Characterizing Deep Gaussian Processes via Nonlinear Recurrence Systems

Anh Tong[1], **Jaesik Choi**[2, 3]

[1] Ulsan National Institute of Science and Technology
[2] Korea Advanced Institute of Science and Technology
[3] INEEJI
anhth@unist.ac.kr, jaesik.choi@kaist.ac.kr

Abstract

Recent advances in Deep Gaussian Processes (DGPs) show the potential to have more expressive representation than that of traditional Gaussian Processes (GPs). However, there exists a pathology of deep Gaussian processes that their learning capacities reduce significantly when the number of layers increases. In this paper, we present a new analysis in DGPs by studying its corresponding nonlinear dynamic systems to explain the issue. Existing work reports the pathology for the squared exponential kernel function. We extend our investigation to four types of common stationary kernel functions. The recurrence relations between layers are analytically derived, providing a tighter bound and the rate of convergence of the dynamic systems. We demonstrate our finding with a number of experimental results.

1 Introduction

Deep Gaussian Process (DGP) (Damianou and Lawrence 2013) is a new promising class of models which are constructed by a hierarchical composition of Gaussian processes. The strength of this model lies in its capacity to have richer representation power from the hierarchical construction and its robustness to overfitting from the probabilistic modeling. Therefore, there have been extensive studies (Hensman and Lawrence 2014; Dai et al. 2016; Bui et al. 2016; Cutajar et al. 2017; Salimbeni and Deisenroth 2017; Havasi, Hernández-Lobato, and Murillo-Fuentes 2018; Salimbeni et al. 2019; Lu et al. 2020; Ustyuzhaninov et al. 2020) contributing to this research area.

There exists a pathology, stating that the increase in the number of layers degrades the learning power of DGP (Duvenaud et al. 2014). That is, the functions produced by DGP priors become flat and cannot fit data. It is important to develop theoretical understanding of this behavior, and therefore to have proper tactics in designing model architectures and parameter regularization to prevent the issue. Existing work (Duvenaud et al. 2014) investigates the Jacobian matrix of a given model which can be analytically interpreted as the product of those in each layer. Based on the connection between the manifold of a function and the spectrum of its Jacobian, the authors show the degree of freedom is reduced significantly at deep layers. Another work (Dunlop et al. 2018) studies the ergodicity of the Markov chain to explain the pathology.

To explain such phenomena, we study a quantity which measures the distance of any two layer outputs. We present a new approach that makes use of the statistical properties of the quantity passing from one layer to another layer. Therefore, our approach accurately captures the relations of the distance quantity between layers. By considering kernel hyperparameters, our method recursively computes the relations of two consecutive layers. Interestingly, the recurrence relations provide a tighter bound than that of (Dunlop et al. 2018) and reveal the rate of convergence to fixed points. Under this unified approach, we further extend our analysis to five popular kernels which are not analyzed yet before. For example, the spectral mixture kernels do not suffer the pathology. We further provide a case study in DGP, showing the connection between our recurrence relations and learning DGPs.

Our contributions in this paper are: (1) we provide a new perspective of the pathology in DGP under the lens of chaos theory; (2) we show that the recurrence relation between layers gives us the rate of convergence to a fixed point; (3) we give a unified approach to form the recurrence relation for several kernel functions including the squared exponential kernel function, the cosine kernel function, the periodic kernel function, the rational quadratic kernel function and the spectral mixture kernel; (4) we justify our findings with numerical experiments. We use the recurrence relations in debugging DGPs and explore a new regularization on kernel hyperparameters to learn zero-mean DGPs. [1]

2 Background

Notation Throughout this paper, we use the boldface as vector or vector-value function. The superscript i.e. $f^{(d)}(\mathbf{x})$ is the d-th dimension of vector-valued function $\boldsymbol{f}(\mathbf{x})$.

2.1 Deep Gaussian Processes

We study DGPs in composition formulation where GP layers are stacked hierarchically. An N-layer DGP is defined as

$$\boldsymbol{f}_N \circ \boldsymbol{f}_{N-1} \circ \cdots \circ \boldsymbol{f}_1(\mathbf{x}),$$

[1] See https://arxiv.org/abs/2010.09301 for an extended version

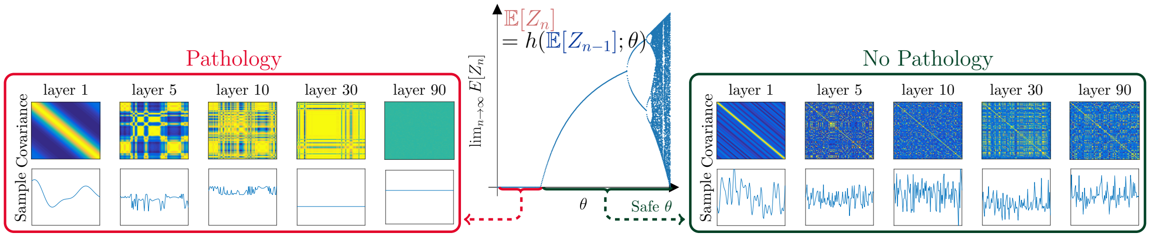

Figure 1: Studying the squared distance, Z_n, between outputs of two consecutive layers. The asymptotic property (middle plot) of the recurrence relation of this quantity between two consecutive layers decides the existence of pathology for a very deep model. Here, θ indicates kernel hyperparameters. The middle plot is the bifurcation plot providing the state of DGP at very deep layer. The pathology is identified by the zero-value region where $\mathbb{E}[Z_n] \to 0$. Note that this bifurcation plot is for illustration purpose only.

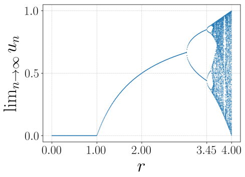

Figure 2: Bifurcation plot of the logistic function $u_n = r u_{n-1}(1 - u_{n-1})$.

where, at layer n, for dimension d, $f_n^{(d)}|\boldsymbol{f}_{n-1} \sim \mathcal{GP}(0, k_n(\cdot,\cdot))$ independently. Note that the GP priors have the mean functions set to zero. The nonzero-mean case is discussed later (Section 4.4). We shorthand $\boldsymbol{f}_n \circ \boldsymbol{f}_{n-1} \circ \cdots \circ \boldsymbol{f}_1(\mathbf{x})$ as $\boldsymbol{f}_n(\mathbf{x})$ and write $k_n(\boldsymbol{f}_{n-1}(\mathbf{x}), \boldsymbol{f}_{n-1}(\mathbf{x}'))$ as $k_n(\mathbf{x}, \mathbf{x}')$. Let m be the number of output of $\boldsymbol{f}_n$. All layers have the same hyperparameters.

Theorem 2.1 ((Dunlop et al. 2018)). *Assume that $k(\mathbf{x},\mathbf{x}')$ is given by the squared exponential kernel function with variance σ^2 and lengthscale ℓ^2 and that the input $\mathbf{x}$ is bounded. Then if $\sigma^2 < \ell^2/m$,*

$$\mathbb{P}(\|\boldsymbol{f}_n(\mathbf{x}) - \boldsymbol{f}_n(\mathbf{x}')\|_2 \xrightarrow[n\to\infty]{} 0 \quad \textit{for all } \mathbf{x}, \mathbf{x}' \in \mathcal{D}) = 1$$

where $\mathbb{P}$ denotes the law of process $\{f_n\}$.

This theorem tells us the criterion that the event of vanishing in output magnitude happens infinitely often with probability 1.

2.2 Analyzing Dynamic Systems with Chaos Theory

Recurrence maps representing dynamic transitions between DGP layers are nonlinear. Studying the dynamic states and convergence properties for nonlinear recurrences is not as well-established as those of linear recurrences. As an example, given a simple nonlinear model like the logistic map: $u_n = r u_{n-1}(1 - u_{n-1})$, its dynamic behaviors can be complicated (May 1976).

Recurrent plots or bifurcation plots have been used to analyze the behavior of chaotic systems. The plots are produced by simulating and recording the dynamic states up to very large time points. This tool allows us to monitor the qualitative changes in a system, illustrating fixed points asymptotically, or possible visited values. Other techniques, e.g. transient chaos (Poole et al. 2016), recurrence relations (Schoenholz et al. 2017) have been used to study deep neural networks.

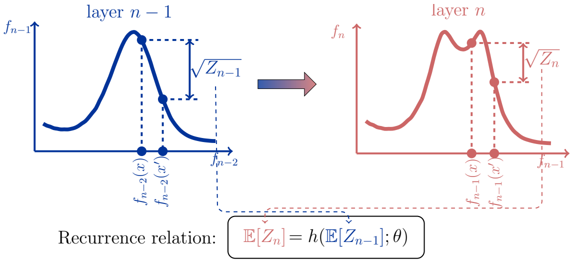

Figure 3: Finding the recurrence relation of the quantity $\mathbb{E}[(f_n(x) - f_n(x'))^2]$ between two consecutive layers.

We take the logistic map as an example to understand a recurrence relation. Figure 2 is the bifurcation plot of the logistic map. This logistic map is used to describe the characteristics of a system which models a population function. We can see that the plot reveals the state of the system, showing whether the population becomes extinct ($0 < r < 1$), stable ($1 < r < 3$), or fluctuating ($r > 3.4$) by seeing the parameter r.

3 Moment-generating Function of Distance Quantity

Throughout this paper, we are interested in quantifying the expectation of the squared Euclidean distance between any two outputs of a layer and thereby study the dynamics of this quantity from a layer to the next layer. Figure 1 shows that we can make use of the found recurrence relations to study the pathology of DGPs.

For any input pair $\mathbf{x}$ and $\mathbf{x}'$, we define such quantity at layer n as $Z_n = \|\boldsymbol{f}_n(\mathbf{x}) - \boldsymbol{f}_n(\mathbf{x}')\|_2^2 = \sum_{d=1}^m \left(f_n^{(d)}(\mathbf{x}) - f_n^{(d)}(\mathbf{x}')\right)^2$. When the previous layer $\boldsymbol{f}_{n-1}$ is given, the difference between any $f_n^{(d)}(\mathbf{x})$ and $f_n^{(d)}(\mathbf{x}')$ is Gaussian,

$$\left(f_n^{(d)}(\mathbf{x}) - f_n^{(d)}(\mathbf{x}')\right)|\boldsymbol{f}_{n-1} \sim \mathcal{N}(0, s_n).$$

Here $s_n = k_n(\mathbf{x},\mathbf{x}) + k_n(\mathbf{x}',\mathbf{x}') - 2k_n(\mathbf{x},\mathbf{x}')$ which is obtained from subtracting two dependent Gaussians. We can normalize the difference between $f_n^{(d)}(\mathbf{x})$ and $f_n^{(d)}(\mathbf{x}')$ by a factor $\sqrt{s_n}$ to obtain the form of standard normal distribution as

$$\frac{(f_n^{(d)}(\mathbf{x}) - f_n^{(d)}(\mathbf{x}'))}{\sqrt{s_n}} | \boldsymbol{f}_{n-1} \sim \mathcal{N}(0,1).$$

Since all dimensions d in a layer are independent, we can say that $\frac{Z_n}{s_n} | \boldsymbol{f}_{n-1} \sim \chi^2_m$, is distributed according to the Chi-squared distribution with m degrees of freedom.

One useful property of the Chi-squared distribution is that the moment-generating function of $\frac{Z_n}{s_n} | \boldsymbol{f}_{n-1}$ can be written in an analytical form, with $t \leq 1/2$,

$$M_{\frac{Z_n}{s_n} | \boldsymbol{f}_{n-1}}(t) = \mathbb{E}\left[\exp\left(t\frac{Z_n}{s_n}\right) | \boldsymbol{f}_{n-1}\right] = (1-2t)^{-m/2}. \tag{1}$$

We shall see that the expectation of the distance quantity Z_n is computed via a kernel function which, in most cases, involves exponentiations. Given that the input of this kernel is governed by a distribution, i.e., χ^2, the moment-generating function becomes convenient to obtain our desired expectations.

Figure 3 depicts our approach to extract a function $h(\cdot)$ which models the recurrence relation between $\mathbb{E}[Z_n]$ and $\mathbb{E}[Z_{n-1}]$. This is also the main theme of this paper.

4 Finding Recurrence Relations

This section presents the formalization of the recurrence relation of $\mathbb{E}[Z_n]$ for each kernel function. We start off with the squared exponential kernel function.

4.1 Squared Exponential Kernel Function

The squared exponential kernel (SE) is defined in the form of

$$\mathrm{SE}(\mathbf{x}, \mathbf{x}') = \sigma^2 \exp\left(-\|\mathbf{x} - \mathbf{x}'\|^2 / 2\ell^2\right). \tag{2}$$

Theorem 4.1 (DGP with SE)**.** *Given a triplet* (m, σ^2, ℓ^2)*,* $m \geq 1$ *such that the following sequence converges to* 0*:*

$$u_n = 2m\sigma^2\left(1 - (1 + u_{n-1}/m\ell^2)^{-m/2}\right), \tag{3}$$

Then, $\mathbb{P}(\|\boldsymbol{f}_n(\mathbf{x}) - \boldsymbol{f}_n(\mathbf{x}')\|_2 \xrightarrow[n\to\infty]{} 0$ *for all* $\mathbf{x}, \mathbf{x}' \in \mathcal{D}) = 1$.

Proof. Note that we do not directly have access to $\mathbb{E}[Z_n]$ but $\mathbb{E}[Z_n | \boldsymbol{f}_{n-1}]$ because of the Markov structure of the DGP construction. Getting $\mathbb{E}[Z_n]$ is done via $\mathbb{E}[Z_n|\boldsymbol{f}_{n-1}]$ where we use the law of total expectation $\mathbb{E}[Z_n] = \mathbb{E}_{\boldsymbol{f}_{n-1}}[\mathbb{E}[Z_n|\boldsymbol{f}_{n-1}]]$.

Now, we study the term $\mathbb{E}[Z_n|\boldsymbol{f}_{n-1}]$:

$$\begin{aligned}\mathbb{E}[Z_n|\boldsymbol{f}_{n-1}] =& \mathbb{E}[\sum_{d=1}^{m}(f_n^{(d)}(x) - f_n^{(d)}(x'))^2 | \boldsymbol{f}_{n-1}] \\ =& 2m\sigma^2 - 2mk_n(\mathbf{x}, \mathbf{x}').\end{aligned} \tag{4}$$

The second equality is followed by $\mathbb{E}[(f_n^{(d)}(\mathbf{x}))^2] = \mathbb{E}[(f_n^{(d)}(\mathbf{x}'))^2] = \sigma^2$ and $\mathbb{E}[f_n^{(d)}(\mathbf{x}) f_n^{(d)}(\mathbf{x}')] = k_n(\mathbf{x}, \mathbf{x}')$.

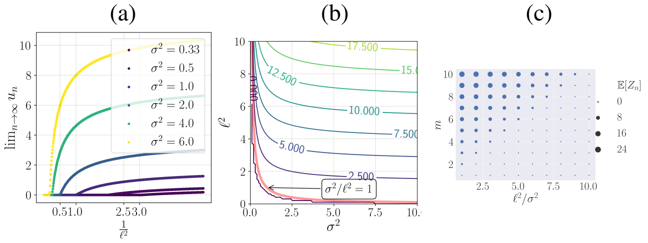

Figure 4: (a): Bifurcation plot of the recurrence relation of SE kernel for $m = 1$. (b): Contour plot of u_n at layer $n = 300$ and $m = 1$. The misalignment between the red line ($\sigma^2/\ell^2 = 1$) and the zero-level contour is due to numerical errors. (c): Increase $m > \sigma^2/\ell^2$ to avoid pathology.

Recall that we write $k_n(\mathbf{x}, \mathbf{x}) = k_n(\boldsymbol{f}_{n-1}(\mathbf{x}), \boldsymbol{f}_{n-1}(\mathbf{x}'))$. By the definition of SE kernel, we have

$$\mathbb{E}[Z_n|\boldsymbol{f}_{n-1}] = 2m\sigma^2\left(1 - \exp\left(-\frac{Z_{n-1}}{2\ell^2}\right)\right).$$

Applying the law of total expectation, we have

$$\mathbb{E}[Z_n] = 2m\sigma^2\left(1 - \mathbb{E}\left[\exp\left(-\frac{Z_{n-1}}{2\ell^2}\right)\right]\right).$$

Again, we can only compute $\mathbb{E}[\exp(-\frac{Z_{n-1}}{2\ell^2})] = \mathbb{E}_{\boldsymbol{f}_{n-2}}[\mathbb{E}[\exp(-\frac{Z_{n-1}}{2\ell^2})|\boldsymbol{f}_{n-2}]$. The expectation will be computed by the formula of the moment-generating function with respect to $\frac{Z_{n-1}}{s_{n-1}}|\boldsymbol{f}_{n-2}$ where $t = -\frac{s_{n-1}}{2\ell^2}$ in Equation (1). Choosing this value also satisfies the condition $t \leq 1/2$. Now, we have

$$\begin{aligned}\mathbb{E}[Z_n] &= 2m\sigma^2\left(1 - \mathbb{E}\left[\left(1 + s_{n-1}/\ell^2\right)^{-m/2}\right]\right) \\ &\leq 2m\sigma^2\left(1 - \left(1 + \mathbb{E}\left[s_{n-1}\right]/\ell^2\right)^{-m/2}\right).\end{aligned} \tag{5}$$

Here, Jensen's inequality is used as $(1+x)^{-a}$ is convex for any $x > 0$. By Equation (4), we have

$$\frac{\mathbb{E}[Z_{n-1}|\boldsymbol{f}_{n-2}]}{m} = 2\sigma^2 - 2k_{n-1}(\mathbf{x}, \mathbf{x}') = s_{n-1}.$$

Replacing s_{n-1} in Equation (5) and applying the law of total expectation for the case of Z_{n-1}, we obtain recurrence relation between layer $n-1$ and layer n is

$$\mathbb{E}[Z_n] \leq 2m\sigma^2\left(1 - \left(1 + \mathbb{E}[Z_{n-1}]/m\ell^2\right)^{-m/2}\right).$$

Using the Markov inequality, for any ϵ, we can bound $\mathbb{P}(Z_n \geq \epsilon) \leq \frac{\mathbb{E}[Z_n]}{\epsilon^2}$.

At this point, u_n defined in Equation (3) is considered as the upper bound of $\mathbb{E}[Z_n]$. We condition that $\{u_n\}$ converges to 0, then $\{\mathbb{E}[Z_n]\}$ converges to 0 as well. By the first Borel-Cantelli lemma, we have $\mathbb{P}(\limsup_{n\to\infty} Z_n \geq \epsilon) = 0$, which leads to the conclusion in the same manners as (Dunlop et al. 2018). □

Analyzing the recurrence Figure 4a illustrates the bifurcation plot of Equation (3) with $m = 1$. The non-zero contour region in Figure 4b tells us that σ^2/ℓ^2 should be smaller than 1 to escape the pathology. When $m > 1$, Figure 4c shows that if $m > \sigma^2/\ell^2$, u_n does not approach to 0, implying the condition to prevent the pathology. This result is consistent with Theorem 2.1 in (Dunlop et al. 2018).

Discussion Note that the relation between $\mathbb{E}[Z_n]$ and $\mathbb{E}[Z_{n-1}]$ presents a tighter bound than existing work (Dunlop et al. 2018). If we construct the recurrence relation based on (Dunlop et al. 2018), $\mathbb{E}[Z_n]$ is bounded by

$$\mathbb{E}[Z_n] \leq \frac{m\sigma^2}{\ell^2}\mathbb{E}[Z_{n-1}]. \tag{6}$$

One can show that $(1+x)^a \geq 1 - ax, a < 0, x > 0$, implying

$$2m\sigma^2(1-(1+\mathbb{E}[Z_{n-1}]/(m\ell^2))^{-m/2}) \leq m\sigma^2\mathbb{E}[Z_{n-1}]/\ell^2.$$

In fact, a numerical experiment shows that our bound of $\mathbb{E}[Z_n]$ is found to be close to the true $\mathbb{E}[Z_n]$ (Section 6.1). That is, we can see the trajectory of $\mathbb{E}[Z_n]$ for every layer of a given model of which the depth is not necessary to be infinitely many.

One can reinterpret the recurrence relation for each dimension d as

$$\mathbb{E}[Z_n^{(d)}] \leq 2\sigma^2\left(1 - \left(1 + \mathbb{E}[Z_{n-1}^{(d)}]/\ell^2\right)^{-m/2}\right),$$

where $\mathbb{E}[Z_n^{(d)}] = \frac{\mathbb{E}[Z_n]}{m}$ with $Z_n^{(d)} = \left(f_n^{(d)}(\mathbf{x}) - f_n^{(d)}(\mathbf{x}')\right)^2$.

A guideline to obtain a recurrence relation Given a specific kernel function, one may follow these steps to acquire the corresponding recurrence relation: (1) considering the form of kernel input where it may be distributed according to either the Chi-squared distribution or its variants (presented in the next sections); (2) checking whether there is a way to represent the kernel function under representations such that statistical properties of kernel inputs are known; (3) caring about the convexity of the function after choosing a proper setting (as we bound the expectation with Jensen's inequality in the proof of Theorem 4.1).

4.2 Cosine Kernel Function

The cosine kernel (COS) function takes inputs as the distance between two points instead of the squared distance like in the case of SE kernel. We will mainly work with $\sqrt{Z_n}$ in this subsection. The cosine kernel function $k(\mathbf{x}, \mathbf{x}') = \text{COS}(\mathbf{x}, \mathbf{x}')$ which is defined as

$$\text{COS}(\mathbf{x}, \mathbf{x}') = \sigma^2 \cos\left(\pi \left\|\mathbf{x} - \mathbf{x}'\right\|_2/p\right).$$

Starting with Equation (4) and using the definition of COS kernel, we have

$$\begin{aligned}\mathbb{E}[Z_n|\boldsymbol{f}_{n-1}] =& 2m\sigma^2 - 2m\sigma^2\cos(\pi\sqrt{Z_{n-1}}/p)\\ =& 2m\sigma^2 - m\sigma^2\exp(i\pi\sqrt{Z_{n-1}}/p)\\ &- m\sigma^2\exp(-i\pi\sqrt{Z_{n-1}}/p).\end{aligned}$$

Here, Euler's formula is used to represent $\cos(\cdot)$ and i is the imaginary unit ($i^2 = -1$). To obtain $\mathbb{E}[Z_n]$, we use the law of total expectation and compute the two following expectations: $\mathbb{E}\left[\exp(i\pi\sqrt{Z_{n-1}}/p)|\boldsymbol{f}_{n-2}\right]$ and $\mathbb{E}\left[\exp(-i\pi\sqrt{Z_{n-1}}/p)|\boldsymbol{f}_{n-2}\right]$. From $\frac{Z_n}{s_n}|\boldsymbol{f}_{n-1} \sim \chi^2_m$, we have $\sqrt{\frac{Z_n}{s_n}}|\boldsymbol{f}_{n-1} \sim \chi_m$, is distributed according to the Chi distribution. This observation follows the first step in the guideline. The characteristic function of the Chi distribution for random variable $\sqrt{\frac{Z_n}{s_n}}|\boldsymbol{f}_{n-1}$ is

$$\begin{aligned}&\varphi_{\sqrt{Z_n/s_n}|\boldsymbol{f}_{n-1}}(t) = \mathbb{E}\left[\exp\left(it\sqrt{Z_n/s_n}\right)\right]\\ &= {}_1F_1(\frac{m}{2}, \frac{1}{2}, \frac{-t^2}{2}) + it\sqrt{2}\frac{\Gamma((m+1)/2)}{\Gamma(m/2)}{}_1F_1(\frac{m+1}{2}, \frac{3}{2}, \frac{-t^2}{2}).\end{aligned}$$

where ${}_1F_1(a, b, z)$ is Kummer's confluent hypergeometric function (see Definition in Appendix A.2). This is considered as the second step in the guideline. Back to our process of finding the recurrence function, we consider the case $\sqrt{\frac{Z_{n-1}}{s_{n-1}}}|\boldsymbol{f}_{n-2} \sim \chi_m$. By choosing $t = \pm\frac{\pi\sqrt{s_{n-1}}}{p}$ for its characteristic function, we can obtain

$$\mathbb{E}[Z_n] = 2m\sigma^2\left(1 - {}_1F_1(\frac{m}{2}, \frac{1}{2}, -\frac{\pi^2}{2p^2}\mathbb{E}[Z_{n-1}|\boldsymbol{f}_{n-2}])\right).$$

This is because the imaginary parts of $\varphi(t = \frac{\pi\sqrt{s_{n-1}}}{p})$ and $\varphi(-\frac{\pi\sqrt{s_{n-1}}}{p})$ are canceled out.

As the third step in the guideline, we perform a sanity check about the convexity of ${}_1F_1$. Only with $m = 1$, ${}_1F_1(\frac{1}{2}, \frac{1}{2}, \frac{-t^2}{2}) = \exp(-\frac{t^2}{2})$ is convex. Our result in this case is restricted to $m = 1$. Now, we can state that the recurrence relation is

$$u_n = 2\sigma^2\left(1 - \exp(-\pi^2 u_{n-1}/2p^2)\right). \tag{7}$$

4.3 Spectral Mixture Kernel Function

In this paper, we consider the spectral mixture (SM) kernel (Wilson and Adams 2013) in one-dimensional case with one mixture:

$$\text{SM}(r) = \exp(-2\pi^2\sigma^2r^2)\cos(2\pi\mu r),$$

where $r = \|x - x'\|_2$, and $\sigma^2, \mu > 0$. We can rewrite this kernel function as $\frac{1}{2}w^2\{\exp(-v^2(r+iu)^2) + \exp(-v^2(r-iu)^2)\}$. Here we simplify the kernel by change in variables as $w^2 = \exp(-\frac{\mu^2}{2\sigma^2})$, $v^2 = 2\pi^2\sigma^2$, and $u = \frac{\mu}{2\pi\sigma^2}$.

With a similar approach, we compute the expectation of $\mathbb{E}[\exp(-v^2(\sqrt{Z_{n-1}} \pm iu)^2)]$. We can identify that $\frac{(\sqrt{Z_{n-1}}\pm iu)^2}{s_{n-1}}|f_{n-2} \sim \chi_1'^2(\lambda)$ is distributed according to a *non-central* Chi-squared distribution of which the moment-generating function is

$$M_{\chi_1'^2}(t;\lambda) = (1-2t)^{-1/2}\text{exp}(\lambda t/(1-2t)),$$

with the noncentrality parameter is $\lambda = -u^2/s_{n-1}$. By choosing an appropriate $t = -v^2 s_{n-1}$, we obtain the recurrence as

$$\mathbb{E}[Z_n] \leq 2(1 - w^2 M_{\chi'^2_1}(t = -v^2\mathbb{E}[Z_{n-1}]; \lambda = -u^2/\mathbb{E}[Z_{n-1}])).$$

Note that the convexity requirement is satisfied. This recurrence relation of SM kernel has one additional exponent term when comparing to that of SE. We provide a precise formula and an extension to the high-dimensional case in Appendix C.

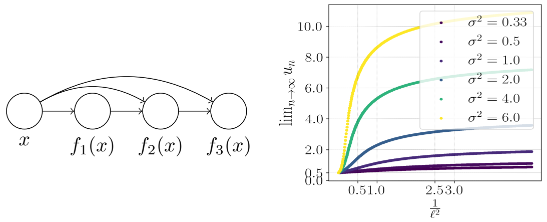

Figure 5: *Left*: Graphical model of input-connected construction suggested by (Neal 1995; Duvenaud et al. 2014). *Right*: The bifurcation plot of input-connected DGP.

4.4 Extension to Non-pathological Cases

We use our approach to analyze two cases including *nonzero-mean* DGPs and *input-connected* DGPs where there is no pathology occurring.

Nonzero-mean DGPs Let $f_n^{(d)}(\mathbf{x}) \sim \mathcal{GP}(\mu_n(\mathbf{x}), k_n(\mathbf{x}, \mathbf{x}'))$ with the mean function $\mu_n(\mathbf{x})$, the difference between two outputs, $(f_n^{(d)}(\mathbf{x}) - f_n^{(d)}(\mathbf{x}')) \sim \mathcal{N}(\nu_n, s_n)$ with $\nu_n = \mu_n(\mathbf{x}) - \mu_n(\mathbf{x}')$. This leads to $\frac{Z_n}{s_n}|\boldsymbol{f}_n \sim \chi'^2_m$, the *non-central* Chi-squared distribution with the non-central parameter $\lambda = m\nu_n^2$.

Since we already provide an analysis involving the non-central Chi-squared distribution with spectral mixture kernels, no pathology of nonzero-mean DGPs can be shown by our analysis (Section 4.3). That is, there is no pathology as $\lambda > 0$. When $\lambda = 0$, this case falls back to zero-mean or constant-mean. Mean functions greatly impact the recurrence relation because λ is inside an exponential function.

To the best of our knowledge, this is the first analytical explanation for the nonexistence of pathology in nonzero-mean DGPs. In practice, there is existing work choosing mean functions (Salimbeni and Deisenroth 2017). (Dunlop et al. 2018) briefly makes a connection between nonzero-mean DGPs and stochastic differential equations. However, there is no clear answer given for this case, yet.

Input-connected DGPs Previously, (Neal 1995; Duvenaud et al. 2014) suggest to make each layer connect to input. The corresponding dynamic system is

$$u_n = 2m\sigma^2(1 - (1 + u_{n-1}/m\ell^2)^{-m/2}) + c,$$

with c is computed from the kernel function taking input data $\mathbf{x}$. By seeing its bifurcation plot in Figure 5, we can reconfirm the solution from (Neal 1995; Duvenaud et al. 2014). That is, u_n converges to the value which is greater than zero, and avoids the pathology. However, the convergence rate of $\mathbb{E}[Z_n]$ stays the same.

5 Analysis of Recurrence Relations

This section explains the condition of hyperparameters that causes the pathology for each kernel function. Then we discuss the rate of convergence for the recurrence functions.

5.1 Identify the Pathology

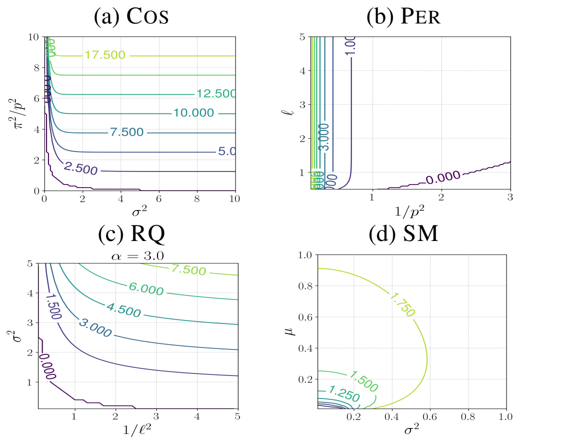

Figure 6: Contour plots of $\mathbb{E}[Z_n]$ at $n = 300$ with respect to four kernel functions.

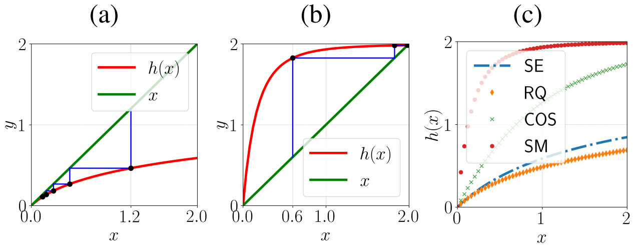

Figure 7: (a-b) Paths to fixed points for two cases: RQ and SM. Iterations of RQ start from $x = 1.2$ and converge to 0. Those of SM start from $x = 0.6$ and converge to a point near 1. (c) Plot of all recurrence functions $h(x)$. Note that x is not input data but plays the role of $\mathbb{E}[Z_n]$.

Table 1 provides the recurrence relations of two more kernel functions: the periodic (Per) kernel function and the rational quadratic (RQ) kernel function. The detailed derivation is in Appendix B and D.

Figure 6 shows contour plots based on our obtained recurrence relations. This will help us identify the pathology for each case. The corresponding bifurcation plots are in Appendix E.

Cos kernel Similar to SE, the condition to escape the pathology is $\pi^2\sigma^2/p^2 > 1$.

Per kernel If we increase ℓ, then we should decrease the periodic length p to prevent the pathology.

RQ kernel The behavior of this kernel resembles that of SE. We also observe that the change in the hyperparameter α does not affect the condition to avoid the pathology (Appendix E, Figure 15).

SM kernel Interestingly, this kernel does *not* suffer the pathology. If (σ^2, μ) goes to $(0, 0)$, $\mathbb{E}[Z_n]$ approaches to 0. However, $\mathbb{E}[Z_n]$ is never equal to 0 since both σ^2 and μ are positive.

5.2 Rate of Convergence

Recall that $h(\cdot)$ is the function modeling the recurrence relation between $\mathbb{E}[Z_n]$ and $\mathbb{E}[Z_{n-1}]$. According to Banach fixed-point theorem (Khamsi 2001), the rate of convergence is decided by the Lipchitz constant of $h(\cdot)$, $L = \sup h'(\cdot)$. The more curved the functions are, the faster the conver-

Rational quadratic (RQ)	$\sigma^2\left(1+\lVert\mathbf{x}-\mathbf{x}'\rVert^2/(2\alpha\ell^2)\right)^{-\alpha}$	$u_n = 2m(1-{}_2F_0(\alpha;\frac{m}{2};\frac{-u_{n-1}}{\alpha\ell^2}))$
Periodic (PER)	$\sigma^2\exp\left(-\frac{2\sin^2\left(\pi\lVert\mathbf{x}-\mathbf{x}'\rVert_2/p\right)}{\ell^2}\right)$	$u_n = \frac{2m\sigma^2}{\ell^2}\left(1-{}_1F_1(\frac{m}{2},\frac{1}{2},-\frac{2\pi^2}{p^2}u_{n-1})\right)$

Table 1: Kernel functions (middle column) and corresponding recurrence relations (right column)

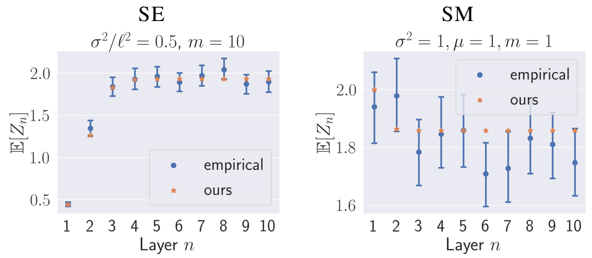

Figure 8: $\mathbb{E}[Z_n]$ computed from recurrence vs. empirical estimation of $\mathbb{E}[Z_n]$ for two kernel functions.

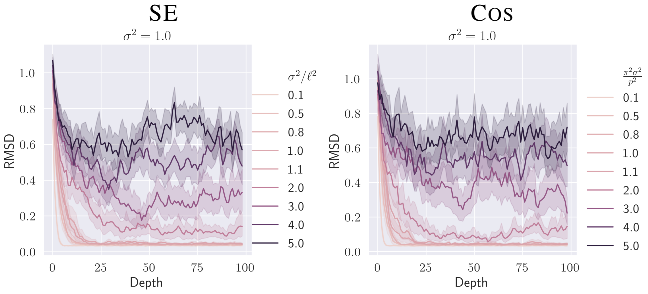

Figure 9: Trace of RMSDs. RMSDs converge to 0 when the pathology occurs.

gence rates are (see Figure 7a and 7b). Figure 7c compares the recurrence relation under the function $h(x)$. Specifically, for SE, the rate of convergence to a fixed point depends on the dimension parameter m. In general, SM has the fastest convergence rate among all. On the other hand, the class of RQ kernels has the slowest rate.

Understanding the convergence rate to a fixed point of recurrence relations can be helpful. For example, if a dynamic system corresponding to a DGP model quickly reaches its fixed point, it may be not necessary to have a very deep model. This can give an intuition for designing architectures in DGP given a kernel.

6 Experimental Results

This section verifies our theoretical claims empirically. Firstly, we investigate the correctness of recurrence relations. Then, we check the condition avoiding pathology. Furthermore, we provide case studies in real-world data sets. All kernels and models are developed based on GPyTorch library (Gardner et al. 2018).

6.1 Correctness of Recurrence Relations

We set up a DGP model with 10 layers with SE kernel. The inputs are $x_0 = 0$ and $x_1 = 1$. We will track the value $Z_n = \lVert \boldsymbol{f}_n(x_0) - \boldsymbol{f}_n(x_1)\rVert_2^2$ for $n = 1 \ldots 10$. Given a kernel $k(x, x')$, we can exactly compute the expectations $\mathbb{E}[Z_n]$. From the model, we collect 2000 samples for each layer n to obtain the empirical expectation of $\mathbb{E}[Z_n]$. Then, we would like to compare the true and empirical estimates. Figure 8 plots the comparisons for SE kernel and SM kernel. This numerical experiment supports the claim that our estimation $\mathbb{E}[Z_n]$ is tight and even close to the true estimation. On the other hand, $\mathbb{E}[Z_n]$ computed based on (Dunlop et al. 2018) in Equation (6) grows exponentially, and cannot fit in Figure 8. The additional plots with different settings of hyperparameter and m can be found in Appendix F (Figure 17 and 18)

6.2 Justifying the Conditions of Pathology

From N_{data} inputs, we generate the outputs of DGPs and measure the root mean squared distance (RMSD) among the outputs $\text{RMSD}(n) = \sqrt{\frac{1}{N_{\text{data}}(N_{\text{data}}-1)}\sum_{i\neq j}\lVert \boldsymbol{f}_n(\mathbf{x}_i) - \boldsymbol{f}_n(\mathbf{x}_j)\rVert_2^2}$. We record this quantity as we increase n. We replicate the procedure 30 times to aggregate the statistics of $\text{RMSD}(n)$. Here, we only consider the case $m = 1$.

SE kernel We set up models in one dimension with inputs of each model in range $(-5, 5)$ with $N_{data} = 100$. The kernel hyperparameter σ^2 is set to 1 while $1/\ell^2$ runs from 0.1 to 5. Figure 9a shows the trace of RMSD computed up to layer 100. When $\sigma^2/\ell^2 > 1$, the models start escaping the pathology.

COS kernel With a similar setup to that of SE, Figure 9b shows that when $\pi^2\sigma^2/p^2 > 1$, the models do not suffer the pathology.

PER kernel Since the PER kernel has three hyperparameters, σ^2, ℓ^2, p, we fix σ^2, and vary ℓ^2 and p. In this case, we collected the RMSDs at layer 100. We then compare the contour plot of these RMSDs with the values of the lower bound of $\mathbb{E}[Z_n]$ computed when n is large. We can find a similarity between Figure 10a and Figure 6b. The lower left of both plots has low values, identified as the region that causes the pathology.

RQ kernel Analogous to PER, only the RMSDs at layer 100 are gathered. We chose two different values of $\alpha = \{0.5, 3\}$, and varying values of σ^2 and ℓ^2. Figure 10c-d shows two contour plots of RMSDs for the two settings of α. Both of the two plots share the same area of which the contour level is close to 0.

SM kernel This kernel shows no sight of pathology (Figure 10b). We can find the similarity between this plot with the contour plot of $\mathbb{E}[Z_n]$ in Figure 6d.

6.3 Using Recurrence Relations in DGPs

Here, we use the recurrence relation as a tool to analyze DGP regression models. We learned the models where

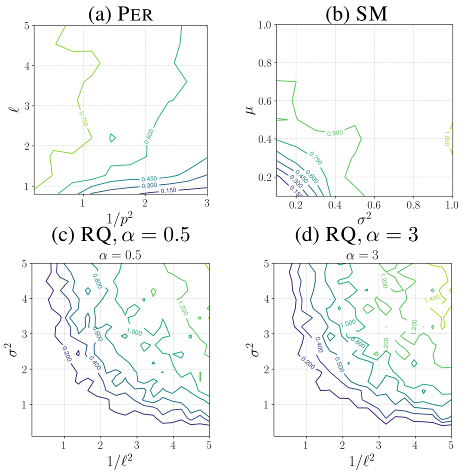
(a) Per (b) SM (c) RQ, $\alpha = 0.5$ (d) RQ, $\alpha = 3$

Figure 10: Contour plots of RMSDs at layer 100 for three kernels: Per, SM and RQ.

the number of layers, N, ranges from 2 to 6 and the number of units per layer, m, is from 2 to 9. We trained our models on Boston housing data set (Dheeru and Karra Taniskidou 2017) and diabetes data set (Efron et al. 2004). For each data set, we train our models with 90% of the data set and hold out the remaining for testing. The inference algorithm is based on (Salimbeni and Deisenroth 2017). We considered two settings: (1) standard zero-mean DGPs with SE kernel; (2) the SE kernel hyperparameters are constrained to avoid pathological regions with $\ell^2 \in (0, c_0 m \sigma^2]$, constraint coefficient $0 < c_0 < 1$.

Figure 11 plots the root mean squared errors (RMSEs) and quantity $\mathbb{E}[Z_n]/\sigma^2$ which describes changes between layers. For the case of standard zero-mean DGPs, we can observe that models can not learn effectively at deeper layers and there are drops in terms of $\mathbb{E}[Z_n]/\sigma^2$ at the last layer. In the case of constraining hyperparameters, we see fewer drops and the results are improved when comparing to non-constrained cases. It seems that the drop pattern of $\mathbb{E}[Z_n]/\sigma^2$ correlates to model performances. We provide detailed figures and an additional result on the diabetes data set with a similar observation in Appendix F.

6.4 High-dimensional Data with Zero-mean DGPs

We test on MNIST data set (LeCun and Cortes 2010) with the two models like previous experiments. The number of units per layer, m, is chosen as $m = 30$. We consider the number of layers, $N = 2, 3, 4$.

The standard zero-mean DGP without any regularization fails to learn from data with accuracy ≈ 10%. This means that the output of this model is just a flat function, making this 10-class classifier have such an accuracy. On the other hand, the constrained zero-mean DGP can alleviate the model performance with accuracy at best 91.21%. Figure 12c provides the results with different settings of c_0.

To have a better understanding of the above models, we visualize the loss landscape (Li et al. 2018) of the two

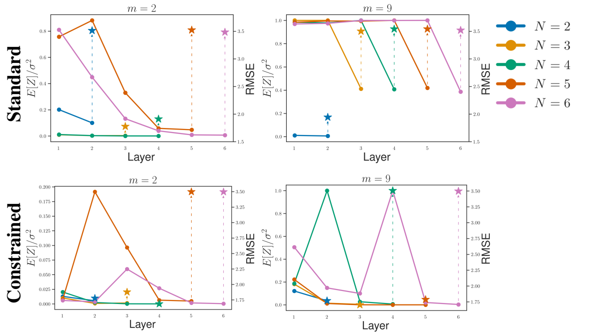

Figure 11: Dual-axis plot of the trajectory $\mathbb{E}[Z_n]/\sigma^2$ with n running from 1 to N and RMSE. Solid lines indicate the trajectories of $\mathbb{E}[Z_n]/\sigma^2$ projected on the left y-axis. Star markers ($\star$) indicate RMSEs projected on the right y-axis. Dashed lines connect the $\mathbb{E}[Z_n]/\sigma^2$ and RMSE of the same N. Here, the constrain coefficient $c_0 = 0.2$.

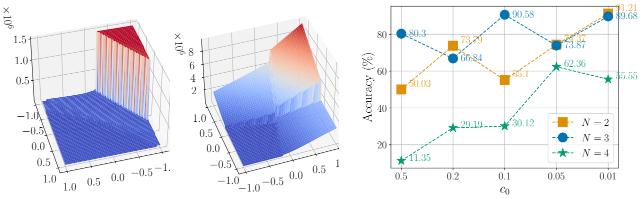
(a) Standard (b) Constrained (c) Accuracy

Figure 12: (a-b) Loss landscape of two models. (c) Classification accuracy with respect to the number of layers, N, and constrain coefficients, c_0.

cases in Figure 12. The standard zero-mean DGP easily falls into unsafe pathological hyperparameters during optimization and cannot escape the unsafe state (see Figure 12a). In contrary, the loss landscape of constrained DGPs (Figure 12b) shows an improved loss surface. However, we note that it still has a flat region where the optimization cannot be improved.

Our result is not as good as the accuracy (98.06%) of nonzero-mean DGPs reported in (Salimbeni and Deisenroth 2017). However, we emphasize that the main contribution of our work is not to demonstrate the classifier performance but to show the importance of incorporating the theoretical insights into practice. This shows that learning zero-mean DGPs is potentially possible.

7 Conclusion

We have presented a new analysis of the existing issue of DGP for a number of kernel functions via analyzing the chaotic properties of corresponding nonlinear systems which models the state of magnitudes between layers. We believe that such analysis can be beneficial in kernel structure discovery tasks (Duvenaud et al. 2013; Ghahramani 2015; Hwang, Tong, and Choi 2016; Tong and Choi 2019) for DGP. Our analysis not only provides a better understanding of the rate of convergence to fixed points but also considers a number of kernel types. Finally, our findings are verified by numerical experiments.

Acknowledgements

This work was supported by Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (No.2019-0-00075, Artificial Intelligence Graduate School Program (KAIST)).

References

Bui, T. D.; Hernández-Lobato, J. M.; Hernández-Lobato, D.; Li, Y.; and Turner, R. E. 2016. Deep Gaussian Processes for Regression Using Approximate Expectation Propagation. In *ICML*, 1472–1481.

Cutajar, K.; Bonilla, E. V.; Michiardi, P.; and Filippone, M. 2017. Random Feature Expansions for Deep Gaussian Processes. In *ICML*, volume 70, 884–893.

Dai, Z.; Damianou, A.; Gonzalez, J.; and Lawrence, N. D. 2016. Variationally Auto-Encoded Deep Gaussian Processes. In *ICLR*.

Damianou, A.; and Lawrence, N. 2013. Deep Gaussian Processes. In *AISTATS*, volume 31, 207–215.

Dheeru, D.; and Karra Taniskidou, E. 2017. UCI Machine Learning Repository. URL http://archive.ics.uci.edu/ml.

Dunlop, M. M.; Girolami, M. A.; Stuart, A. M.; and Teckentrup, A. L. 2018. How Deep Are Deep Gaussian Processes? *Journal of Machine Learning Research* 19(54): 1–46.

Duvenaud, D.; Lloyd, J. R.; Grosse, R.; Tenenbaum, J. B.; and Ghahramani, Z. 2013. Structure Discovery in Nonparametric Regression through Compositional Kernel Search. In *ICML*, 1166–1174.

Duvenaud, D. K.; Rippel, O.; Adams, R. P.; and Ghahramani, Z. 2014. Avoiding pathologies in very deep networks. In *AISTATS*, 202–210.

Efron, B.; Hastie, T.; Johnstone, I.; and Tibshirani, R. 2004. Least angle regression. *Annals of Statistics* 32: 407–499.

Gardner, J. R.; Pleiss, G.; Bindel, D.; Weinberger, K. Q.; and Wilson, A. G. 2018. GPyTorch: Blackbox Matrix-Matrix Gaussian Process Inference with GPU Acceleration. In *NeurIPS*.

Ghahramani, Z. 2015. Probabilistic machine learning and artificial intelligence. *Nature* 521(7553): 452–459.

Havasi, M.; Hernández-Lobato, J. M.; and Murillo-Fuentes, J. J. 2018. Inference in Deep Gaussian Processes using Stochastic Gradient Hamiltonian Monte Carlo. In *NeurIPS*, 7517–7527.

Hensman, J.; and Lawrence, N. D. 2014. Nested Variational Compression in Deep Gaussian Processes. *arXiv preprint arXiv:1412.1370* .

Hwang, Y.; Tong, A.; and Choi, J. 2016. Automatic Construction of Nonparametric Relational Regression Models for Multiple Time Series. In *ICML*, 3030–3039.

Khamsi, M. 2001. An Introduction to Metric Spaces and Fixed Point Theory .

LeCun, Y.; and Cortes, C. 2010. MNIST handwritten digit database URL http://yann.lecun.com/exdb/mnist/.

Li, H.; Xu, Z.; Taylor, G.; Studer, C.; and Goldstein, T. 2018. Visualizing the Loss Landscape of Neural Nets. In *NeurIPS*.

Lu, C.; Yang, S. C.; Hao, X.; and Shafto, P. 2020. Interpretable Deep Gaussian Processes with Moments. In Chiappa, S.; and Calandra, R., eds., *AISTATS*.

May, R. M. 1976. Simple mathematical models with very complicated dynamics. *Nature* 261(5560): 459–467.

Neal, R. M. 1995. Bayesian Learning for Neural Networks. *PhD thesis* .

Poole, B.; Lahiri, S.; Raghu, M.; Sohl-Dickstein, J.; and Ganguli, S. 2016. Exponential expressivity in deep neural networks through transient chaos. In *NeurIPS*, 3360–3368.

Rasmussen, C. E.; and Williams, C. K. I. 2005. *Gaussian Processes for Machine Learning (Adaptive Computation and Machine Learning)*. The MIT Press.

Salimbeni, H.; and Deisenroth, M. 2017. Doubly stochastic variational inference for deep gaussian processes. In *NeurIPS*.

Salimbeni, H.; Dutordoir, V.; Hensman, J.; and Deisenroth, M. 2019. Deep Gaussian Processes with Importance-Weighted Variational Inference. In *ICML*, 5589–5598.

Schoenholz, S. S.; Gilmer, J.; Ganguli, S.; and Sohl-Dickstein, J. 2017. Deep Information Propagation. In *ICLR*.

Tong, A.; and Choi, J. 2019. Discovering Latent Covariance Structures for Multiple Time Series. In *ICML*, 6285–6294.

Ustyuzhaninov, I.; Kazlauskaite, I.; Kaiser, M.; Bodin, E.; Campbell, N. D. F.; and Ek, C. H. 2020. Compositional uncertainty in deep Gaussian processes. In *UAI*, 206.

Wilson, A. G.; and Adams, R. P. 2013. Gaussian Process Kernels for Pattern Discovery and Extrapolation. In *ICML*, 1067–1075.