but in the sparse scenario. We introduce the collapsed bound introduced in Titsias (2009):

$$\mathcal{L}_{Titsias} = \log \mathcal{N}\left(y \mid Q_{ff} + \sigma^2_{noise}\mathbb{I}_n\right) - \frac{1}{2\sigma^2_{noise}} Tr\left[K_{ff} - Q_{ff}\right] \tag{81}$$

$$\propto -\underbrace{\frac{1}{2}\log \mid Q_{ff} + \sigma^2_{noise}\mathbb{I}_n \mid}_{complexity\ penalty} - \underbrace{\frac{1}{2} y^\top \left(Q_{ff} + \sigma^2_{noise}\mathbb{I}_n\right)^{-1} y}_{data\ fit} - \underbrace{\frac{1}{2\sigma^2_{noise}} Tr\left[K_{ff} - Q_{ff}\right]}_{trace\ term} \tag{82}$$

$$\propto -\frac{N}{2}\sigma^2 - \frac{1}{2}\log \mid \tilde{Q_{ff}} + \tilde{\sigma^2_{noise}}\mathbb{I}_n \mid -\frac{1}{2\sigma^2} y^\top \left[\tilde{Q_{ff}} + \tilde{\sigma^2_{noise}}\mathbb{I}_n\right] y \tag{83}$$
$$- \frac{1}{2\tilde{\sigma^2_{noise}}} Tr\left[\tilde{K_{ff}} - \tilde{Q_{ff}}\right]$$

, where we have used again the following notation for kernel terms $k\left(\cdot,\cdot\right) = \sigma^2 k\left(\tilde{\cdot},\cdot\right)$ and $\sigma^2_{noise} = \sigma^2\tilde{\sigma^2_{noise}}$. To obtain predictions at testing time under this framework we can make use of the optimal $q(U)$ being given by the following first two moments:

$$m(U^*) = \sigma^{-2}_{noise} K_{uu}\left[K_{uu} + \sigma^{-2}_{noise} K_{uf} K_{fu}\right]^{-1} K_{uf} y \tag{84}$$

$$v(U^*) = K_{uu}\left[K_{uu} + \sigma^{-2}_{noise} K_{uf} K_{fu}\right]^{-1} K_{uu} \tag{85}$$

, which we can plug in to standard SVGP predictive equations (equations (31) and (32)).

We adapt the derivation in Ober et al. (2021) to our framework at hand:

$$\frac{\partial \mathcal{L}_{Titsias}}{\partial \sigma^2} = \frac{\partial -\frac{N}{2}\sigma^2 - \frac{1}{2}\log \mid \tilde{Q_{ff}} + \tilde{\sigma^2_{noise}}\mathbb{I}_n \mid -\frac{1}{2\sigma^2} y^\top \left[\tilde{Q_{ff}} + \tilde{\sigma^2_{noise}}\mathbb{I}_n\right]^{-1} y - \frac{1}{2\tilde{\sigma^2_{noise}}}\left[\tilde{K_{ff}} - \tilde{Q_{ff}}\right]}{\partial \sigma^2} \tag{86}$$

$$= -\frac{N}{2\sigma^2} + \frac{1}{2\sigma^4} y^\top \left[\tilde{Q_{ff}} + \tilde{\sigma^2_{noise}}\mathbb{I}_n\right]^{-1} y \tag{87}$$

Hence, if we set the derivative to 0, then we obtain that $\sigma^2 = \frac{1}{N} y^\top \left[\tilde{Q_{ff}} + \tilde{\sigma^2_{noise}}\mathbb{I}\right]^{-1} y$, which if we input it into the data fit term it results in $\frac{N}{2}$, similar to the non-sparse scenario analyzed in Ober et al. (2021). The difference between the sparse and non-sparse framework is that after convergence in the data fit term, the model now has to achieve over-correlation in Q_{ff}, while still minimizing $K_{ff} - Q_{ff}$.

3.6 Pooling operations on stochastic layers

Previous work that dealt with combining GP with convolutional architectures Dutordoir et al. (2020); Kumar et al. (2018); Blomqvist et al. (2018) have used in their experiments simple architectures involving a couple of stacked layers. In this paper, we propose to experiment with more modern architectures such as DenseNet Huang et al. (2017) or ResNet He et al. (2016). However, both these architectures include pooling layers such as average pooling, which for Euclidean data is a straightforward operation since we have a naturally

induced metric. Since we are using stochastic layers that operate in the space of Gaussian distributions, this introduces some complications as it is not desirable to sample from the stochastic layers, subsequently applying the Euclidean space average pooling operation. Nevertheless, in the remainder of this subsection we show a simple method for replicating average pooling in Wasserstein space by using Wasserstein barycentres (Agueh and Carlier, 2011).

We consider probability measures $\mu_1, ..., \mu_k$ and fixed weights $\theta_1, ..., \theta_k$ that are positive real numbers such that $\sum_{k=1}^{K} \theta_k = 1$. For $\nu \in \mathbb{P}_2(\mathbf{R}^d)$, where $\mathbb{P}_2$ is the set of Borel probabilities on $\mathbb{R}$ with finite second moment and absolutely continuous with respect to Lebesque measures, we consider the following functionals:

$$\mathbf{V}(\nu) = \sum_{k=1}^{K} \theta_k \mathbf{W}_2^2(\mu, \mu_k) \tag{88}$$

$$\mathbb{V}(\tilde{\mu}) = \min_{\mu \in \mathbb{P}_2} \mathbb{V}(\mu) \tag{89}$$

, where $\mathbb{V}(\tilde{\mu})$ is defined as the barycentre with respect to the Wasserstein-2 distance of the set of probabilities $\{\mu_1, ..., \mu_k\}$. Intuitively, barycentres can be seen as the equivalent of averaging in Euclidean space, while still maintaining the geometric properties of the distributions at hand.

Theorem 9 (Theorem 4.2. in Álvarez-Esteban et al. (2016)) *Assume $\Sigma_1, ..., \Sigma_K$ are symmetric positive semidefinite matrices, with at least one of them positive definite. We take $S_0 \in \mathbb{M}_{d\times d}^{+}$ and define:*

$$S_{n+1} = S_n^{-1/2} (\sum_{k=1}^{K} \theta_k (S_n^{1/2} \Sigma_k S_n^{1/2})^{1/2})^2 S_n^{-1/2} \tag{90}$$

If $\mathbb{N}(0, \Sigma_0)$ is the barycenter of $\mathbb{N}(0, \Sigma_1), ..., \mathbb{N}(0, \Sigma_K)$, then $W_2^2(\mathbb{N}(0, S_n), \mathbb{N}(0, \Sigma_0) \to 0$ as $n \to \infty$.

Remark 10 *In the case of computing the barycentre of univariate Gaussian measures, the iterative algorithm converges in one iteration to $\Sigma_0 = \left(\sum_{k=1}^{K} \theta_k \Sigma_k^{\frac{1}{2}}\right)^2$. This provides us with a deterministic and single step equation to downsample stochastic layers, where we can additionally calculate the mean of the barycentre by $\sum_{k=1}^{K} \theta_k m_k$, where $\{m_1, \cdots, m_K\}$ represent the first moments of the respective distributions.*

4. DistGP Layer Networks & OOD detection

An outlier can be defined in various ways (Ruff et al., 2021). In this paper we follow the most basic one, namely *"An anomaly is an observation that deviates considerably from some concept of normality."* More concretely, it can be formalised as follows: our data resides in $X \in \mathbb{R}^D$, an anomaly/outlier is a data point $x \in X$ that lies in a low probability region

under $\mathcal{P}$ such that the set of anomalies/outliers is defined as $A = \{x \in X | p(x) \leq \xi\},\ \xi \geq 0$, with ξ is a threshold under which we consider data points to deviate sufficiently from what normality constitutes.

Influence of enforced Lipschitz condition. We aim to visually assess if the Lipschitz condition imposed via Proposition 5 negatively influences the predictive capabilities. We use a standard neural network architecture with two hidden layers with 5 dimensions each, with the affine embeddings operations described in equations (72) and (73) being replaced by a non-convolutional dense layer. From Figure 8 we can notice that imposing a unitary Lipschitz constant does not result in the over-regularization of the predictive mean. A slight smoothing effect on the predictive mean can be noticed in output space. Moreover, for the Lipschitz constrained version we can discern a better fit of the data manifold in terms of distributional variance, with a noticeable difference in the second hidden layer.

Over-correlation in latent space. We aim to understand whether the over-correlation phenomenon occurs for our model. We consider a standard neural network architecture with two hidden layers with 50 hidden units per layer. From Figure 9 we can notice that for DKL, the sparse framework does remove any unwanted over-correlations in the final hidden layer latent space. In the unconstrained model, there is a notion of locality in the final hidden layer latent space, albeit of a lower degree compared to the DKL model. With regards to OOD detection, of utmost importance is the fact that regions outside the training set manifold have a correlation value of 0. Perhaps unsurprisingly, introducing a unitary Lipschitz constraint resulted in an increased correlation in the latent space, alongside a smoother predictive mean.

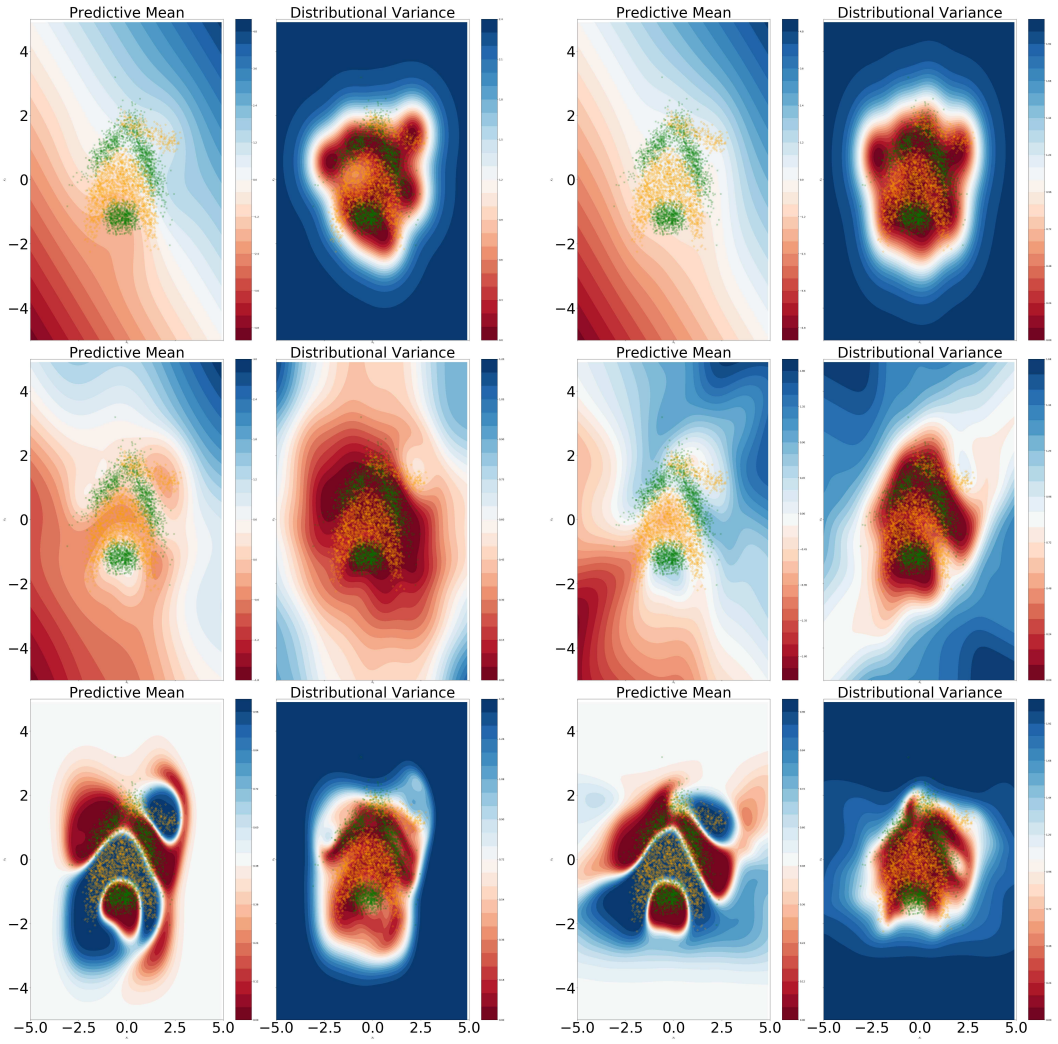
4.1 Reliability of *in-between* uncertainty estimates

We are interested to test our newly introduced module in scenarios where *in-between* uncertainty can fail. For this we use the "snelson" dataset, with the training set taken to comprise the intervals between 0.0 and 2.0, respectively 4.0 and 6.5. Thereby, in an ideal scenario we would expect our model to offer high distributional uncertainty estimates between 2.0 and 4.0, which constitutes our *in-between* region. To benchmark our approach, we compare it to a collapsed SGPR as defined in Titsias (2009).

From Figure 10 we can observe that the behaviour is strikingly similar between a collapsed SGPR and a three layer DistGP-Layers network.

Reliability of *within-data* uncertainty estimates. Within-data uncertainty or more conveniently epistemic uncertainty is responsible to detect regions of the input space where the variance in the model parameters, in this case of U, can be further reduced if we add more data points in said input regions. To test if our newly introduced module can provide reliable within-data uncertainty estimates one can proceed to subsample a dataset (as done in Figure 11 subsampling in the interval $[0, 2.5]$), with the intended effect being of an increase in within-data uncertainty across the input region where we subsampled.

From Figure 11 we can see that despite the low number of training points, it did not result in over-fitting, with our model exhibiting a relatively smooth predictive mean. Moreover, in comparison to Figure 9 we can notice that the within-data uncertainty has substantially increased in the $[0, 2.5]$ interval.

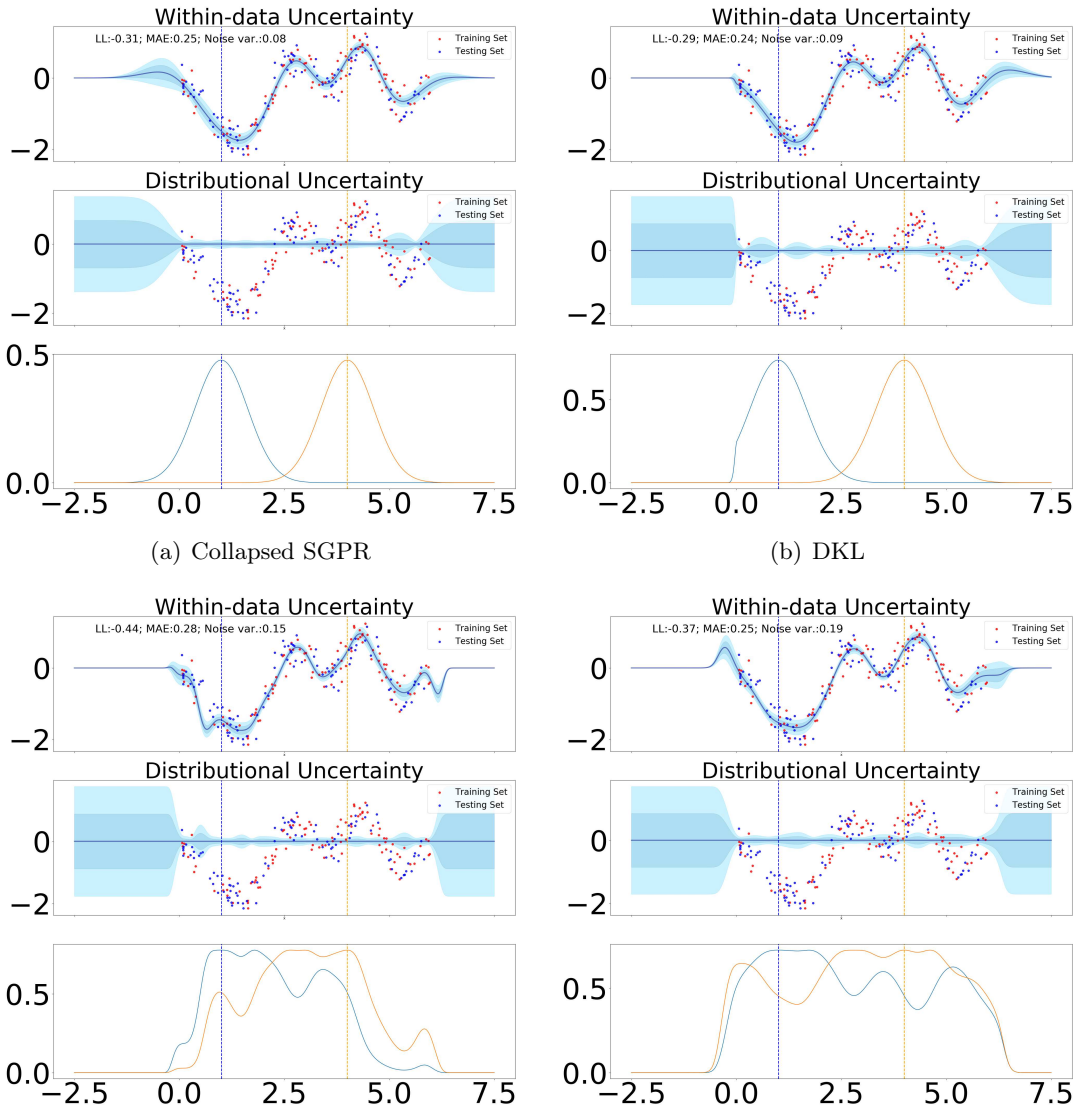


(a) DistGP-NN without Lipschitz constraint

(b) DistGP-NN with Lipschitz constraint

Figure 8: Layer-wise predictive moments of DistGP-NN models (with or without unitary Lipschitz constraints) trained on toy binary classification dataset.

MNIST and CIFAR10. We compare our approach on the standard image classification benchmarks of MNIST Lecun et al. (1998) and CIFAR-10 Krizhevsky (2009), which have standard training and test folds to facilitate direct performance comparisons. MNIST contains 60,000 training examples of 28×28 sized grayscale images of 10 hand-drawn digits, with a separate 10,000 validation set. CIFAR-10 contains 50,000 training examples of RGB colour images of size 32×32 from 10 classes, with 5,000 images per class. We preprocess the images such that the input is normalized to be between 0 and 1. We compare our model primarily against the original shallow Convolutional Gaussian process Van der Wilk et al. (2017) and Deep Convolutional Gaussian Process (DeepConvGP) Blomqvist et al. (2018). In terms of model architectures, we have used a standard stacked convolutional approach, with the model



(a) Collapsed SGPR

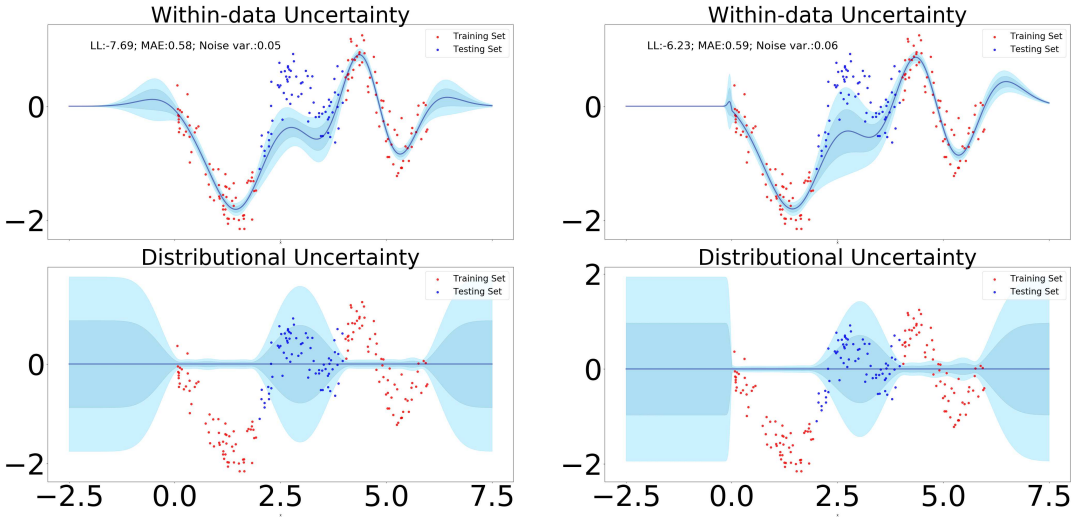
(b) DKL

(c) DistGP-NN Not Constrained

(d) DistGP-NN Lipschitz Constraint

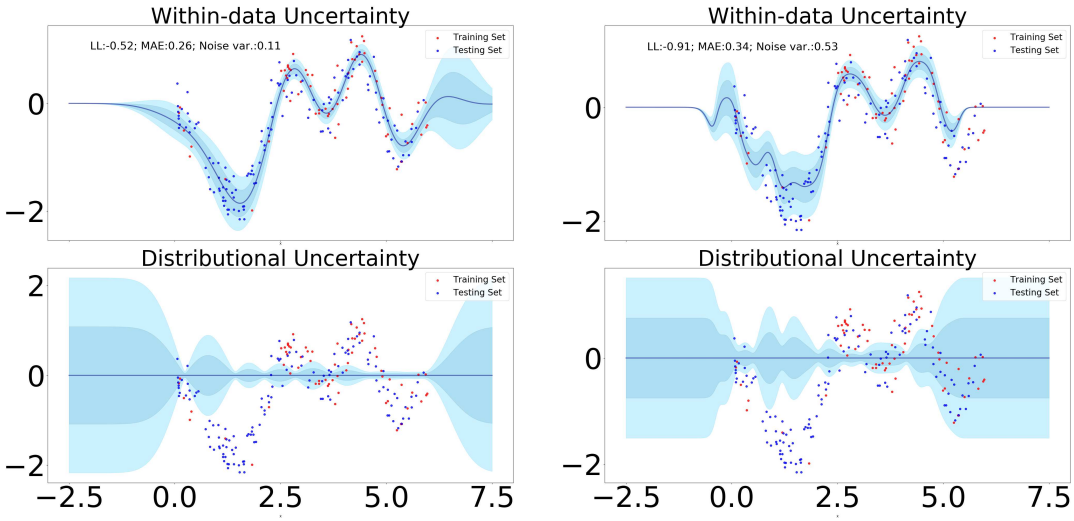
Figure 9: Top row: Predictive mean and variance of parametric part of SGP; **Middle row**: Predictive mean and variance of non-parametric part of SGP; **Bottom row**: Kernel evaluated across whole input span with respect to -2.0 (blue) and 2.0 (orange).

entitled "DistGP-DeepConv" consisting of 64 hidden units for the "Convolutionally Warped DistGP" part of the module, respectively, 5 hidden units for the "DistGP activation-function" part. For the DeepConvGP, we used 64 hidden units at each hidden layer. All models use a stride of 2 at the first layer. In all experiments we use 250 inducing points at each layer. Lastly, we also devised 18 hidden layers size versions of the ResNet (He et al., 2016) and DenseNet (Huang et al., 2017) architectures.



(a) Collapsed SGPR (b) DistGP-NN

Figure 10: Reliability of *in-between* uncertainty. Top row: Predictive mean and variance of parametric part of SGP; Bottom row: Predictive mean and variance of non-parametric part of SGP.



(a) Collapsed SGPR (b) DistGP-NN

Figure 11: Reliability of *within-data* uncertainty. Top row: Predictive mean and variance of parametric part of SGP; Bottom row: Predictive mean and variance of non-parametric part of SGP.

Table 1 shows the classification accuracy on MNIST and CIFAR-10 for different Convolutional GP models. Compared to other convolutional GP approaches, our method achieves superior classification accuracy compared to DeepConvGP (Blomqvist et al., 2018). We

Convolutional GP models	Hidden Layers	MNIST	CIFAR-10
ConvGP	0	98.83	64.6
DeepConvGP	1	98.38	58.65
DeepConvGP	2	99.24	73.85
DeepConvGP	3	99.44	75.89
DistGP-DeepConv	1	99.01	70.12
DistGP-DeepConv	2	99.43	76.54
DistGP-DeepConv	3	99.67	78.49
DistGP-ResNet-18	18	99.52	74.56
DistGP-DenseNet	18	99.75	75.29
Hybrid NN-GP models	Hidden Layers	MNIST	CIFAR-10
Deep Kernel Learning	5	99.2	77.0
GPDNN	40	99.95	93.0

Table 1: Performance on MNIST and CIFAR-10. Deep Kernel Learning are the set of models from Wilson et al. (2016), whereas GPDNN are the set of models published in Bradshaw et al. (2017). Other results than our method are taken from the respective publications

find that for our method, adding more layers increases the performance significantly. This observation is only available for a couple of stacked layers, as the results from our ResNet and DenseNet variants do not support this assertion. The GPDNN models introduced in Bradshaw et al. (2017) are nonetheless close to state of the art on CIFAR10 but also using a variant of DenseNet (Huang et al., 2017) as the building blocks for their GP classifier.

Outlier detection on different fonts of digits. We test if DistGP-DeepConv models outperform OOD detection models from literature such as DUQ (van Amersfoort et al., 2020), OVA-DM (Padhy et al., 2020) and OVVNI (Franchi et al., 2020). In these experiments we assess the capacity of our model to detect domain shift by training it on MNIST and looking at the uncertainty measures computed on the testing set of MNIST and the entire NotMNIST dataset (Bulatov, 2011), respectively SVHN (Netzer et al., 2011). The hypothesis is that we ought to see both higher predictive entropy and differential entropy for distributional uncertainty (respectively higher OOD measures specific to each of the baseline models) for the digits stemming from a wide array of fonts present in NotMNST as none of the fonts are handwritten, respectively the digit fonts in SVHN exhibit different backgrounds, orientations besides not being handwritten.

From Table 2 we can observe that generally all models exhibit a shift in their uncertainty measure between MNIST and notMNIST, with the notable exception of OVNNI which barely manages to better separate the two datasets compared to a random guess. Moreover, OVA-DM manages to completely separate the two datasets with the caveat that it obtains lower predictive entropy for MNIST vs. notMNIST compared to DistGP-DeepConv. The latter achieves similar results to DUQ, with the added benefit of a higher degree of separation

Model	MNIST vs. NotMNIST AUC		MNIST vs. SVHN AUC	
AUC	Pred. Entropy	OOD measure	Pred. Entropy	OOD measure
DistGP-DeepConv	**0.92**	0.82	**0.95**	0.98
0VA-DM	0.73	**1.0**	0.70	**1.0**
OVNNI	0.68	0.55	0.56	0.81
DUQ	0.82	0.81	0.65	0.74

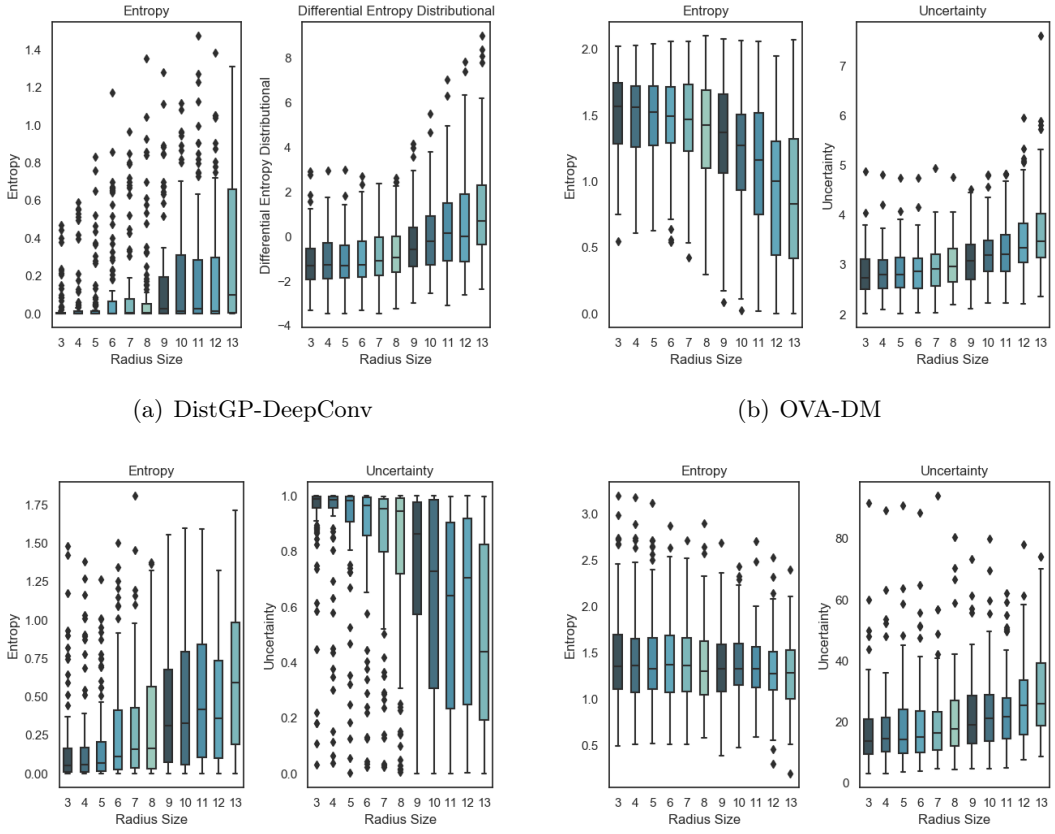
Table 2: OOD detection results. Performance of OOD detection based on predictive entropy and distributional differential entropy (for baseline OOD models each has a different OOD measure). Models are trained on MNIST (normative data).

using predictive entropy. In the case of SVHN we can observe similar patterns to notMNIST, with OVA-DM and DistGP-DeepConv managing to almost separate the two datasets (MNIST vs. SVHN) by inspecting their uncertainty measure, again with the caveat for OVA-DM that it exhibits lower predictive entropy for SVHN in comparison to MNIST.

Sensitivity to input perturbations. MorphoMNIST (Castro et al., 2018) enables the systematic deformation of MNIST digits using morphological operations. We use MorphoMNIST to better understand the outlier detection capabilities of each method by exposing them to increasingly deformed samples. We use the first 500 MNIST digits in the testing set to generate new images with controlled morphological deformations. We use the swelling deformation with a strength of 3 and increasing radius from 3 to 14. Our hypothesis is that the predictive entropy should increase as the deformation is increased, alongside with the distributional differential entropy, which is a measure of the overall uncertainty in the logit space. This is motivated by the fact that the newly obtained images from MorphoMNIST are outside of the data manifold, which is different from the concept of having high uncertainty as expressed by entropy upon seeing a difficult digit to classify. In this case we would expect high entropy but low differential entropy.

All models are able to pick up on the shift in the data manifold as swelling is applied to the original digits, with the model-specific uncertainty measure steadily increasing (for OVNNI, a decrease in the measure translates to higher uncertainty) as increasing deformation is applied. However, for OVA-DM and DUQ the predictive entropy is stable or actually decreases as more deformation is applied, which is in contrast to what one would expect (Figure 12).

To further assess the sensitivity to input perturbations of our methods, we employ the experiments introduced in Gal and Ghahramani (2016a) by successively rotating digits from MNIST. We expect to see an increase in both predictive entropy and distributional differential entropy as digits are rotated. For our experiment we rotate digit 6. When the digit is rotated by around 180 degrees the entropy and differential entropy should revert back closer to initial levels, as it will resembles digit 9.



(a) DistGP-DeepConv

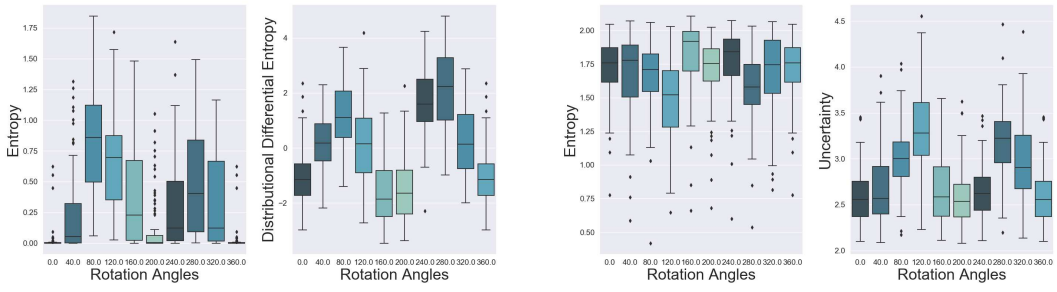
(b) OVA-DM

(c) OVNNI

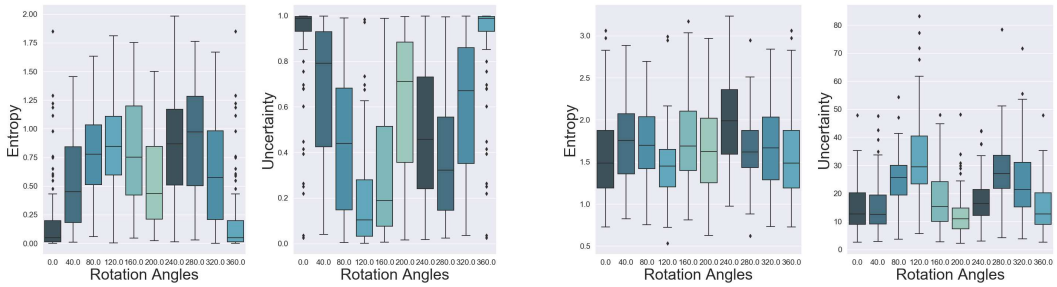
(d) DUQ

Figure 12: Predictive entropy and model-specific uncertainty measure for varying models as swelling of increasing radius is applied on MNIST digits. Higher values of uncertainty measure indicate outlier status, expect for OVNNI where the inverse is true. Results are shown for 3 hidden layers with DistGP-DeepConv dimensionality being set to 5, whereas the capacity of the convolutionally warped DistGPs was set to 12, whereas for OOD models we use 128 hidden units at each layer.

From Figure 13 we can notice that all models exhibit an increase (decrease for OVNNI translates into higher uncertainty) in their specific uncertainty measures for rotation angles between 40 and 160, respectively between 240 and 320 degrees. In terms of predictive entropy, we can discern relatively stable and highly overlapping values for OVA-DM and DUQ, whereas for DistGP-DeepConv and OVNNI we can observe a clear pattern of increases and decreases as what was originally a 6 becomes a 9.



(a) DistGP-DeepConv (b) OVA-DM



(c) OVNNI (d) DUQ

Figure 13: Predictive entropy and model-specific uncertainty measure for varying models as varying degrees of rotation is applied to digit 6. Higher values of uncertainty measure indicate outlier status, expect for OVNNI where the inverse is true. Results are shown for 3 hidden layers with DistGP-DeepConv dimensionality being set to 5, whereas the capacity of the convolutionally warped DistGPs was set to 12, whereas for OOD models we use 128 hidden units at each layer.

5. DistGP-based Segmentation Network & OOD Detection in Medical Imaging

The above introduced modules in Sec. 3.2 can be used to construct a convolutional network that benefits from properties of DistGP. Specifically, we construct a 3D network for segmenting volumetric medical images, which is depicted in Figure 14 (top). It consists of a convolved GP layer, followed by two measure-preserving DistGP layers. Each hidden layer uses filters of size 5×5×5. To increase the model's receptive field, in the second layer we use convolution dilated by 2. We use 250 inducing points and 2 channels for the DistGP "activation functions". The affine operators project the stochastic patches into a 12 dimensional space. The size of the network is limited by computational requirements for GP-based layers, which is an active research area. Like regular convolutional nets, this model can process input of arbitrary size but GPU memory requirement increases with input size. We here provide input of size 32^3 to the model, which then segments the central 16^3 voxels. To segment a whole scan we divide it into tiles and stitch together the segmentations.

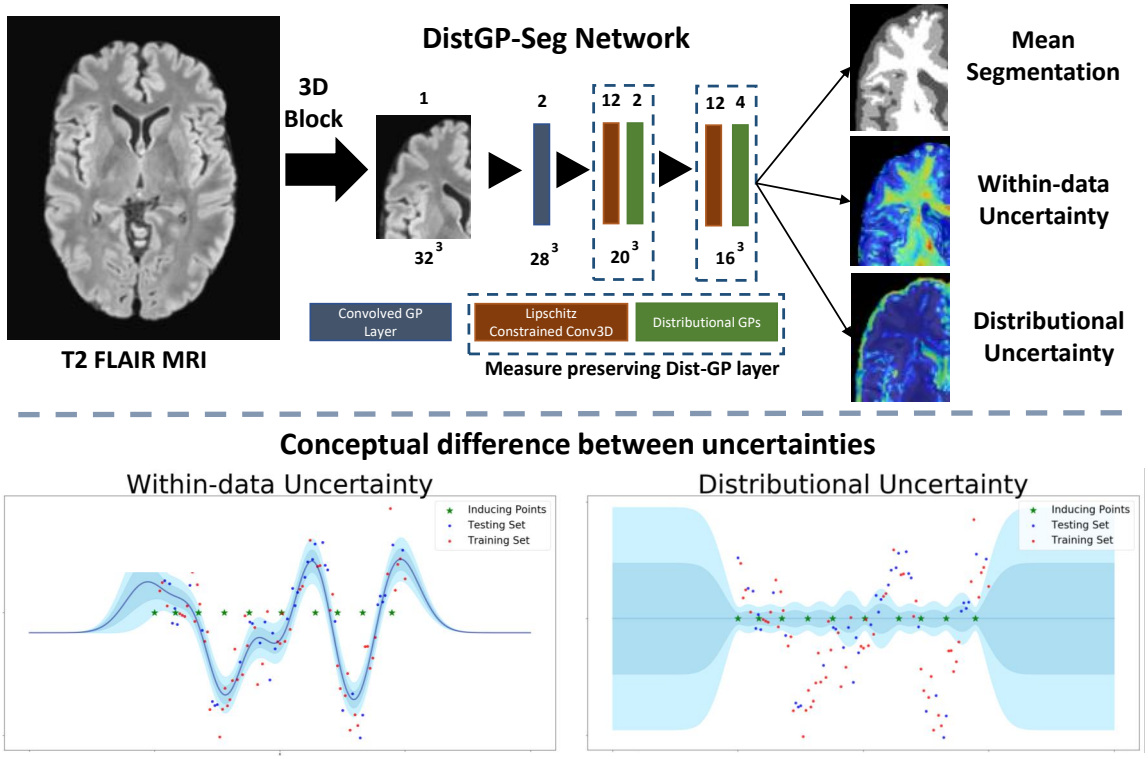


Figure 14: Top: Schematic of proposed DistGP activated segmentation net. Above and below each layer we show the number of channels and their dimension respectively. **Bottom**: Visual depiction of the two uncertainties in DistGP after fitting a toy regression dataset. Hyperparameters and variational approximate posteriors are optimized. Distributional uncertainty increases outside the manifold of training data and is therefore useful for OOD detection.

5.1 Evaluation on Brain MRI

In this section we evaluate our method alongside recent OOD models (van Amersfoort et al., 2020; Franchi et al., 2020; Padhy et al., 2020), assessing their capabilities to reach segmentation performance comparable to well-established deterministic models and whether they can accurately detect outliers.

5.1.1 Data and pre-processing

For evaluation we use publicly available datasets:

1) Brain MRI scans from the UKBB study (Alfaro-Almagro et al., 2018), which contains scans from nearly 15,000 subjects. We selected for training and evaluation the bottom 10% percentile in terms of white matter hypointensities with an equal split between training and testing. All subjects have been confirmed to be normal by radiological assessment. Segmentation of brain tissue (CSF,GM,WM) has been obtained with SPM12.

2) MRI scans of 285 patients with gliomas from BraTS 2017 (Bakas et al., 2017). All classes are fused into a *tumor* class, which we will use to quantify OOD detection performance.

In what follows, we use only the FLAIR sequence to perform the brain tissue segmentation task and OOD detection of tumors, as this MRI sequence is available for both UKBB and BraTS. All FLAIR images are pre-processed with skull-stripping, N4 bias correction, rigid registration to MNI152 space and histogram matching between UKBB and BraTS. Finally, we normalize intensities of each scan via linear scaling of its minimum and maximum intensities to the [-1,1] range.

5.1.2 Brain tissue segmentation on normal MRI scans

Model	Hidden Layers	DICE CSF	DICE GM	DICE WM
OVA-DM (Padhy et al., 2020)	3	0.72	0.79	0.77
OVNNI (Franchi et al., 2020)	3	0.66	0.77	0.73
DUQ (van Amersfoort et al., 2020)	3	0.745	0.825	0.781
DistGP-Seg (ours)	3	0.829	0.823	0.867
U-Net	3 scales	0.85	0.89	0.86

Table 3: Performance on UK Biobank in terms of Dice scores per tissue.

Task: We train and test our model on the task of segmenting brain tissue of healthy UKBB subjects. This corresponds to the within-data manifold in our setup.

Baselines: We compare our model with recent Bayesian approaches for enabling task-specific models (such as image segmentation) to perform uncertainty-based OOD detection (van Amersfoort et al., 2020; Franchi et al., 2020; Padhy et al., 2020). For fair comparison, we use these methods in an architecture similar to ours (Figure 14), except that each layer is replaced by standard convolutional layer, each with 256 channels, LeakyRelu activations, and dilation rates as in ours. We also compare these Bayesian methods with a well-established deterministic baseline, a U-Net with 3 scales (down/up-sampling) and 2 convolution layers per scale in encoder and 2 in decoder (total 12 layers).

Results: Table 3 shows that DistGP-Seg surpasses other Bayesian methods with respect to Dice score for all tissue classes. Our method approaches the performance of the deterministic U-Net, which has a much larger architecture and receptive field. We emphasize this has not been previously achieved with GP-based architectures, as their size (e.g., number of layers) is limited due to computational requirements. This supports the potential of DistGP, which is bound to be further unlocked by advances in scaling GP-based models.

5.1.3 Outlier detection in MRI scans with tumors

Task: The previous task of brain tissue segmentation on UKBB serves as a proxy task for learning normative patterns with our network. Here, we apply this pre-trained network on BRATS scans with tumors. We expect the region surrounding the tumor and other related pathologies, such as squeezed brain parts or shifted ventricles, to be highlighted with

Model	DICE FPR=0.1	DICE FPR=0.5	DICE FPR=1.0	DICE FPR=5.0
OVA-DM (Padhy et al., 2020)	0.382	0.428	0.457	0.410
OVNNI (Franchi et al., 2020)	≤ 0.001	≤ 0.001	≤ 0.001	≤ 0.001
DUQ (van Amersfoort et al., 2020)	0.068	0.121	0.169	0.182
DistGP-Seg (ours)	0.512	0.571	0.532	0.489
VAE-LG (Chen et al., 2019)	0.259	0.407	0.448	0.303
AAE-LG (Chen et al., 2019)	0.220	0.395	0.418	0.302

Table 4: Performance comparison of Dice for detecting outliers on BraTS for different thresholds obtained from UKBB.

higher distributional uncertainty, which is the OOD measure for the Bayesian deep learning models. To evaluate quality of OOD detection at a pixel level, we follow the procedure in Chen et al. (2019), for example to get the 5.0% False Positive Ratio threshold value we compute the 95% percentile of distributional variance on the testing set of UKBB, taking into consideration that there is no outlier tissue there. Subsequently, using this value we threshold the distributional variance heatmaps on BraTS, with tissue having a value above the threshold being flagged as an outlier. We then quantify the overlap of the pixels detected as outliers (over the threshold) with the ground-truth tumor labels by computing the Dice score between them.

Results: Table 4 shows the results from our experiments with DistGP and compared Bayesian deep learning baselines. We also provide performance of reconstruction-based OOD detection models as reported in Chen et al. (2019) for similar experimental setup. DistGP-Seg surpasses its Bayesian deep learning counterparts, as well as reconstructed-based models. In Figure 15 we provide representative results from the methods we implemented for qualitative assessment. Moreover, although BRATS does not provide labels for WM/GM/CSF tissues hence we cannot quantify how well these tissues are segmented, visual assessment shows our method compares favorably to compared counterparts.

In Figure 16 we plotted the different differential entropy measures based on BRATS scans by overlying their tumor labels on the obtained uncertainties from our model. We can notice that tumor tissue is highlighted with higher inside and outside of the data manifold uncertainty compared to healthy tissue. More detailed plots are available in Appendix C.

6. Discussion & Conclusion

We have introduced a novel Bayesian convolutional layer with Lipschitz continuity that is capable of reliably propagating uncertainty. We have shown on a wide array of general OOD detection tasks that it surpasses other OOD models from literature, while also offering an increase in accuracy compared to counterpart architectures based solely on Euclidean space SVGPs (Blomqvist et al., 2018). General criticism surrounding deep and convolutional GP

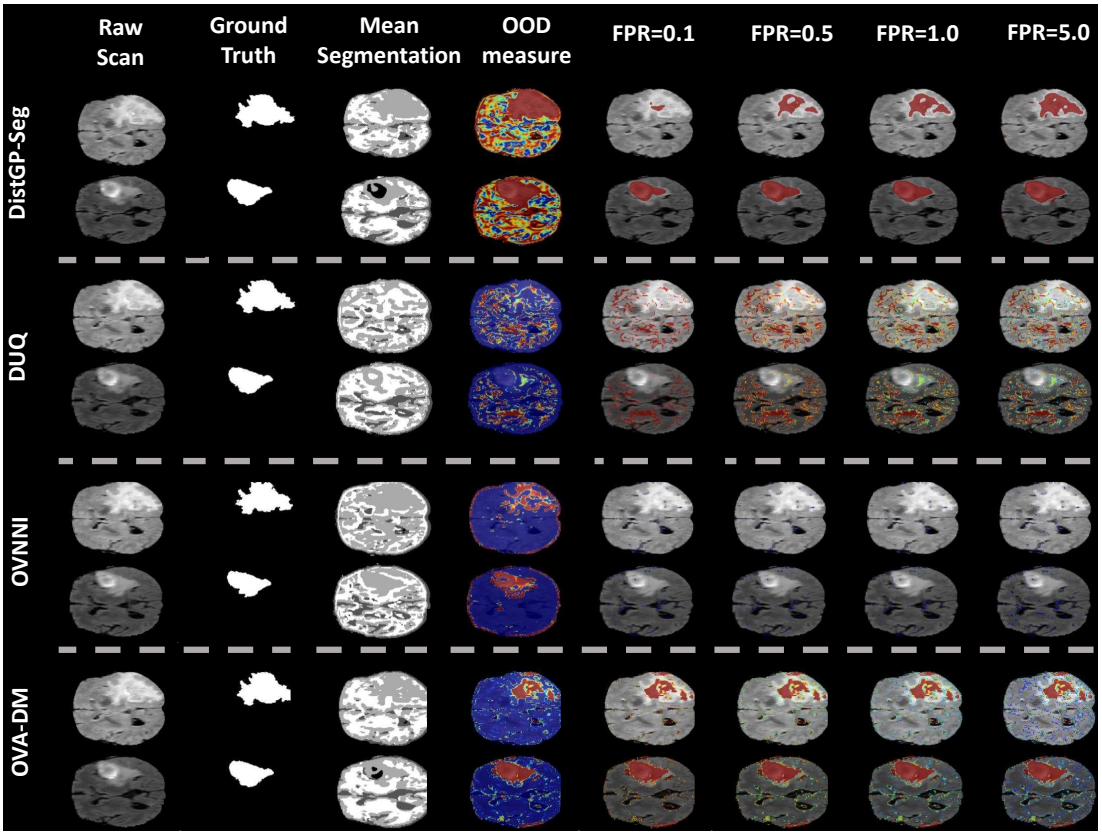
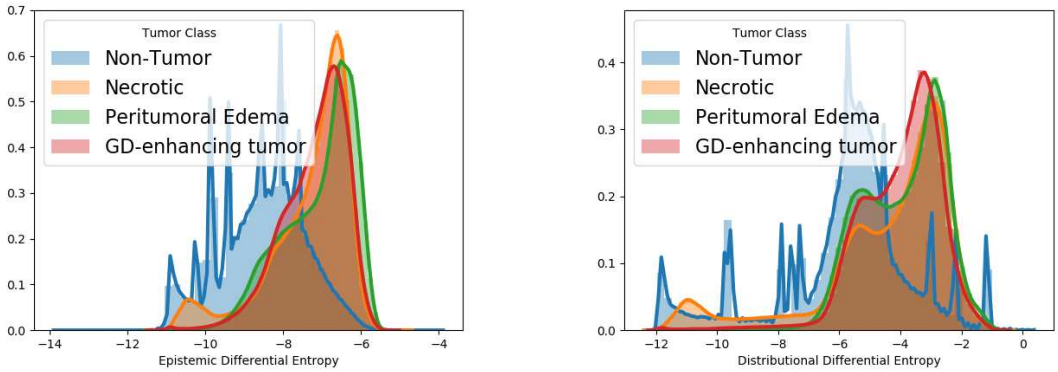


Figure 15: Comparison between models in terms of voxel-level outlier detection of tumors on BRATS scans. Mean segmentation represents the hard segmentation of brain tissues. OOD measure is the quantification of uncertainty for each model, using their own procedure. Higher values translate to appartenance to outlier status, whereas for OVNNI it is the converse. OOD measures have been normalized to be between 0 and 1 for each model in part.

involves the issue of under-performance compared to other Bayesian deep learning techniques, and especially compared to deterministic networks. Our experiments demonstrate that our 3-layers model, size limited due to computational cost, is capable of approaching the performance of a U-Net, an architecture with a much larger receptive field. Further advances in computational efficient GP-based models, an active area of research, will enable our model to scale further and unlock its full potential. Importantly, we showed that our DistGP-Seg network offers better uncertainty estimates for OOD detection than state-of-the-art OOD detection models, and also surpasses some recent unsupervised reconstruction-based deep learning models for identifying outliers corresponding to pathology on brain scans.

This framework can also be used for regression and classification tasks within a medical imaging context, facilitating the adoption of deep learning in clinical settings thanks to enhanced accountability in predictions. For example, parts of scans flagged with high distributional uncertainty can be sent back for inspection and quality control. To support



(a) (b)

Figure 16: Comparison in terms of voxel-level epistemic and distributional differential entropy between non-tumor tissues and different tumor gradations from subjects in the BRATS dataset.

our claim, we have included additional results on flagging white matter hyperintensities as outliers (see Appendix D), respectively retina pathologies (see Appendix E).

Our results indicate that OOD methods that do not take into account distances in latent space, such as OVNNI, tend to fail in detecting outliers, whereas OVA-DM and DUQ that make predictions based on distances in the last layer perform better. Our model utilises distances at every hidden layer, thus allowing the notion of outlier to evolve gradually through the depth of our network. This difference can be noticed in the smoothness of OOD measure for our model in comparison to other methods in Figure 15. Furthermore, the issue of *feature collapse* (van Amersfoort et al., 2020) in deep networks can be precisely controlled due to the mathematical underpinnings of our proposed network, enabling us to assess the scenarios when this happens by simple equations. Additionally, we have shown that despite the possibility of achieving over-correlation in the latent space via the loss function, that this does not happen in practice.

A drawback of our study resides in the small architecture used on medical imaging scans. Extending our "measure preserving DistGP" module to larger architectures such as U-Net for segmentation or modern CNNs for whole-image prediction tasks remains a prospective research avenue fuelled by advances in scalability of SGP. Moreover, our experiments involving more complicated architectures, such as ResNet or DenseNet for standard multi-class classification, have not managed to surpass in accuracy a far less complex model with only 3 hidden layers. A plausible reason behind this under-fitting resides in the factorized approximate posterior formulation, which was shown to negatively affect predictive performance compared to MCMC inference schemes (Havasi et al., 2018). We posit that using alternative inference frameworks (Ustyuzhaninov et al., 2019) whereby we impose correlations between layers might alleviate this issue. Moreover, the lack of added representational capacity upon adding new layers raises some further questions regarding what are optimal architectures for hierarchical GPs, what inductive biases do they need or how to properly initialize them to facilitate adequate

training. Additionally, our comparison with respect to reconstruction based approaches towards OOD detection was not complete as it did not include a comprehensive list of recent models (Dey and Hong, 2021; Pinaya et al., 2021; Schlegl et al., 2019; Baur et al., 2018). However, comparing our proposed model with reconstruction based approaches was not our intended goal for this paper, the main aim being to compare with models which can provide accurate predictive results alongside OOD detection capabilities at the same time. Another limitation of our work is the training speed for our proposed module, with matrix inversion operations and log determinants being required at each layer. Future work should consider matrix inversion free inference techniques for GPs (van der Wilk et al., 2020).

In conclusion, our work shows that incorporating DistGP in convolutional architectures provides both competitive performance and reliable uncertainty quantification in medical image analysis alongside general OOD tasks, opening up a new direction of research.

Acknowledgments

SGP is funded by an EPSRC Centre for Doctoral Training studentship award to Imperial College London. KK is funded by the UKRI London Medical Imaging & Artificial Intelligence Centre for Value Based Healthcare. BG received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement No 757173, project MIRA, ERC-2017-STG). DJS is supported by the NIHR Biomedical Research Centre at Imperial College Healthcare NHS Trust and the UK Dementia Research Institute (DRI) Care Research and Technology Centre. JHC acknowledges funding from UKRI/MRC Innovation Fellowship (MR/R024790/2).

Ethical Standards

The work follows appropriate ethical standards in conducting research and writing the manuscript, following all applicable laws and regulations regarding treatment of human subjects.

Conflicts of Interest

BG has received grants from European Commission and UK Research and Innovation Engineering and Physical Sciences Research Council, during the conduct of this study; and is Scientific Advisor for Kheiron Medical Technologies and Advisor and Scientific Lead of the HeartFlow-Imperial Research Team. JHC is a shareholder in and Scientific Advisor to BrainKey and Claritas Healthcare, both medical image analysis software companies.

References

Martial Agueh and Guillaume Carlier. Barycenters in the wasserstein space. *SIAM Journal on Mathematical Analysis*, 43(2):904–924, 2011.

Fidel Alfaro-Almagro, Mark Jenkinson, Neal K Bangerter, Jesper LR Andersson, Ludovica Griffanti, Gwenaëlle Douaud, Stamatios N Sotiropoulos, Saad Jbabdi, Moises Hernandez-Fernandez, Emmanuel Vallee, et al. Image processing and quality control for the first 10,000 brain imaging datasets from uk biobank. *Neuroimage*, 166:400–424, 2018.

Pedro C Álvarez-Esteban, E Del Barrio, JA Cuesta-Albertos, and C Matrán. A fixed-point approach to barycenters in wasserstein space. *Journal of Mathematical Analysis and Applications*, 441(2):744–762, 2016.

Alexander Amini, Wilko Schwarting, Ava Soleimany, and Daniela Rus. Deep evidential regression. *arXiv preprint arXiv:1910.02600*, 2019.

François Bachoc, Fabrice Gamboa, Jean-Michel Loubes, and Nil Venet. A gaussian process regression model for distribution inputs. *IEEE Transactions on Information Theory*, 64 (10):6620–6637, 2017.

Spyridon Bakas, Hamed Akbari, Aristeidis Sotiras, Michel Bilello, Martin Rozycki, Justin S Kirby, John B Freymann, Keyvan Farahani, and Christos Davatzikos. Advancing the cancer genome atlas glioma mri collections with expert segmentation labels and radiomic features. *Scientific data*, 4:170117, 2017.

Christian F Baumgartner, Kerem C Tezcan, Krishna Chaitanya, Andreas M Hötker, Urs J Muehlematter, Khoschy Schawkat, Anton S Becker, Olivio Donati, and Ender Konukoglu. Phiseg: Capturing uncertainty in medical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 119–127. Springer, 2019.

Christoph Baur, Benedikt Wiestler, Shadi Albarqouni, and Nassir Navab. Deep autoencoding models for unsupervised anomaly segmentation in brain mr images. In *International MICCAI Brainlesion Workshop*, pages 161–169. Springer, 2018.

Kenneth Blomqvist, Samuel Kaski, and Markus Heinonen. Deep convolutional gaussian processes. *arXiv preprint arXiv:1810.03052*, 2018.

John Bradshaw, Alexander G de G Matthews, and Zoubin Ghahramani. Adversarial examples, uncertainty, and transfer testing robustness in gaussian process hybrid deep networks. *arXiv preprint arXiv:1707.02476*, 2017.

Wessel Bruinsma, Eric Perim, William Tebbutt, Scott Hosking, Arno Solin, and Richard Turner. Scalable exact inference in multi-output gaussian processes. In *International Conference on Machine Learning*, pages 1190–1201. PMLR, 2020.

Yaroslav Bulatov. Notmnist dataset. *Google (Books/OCR), Tech. Rep.[Online]. Available: http://yaroslavvb. blogspot. it/2011/09/notmnist-dataset. html*, 2, 2011.

Daniel C Castro, Jeremy Tan, Bernhard Kainz, Ender Konukoglu, and Ben Glocker. Morpho-mnist: Quantitative assessment and diagnostics for representation learning. *arXiv preprint arXiv:1809.10780*, 2018.

Bertrand Charpentier, Daniel Zügner, and Stephan Günnemann. Posterior network: Uncertainty estimation without ood samples via density-based pseudo-counts. *arXiv preprint arXiv:2006.09239*, 2020.

Bertrand Charpentier, Oliver Borchert, Daniel Zügner, Simon Geisler, and Stephan Günnemann. Natural posterior network: Deep bayesian predictive uncertainty for exponential family distributions. *arXiv preprint arXiv:2105.04471*, 2021.

Xiaoran Chen, Nick Pawlowski, Ben Glocker, and Ender Konukoglu. Unsupervised lesion detection with locally gaussian approximation. In *International Workshop on Machine Learning in Medical Imaging*, pages 355–363. Springer, 2019.

Xiaoran Chen, Nick Pawlowski, Ben Glocker, and Ender Konukoglu. Normative ascent with local gaussians for unsupervised lesion detection. *Medical Image Analysis*, page 102208, 2021.

Alicia Curth, Patrick Thoral, Wilco van den Wildenberg, Peter Bijlstra, Daan de Bruin, Paul WG Elbers, and Mattia Fornasa. Transferring clinical prediction models across hospitals and electronic health record systems. In *PKDD/ECML Workshops (1)*, pages 605–621, 2019.

Steffen Czolbe, Kasra Arnavaz, Oswin Krause, and Aasa Feragen. Is segmentation uncertainty useful? In *International Conference on Information Processing in Medical Imaging*, pages 715–726. Springer, 2021.

Andreas Damianou and Neil Lawrence. Deep gaussian processes. In *Artificial Intelligence and Statistics*, pages 207–215, 2013.

Francesco D'Angelo and Vincent Fortuin. Repulsive deep ensembles are bayesian. *arXiv preprint arXiv:2106.11642*, 2021.

Raunak Dey and Yi Hong. Asc-net: Adversarial-based selective network for unsupervised anomaly segmentation. In Marleen de Bruijne, Philippe C. Cattin, Stéphane Cotin, Nicolas Padoy, Stefanie Speidel, Yefeng Zheng, and Caroline Essert, editors, *Medical Image Computing and Computer Assisted Intervention – MICCAI 2021*, pages 236–247, Cham, 2021. Springer International Publishing. ISBN 978-3-030-87240-3.

D. Dowson and B. Landau. The fréchet distance between multivariate normal distributions. *Journal of Multivariate Analysis*, 12:450–455, 1982.

Matthew M Dunlop, Mark A Girolami, Andrew M Stuart, and Aretha L Teckentrup. How deep are deep gaussian processes? *Journal of Machine Learning Research*, 19(54):1–46, 2018.

Vincent Dutordoir, Mark van der Wilk, Artem Artemev, Marcin Tomczak, and James Hensman. Translation insensitivity for deep convolutional gaussian processes. *arXiv preprint arXiv:1902.05888*, 2019.

Vincent Dutordoir, Mark van der Wilk, Artem Artemev, and James Hensman. Bayesian image classification with deep convolutional gaussian processes. In Silvia Chiappa and Roberto Calandra, editors, *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, volume 108 of *Proceedings of Machine Learning Research*, pages 1529–1539. PMLR, 26–28 Aug 2020. URL `http://proceedings.mlr.press/v108/dutordoir20a.html`.

Andrew YK Foong, Yingzhen Li, José Miguel Hernández-Lobato, and Richard E Turner. 'in-between'uncertainty in bayesian neural networks. *arXiv preprint arXiv:1906.11537*, 2019.

Gianni Franchi, Andrei Bursuc, Emanuel Aldea, Severine Dubuisson, and Isabelle Bloch. One versus all for deep neural network incertitude (ovnni) quantification. *arXiv preprint arXiv:2006.00954*, 2020.

Yarin Gal and Zoubin Ghahramani. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *ICML*, 2016a.

Yarin Gal and Zoubin Ghahramani. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *international conference on machine learning*, pages 1050–1059. PMLR, 2016b.

Adrià Garriga-Alonso, Laurence Aitchison, and Carl Edward Rasmussen. Deep convolutional networks as shallow gaussian processes. *arXiv preprint arXiv:1808.05587*, 2018.

Agathe Girard. *Approximate methods for propagation of uncertainty with Gaussian process models*. University of Glasgow (United Kingdom), 2004.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. On calibration of modern neural networks. *arXiv preprint arXiv:1706.04599*, 2017.

Danijar Hafner, Dustin Tran, Timothy Lillicrap, Alex Irpan, and James Davidson. Noise contrastive priors for functional uncertainty. In *Uncertainty in Artificial Intelligence*, pages 905–914. PMLR, 2020.

Marton Havasi, José Miguel Hernández Lobato, and Juan José Murillo Fuentes. Inference in deep gaussian processes using stochastic gradient hamiltonian monte carlo. *arXiv preprint arXiv:1806.05490*, 2018.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Identity mappings in deep residual networks. In *European conference on computer vision*, pages 630–645. Springer, 2016.

Dan Hendrycks and Kevin Gimpel. A baseline for detecting misclassified and out-of-distribution examples in neural networks. *arXiv preprint arXiv:1610.02136*, 2016.

Christian Henning, Francesco D'Angelo, and Benjamin F Grewe. Are bayesian neural networks intrinsically good at out-of-distribution detection? *arXiv preprint arXiv:2107.12248*, 2021.

James Hensman, Alexander Matthews, and Zoubin Ghahramani. Scalable variational gaussian process classification. In *Artificial Intelligence and Statistics*, pages 351–360. PMLR, 2015.

Matthew D Hoffman, David M Blei, Chong Wang, and John Paisley. Stochastic variational inference. *The Journal of Machine Learning Research*, 14(1):1303–1347, 2013.

AD Hoover, Valentina Kouznetsova, and Michael Goldbaum. Locating blood vessels in retinal images by piecewise threshold probing of a matched filter response. *IEEE Transactions on Medical imaging*, 19(3):203–210, 2000.

Shi Hu, Daniel Worrall, Stefan Knegt, Bas Veeling, Henkjan Huisman, and Max Welling. Supervised uncertainty quantification for segmentation with multiple annotations. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 137–145. Springer, 2019.

Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4700–4708, 2017.

Shungo Imai, Yoh Takekuma, Hitoshi Kashiwagi, Takayuki Miyai, Masaki Kobayashi, Ken Iseki, and Mitsuru Sugawara. Validation of the usefulness of artificial neural networks for risk prediction of adverse drug reactions used for individual patients in clinical practice. *Plos one*, 15(7):e0236789, 2020.

Jörn-Henrik Jacobsen, Jens Behrmann, Richard Zemel, and Matthias Bethge. Excessive invariance causes adversarial vulnerability. *arXiv preprint arXiv:1811.00401*, 2018.

Alex Kendall, Vijay Badrinarayanan, and Roberto Cipolla. Bayesian segnet: Model uncertainty in deep convolutional encoder-decoder architectures for scene understanding. *arXiv preprint arXiv:1511.02680*, 2015.

Durk P Kingma, Tim Salimans, and Max Welling. Variational dropout and the local reparameterization trick. In *Advances in neural information processing systems*, pages 2575–2583, 2015.

Simon Kohl, Bernardino Romera-Paredes, Clemens Meyer, Jeffrey De Fauw, Joseph R Ledsam, Klaus Maier-Hein, SM Ali Eslami, Danilo Jimenez Rezende, and Olaf Ronneberger. A probabilistic u-net for segmentation of ambiguous images. In *Advances in Neural Information Processing Systems*, pages 6965–6975, 2018.

Alex Krizhevsky. Learning multiple layers of features from tiny images. Technical report, 2009.

Vinayak Kumar, Vaibhav Singh, PK Srijith, and Andreas Damianou. Deep gaussian processes with convolutional kernels. *arXiv preprint arXiv:1806.01655*, 2018.

Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. *Advances in neural information processing systems*, 30, 2017.

Yann Lecun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. In *Proceedings of the IEEE*, pages 2278–2324, 1998.

Jaehoon Lee, Yasaman Bahri, Roman Novak, Samuel S Schoenholz, Jeffrey Pennington, and Jascha Sohl-Dickstein. Deep neural networks as gaussian processes. *arXiv preprint arXiv:1711.00165*, 2017.

Shiyu Liang, Yixuan Li, and Rayadurgam Srikant. Enhancing the reliability of out-of-distribution image detection in neural networks. *arXiv preprint arXiv:1706.02690*, 2017.

Jeremiah Zhe Liu, Zi Lin, Shreyas Padhy, Dustin Tran, Tania Bedrax-Weiss, and Balaji Lakshminarayanan. Simple and principled uncertainty estimation with deterministic deep learning via distance awareness. *arXiv preprint arXiv:2006.10108*, 2020.

Andrey Malinin and Mark Gales. Predictive uncertainty estimation via prior networks. In *Advances in Neural Information Processing Systems*, pages 7047–7058, 2018.

Gustav Mårtensson, Daniel Ferreira, Tobias Granberg, Lena Cavallin, Ketil Oppedal, Alessandro Padovani, Irena Rektorova, Laura Bonanni, Matteo Pardini, Milica G Kramberger, et al. The reliability of a deep learning model in clinical out-of-distribution mri data: a multicohort study. *Medical Image Analysis*, 66:101714, 2020.

Patrick McClure, Nao Rho, John A Lee, Jakub R Kaczmarzyk, Charles Y Zheng, Satrajit S Ghosh, Dylan M Nielson, Adam G Thomas, Peter Bandettini, and Francisco Pereira. Knowing what you know in brain segmentation using bayesian deep neural networks. *Frontiers in neuroinformatics*, 13:67, 2019.

Andrew McHutchon. Differentiating gaussian processes. *Cambridge (ed.)*, 2013.

Dimitrios Milios, Raffaello Camoriano, Pietro Michiardi, Lorenzo Rosasco, and Maurizio Filippone. Dirichlet-based gaussian processes for large-scale calibrated classification. *arXiv preprint arXiv:1805.10915*, 2018.

Thomas P Minka. Expectation propagation for approximate bayesian inference. *arXiv preprint arXiv:1301.2294*, 2013.

Miguel Monteiro, Loïc Le Folgoc, Daniel Coelho de Castro, Nick Pawlowski, Bernardo Marques, Konstantinos Kamnitsas, Mark van der Wilk, and Ben Glocker. Stochastic segmentation networks: Modelling spatially correlated aleatoric uncertainty. *arXiv preprint arXiv:2006.06015*, 2020.

Tanya Nair, Doina Precup, Douglas L Arnold, and Tal Arbel. Exploring uncertainty measures in deep networks for multiple sclerosis lesion detection and segmentation. *Medical image analysis*, 59:101557, 2020.

Radford M Neal. *Bayesian learning for neural networks*, volume 118. Springer Science & Business Media, 2012.

Radford M Neal et al. Mcmc using hamiltonian dynamics. *Handbook of markov chain monte carlo*, 2(11):2, 2011.

Yuval Netzer, Tao Wang, Adam Coates, Alessandro Bissacco, Bo Wu, and Andrew Y Ng. Reading digits in natural images with unsupervised feature learning. 2011.

Trung V Nguyen, Edwin V Bonilla, et al. Collaborative multi-output gaussian processes. In *UAI*, pages 643–652. Citeseer, 2014.

Sebastian W Ober, Carl E Rasmussen, and Mark van der Wilk. The promises and pitfalls of deep kernel learning. *arXiv preprint arXiv:2102.12108*, 2021.

Shreyas Padhy, Zachary Nado, Jie Ren, Jeremiah Liu, Jasper Snoek, and Balaji Lakshminarayanan. Revisiting one-vs-all classifiers for predictive uncertainty and out-of-distribution detection in neural networks. *arXiv preprint arXiv:2007.05134*, 2020.

Nick Pawlowski, Andrew Brock, Matthew CH Lee, Martin Rajchl, and Ben Glocker. Implicit weight uncertainty in neural networks. *arXiv preprint arXiv:1711.01297*, 2017.

Walter Hugo Lopez Pinaya, Petru-Daniel Tudosiu, Robert Gray, Geraint Rees, Parashkev Nachev, Sébastien Ourselin, and M. Jorge Cardoso. Unsupervised brain anomaly detection and segmentation with transformers. In Mattias Heinrich, Qi Dou, Marleen de Bruijne, Jan Lellmann, Alexander Schläfer, and Floris Ernst, editors, *Proceedings of the Fourth Conference on Medical Imaging with Deep Learning*, volume 143 of *Proceedings of Machine Learning Research*, pages 596–617. PMLR, 07–09 Jul 2021. URL `https://proceedings.mlr.press/v143/pinaya21a.html`.

Sebastian Popescu, David Sharp, James Cole, and Ben Glocker. Hierarchical gaussian processes with wasserstein-2 kernels. *arXiv preprint arXiv:2010.14877*, 2020.

Prasanna Porwal, Samiksha Pachade, Ravi Kamble, Manesh Kokare, Girish Deshmukh, Vivek Sahasrabuddhe, and Fabrice Meriaudeau. Indian diabetic retinopathy image dataset (idrid): a database for diabetic retinopathy screening research. *Data*, 3(3):25, 2018.

Joaquin Quinonero-Candela and Carl Edward Rasmussen. A unifying view of sparse approximate gaussian process regression. *The Journal of Machine Learning Research*, 6:1939–1959, 2005.

Mihaela Rosca, Theophane Weber, Arthur Gretton, and Shakir Mohamed. A case for new neural network smoothness constraints. 2020.

Lukas Ruff, Jacob R Kauffmann, Robert A Vandermeulen, Grégoire Montavon, Wojciech Samek, Marius Kloft, Thomas G Dietterich, and Klaus-Robert Müller. A unifying review of deep and shallow anomaly detection. *Proceedings of the IEEE*, 2021.

Hugh Salimbeni and Marc Deisenroth. Doubly stochastic variational inference for deep gaussian processes. In *Advances in Neural Information Processing Systems*, pages 4588–4599, 2017.

Thomas Schlegl, Philipp Seeböck, Sebastian M Waldstein, Georg Langs, and Ursula Schmidt-Erfurth. f-anogan: Fast unsupervised anomaly detection with generative adversarial networks. *Medical image analysis*, 54:30–44, 2019.

Jure Sokolić, Raja Giryes, Guillermo Sapiro, and Miguel RD Rodrigues. Robust large margin deep neural networks. *IEEE Transactions on Signal Processing*, 65(16):4265–4280, 2017.

Joes Staal, Michael D Abràmoff, Meindert Niemeijer, Max A Viergever, and Bram Van Ginneken. Ridge-based vessel segmentation in color images of the retina. *IEEE transactions on medical imaging*, 23(4):501–509, 2004.

EM Sweeney, RT Shinohara, CD Shea, DS Reich, and Ciprian M Crainiceanu. Automatic lesion incidence estimation and detection in multiple sclerosis using multisequence longitudinal mri. *American Journal of Neuroradiology*, 34(1):68–73, 2013.

Xiaoli Tang. The role of artificial intelligence in medical imaging research. *BJR| Open*, 2(1): 20190031, 2019.

Michalis Titsias. Variational learning of inducing variables in sparse gaussian processes. In *Artificial Intelligence and Statistics*, pages 567–574, 2009.

Ivan Ustyuzhaninov, Ieva Kazlauskaite, Markus Kaiser, Erik Bodin, Neill DF Campbell, and Carl Henrik Ek. Compositional uncertainty in deep gaussian processes. *arXiv preprint arXiv:1909.07698*, 2019.

Joost van Amersfoort, Lewis Smith, Yee Whye Teh, and Yarin Gal. Simple and scalable epistemic uncertainty estimation using a single deep deterministic neural network. *arXiv preprint arXiv:2003.02037*, 2020.

Joost van Amersfoort, Lewis Smith, Andrew Jesson, Oscar Key, and Yarin Gal. On feature collapse and deep kernel learning for single forward pass uncertainty. *arXiv preprint arXiv:2102.11409*, 2021.

Mark Van der Wilk, Carl Edward Rasmussen, and James Hensman. Convolutional gaussian processes. In *Advances in Neural Information Processing Systems*, pages 2849–2858, 2017.

Mark van der Wilk, ST John, Artem Artemev, and James Hensman. Variational gaussian process models without matrix inverses. In *Symposium on Advances in Approximate Bayesian Inference*, pages 1–9. PMLR, 2020.

Tsui-Wei Weng, Huan Zhang, Pin-Yu Chen, Jinfeng Yi, Dong Su, Yupeng Gao, Cho-Jui Hsieh, and Luca Daniel. Evaluating the robustness of neural networks: An extreme value theory approach. *arXiv preprint arXiv:1801.10578*, 2018.

Christopher KI Williams and David Barber. Bayesian classification with gaussian processes. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(12):1342–1351, 1998.

Andrew Gordon Wilson and Pavel Izmailov. Bayesian deep learning and a probabilistic perspective of generalization. *arXiv preprint arXiv:2002.08791*, 2020.

Andrew Gordon Wilson, Zhiting Hu, Ruslan Salakhutdinov, and Eric P Xing. Deep kernel learning. In *Artificial Intelligence and Statistics*, pages 370–378, 2016.

S Kevin Zhou, Hayit Greenspan, Christos Davatzikos, James S Duncan, Bram Van Ginneken, Anant Madabhushi, Jerry L Prince, Daniel Rueckert, and Ronald M Summers. A review of deep learning in medical imaging: Imaging traits, technology trends, case studies with progress highlights, and future promises. *Proceedings of the IEEE*, 2021.

Appendix A. Proving Lipschitz bounds in a DistGP layer

We here prove Propositions 3 and 5 of Sec. 3.3.

Lemmas on p-norms/ We have the following relations between norms : $\|x\|_2 \leq \|x\|_1$ and $\|x\|_1 \leq \sqrt{D}\|x\|_2$. Will be used for the proof of Proposition 2.

Proof of Proposition 3. Throughout this proof we shall refer to the first two moments of a Gaussian distribution by $m(\cdot)$, $v(\cdot)$. Explicitly writing the Wasserstein-2 distances of the inequality we get:

$$|m(F(\mu)) - m(F(\nu))|^2 + |v(F(\mu)) - v(F(\nu)|^2 \leq L|m_1 - m_2|^2 + |\Sigma_1 - \Sigma_2|^2 \tag{91}$$

We focus on the mean part and applying Cauchy–Schwarz we get the following inequality:

$$|[K_{\mu u} - K_{\nu u}]\, K_{uu}^{-1} m|^2 \leq \|K_{\mu u} - K_{\nu u}\|_2^2 \|K_{uu}^{-1} m\|_2^2 \tag{92}$$

To simplify the problem and without loss of generality we consider U_z to be a sufficient statistic for the set of inducing points Z. Expanding the first term of the r.h.s. we get:

$$\left[\sigma^2 \exp \frac{-W_2(\mu, U_z)}{l^2} - \sigma^2 \exp \frac{-W_2(\nu, U_z)}{l^2}\right]^2 \tag{93}$$

We assume $\nu = \mu + h$, where $h \sim \mathcal{N}(|m_1 - m_2|, |\Sigma_1 - \Sigma_2|)$ and μ is a high density point in the data manifold, hence $W_2(\mu - U_z) = 0$. We denote $m(h)^2 + var(h)^2 = \lambda$. Considering the general equality $\log(x - y) = \log(x) + \log(y) + \log(\frac{1}{y} - \frac{1}{x})$ and applying it to our case we get that:

$$\log \mid m(F(\mu)) - m(F(\nu)) \mid^2 \leq \log \left[\sigma^2 - \sigma^2 \exp \frac{-\lambda}{l^2}\right]^2 \tag{94}$$

$$\leq 2 \log \sigma^2 - 2\frac{\lambda}{l^2} + 2 \log \left[\exp \frac{\lambda}{l^2} - 1\right] \tag{95}$$

$$\leq 2 \log \left[\sigma^2 \exp \frac{\lambda}{l^2}\right] \tag{96}$$

We have the general inequality $\exp x \leq 1 + x + x^2$ for $x \leq 1.79$, which for $0 \leq x \leq 1$ can be modified as $\exp x \leq 1 + 2x$. Applying this new inequality and taking the exponential we now obtain:

$$\mid m(F(\mu)) - m(F(\nu)) \mid^2 \leq \left[\sigma^2 + 2\sigma^2 \frac{\lambda}{l^2}\right]^2 \tag{97}$$

$$\leq \sigma^4 + \sigma^4 \frac{\lambda}{l^2} + 4\sigma^4 \frac{(\lambda)^2}{l^4} \tag{98}$$

$$\leq 16\sigma^4 \frac{\lambda}{l^2} \tag{99}$$

where the last inequality follows from the ball constraints made in the definition. We now move to the variance components of the Lipschitz bound, we notice that

$$\mid v(F(\mu))^{\frac{1}{2}} - v(F(\nu))^{\frac{1}{2}} \mid^2 \leq \mid v(F(\mu))^{\frac{1}{2}} - v(F(\nu))^{\frac{1}{2}} \mid\mid v(F(\mu))^{\frac{1}{2}} + v(F(\nu))^{\frac{1}{2}} \mid \tag{100}$$

$$\leq \mid v(F(\mu)) - v(F(\nu)) \mid \tag{101}$$

which after applying Cauchy–Schwarz results in an upper bound of the form:

$$\|K_{\mu,U_z} - K_{\nu,U_z}\|_2^2 \|K_{U_z}^{-1}(K_{U_z-S})K_{U_z}^{-1}\|_2 \tag{102}$$

Using that $\|K_{\mu,U_z} - K_{\nu,U_z}\|_2^2 \leq \frac{16\sigma^4\lambda}{l^2}$ we obtain that:

$$|v(F(\mu)) - v(F(\nu))| \leq \frac{16\sigma^4\lambda}{l^2}\|K_{U_z}^{-1}(K_{U_z} - S)K_{U_z}^{-1}\|_2 \tag{103}$$

Now taking into consideration both the upper bounds on the mean and variance components we arrive at the desired Lipschitz constant.

Proof of Proposition 5. Using the definition for Wasserstein-2 distances for the l.h.s of the inequality, we can re-express as follows:

$$W_2\left(f(\mu), f(\nu)\right) \leq \|m_1 A - m_2 A\|_2^2 + \|(\sigma_1^2 A^2)^{1/2} - (\sigma_2^2 A^2)^{1/2}\|_F^2 \tag{104}$$

which after rearranging terms and noticing that inside the Frobenius norm we have scalars, becomes:

$$W_2\left(f(\mu), f(\nu)\right) \leq \|(m_1 - m_2)A\|_2^2 + [\sigma_1^2 A^2)^{1/2} - (\sigma_2^2 A^2)^{1/2}]^2 \tag{105}$$

We can now apply the Cauchy–Schwarz inequality for the part involving means and multiplying the right hand side with $\sqrt{C}$, which represents the number of channels, we get:

$$\|(m_1 - m_2)A\|_2^2 + [\sigma_1^2 A^2)^{1/2} - (\sigma_2^2 A^2)^{1/2}]^2 \leq \|m_1 - m_2\|_2^2 \sqrt{C}\|A\|_2^2 + \sqrt{C}[\sigma_1^2 A^2)^{1/2} - (\sigma_2^2 A^2)^{1/2}]^2 \tag{106}$$

We can notice that the Lipschitz constant for the component involving mean terms is $\sqrt{C}\|A\|_2^2$. Hence, we try to prove that the same L is also available for the variance terms component. Hence, we can affirm that:

$$L = \sqrt{C}\|A\|_2^2 \leftrightarrow \sqrt{C}[\sigma_1^2 A^2)^{1/2} - (\sigma_2^2 A^2)^{1/2}]^2 \leq [\sigma_1 - \sigma_2]^2 \sqrt{C}\|A\|_2^2 \tag{107}$$

By virtue of Cauchy–Schwarz we have the following inequality $\sqrt{C}[\sigma_1 A - \sigma_2 A]^2 \leq [\sigma_1 - \sigma_2]^2 \sqrt{C}\|A\|_2^2$. Hence the aforementioned if and only if statement will hold if we prove that

$$\sqrt{C}\left[(\sigma_1^2 A^2)^{\frac{1}{2}} - (\sigma_2^2 A^2)^{\frac{1}{2}}\right]^2 \leq \sqrt{C}\left[\sigma_1 A - \sigma_2 A\right]^2 \tag{108}$$

which after expressing in terms of norms becomes:

$$\sqrt{C}\left[\|\sigma_1 A\|_2 - \|\sigma_2 A\|_2\right]^2 \leq \sqrt{C}\left[\|\sigma_1 A\|_1 - \|\sigma_2 A\|_1\right]^2 \tag{109}$$

Expanding the square brackets gives:

$$\sqrt{C}\left[\|\sigma_1 A\|_2^2 + \|\sigma_2 A\|_2^2 - 2\|\sigma_1 A\|_2\|\sigma_2 A\|_2\right] \leq \sqrt{C}\left[\|\sigma_1 A\|_1^2 + \|\sigma_2 A\|_1^2 - 2\|\sigma_1 A\|_1\|\sigma_2 A\|_1\right] \tag{110}$$

This inequality holds by applying the p-norm lemma, thereby the if and only if statement is satisfied. Consequently, the Lipschitz constant is $\sqrt{C}\|A\|_2^2$.

Appendix B. Deriving function contraction requirements in DistGP Layers

We here prove Proposition 7 of Sec. 3.4.

Proof of Proposition 7. We are interested in determining the specific scenarios in which the function space collapses to constant values. Hence we explicitly write $\mathbb{E}\left[\|u_l(x) - u_l(x^*)\|_2^2|\ u_{l-1}\right]$ as:

$$= \sum_{j=1}^{D_l} \mathbb{E}\left[\|u_l^j(x) - u_l^j(x^*)\|_2^2|\ u_{l-1}\right] \tag{111}$$

$$= \sum_{j=1}^{D_l} \mathbb{E}\left[\left(u_l^j(x)\right)^2 \mid u_{l-1}\right] - 2\mathbb{E}\left[u_l^j(x)u_l^j(x^*) \mid u_{l-1}\right] + \mathbb{E}\left[\left(u_l^j(x^*)\right)^2 \mid u_{l-1}\right] \tag{112}$$

$$= \sum_{j=1}^{D_l} \sigma_l^2 + m_l^2(x) - 2m_l(x)m_l(x^*) - 2k^{W_2}\left[\mu_l(x), \mu_l(x^*)\right] + \sigma_l^2 + m_l^2(x^*) \tag{113}$$

$$= \sum_{j=1}^{D_l} \left[m_l^j(x) - m_l^j(x^*)\right]^2 + 2\sigma_l^2 - 2\sigma_l^2 \exp - \left[\frac{|\ m_{l-1}^j(x) - m_{l-1}^j(x^*)\ |^2}{2l_l^2}\right] \tag{114}$$

, where in the last equation we have ignored the variance part of the Wasserstein-2 kernel since the two variance terms are equal. We make use of the following inequality $1 - \exp -x \leq x$ for $x \geq 0$ and equality only in the case that $x = 0$, resulting in the following upper bound:

$$\mathbb{E}\left[\|u_l(x) - u_l(x^*)\|_2^2|u_{l-1}\right] \leq \sum_{j=1}^{D_l} \left[m_l^j(x) - m_l^j(x^*)\right]^2 + \sigma_l^2 \frac{|\ m_{l-1}^j(x) - m_{l-1}^j(x^*)\ |^2}{l_l^2} \tag{115}$$

We can now view the previously defined operator $\overline{m_l(x)W_l}$ as an inner product in vector space between a tiled version of $m_l(x)$ and a normalised version of W_l, more specifically:

$$\langle [m_l(x), \cdots, m_l(x)]\,, \left[\frac{W_{l,1}}{m_l}, \cdots, \frac{W_{l,m_l D_{l-1}}}{m_l}\right]\rangle = \overline{m_l(x)W_l} \tag{116}$$

where m_l is the number of dimensions caused by the affine embedding function Ψ_l in the l-th layer of the hierarchy.

Our current goal is to relate $m_l^j(\cdot)$ to $m_{l-1}^j(\cdot)$. We can now apply Cauchy-Schwarz to:

$$|\ m_l^j(x) - m_l^j(x^*)\ |^2 = |\ \overline{m_{l-1}(x)W_l} - \overline{m_{l-1}(x^*)W_l}\ |^2 \tag{117}$$

$$= |\ \langle [m_{l-1}(x) - m_{l-1}(x^*), \cdots, m_{l-1}(x) - m_{l-1}(x^*)]\,, \left[\frac{W_{l,1}}{m_l}, \cdots, \frac{W_{l,m_l D_{l-1}}}{m_l}\right]\rangle\ |^2 \tag{118}$$

$$\leq D_{l-1}m_l\left[m_{l-1}(x) - m_{l-1}(x^*)\right]^2 * \langle \tilde{W}_l, \tilde{W}_l\rangle \tag{119}$$

where in the last line we denoted $\tilde{W}_l = [\frac{W_{l,1}}{D_l}, \cdots, \frac{W_{l,D_{l-1}m_l}}{m_l}]$ to avoid cluttering.

We can now apply the previous result to equation (115):

$$\mathbb{E}\left[\|u_l(x) - u_l(x^*)\|_2^2 || \; u_{l-1}\right] \leq \sum_{j=1}^{D_l} m_l D_{l-1} \left(m_{l-1}(x) - m_{l-1}(x^*)\right)^2 * \langle \tilde{W}_l, \tilde{W}_l \rangle \tag{120}$$

$$+ \frac{\sigma_l^2}{l_l^2} \mid m_{l-1}^j(x) - m_{l-1}^j(x^*) \mid^2$$

$$\leq \sum_{j=1}^{D_l} \left[m_l D_{l-1} * \langle \tilde{W}_l, \tilde{W}_l \rangle + \frac{\sigma_l^2}{l_l^2} \right] \mid m_{l-1}^j(x) - m_{l-1}^j(x^*) \mid^2 \tag{121}$$

We can now recursively apply the previously derived Cauchy-Schwarz based inequality to obtain:

$$\mathbb{E}\left[|| \; u_l(x) - u_l(x^*) \; ||_2^2 \; | \{u_l\}_{l=1}^{l-1} \right] \leq \left[m_l D_{l-1} * \langle \tilde{W}_l, \tilde{W}_l \rangle + \frac{\sigma_l^2}{2l_l^2} \right] \prod_{l=1}^{l-1} D_l m_l D_{l-1} \langle \tilde{W}_l, \tilde{W}_l \rangle \tag{122}$$

$$[m_1(x) - m_1(x^*)]^2$$

By Markov's inequality, for any $\epsilon > 0$ we have that:

$$P\left(|| \; u_{l+1}(x) - u_{l+1}(x^*) \; ||_2 \geq \epsilon \right) \leq \frac{1}{\epsilon^2} \left[m_l D_{l-1} * \langle \tilde{W}_l, \tilde{W}_l \rangle + \frac{\sigma_l^2}{2l_l^2} \right] \prod_{l=1}^{l-1} D_l m_l D_{l-1} \langle \tilde{W}_l, \tilde{W}_l \rangle \tag{123}$$

$$[m_1(x) - m_1(x^*)]^2$$

Then, only in the case that $\left[m_l D_{l-1} * \langle \tilde{W}_l, \tilde{W}_l \rangle + \frac{\sigma_l^2}{2l_l^2} \right] \leq 1$ and $D_l m_l D_{l-1} \langle \tilde{W}_l, \tilde{W}_l \rangle \leq$ is satisfied for intermediate layers, we can apply the first Borel-Cantelli lemma to obtain:

$$P\left(\cap_{l=1}^{\infty} \cup_{m=l}^{\infty} || \; u_m(x) - u_m(x^*) \; ||_2 \geq \epsilon \right) = 0 \tag{124}$$

Lastly, we can express the following:

$$P\left(|| \; u_n(x) - u_n(x^*) \; ||_2 \to 0 \right) = P\left(\cap_{k=1}^{\infty} \cup_{l=1}^{\infty} \cap_{m=l}^{\infty} || \; u_m(x) - u_m(x^*) \; ||_2 \leq \frac{1}{k} \right) \tag{125}$$

$$= 1 - P\left(\cup_{k=1}^{\infty} \cap_{l=1}^{\infty} \cup_{m=l}^{\infty} || \; u_m(x) - u_m(x^*) \; ||_2 \geq \frac{1}{k} \right) \tag{126}$$

$$\geq 1 - \sum_{k=1}^{\infty} P\left(\cap_{l=1}^{\infty} \cup_{m=l}^{\infty} || \; u_m(x) - u_m(x^*) \; ||_2 \geq \frac{1}{k} \right) = 1 \tag{127}$$

From which we obtain the proof of our proposition, respectively $P\left(|| \; u_n(x) - u_n(x^*) \; ||_2 \to 0 \right) = 1$

Appendix C. Outlier detection in MRI scans with Tumors.

Remarks. We provide additional plots for the task investigated in sec. 5.1.3 for DistGP-Seg and OVA-DM as they were the only models to provide decent outlier detection capabilities. We refer the reader to Figures 17 and 18. From case study A, we can see that OVA-DM is over-segmenting across all FPR levels almost randomly from outside the tumor area, whereas DistGP-Seg is over-segmenting at $FPR = \{1.0, 5.0\}$ areas around the margins of the ventricles. For case study B, at $FPR = \{0.5, 1.0, 5.0\}$ OVA-DM seems to be under-segmenting the tumor in comparison to DistGP-Seg. The same observation can be made again for case study C. Lastly, for case study D DistGP-Seg seems to be under-segmenting for $FPR = \{0.1, 0.5\}$.

Appendix D. Outlier detection in MRI scans with WMH.

Data and pre-processing. Brain MRI scans from the 2015 Longitudinal Multiple Sclerosis Lesion Segmentation Challenge Sweeney et al. (2013) which comprises of FLAIR, PD, T2-weighted, and T1- weighted volumes from a total of 110 MR imaging studies (11 longitudinal studies each of 10 subjects). All participants gave written consent and were scanned as part of an institutional review board approved natural history protocol. For the purposes of the task at hand, we only use the baseline FLAIR scans. All FLAIR images are pre-processed with skull-stripping, N4 bias correction, rigid registration to MNI152 space and histogram matching between UKBB and BraTS. Finally, we normalize intensities of each scan via linear scaling of its minimum and maximum intensities to the [-1,1] range.

Remarks. The task of detecting white matter hyperintensities (WMH) is considerably more difficult than detecting tumors, the latter usually presenting itself as a large blob, whereas the former constitutes of multiple non-contiguous areas of varying shapes. From Figure 19 we can notice that the large connected WMH regions are reliably detected as outliers, with smaller disconnected WMH regions being only in some cases outlined. Another issue is over-segmentation, as seen in case study C.

Appendix E. Evaluation on Retina scans.

Data and pre-processing. **DRIVE:** The Digital Retinal Images for Vessel Extraction (DRIVE) dataset Staal et al. (2004) is a dataset for retinal vessel segmentation. It consists of a total of JPEG 40 color fundus images; including 7 abnormal pathology cases. The images were obtained from a diabetic retinopathy screening program in the Netherlands. The images were acquired using Canon CR5 non-mydriatic 3CCD camera with FOV equals to 45 degrees. Each image resolution is 584*565 pixels with eight bits per color channel (3 channels).

The set of 40 images was equally divided into 20 images for the training set and 20 images for the testing set. Inside both sets, for each image, there is circular field of view (FOV) mask of diameter that is approximately 540 pixels. Inside training set, for each image, one manual segmentation by an ophthalmological expert has been applied. Inside testing set, for each image, two manual segmentations have been applied by two different observers, where the first observer segmentation is accepted as the ground-truth for performance evaluation.

STARE: STructured Analysis of the Retina (STARE) database Hoover et al. (2000) was created by scanning and digitizing the retinal image photographs. Hence, the image quality of this database is less than the other public databases. The STARE dataset comprises 97 images (59 AMD and 38 normal) taken using a fundus camera (TOPCON TRV-50; Topcon Corp., Tokyo, Japan) at a 35° field and with a resolution of 605×700 pixels. Its retina scans are from subjects suffering from the following retina pathologies: Hollenhorst Emboli Branch Retinal Artery Occlusion, Cilio-Retinal Artery Occlusion, Branch Retinal Vein Occlusion, Central Retinal Vein Occlusion, Hemi-Central Retinal Vein Occlusion, Background Diabetic Retinopathy, Proliferative Diabetic Retinopathy, Arteriosclerotic Retinopathy, Hypertensive Retinopathy, Coat's, Macroaneurism, Choroidal Neovascularization.

IDRID: The Indian Diabetic Retinopathy Image Dataset (IDRID) dataset Porwal et al. (2018), is a publicly available retinal fundus image database consisting of 516 images categorised into two parts: retina images with the signs of Diabetic Retinopathy and/or Diabetic Macular Edema; normal retinal images. Images were acquired using a Kowa VX-10a digital fundus camera with 50° field of view (FOV). The images have resolution of 4288×2848 pixels and are stored in jpg file format. We have pre-processed these images to match the FOV and resolution of DRIVE.

Task. We train a similar DistGP-Seg architecture (see sec. 5) adapted for 2D data on DRIVE (normative data) to segment blood vessels, subsequently at testing time we use STARE and IDRID (OOD data) to segment blood vessels in the presence of previouslt unseen pathologies on Retina scans.

Blood vessel segmentation on normal Retina scans. From Figure 20 we can observe that DistGP-Seg manages to correctly segment the blood vessels, while both distributional and within-data uncertainty are relatively low, which is to be expected as these testing examples represent in-distribution data.

Outlier detection of Retina pathologies. From Figure 21 we can observe that DistGP-Seg manages to correctly identity the vast majority of soft and hard exudates as outliers.

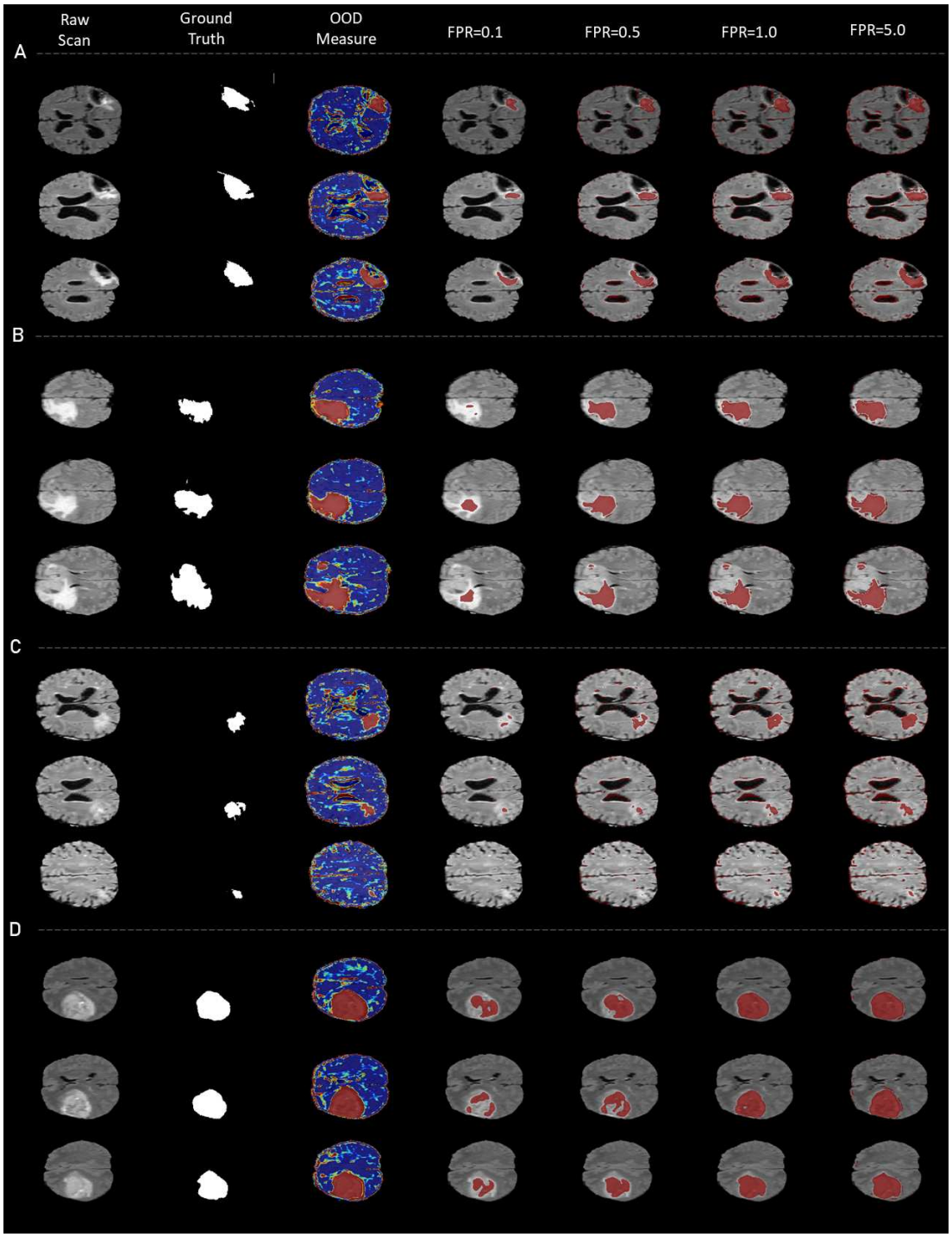


Figure 17: Detailed segmentation output for DistGP-Seg on BRATS. Mean segmentation represents the hard segmentation of brain tissues. OOD measure is the quantification of uncertainty for each model, using their own procedure. Higher values translate to appartenance to outlier status.

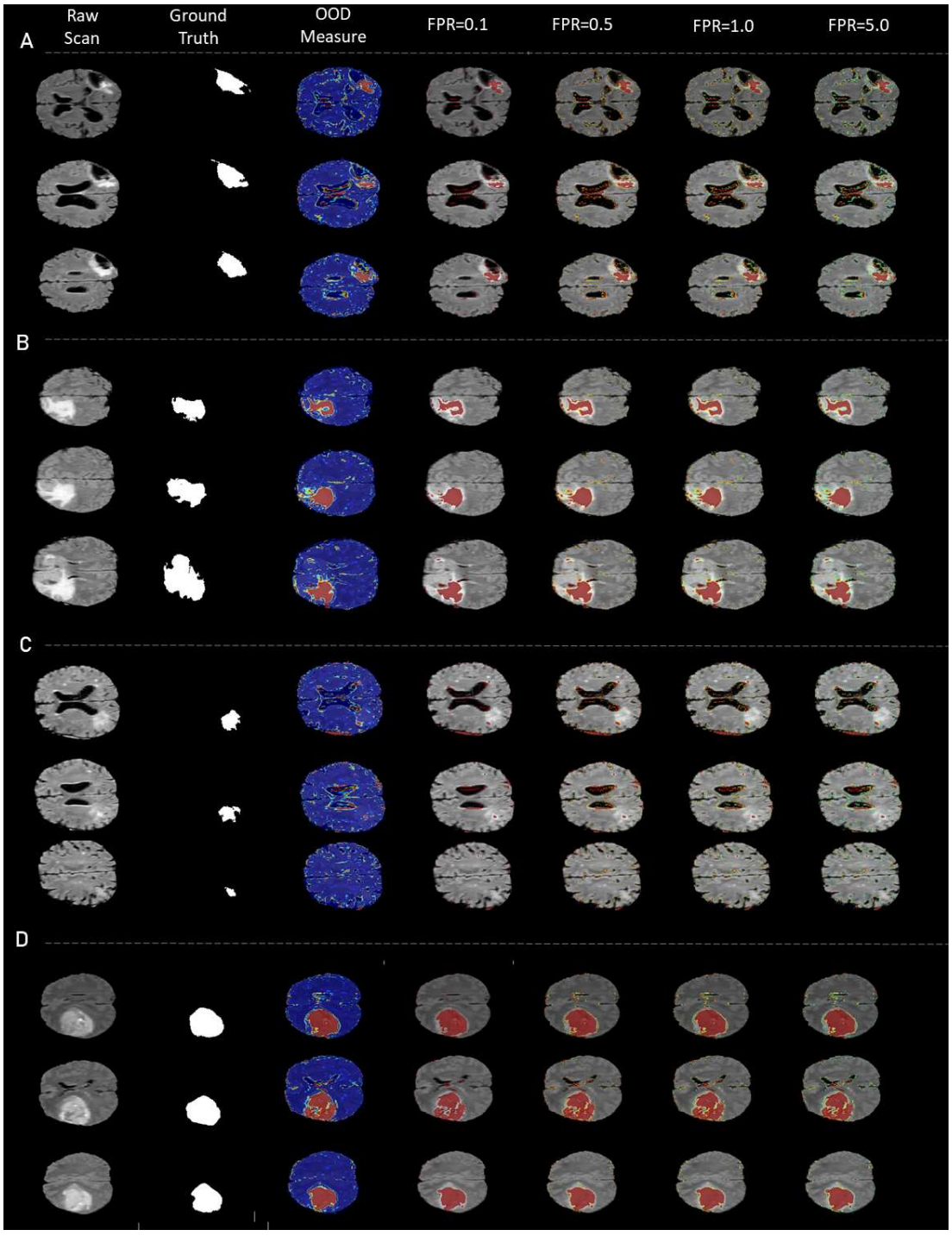


Figure 18: Detailed segmentation output for OVA-DM on BRATS. Mean segmentation represents the hard segmentation of brain tissues. OOD measure is the quantification of uncertainty for each model, using their own procedure. Higher values translate to appartenance to outlier status.

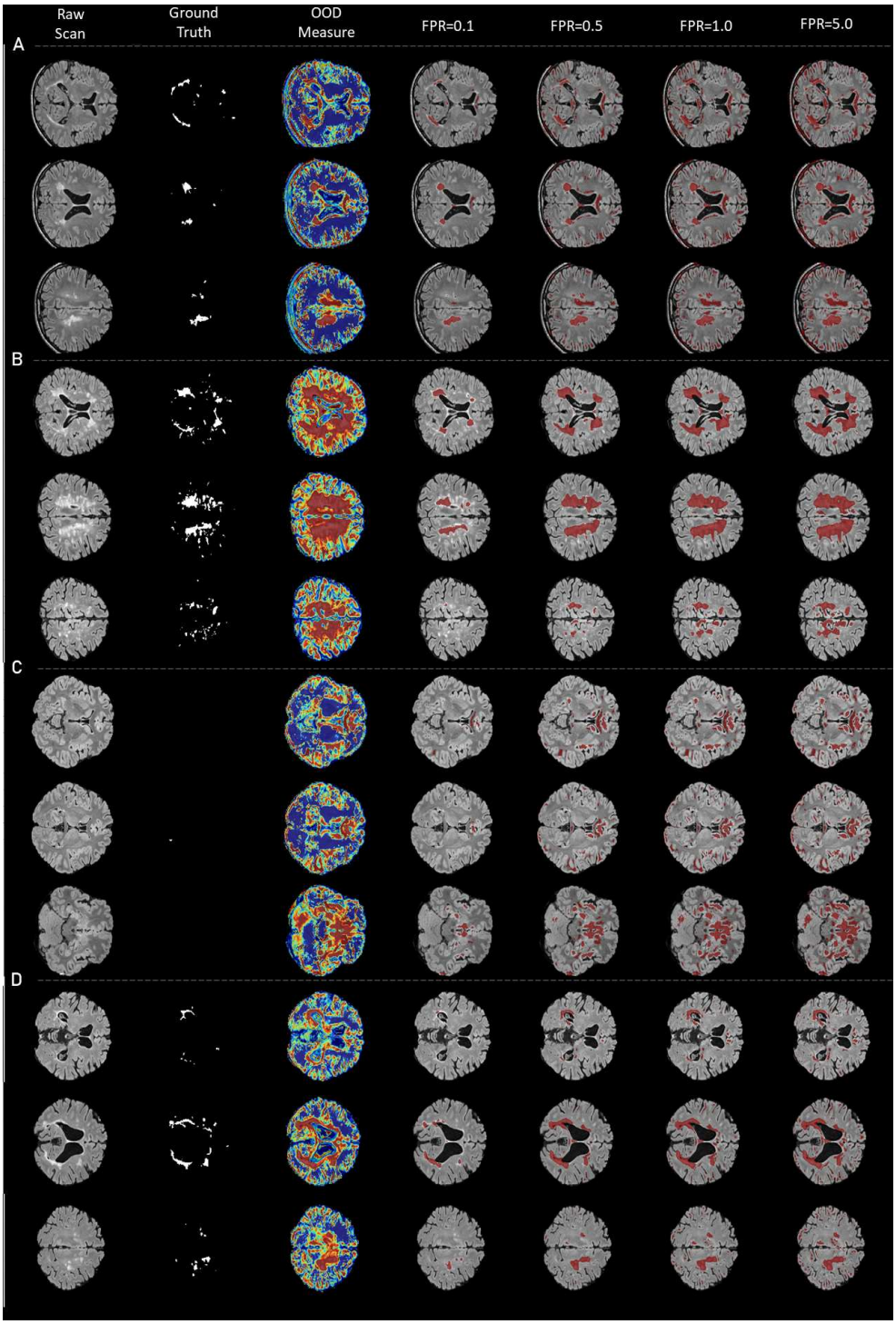


Figure 19: Detailed segmentation output for DistGP-Seg on WMH dataset. Mean segmentation represents the hard segmentation of brain tissues. OOD measure is the quantification of uncertainty for each model, using their own procedure. Higher values translate to appartenance to outlier status.

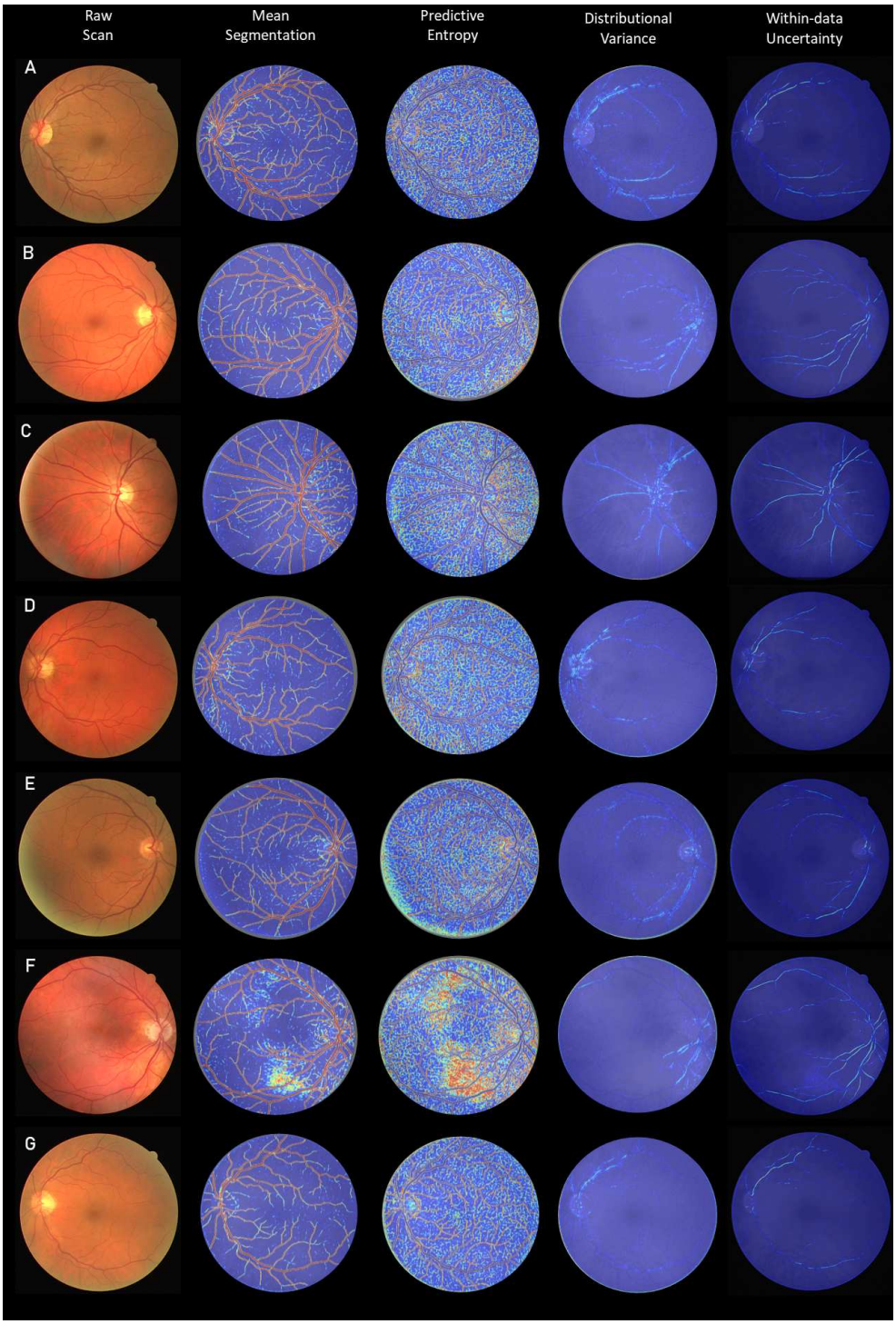


Figure 20: Detailed segmentation output for DistGP-Seg DRIVE dataset. Mean segmentation represents the hard segmentation of brain tissues. OOD measure is the quantification of uncertainty for each model, using their own procedure. Higher values translate to appartenance to outlier status.

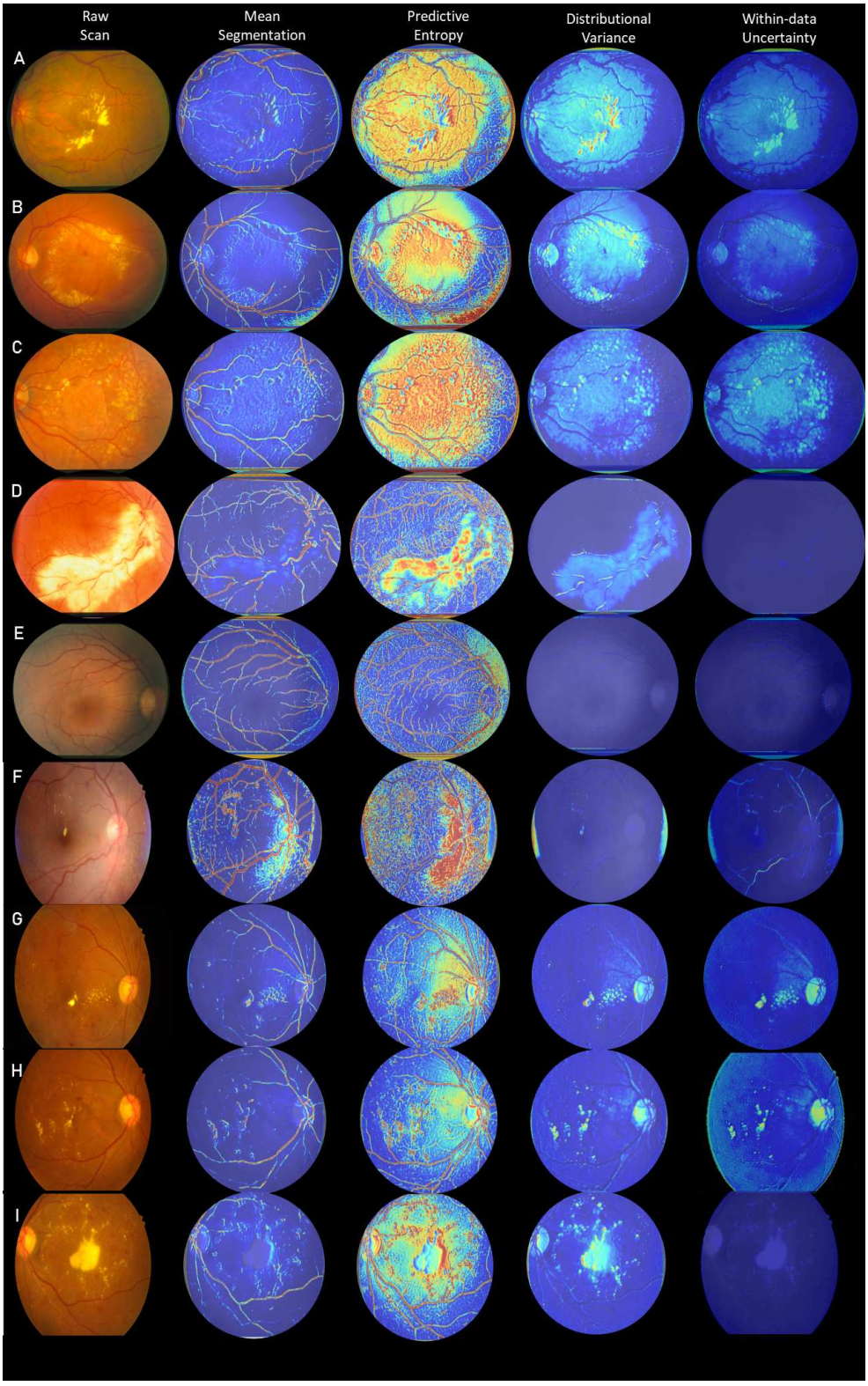


Figure 21: Detailed segmentation output for DistGP-Seg on STARE/IDRID datasets. Mean segmentation represents the hard segmentation of brain tissues. OOD measure is the quantification of uncertainty for each model, using their own procedure. Higher values translate to appartenance to outlier status. Case studies A-E originate from STARE, whereas the remainder from IDRID.