3rd Symposium on Advances in Approximate Bayesian Inference, 2020 1–34

The Gaussian Neural Process

Wessel P. Bruinsma *University of Cambridge and Invenia Labs* wpb23@cam.ac.uk
James Requeima *University of Cambridge and Invenia Labs* jrr41@cam.ac.uk
Andrew Y. K. Foong *University of Cambridge* ykf21@cam.ac.uk
Jonathan Gordon *University of Cambridge* jg801@cam.ac.uk
Richard E. Turner *University of Cambridge* ret26@cam.ac.uk

Abstract

Neural Processes (NPs; Garnelo et al., 2018a,b) are a rich class of models for meta-learning that map data sets directly to predictive stochastic processes. We provide a rigorous analysis of the standard maximum-likelihood objective used to train conditional NPs. Moreover, we propose a new member to the Neural Process family called the Gaussian Neural Process (GNP), which models predictive correlations, incorporates translation equivariance, provides universal approximation guarantees, and demonstrates encouraging performance.
Keywords: Meta-Learning, Neural Processes, Gaussian Processes

1. Introduction

Neural Processes (NPs; Garnelo et al., 2018a,b) use neural networks to directly parameterise and learn a map from observed data to posterior predictive distributions of a stochastic process. In this work, we provide two contributions to the NP framework.

Our first contribution is a rigorous analysis of the standard maximum-likelihood (ML) objective used to train conditional NP models. In particular, we relate the objective to the KL divergence between stochastic processes (d. G. Matthews et al., 2016), which we call a *functional* KL. For a ground truth $\mathbb{P}$ and approximating process $\mathbb{Q}$, learning procedures that minimise a functional KL have previously been investigated (Sun et al., 2018; Shi et al., 2019; Ma et al., 2018), but these works leave important questions about finiteness of the objective and existence/uniqueness of its minimiser unanswered. In this work, we consider the objective $\mathrm{KL}(\mathbb{P}, \mathbb{Q})$. In a well-defined and rigorous setup, we demonstrate that the ML objective can be interpreted as a well-behaved relaxation of this functional objective.

Our second contribution addresses the inability of conditional NPs (CNPs; Garnelo et al., 2018a) to model correlations and produce coherent samples. Several authors propose to overcome this limitation by introducing a latent variable (Garnelo et al., 2018b; Kim et al., 2019; Foong et al., 2020). Unfortunately, this renders the likelihood intractable, complicating learning and evaluation. Building on the ConvCNP (Gordon et al., 2020), we introduce the *Gaussian* NP (GNP), a novel member of the NP family that incorporates *translation equivariance* and models the predictive distributions directly with Gaussian processes (GPs; Rasmussen and Williams, 2006). The GNP allows for correlations in the predictive distribution whilst admitting a closed-form likelihood. Moreover, like the ConvCNP, the GNP provides universal approximation guarantees, which we showcase by providing empirical evidence that the GNP can recover the *prediction map* of a ground-truth Gaussian process in terms of likelihood and prior covariance function.

© W. P. Bruinsma, J. Requeima, A. Y. K. Foong, J. Gordon & R. E. Turner.

2. A Practical Objective for Meta-Learning with Gaussian Processes

A detailed description of the notation and terminology used in this section can be found in App A. The statements and proofs of all theorems are deferred to Apps B to D.

Problem setup: Let f be a ground-truth stochastic process. In the meta-learning setup, we aim to make multiple predictions for f based on a collection of observed data sets $(D_i)_{i=1}^N$ drawn from f. With access to f, these predictions are given by the posteriors over f given $(D_i)_{i=1}^N$. We can view prediction as a map from observed data sets $\mathcal{D}$ to posteriors over f. This map is called the *posterior prediction map* $\pi_f \colon \mathcal{D} \to \mathcal{P}$ (Foong et al., 2020). Our goal is to learn a *Gaussian* approximation $\tilde{\pi} \colon \mathcal{D} \to \mathcal{P}_{\mathrm{G}}$ of π_f (Def C.4) that approximates the posteriors over f with Gaussian processes. Note that a Gaussian approximation of the posterior prediction map is distinctly different from learning a Gaussian approximation of the prior f: the only requirement on $\tilde{\pi}$ is that $\tilde{\pi}(D)$ is a Gaussian process for all $D \in \mathcal{D}$; in particular, these GPs are not constrained to be posteriors obtained from a fixed prior, which means that learning $\tilde{\pi}$ enjoys significantly more flexibility. In fact, this setup is strictly more flexible than the originally proposed CNP (Garnelo et al., 2018a), as the CNP can be viewed as a map $\mathcal{D} \to \mathcal{P}_{\mathrm{G}}$ that does not model correlations.

Functional objective: We directly define our approximation $\tilde{\pi}$ of π_f: for every $D \in \mathcal{D}$, approximate $\pi_f(D)$ with a Gaussian process μ:

$$\tilde{\pi}(D) = \arg\min_{\mu \in \mathcal{P}_{\mathrm{G}}} \mathrm{KL}(\pi_f(D), \mu). \tag{1}$$

Under reasonable regularity conditions and the assumption that there exists some non-degenerate Gaussian process $\mu_{\mathrm{G}} \in \mathcal{P}_{\mathrm{G}}$ such that $\mathrm{KL}(\pi_f(D), \mu_{\mathrm{G}}) < \infty$, this minimiser exists and is unique (Cor B.1). However, such a Gaussian process μ_{G} may not exist. Moreover, even if the minimiser $\tilde{\pi}(D)$ exists and is unique, meaning that (1) is finite at $\tilde{\pi}(D)$, there may not exist a ball of approximations around $\tilde{\pi}(D)$ for which the objective (1) is finite; in that case, the minimiser $\tilde{\pi}(D)$ cannot be approximated by minimising (1). For example, suppose that $\pi_f(D) = \mathcal{GP}(0, k)$ where $k(t, t') = \exp(-\frac{1}{2}(t - t')^2)$. Set $\mu_{\sigma^2} = \mathcal{GP}(0, \sigma^2 k)$. Then a quick computation shows that $\mathrm{KL}(\pi_f(D), \mu_{\sigma^2}) = \infty$ for all $\sigma^2 \neq 1$. Hence, we cannot recover the true variance $\sigma^2 = 1$ by initialising μ_{σ^2} with some reasonable $\sigma^2 > 0$ and minimising $\mathrm{KL}(\pi_f(D), \mu_{\sigma^2})$, because the objective is infinite for all but the true value of σ^2.

Relaxation: To work around the potential absence of a minimiser, we take a pragmatic stance and instead simply approximate the *finite-dimensional distributions* (f.d.d.s):

$$\tilde{\pi}^{\mathbf{x}}(D) = \arg\min_{\mu^{\mathbf{x}} \in \mathcal{P}_{\mathrm{G}}^{|\mathbf{x}|}} \mathrm{KL}(P_{\mathbf{x}} \pi_f(D), \mu^{\mathbf{x}}) \quad \text{for all finite index sets } \mathbf{x}, \tag{2}$$

where $P_{\mathbf{x}} f = (f(x_1), \ldots, f(x_{|\mathbf{x}|}))$ for $f \in \mathcal{Y}^{\mathcal{X}}$ is the projection onto the index set $\mathbf{x}$. Under reasonable regularity conditions and the assumption that, for all finite index sets $\mathbf{x}$, there exists an appropriate $|\mathbf{x}|$-dimensional Gaussian distribution $\mu_{\mathrm{G}}^{\mathbf{x}}$ such that $\mathrm{KL}(P_{\mathbf{x}} \pi_f(D), \mu_{\mathrm{G}}^{\mathbf{x}}) < \infty$, these minimisers exist and are unique (Prop B.1). This condition is much milder than that for (1): it is satisfied for any appropriate $\mu_{\mathrm{G}}^{\mathbf{x}}$ if the differential entropy of $P_{\mathbf{x}} \pi_f(D)$ is finite. Crucially, it turns out that (2) gives rise to a consistent collection of f.d.d.s (Prop B.1) and therefore uniquely defines an approximating process $\tilde{\pi}(D)$ satisfying $P_{\mathbf{x}} \tilde{\pi}(D) = \tilde{\pi}^{\mathbf{x}}(D)$ for all finite index sets $\mathbf{x}$. Moreover, if a solution to (1) exists, then it will be equal to $\tilde{\pi}(D)$ (Prop B.1 and Cor B.1). Therefore, (2) defines a relaxation of (1) that can be used in many

cases where a solution to (1) does not exist. The solution to (1) and (2), if it exists, is given by the *moment-matched Gaussian process*: the Gaussian process obtained by taking the mean function and covariance function of $\pi_f(D)$; see also Ma et al. (2018).

Approximable objective: The workaround (2) solves the problem of *existence*. However, there is still the problem of *approximability*: (1) cannot always be minimised to approximate the minimiser, if one exists. We therefore define another objective, one that is always finite and consequently can always be minimised to approximate the solution to (2). This objective is obtained by averaging (2) over index sets of a fixed size:

$$\tilde{\pi}(D) = \arg\min_{\mu \in \overline{\mathcal{P}}_{\mathrm{G}}} \mathbb{E}_{p(\mathbf{x})}[\mathrm{KL}(P_{\mathbf{x}}\pi_f(D), P_{\mathbf{x}}\mu)] \tag{3}$$

where $p(\mathbf{x})$ is a Borel distribution with full support over all index sets of a fixed size $n \geq 2$. (See Def C.1 for the definition of $\overline{\mathcal{P}}_{\mathrm{G}}$.) This objective is well defined (Prop D.1). If (i) the mean and covariances functions of $\pi_f(D)$ and μ exist and are uniformly bounded by $M > 0$ and (ii) the processes $\pi_f(D)$ and μ are noisy (Def C.1), then Prop D.3 shows that the objective is finite and consequently suitable for optimisation; Prop D.4 shows that the minimisers of (2) and (3) are equal. A useful feature of (3) is that it averages over index sets of a fixed size $n \geq 2$, unlike previous objectives, *e.g.* Prop 1 by Foong et al. (2020), which requires an average over index sets of all sizes.

Practical objective: We further average (3) over an appropriate selection of data sets, which formulates a single objective that captures the total approximation error of $\tilde{\pi}$:

$$\tilde{\pi} = \arg\min_{\pi \in \overline{\mathcal{M}}_{\mathrm{G}}^{\text{f.d.d.}}} \mathbb{E}_{p(D)p(\mathbf{x})}[\mathrm{KL}(P_{\mathbf{x}}\pi_f(D), P_{\mathbf{x}}\pi(D))] \tag{4}$$

where $p(D)$ is a Borel distribution with full support over a collection of data sets $\tilde{\mathcal{D}} \subseteq \mathcal{D}$ that is open and bounded (Def C.5). (See Def C.6 for the definition of $\overline{\mathcal{M}}_{\mathrm{G}}^{\text{f.d.d.}}$.) This objective is also well defined (Prop D.2). Under conditions similar to the conditions for (3), (4) is finite and the minimisers of (2) and (4) are equal (Props D.5 and D.6). We therefore propose to learn $\tilde{\pi}$ though minimising (4). In practice, we optimise a Monte Carlo approximation of (4). Let $(D_i)_{i=1}^N \subseteq \mathcal{D}$ be a collection of data sets, all sampled from f and split up $D_i = (D_i^{(\mathrm{c})}, D_i^{(\mathrm{t})})$ into *context sets* $D_i^{(\mathrm{c})}$ and *target sets* $D_i^{(\mathrm{t})} = (\mathbf{x}_i^{(\mathrm{t})}, \mathbf{y}_i^{(\mathrm{t})})$ (Vinyals et al., 2016; Ravi and Larochelle, 2017). We then maximise

$$\tilde{\pi} \approx \arg\max_{\pi \in \overline{\mathcal{M}}_{\mathrm{G}}^{\text{f.d.d.}}} \frac{1}{N}\sum_{i=1}^N \log \mathcal{N}(\mathbf{y}_i^{(\mathrm{t})} \,|\, \mathbf{m}_i, \mathbf{K}_i) \quad \text{with} \quad P_{\mathbf{x}_i^{(\mathrm{t})}}\pi(D_i^{(\mathrm{c})}) = \mathcal{N}(\mathbf{m}_i, \mathbf{K}_i), \tag{5}$$

where we ignore irrelevant additive constants that do not depend on π. This objective is exactly the standard maximum likelihood objective that is used to train conditional NP models (Garnelo et al., 2018a; Gordon et al., 2020). Analysis of the minimising procedure is difficult and depends on the details of the particular algorithm. What we can say, however, is that a minimising sequence either diverges or converges to the right limit; and, under certain conditions, a minimising sequence always has a convergent subsequence (Prop D.7).

3. The Gaussian Neural Process

Having defined a suitable objective, to learn the approximation $\tilde{\pi}$ in practice, we proceed to generally parametrise $\tilde{\pi}$. In this paper, we confine ourselves to *stationary* ground truths f.

Translation equivariance: Foong et al. (2020) show that stationarity of f is equivalent to *translation equivariance* (TE) of the posterior prediction map π_f of f: for

all $D \in \mathcal{D}$, $\mathsf{T}_\tau \pi_f(D) = \pi_f(D + \tau)$ for all $\tau \in \mathcal{X}$ where $\mathsf{T}_\tau f = f(\bullet - \tau)$ is the *shifting operator*, $\mathsf{T}_\tau \pi_f(D)$ is the measure $\pi_f(D)$ pushed through T_τ, and $D + \tau = (\mathbf{x}, \mathbf{y}) + \tau = ((x_1 + \tau, \ldots, x_{|\mathbf{x}|} + \tau), \mathbf{y})$. If π_f is TE, then it is reasonable to restrict our approximation $\tilde{\pi}$ to also be TE. Incorporating translation equivariance directly into the model has been shown to yield large improvements in generalisation capability, parameter efficiency, and predictive performance (Gordon et al., 2020; Foong et al., 2020). Denote $\tilde{\pi}(D) = \mathcal{GP}(m(D), k(D))$ where $m \colon \mathcal{D} \to C(\mathcal{X}, \mathcal{Y})$ is the TE *mean map* of our approximation $\tilde{\pi}$ and $k \colon \mathcal{D} \to C^{\text{p.s.d.}}(\mathcal{X}^2, \mathcal{Y})$ is the TE *kernel map*; see App E for more details.

Universal parametrisation of mean and kernel: For the mean map m, we use a ConvDeepSet architecture (used in the ConvCNP, Gordon et al., 2020), which can approximate any translation-equivariant map from data sets to continuous mean functions (Thm 1 by Gordon et al., 2020). Unfortunately, as we explain in App E.1, the ConvDeepSet architecture is not directly applicable to the kernel map k. In App E, we modify the ConvDeepSet architecture to make it suitable for the kernel. This architecture has universal approximation guarantees similar to the ConvDeepSet (Thm E.1) and thus completes a general approximate parametrisation of $\tilde{\pi}$. Intuitively, the architecture works as follows. A covariance function is a function $\mathcal{X} \times \mathcal{X} \to \mathbb{R}$ and can therefore be interpreted as an *image* (*e.g.*, imagine that $\mathcal{X} = \{1, \ldots, n\}$). Whereas the ConvCNP generates the mean by embedding the data in a 1D array and passing it through 1D convolutions, the architecture for the kernel similarly embeds the data in a 2D *image* and passes it through 2D convolutions. Let $D^{(\text{c})}$ be a context set and $\mathbf{x}^{(\text{t})}$ inputs of a target set. Then the covariance matrix $\mathbf{K}^{(\text{t})}$ at the target points $\mathbf{x}^{(\text{t})}$ is generated as follows:

$$❶\ \mathbf{H} = \text{enc}(D^{(\text{c})}, \mathbf{Z}), \quad ❷\ \mathbf{K} = \Pi^{\text{p.s.d.}}\, \text{CNN}(\mathbf{H}), \quad ❸\ \mathbf{K}^{(\text{t})} = \text{dec}(\mathbf{K}, \mathbf{x}^{(\text{t})}) : \tag{6}$$

❶ $\mathbf{H} = \text{enc}(D^{(\text{c})}, \mathbf{Z})$ maps the target set $D^{(\text{c})}$ to an *encoding* $\mathbf{H} \in \mathbb{R}^{M \times M \times 3}$ at a prespecified grid $\mathbf{Z} \in \mathbb{R}^{M \times M}$ for some $M \in \mathbb{N}$ (*c.f.* the discretisation in the ConvCNP (Gordon et al., 2020)), comprising a *data channel* $\mathbf{H}_{::1}$ (*c.f.* the data channel in the ConvDeepSet), *density channel* $\mathbf{H}_{::2}$ (*c.f.* the density channel in the ConvDeepSet), and *source channel* $\mathbf{H}_{::3} = \mathbf{I}$ (not present in the ConvDeepSet; see App E.2);

❷ $\mathbf{K} = \Pi^{\text{p.s.d.}}\, \text{CNN}(\mathbf{H})$ passes the encoding $\mathbf{H}$ through a CNN, producing an $M \times M$ matrix, and projects this matrix with $\Pi^{\text{p.s.d.}}$ onto the nearest positive semi-definite (PSD) matrix with respect to the Frobenius norm (Higham, 1988); and

❸ $\mathbf{K}^{(\text{t})} = \text{dec}(\mathbf{K}, \mathbf{x}^{(\text{t})})$ finally interpolates the obtained PSD matrix $\mathbf{K}$ to the desired covariances $\mathbf{K}^{(\text{t})}$ for the target inputs $\mathbf{x}^{(\text{t})}$.

The architecture and the precise definitions of enc and dec are described in more detail in App E.2. In our experiments, contrary to the description above and App E.2, we substitute ❷ with the simpler operation $\mathbf{K} = \text{CNN}(\mathbf{H})\, \text{CNN}(\mathbf{H})^\mathsf{T}$, which also guarantees positive semi-definiteness. It is unclear whether this substitution limits the expressivity of the resulting architecture or interferes with translation equivariance. We leave an investigation of the implementation of ❷ with $\Pi^{\text{p.s.d.}}$ for future work.

Source channel: A novel aspect of the architecture (6) is the *source channel* $\mathbf{H}_{::3} = \mathbf{I}$, which is simply the identity matrix and not present in the ConvDeepSet architecture.

	EQ	Matérn–$\frac{5}{2}$	Weakly Per.	Sawtooth	Mixture
GP (truth)	$0.70 \pm 4.8\text{E}{-3}$	$0.31 \pm 4.8\text{E}{-3}$	$-0.32 \pm 4.3\text{E}{-3}$	n/a	n/a
GNP	$\mathbf{0.70} \pm 5.0\text{E}{-3}$	$\mathbf{0.30} \pm 5.0\text{E}{-3}$	$\mathbf{-0.47} \pm 5.0\text{E}{-3}$	0.42 ± 0.01	$\mathbf{0.10} \pm 0.02$
ConvNP	-0.46 ± 0.01	$-0.67 \pm 9.0\text{E}{-3}$	$-1.02 \pm 6.0\text{E}{-3}$	$\mathbf{1.20} \pm 7.0\text{E}{-3}$	-0.50 ± 0.02
ANP	-0.61 ± 0.01	-0.75 ± 0.01	$-1.19 \pm 5.0\text{E}{-3}$	$0.34 \pm 7.0\text{E}{-3}$	-0.69 ± 0.02
GP (truth, no corr.)	-0.81 ± 0.01	-0.93 ± 0.01	$-1.18 \pm 7.0\text{E}{-3}$	n/a	n/a
ConvCNP	-0.80 ± 0.01	-0.95 ± 0.01	$-1.20 \pm 7.0\text{E}{-3}$	0.55 ± 0.02	-0.93 ± 0.02

Table 1: Likelihoods for the 1D regression experiments for interpolation inside the training range. Highlights best performance. Only the ConvCNP does not model correlations. The errors are 95%-confidence intervals. See App F.

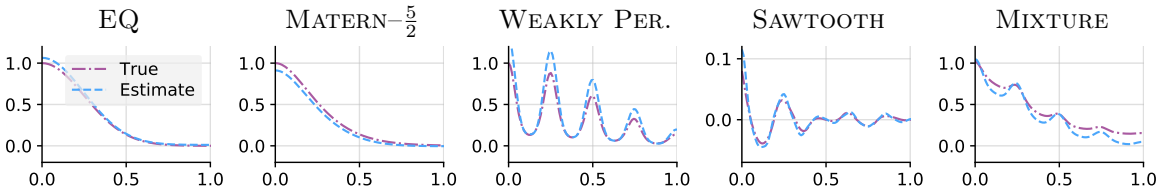

Figure 1: Learned and true stationary prior covariance functions in the 1D experiments. The prior covariance function can be extracted from the model by taking $D^{(\mathrm{c})} = \varnothing$.

Intuitively, the source channel allows the architecture to "start out" with a stationary prior with covariance $\mathbf{I}$, corresponding to white noise, and then pass it through a CNN to modulate this prior to introduce correlations inferred from the context set. *C.f.*, any Gaussian process can be sampled from by first sampling white noise and then convolving with an appropriate filter; the kernel architecture is a nonlinear generalisation of this procedure.

The Gaussian Neural Process: The ConvDeepSet for the mean mapping m and the above architecture for the kernel mapping k form a model that we call the Gaussian Neural Process (GNP). The GNP depends on some parameters θ, *e.g.* weights and biases for the CNNs. To train these parameters θ, we maximise (5). See App E.3 for more details.

4. Experiments

We evaluate the Gaussian Neural Process on synthetic 1D regression experiments. We follow the experimental setup from Foong et al. (2020); see App F for more details. In line with the theoretical analysis from Sec 2, but unlike Foong et al. (2020), we contaminate all data samples with $\mathcal{N}(0, 0.05^2)$-noise. Code that implements the GNP and reproduces all experiments can be found at `https//github.com/wesselb/NeuralProcesses.jl`.

Tab 1 shows the performance of the ground-truth GP (where applicable), the ground-truth GP without correlations, the GNP, the ConvCNP (Gordon et al., 2020), the ConvNP (Foong et al., 2020), and the ANP (Kim et al., 2019) in an interpolation setup; Fig 2 shows samples of the learned models. The GNP significantly outperforms all other models on all tasks except Sawtooth. On EQ, and Matérn–$\frac{5}{2}$, the GNP even achieves parity with the ground-truth GP. Moreover, the GNP is the best performing model on Mixture, which is highly non-Gaussian, demonstrating that the GNP can successfully approximate non-Gaussian processes. Fig 1 shows the stationary prior covariance functions learned by

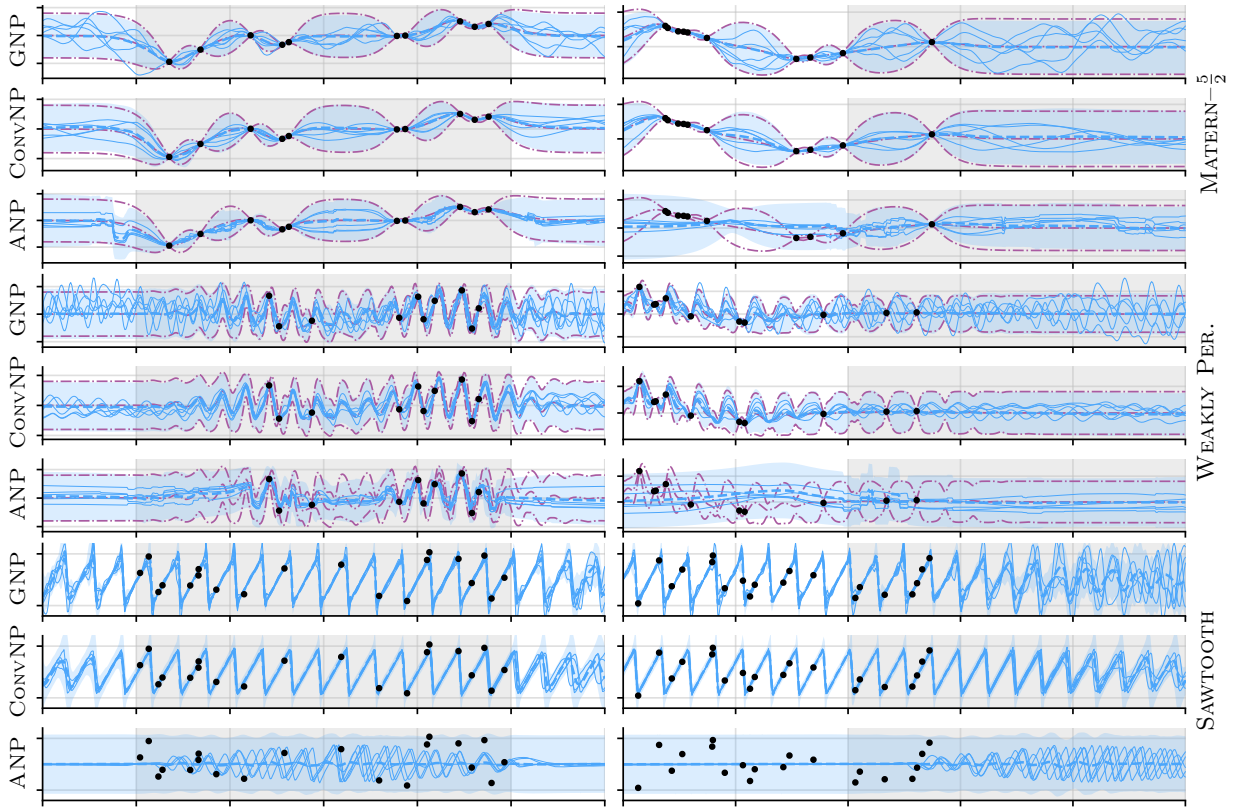

Figure 2: Sample predictions of models trained in the 1D experiments. Grey regions indicate where the models were trained; dashed purple lines show optimal predictions by the ground truth.

the GNP, which are close to the ground-truth on the GP tasks; this, together with the likelihood numbers, empirically validates the ability of the GNP to recover the prediction map of a ground-truth GP. Note that the learned covariance functions also match the truth for the non-Gaussian tasks Sawtooth and Mixture. For Sawtooth, the likelihood of the GNP is worse than the ConvCNP and only improves over the ANP, which demonstrates that, on certain non-Gaussian tasks, non-Gaussian approximations, like the ConvNP, can offer substantially better performance. On the most expensive tasks (Sawtooth and Mixture), an epoch for the GNP took roughly six times longer than any other model; see Tab F.2. More results are in App F, including results for setups that test generalisation and extrapolation performance; like the ConvCNP and ConvNP, due to translation equivariance, these results shows that the GNP demonstrates excellent ability to generalise.

5. Conclusion and Future Work

We have provided a rigorous analysis of the standard ML objective used to train NPs. Moreover, we propose a new member to the Neural Process family called the Gaussian Neural Process (GNP), which incorporates translation equivariance, provides universal approximation guarantees, and demonstrates encouraging performance in preliminary experiments. In future work, we aim to investigate incorporating $\Pi^{\text{p.s.d.}}$ (Sec 3) directly into the architecture.

References

A. G. d. G. Matthews, J. Hensman, R. E. Turner, and Z. Ghahramani. On sparse variational methods and the Kullback-Leibler divergence between stochastic processes. In Aarti Singh and Jerry Zhu, editors, *International Conference on Artificial Intelligence and Statistics 22*, volume 54 of *Proceedings of Machine Learning Research*. Proceedings of Machine Learning Research, Apr 2016.

James Dugundji. An extension of Tietze's theorem. *Pacific Journal of Mathematics*, 1(3): 353–367, 1951.

Aasa Feragen. Characterization of equivariant ANEs, 2006. Licentiate thesis.

Andrew Y. K. Foong, Wessel P. Bruinsma, Jonathan Gordon, Yann Dubois, James Requeima, and Richard E. Turner. Meta-learning stationary stochastic process prediction with convolutional neural processes. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems 33*. Curran Associates, Inc., Jul 2020.

M. Garnelo, D. Rosenbaum, C. J. Maddison, T. Ramalho, D. Saxton, M. Shanahan, Y. Whye Teh, D. J. Rezende, and S. M. A. Eslami. Conditional neural processes. In Jennifer Dy and Andreas Krause, editors, *International Conference on Machine Learning 35*, volume 80 of *Proceedings of Machine Learning Research*. Proceedings of Machine Learning Research, Jul 2018a.

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

Jonathan Gordon, Wessel P. Bruinsma, Andrew Y. K. Foong, James Requeima, Yann Dubois, and Richard E. Turner. Convolutional conditional neural processes. *International Conference on Learning Representations (ICLR), 8th*, Oct 2020. URL https://openreview.net/forum?id=Skey4eBYPS.

Nicholas J. Higham. Computing a nearest symmetric positive semidefinite matrix. *Linear Algebra and Its Applications*, 103:103–118, 1988.

H. Kim, A. Mnih, J. Schwarz, M. Garnelo, A. Eslami, D. Rosenbaum, O. Vinyals, and Y. Whye Teh. Attentive neural processes. In *International Conference on Learning Representations 7*, Jan 2019.

D. P. Kingma and J. Ba. ADAM: A method for stochastic optimization. In *International Conference on Learning Representations 3*, Dec 2015.

Chao Ma, Yingzhen Li, and José Miguel Hernández-Lobato. Variational implicit processes. In Yarin Gal, José Miguel Hernández-Lobato, Christos Louizos, Andrew G. Wilson, Zoubin Ghahramani, Kevin Murphy, and Max Welling, editors, *Advances in Neural Information Processing Systems 31*, Dec 2018.

Edward C. Posner. Random coding strategies for minimum entropy. *IEEE Transactions on Information Theory*, 21(4), Jul 1975. ISSN 0018-9448. doi: 10.1109/TIT.1975.1055416.

Carl Edward Rasmussen and Christopher K. I. Williams. *Gaussian Processes for Machine Learning*. MIT Press, 2006.

Sachin Ravi and Hugo Larochelle. Optimization as a model for few-shot learning. In *International Conference on Learning Representations 5*, Apr 2017.

J. Shi, M. Emtiyaz Khan, and J. Zhu. Scalable training of inference networks for Gaussian-process models. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *International Conference on Machine Learning 36*, volume 97 of *Proceedings of Machine Learning Research*. Proceedings of Machine Learning Research, May 2019.

Shengyang Sun, Guodong Zhang, Jiaxin Shi, and Roger Grosse. Functional variational Bayesian neural networks. In Yarin Gal, José Miguel Hernández-Lobato, Christos Louizos, Andrew G. Wilson, Zoubin Ghahramani, Kevin Murphy, and Max Welling, editors, *Advances in Neural Information Processing Systems 31*, Dec 2018.

Oriol Vinyals, Charles Blundell, Timothy Lillicrap, Koray Kavukcuoglu, and Daan Wierstra. Matching networks for one shot learning. In D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems 29*. Curran Associates, Inc., Jun 2016.

D. Yarotsky. Universal approximations of invariant maps by neural networks. *arXiv preprint arXiv:1804.10306*, Apr 2018.

Appendix A. Notation and Terminology

Vectors and matrices: Denote vectors $\mathbf{x}$ with boldface lowercase letters and matrices $\mathbf{A}$ with boldface uppercase letters. For a vector $\mathbf{x}$, let $|\mathbf{x}|$ be its length. For two matrices $\mathbf{A}$ and $\mathbf{B}$, the notation $\mathbf{A} \succ \mathbf{B}$ means that $\mathbf{A} - \mathbf{B}$ is strictly positive definite.

Observations and data sets: Let $\mathcal{X} = \mathbb{R}$ be the space of *inputs* and $\mathcal{Y} = \mathbb{R}$ be the space of *outputs*. Call a tuple $(x, y) \in \mathcal{X} \times \mathcal{Y}$ an *observation*. Let $\mathcal{D}_n = (\mathcal{X} \times \mathcal{Y})^n$ be the space of all collections of n observations, and let $\mathcal{D} = \bigcup_{n=0}^{\infty} \mathcal{D}_n$ be the space of all finite collections of observations. Call an element $D \in \mathcal{D}$ a *data set*. Note that $\varnothing \in \mathcal{D}$. If $D = \{(x_i, y_i)\}_{i=1}^n \in \mathcal{D}_n$, then denote $D = (\mathbf{x}, \mathbf{y})$ where $\mathbf{x} = (x_1, \ldots, x_n)$ and $\mathbf{y} = (y_1, \ldots, y_n)$. Endow $\mathcal{D}$ with the metric

$$d_{\mathcal{D}}(D_1, D_2) = \begin{cases} \|\mathbf{x}_1 - \mathbf{x}_2\|_2 + \|\mathbf{y}_1 - \mathbf{y}_2\|_2 & \text{if } |\mathbf{x}_1| = |\mathbf{x}_2|, \\ \infty & \text{otherwise} \end{cases} \tag{A.1}$$

where $D_1 = (\mathbf{x}_1, \mathbf{y}_1)$ and $D_2 = (\mathbf{x}_2, \mathbf{y}_2)$.

Probability distributions: For all $n \in \mathbb{N}$, let $\mathcal{P}^n$ be the collection of all distributions on $\mathbb{R}^n$, and let $\mathcal{P}_{\mathrm{G}}^n$ be the collection of all such Gaussian distributions. Let $C(\mathcal{X}, \mathcal{Y})$ be the collection of continuous functions $\mathcal{X} \to \mathcal{Y}$ endowed with the metric

$$d_C(f, g) = \sum_{n=1}^{\infty} 2^{-n} \sup_{x \in [-n,n]} (|f(x) - g(x)| \wedge 1), \tag{A.2}$$

which makes it a complete and separable metric space. This metric metricises the topology of *compact convergence*. Let $\mathcal{P}$ be the space of all probability measures on $(\mathcal{Y}^{\mathcal{X}}, \mathcal{B}^{\mathcal{X}})$ where $\mathcal{B}$ is the usual Borel σ-algebra on $\mathcal{Y}$ and $\mathcal{B}^{\mathcal{X}}$ is the cylindrical σ-algebra on $C(\mathcal{X}, \mathcal{Y})$. Similarly, let $\mathcal{P}_{\mathrm{G}}$ be the collection of all such Gaussian measures. Say that a distribution has *full support* if every open set has positive measure. Let $I = \bigcup_{n=1}^{\infty} \mathcal{X}^n$ be the collection of all finite index sets or equivalently all finite sets of inputs. For $\mathbf{x} \in I$, let $P_{\mathbf{x}}$ be the projection on the coordinates $\mathbf{x}$: $P_{\mathbf{x}} f = (f(\mathbf{x}_1), \ldots, f(\mathbf{x}_{|\mathbf{x}|}))$. For $(\mu_i)_{i \geq 1} \subseteq \mathcal{P}$ and $\mu \in \mathcal{P}$, say that $(\mu_i)_{i \geq 1}$ converges weakly to μ, denoted $\mu_i \rightharpoonup \mu$, if $\mu_i(L) \to \mu(L)$ for all $L \colon C(\mathcal{X}, \mathcal{Y}) \to \mathbb{R}$ continuous and bounded. (Recall that $\mu(L)$ denotes $\mathbb{E}_{\mu}[L]$.)

Lebesgue class: For all $n \in \mathbb{N}$, let $\mathcal{P}_{\lambda}^n \subseteq \mathcal{P}^n$ be the collection of all distributions on $\mathbb{R}^n$ that admit a density with respect to the Lebesgue measure. Let $\mathcal{P}_{\lambda} \subseteq \mathcal{P}$ be the collection of processes where every finite-dimensional distribution admits a density with respect to the Lebesgue measure.

Degeneracy: Call a distribution $\mu \in \mathcal{P}^n$ *non-degenerate* if it has a covariance matrix and the covariance matrix is strictly positive definite. Similarly, call a measure $\mu \in \mathcal{P}$ non-degenerate if every finite-dimensional distribution has a covariance matrix and all those covariance matrices are strictly positive definite.

Gaussianisation: For a distribution $\mu \in \mathcal{P}^n$, let $\mathcal{N}(\mu) \in \mathcal{P}_{\mathrm{G}}^n$ be the n-dimensional Gaussian distribution with mean equal to the mean of μ and covariance matrix equal to the covariance matrix of μ, assuming that the latter exists. Similarly, for a measure $\mu \in \mathcal{P}$, let $\mathcal{GP}(\mu) \in \mathcal{P}_{\mathrm{G}}$ be the Gaussian process with mean function equal to the mean function of μ and covariance function equal to the covariance function of μ, assuming that the latter exists.

Appendix B. The Gaussian Divergence

For two measures $\mu \in \mathcal{P}$ and $\nu \in \mathcal{P}$, their Kullback–Leibler divergence enjoys *positive definiteness*: $\mathrm{KL}(\mu, \nu) \geq 0$ with equality if and only if $\mu = \nu$. However, when restricting to only Gaussian measures $\nu \in \mathcal{P}_{\mathrm{G}}$, this property cannot be used anymore, because even the best Gaussian approximation ν may not achieve $\mathrm{KL}(\mu, \nu) = 0$. To get around this, we define a divergence induced by the Kullback–Leibler divergence called the *Gaussian divergence*.

For an arbitrary probability distribution μ on $\mathbb{R}^n$ and a Gaussian distribution ν on $\mathbb{R}^n$, define their Gaussian divergence $\mathrm{G}(\mu, \nu)$ by

$$\mathrm{G}(\mu, \nu) = \begin{cases} \mathrm{KL}(\mu, \nu) - \inf_{\xi \in \mathcal{P}_{\mathrm{G}}^n} \mathrm{KL}(\mu, \xi) & \text{if } \mathrm{KL}(\mu, \xi) < \infty \text{ for some } \xi \in \mathcal{P}_{\mathrm{G}}^n, \\ \infty & \text{otherwise.} \end{cases} \tag{B.1}$$

Let $\mathbf{m}_1$ and $\mathbf{m}_2$ be the means of μ and ν respectively and let $\mathbf{K}_1$ and $\mathbf{K}_2$ be their covariances. Assume that μ has a density with respect to the Lebesgue measure and that $\mathbf{K}_1 \succ 0$ and $\mathbf{K}_2 \succ 0$. Then a quick computation shows that

$$\mathrm{G}(\mu, \nu) = \frac{1}{2} d_{\mathbf{K}}(\mathbf{K}_1, \mathbf{K}_2) + \frac{1}{2} d_{\mathbf{m}}(\mathbf{m}_1, \mathbf{m}_2; \mathbf{K}_2) = \mathrm{KL}(\mathcal{N}(\mu), \nu) \tag{B.2}$$

where $\mathcal{N}(\mu) = \mathcal{N}(\mathbf{m}_1, \mathbf{K}_1)$ and

$$d_{\mathbf{m}}(\mathbf{m}_1, \mathbf{m}_2; \mathbf{K}_2) = (\mathbf{m}_1 - \mathbf{m}_2)^{\mathsf{T}} \mathbf{K}_2^{-1} (\mathbf{m}_1 - \mathbf{m}_2) \geq 0, \tag{B.3}$$

$$d_{\mathbf{K}}(\mathbf{K}_1, \mathbf{K}_2) = \log \frac{|\mathbf{K}_2|}{|\mathbf{K}_1|} + \mathrm{tr}(\mathbf{K}_2^{-1} \mathbf{K}_1) - n \geq 0. \tag{B.4}$$

Since the right-hand side of (B.2) is a Kullback–Leibler divergence, the Gaussian divergence inherits positive definiteness: if $\mu \in \mathcal{P}_\lambda^n$ and $\nu \in \mathcal{P}_{\mathrm{G}}^n$ are non-degenerate, then $\mathrm{G}(\mu, \nu) \geq 0$ with equality if and only if $\nu = \mathcal{N}(\mu)$. The Gaussian divergence inherits more properties from the Kullback–Leibler divergence. An important inherited property that we will make use of is *monotonicity*. Let $S \colon \mathbb{R}^n \to \mathbb{R}^m$ where $m \leq n$ be a projection onto a subset of the coordinates. Then it is true that $\mathrm{G}(S(\mu), S(\nu)) \leq \mathrm{G}(\mu, \nu)$.

Thm B.1 (Sun et al. (2018)): Let $\mu \in \mathcal{P}$ and $\nu \in \mathcal{P}$. Then

$$\mathrm{KL}(\mu, \nu) = \sup_{\mathbf{x} \in I} \mathrm{KL}(P_{\mathbf{x}} \mu, P_{\mathbf{x}} \nu). \tag{B.5}$$

We take inspiration from the above characterisation of the Kullback–Leibler divergence between stochastic processes by Sun et al. (2018) to extend the definition of the Gaussian divergence over distributions on $\mathbb{R}^n$ to arbitrary probability measures in $\mathcal{P}$ by taking the supremum over index sets:

$$\mathrm{G}(\mu, \nu) = \sup_{\mathbf{x} \in I} \mathrm{G}(P_{\mathbf{x}} \mu, P_{\mathbf{x}} \nu). \tag{B.6}$$

Suppose that μ and ν are non-degenerate and that every finite-dimensional distribution of μ has a density with respect to the Lebesgue measure. Then (B.2) shows the equality

$$\mathrm{G}(P_{\mathbf{x}} \mu, P_{\mathbf{x}} \nu) = \mathrm{KL}(\mathcal{N}(P_{\mathbf{x}} \mu), P_{\mathbf{x}} \nu) = \mathrm{KL}(P_{\mathbf{x}} \mathcal{GP}(\mu), P_{\mathbf{x}} \nu). \tag{B.7}$$

Therefore, taking the supremum over $\mathbf{x} \in I$, we find that $\mathrm{G}(\mu, \nu) = \mathrm{KL}(\mathcal{GP}(\mu), \nu)$, *c.f.* (B.2). The right-hand side is a Kullback–Leibler divergence. This means that the Gaussian divergence between processes also inherits positive definiteness: if $\mu \in \mathcal{P}_\lambda$ and $\nu \in \mathcal{P}_\mathrm{G}$ are non-degenerate, then $\mathrm{G}(\mu, \nu) \geq 0$ with equality if and only if $\nu = \mathcal{GP}(\mu)$.

Let $\mu \in \mathcal{P}_\lambda^n$ and $\nu \in \mathcal{P}_\mathrm{G}^n$ be non-degenerate. Then (B.2) and the definition (B.1) show that

$$\mathrm{KL}(\mu, \mathcal{N}(\mu)) = \inf_{\xi \in \mathcal{P}_\mathrm{G}^n} \mathrm{KL}(\mu, \xi). \tag{B.8}$$

Therefore, if $\mu \in \mathcal{P}_\lambda^n$ and $\nu \in \mathcal{P}_\mathrm{G}^n$ are non-degenerate, then we can also write

$$\mathrm{G}(\mu, \nu) = \mathrm{KL}(\mu, \nu) - \mathrm{KL}(\mu, \mathcal{N}(\mu)). \tag{B.9}$$

Since $\mathrm{KL}(\mu, \mathcal{N}(\mu))$ is a constant that does not depend on ν, we have the following result.

Prop B.1: Let $\mu \in \mathcal{P}_\lambda^n$ be non-degenerate and assume that, for every $\mathbf{x} \in I$, there exists some non-degenerate $\mu_\mathrm{G}^\mathbf{x} \in \mathcal{P}_\mathrm{G}^{|\mathbf{x}|}$ such that $\mathrm{KL}(P_\mathbf{x}\mu, \mu_\mathrm{G}^\mathbf{x}) < \infty$. Then

$$\operatorname*{arg\,min}_{\nu \in \mathcal{P}_\mathrm{G}^{|\mathbf{x}|}} \mathrm{KL}(P_\mathbf{x}\mu, \nu) = \operatorname*{arg\,min}_{\nu \in \mathcal{P}_\mathrm{G}^{|\mathbf{x}|}} \mathrm{G}(P_\mathbf{x}\mu, \nu) = \mathcal{N}(P_\mathbf{x}\mu) = P_\mathbf{x}\,\mathcal{GP}(\mu) \quad \text{for all } \mathbf{x} \in I. \tag{B.10}$$

Proof: Follows from (B.9) if $\mathrm{KL}(P_\mathbf{x}\mu, \mathcal{N}(P_\mathbf{x}\mu)) < \infty$ for all $\mathbf{x} \in I$, which, in turn, follows from (B.8) and the assumption. ■

Equality (B.9) was key in the proof of Prop B.1. We proceed to develop a similar expression for the Gaussian divergence between processes: Thm B.2. We first address the issue that $\mathrm{KL}(\mu, \mathcal{GP}(\mu))$ may be infinite.

Prop B.2: Let $\mu \in \mathcal{P}_\lambda$ and $\nu \in \mathcal{P}_\mathrm{G}$ be non-degenerate. Then $\mathrm{KL}(\mu, \mathcal{GP}(\mu)) \leq \mathrm{KL}(\mu, \nu)$.

Proof: Let $\mathbf{x} \in I$. Using that $\mathrm{G}(P_\mathbf{x}\mu, P_\mathbf{x}\,\mathcal{GP}(\mu)) = 0$,

$$\mathrm{KL}(P_\mathbf{x}\mu, P_\mathbf{x}\,\mathcal{GP}(\mu)) = \inf_{\xi \in \mathcal{P}_\mathrm{G}^n} \mathrm{KL}(P_\mathbf{x}\mu, \xi). \tag{B.11}$$

Therefore,

$$\mathrm{KL}(P_\mathbf{x}\mu, P_\mathbf{x}\,\mathcal{GP}(\mu)) \leq \mathrm{KL}(P_\mathbf{x}\mu, P_\mathbf{x}\nu), \tag{B.12}$$

and take the supremum over $\mathbf{x} \in I$ to conclude. ■

As as consequence, $\mathrm{KL}(\mu, \mathcal{GP}(\mu)) < \infty$ if and only if there exits some non-degenerate $\mu_\mathrm{G} \in \mathcal{P}_\mathrm{G}$ such that $\mathrm{KL}(\mu, \mu_\mathrm{G}) < \infty$.

Thm B.2: Let $\mu \in \mathcal{P}_\lambda$ be non-degenerate and assume that there exits some non-degenerate $\mu_\mathrm{G} \in \mathcal{P}_\mathrm{G}$ such that $\mathrm{KL}(\mu, \mu_\mathrm{G}) < \infty$. Let $\nu \in \mathcal{P}_\mathrm{G}$ be non-degenerate. Then $\mathrm{KL}(\mu, \nu)$ and $\mathrm{G}(\mu, \nu)$ differ by a finite constant that only depends on μ:

$$\mathrm{KL}(\mu, \nu) - \mathrm{KL}(\mu, \mathcal{GP}(\mu)) = \mathrm{G}(\mu, \nu). \tag{B.13}$$

Proof: By the assumptions and Prop B.2, $\mathrm{KL}(\mu, \mathcal{GP}(\mu)) < \infty$. Let $\mathbf{x} \in I$. Note that

$$\mathrm{KL}(P_{\mathbf{x}}\mu, P_{\mathbf{x}}\nu) - \mathrm{KL}(\mu, \mathcal{GP}(\mu)) \leq \mathrm{G}(P_{\mathbf{x}}\mu, P_{\mathbf{x}}\nu). \tag{B.14}$$

Take the supremum over $\mathbf{x} \in I$ to find "$\leq$".

For the reverse inequality, write

$$\inf_{\mathbf{x} \in I} (\mathrm{KL}(\mu, \nu) - \mathrm{KL}(P_{\mathbf{x}}\mu, P_{\mathbf{x}}\,\mathcal{GP}(\mu)) = \mathrm{KL}(\mu, \nu) - \mathrm{KL}(\mu, \mathcal{GP}(\mu)). \tag{B.15}$$

Let $\mathbf{x} \in I$. For every $\varepsilon > 0$, can find a $\mathbf{x}' \in I$ such that

$$\mathrm{KL}(\mu, \nu) - \mathrm{KL}(P_{\mathbf{x}'}\mu, P_{\mathbf{x}'}\mathcal{GP}(\mu)) < \mathrm{KL}(\mu, \nu) - \mathrm{KL}(\mu, \mathcal{GP}(\mu)) + \varepsilon. \tag{B.16}$$

Then, by monotonicity of the Kullback–Leibler divergence,

$$\mathrm{KL}(\mu, \nu) - \mathrm{KL}(P_{\mathbf{x}\cup\mathbf{x}'}\mu, P_{\mathbf{x}\cup\mathbf{x}'}\mathcal{GP}(\mu)) < \mathrm{KL}(\mu, \nu) - \mathrm{KL}(\mu, \mathcal{GP}(\mu)) + \varepsilon. \tag{B.17}$$

Therefore, bounding $\mathrm{KL}(P_{\mathbf{x}\cup\mathbf{x}'}\mu, P_{\mathbf{x}\cup\mathbf{x}'}\nu) \leq \mathrm{KL}(\mu, \nu)$,

$$\mathrm{G}(P_{\mathbf{x}\cup\mathbf{x}'}\mu, P_{\mathbf{x}\cup\mathbf{x}'}\nu) < \mathrm{KL}(\mu, \nu) - \mathrm{KL}(\mu, \mathcal{GP}(\mu)) + \varepsilon. \tag{B.18}$$

Hence, by monotonicity of the Gaussian divergence,

$$\mathrm{G}(P_{\mathbf{x}}\mu, P_{\mathbf{x}}\nu) < \mathrm{KL}(\mu, \nu) - \mathrm{KL}(\mu, \mathcal{GP}(\mu)) + \varepsilon. \tag{B.19}$$

Let $\varepsilon \downarrow 0$ and take the supremum over $\mathbf{x} \in I$ to conclude. ■

Cor B.1: Let $\mu \in \mathcal{P}_\lambda$ be non-degenerate and assume that there exits some non-degenerate $\mu_{\mathrm{G}} \in \mathcal{P}_{\mathrm{G}}$ such that $\mathrm{KL}(\mu, \mu_{\mathrm{G}}) < \infty$. Then

$$\operatorname*{arg\,min}_{\nu \in \mathcal{P}_{\mathrm{G}}} \mathrm{KL}(\mu, \nu) = \operatorname*{arg\,min}_{\nu \in \mathcal{P}_{\mathrm{G}}} \mathrm{G}(\mu, \nu) = \mathcal{GP}(\mu). \tag{B.20}$$

Appendix C. Noisy Processes and Prediction Maps

C.1. Noisy Processes

Def C.1 (Noisy Process): Call a stochastic process f *noisy with noise variance* $\sigma^2 > 0$ if it a sum $f = f^{\mathrm{s}} + f^{\mathrm{n}}$ of two independent processes f^{s} and f^{n} where f^{s} is a continuous process called the *smooth part* and f^{n} is such that $P_{\mathbf{x}} f^{\mathrm{n}} \sim \mathcal{N}(\mathbf{0}, \sigma^2 \mathbf{I})$ for all $\mathbf{x} \in I$ and called the *noisy part*. We indicate that a collection of processes is noisy by adding a bar $\overline{}$: $\overline{\mathcal{P}}$ and $\overline{\mathcal{P}}_{\mathrm{G}}$.

Before anything else, we check that the definition is well posed.

Prop C.1 (Noisy Processes are Well Defined): Consider two noisy processes $f_1 = f_1^{\mathrm{s}} + f_1^{\mathrm{n}}$ and $f_2 = f_2^{\mathrm{s}} + f_2^{\mathrm{n}}$. If $f_1 \overset{\mathrm{d}}{=} f_2$, then $f_1^{\mathrm{s}} \overset{\mathrm{d}}{=} f_2^{\mathrm{s}}$ and $f_1^{\mathrm{n}} \overset{\mathrm{d}}{=} f_2^{\mathrm{n}}$.

Proof: Let $\mathbf{x} \in I$ have all distinct elements. We show that $f_1^{\mathrm{s}}(\mathbf{x}) \overset{\mathrm{d}}{=} f_2^{\mathrm{s}}(\mathbf{x})$ and $f_1^{\mathrm{n}}(\mathbf{x}) \overset{\mathrm{d}}{=} f_2^{\mathrm{n}}(\mathbf{x})$. For $i = 1, 2$, set

$$\mathbf{y}^{(i,n)} = \frac{1}{n} \sum_{i=1}^{n} \begin{bmatrix} f_i(x_1 - 2^{-n}) \\ \vdots \\ f_i(x_{|\mathbf{x}|} - 2^{-n}) \end{bmatrix} = \frac{1}{n} \sum_{i=1}^{n} \begin{bmatrix} f_i^{\mathrm{s}}(x_1 - 2^{-n}) \\ \vdots \\ f_i^{\mathrm{s}}(x_{|\mathbf{x}|} - 2^{-n}) \end{bmatrix} + \frac{1}{n} \sum_{i=1}^{n} \begin{bmatrix} f_i^{\mathrm{n}}(x_1 - 2^{-n}) \\ \vdots \\ f_i^{\mathrm{n}}(x_{|\mathbf{x}|} - 2^{-n}) \end{bmatrix}. \tag{C.1}$$

By continuity of f^{s} and the Strong Law of Large Numbers, for $i = 1, 2$,

$$(\mathbf{y}^{(i,n)}, f_i(\mathbf{x})) \to (f_i^{\mathrm{s}}(\mathbf{x}), f_i(\mathbf{x})) \quad \text{as } n \to \infty \tag{C.2}$$

almost surely and hence weakly. But we assumed that $f_1 \overset{\mathrm{d}}{=} f_2$, so

$$(\mathbf{y}^{(1,n)}, f_1(\mathbf{x})) \overset{\mathrm{d}}{=} (\mathbf{y}^{(2,n)}, f_2(\mathbf{x})) \quad \text{for all } n \geq 1. \tag{C.3}$$

Therefore, since weak limits are unique,

$$(f_1^{\mathrm{s}}(\mathbf{x}), f_1(\mathbf{x})) \overset{\mathrm{d}}{=} (f_2^{\mathrm{s}}(\mathbf{x}), f_2(\mathbf{x})). \tag{C.4}$$

Finally, using that $(\mathbf{y}_1, \mathbf{y}_2) \mapsto (\mathbf{y}_1, \mathbf{y}_2 - \mathbf{y}_1)$ is continuous, hence measurable, we have

$$(f_1^{\mathrm{s}}(\mathbf{x}), f_1^{\mathrm{n}}(\mathbf{x})) \overset{\mathrm{d}}{=} (f_2^{\mathrm{s}}(\mathbf{x}), f_2^{\mathrm{n}}(\mathbf{x})). \tag{C.5}$$

In particular, this means that $f_1^{\mathrm{s}}(\mathbf{x}) \overset{\mathrm{d}}{=} f_2^{\mathrm{s}}(\mathbf{x})$ and $f_1^{\mathrm{n}}(\mathbf{x}) \overset{\mathrm{d}}{=} f_2^{\mathrm{n}}(\mathbf{x})$. ■

Note that noise variance $\sigma^2 = 0$ is not allowed. Moreover, note that noisy process do not have measurable sample paths in general. They are very irregular objects that can only be worked with at a countable collection of indices. From a theoretical perspective, this is undesirable and raises concerns. However, from a practical perspective, we will only ever work at a finite collection of indices; and at every finite collection indices, noisy processes do provide the right model, because in most practical applications observations are contaminated with a little bit of noise.

Prop C.2 (Noisy Processes are Lebesgue Class): Let $f \sim \mu \in \overline{\mathcal{P}}$ be a noisy process with noise variance $\sigma^2 > 0$. Let $\mathbf{x} \in I$ be an index set. Then $P_{\mathbf{x}}\mu$ has the following density with respect to the Lebesgue measure:

$$p_f^{\mathbf{x}}(\mathbf{y}) = \mathbb{E}_{f^{\mathrm{s}}}[\mathcal{N}(\mathbf{y} \mid f^{\mathrm{s}}(\mathbf{x}), \sigma^2\mathbf{I})]. \tag{C.6}$$

Consequently, $\overline{\mathcal{P}} \subseteq \mathcal{P}_\lambda$.

Proof: Let $B \subseteq \mathcal{Y}^{|\mathbf{x}|}$ be a Borel set. Then

$$\begin{aligned}
\int_B p_f^{\mathbf{x}}(\mathbf{y})\,\mathrm{d}\mathbf{y} &\overset{\text{(i)}}{=} \mathbb{E}_{f^{\mathrm{s}}}[\mathbb{P}_{f^{\mathrm{n}}}(f^{\mathrm{n}}(\mathbf{x}) \in B - f^{\mathrm{s}}(\mathbf{x}))] && \text{(C.7)}\\
&= \mathbb{E}_{f^{\mathrm{s}}}[\mathbb{E}_{f^{\mathrm{n}}}[\mathbb{1}_B(f^{\mathrm{s}}(\mathbf{x}) + f^{\mathrm{n}}(\mathbf{x}))]] && \text{(C.8)}\\
&= \mathbb{P}(f^{\mathrm{s}}(\mathbf{x}) + f^{\mathrm{n}}(\mathbf{x}) \in B), && \text{(C.9)}
\end{aligned}$$

using in (i) that $f^{\mathrm{n}}(\mathbf{x}) \sim \mathcal{N}(\mathbf{0}, \sigma^2\mathbf{I})$. ■

Prop C.3 (Equality of Noisy Processes): Let $\mu \in \overline{\mathcal{P}}$ and $\nu \in \overline{\mathcal{P}}$ be two noisy processes with noise variances $\sigma_\mu^2 > 0$ and $\sigma_\nu^2 > 0$. Let $\tilde{I} \subseteq I$ be dense in I. If $P_{\mathbf{x}}\mu = P_{\mathbf{x}}\nu$ for all $\mathbf{x} \in \tilde{I}$, then $\mu = \nu$. Moreover, if μ and ν are Gaussian, then it suffices to consider $\tilde{I} \subseteq \mathcal{X}^n$ dense in $\mathcal{X}^n$ for any fixed $n \geq 2$.

Proof: Let $\mathbf{x} \in I$. Use density of $\tilde{I}$ to extract a sequence $(\mathbf{x}_i)_{i\geq 1} \subseteq \tilde{I}$ convergent to $\mathbf{x}$. Consider $B \subseteq \mathcal{Y}^{|\mathbf{x}|}$ open. Let $\mathbf{n}_\mu \sim \mathcal{N}(\mathbf{0}, \sigma_\mu^2\mathbf{I})$ and $\mathbf{n}_\nu \sim \mathcal{N}(\mathbf{0}, \sigma_\nu^2\mathbf{I})$. Then

$$\begin{aligned}
P_{\mathbf{x}}\mu(B) &= \mathbb{E}_\mu[\mathbb{1}_B(f(\mathbf{x}) + \mathbf{n}_\mu)] && \text{(C.10)}\\
&= \lim_{i\to\infty} \mathbb{E}_\mu[\mathbb{1}_B(f(\mathbf{x}_i) + \mathbf{n}_\mu)] && \text{(C.11)}\\
&= \lim_{i\to\infty} \mathbb{E}_\nu[\mathbb{1}_B(f(\mathbf{x}_i) + \mathbf{n}_\nu)] && \text{(C.12)}\\
&= \mathbb{E}_\nu[\mathbb{1}_B(f(\mathbf{x}) + \mathbf{n}_\nu)] && \text{(C.13)}\\
&= P_{\mathbf{x}}\nu(B). && \text{(C.14)}
\end{aligned}$$

Since B was an arbitrary open set, we conclude that $P_{\mathbf{x}}\mu = P_{\mathbf{x}}\nu$.

If μ and ν are Gaussian, then all finite-dimensional distributions are equal if and only if all means and covariances are equal. And all means and covariances are equal if all finite-dimensional distributions of any fixed size $n \geq 2$ are equal, so it suffices to consider $\tilde{I} \subseteq \mathcal{X}^n$ dense in $\mathcal{X}^n$ for any fixed $n \geq 2$. ■

Since noisy processes do not produce measurable sample paths, we need to define what weak convergence means.

Def C.2 (Weak Convergence of Noisy Processes): Say that a sequence $(f_i)_{i\geq 1}$ of noisy processes with noise variances $(\sigma_i^2)_{i\geq 1}$ weakly converges to a noisy process f with noise variance σ^2 if $f_i^{\mathrm{s}} \rightharpoonup f^{\mathrm{s}}$ and $\sigma_i^2 \to \sigma^2$.

Note that there is no problem with weak convergence of finite-dimensional distributions of noisy processes. With the above definition, it is easy to check that $f_i \rightharpoonup f$ implies that $P_{\mathbf{x}} f_i \rightharpoonup P_{\mathbf{x}} f$ for all $\mathbf{x} \in I$ where the latter weak convergence is in the usual sense.

In the posterior of a noisy process, the noisy part is renewed.

Def C.3 (Posterior of Noisy Process): The posterior of a noisy process f given some data $D = (\mathbf{x}, \mathbf{y}) \in \mathcal{D}$ is defined by the sum $f^{\mathrm{s}}_D + \hat{f}^{\mathrm{n}}$ where f^{s}_D is the posterior of f^{s} given that $f(\mathbf{x}) = \mathbf{y}$ and $\hat{f}^{\mathrm{n}}$ is an independent copy of f^{n}.

If $f^{\mathrm{s}} \sim \mu^{\mathrm{s}} \in \mathcal{P}$ and $f^{\mathrm{s}}_D \sim \mu^{\mathrm{s}}_D \in \mathcal{P}$, then one can show that

$$\frac{\mathrm{d}\mu^{\mathrm{s}}_D}{\mathrm{d}\mu^{\mathrm{s}}}(f^{\mathrm{s}}) = \frac{\mathcal{N}(\mathbf{y}; f^{\mathrm{s}}(\mathbf{x}), \sigma^2 \mathbf{I})}{\mathbb{E}_{f^{\mathrm{s}}}[\mathcal{N}(\mathbf{y}; f^{\mathrm{s}}(\mathbf{x}), \sigma^2 \mathbf{I})]}. \tag{C.15}$$

Note that $\mathbb{E}_{f^{\mathrm{s}}}[\mathcal{N}(\mathbf{y}; f^{\mathrm{s}}(\mathbf{x}), \sigma^2 \mathbf{I})] > 0$, for otherwise $|f^{\mathrm{s}}(x)| = \infty$ almost surely for some $x \in \mathcal{X}$. Denote

$$\pi_f(D) = \mu^{\mathrm{s}}_D \quad \text{and} \quad \pi'_f(D) = \frac{\mathrm{d}\mu^{\mathrm{s}}_D}{\mathrm{d}\mu^{\mathrm{s}}}. \tag{C.16}$$

The map $\pi_f \colon \mathcal{D} \to \overline{\mathcal{P}}$ is called the *posterior prediction map* of f.

C.2. Prediction Maps

Def C.4 (Prediction Map): A map $\pi \colon \mathcal{D} \to \mathcal{P}$ is called a *prediction map*. Call a prediction map *Gaussian* if it maps to Gaussian processes and *noisy* if it maps to noisy processes.

Prop C.4 (Continuity in the Data): With probability one, the map $D \mapsto \pi'_f(D)(f^{\mathrm{s}})$ is continuous. We call this property *continuity in the data*.

Proof: Follows from continuity of f^{s} and

$$\mathcal{N}(\mathbf{y}; f^{\mathrm{s}}(\mathbf{x}), \sigma^2 \mathbf{I}) \leq (2\pi\sigma^2)^{-\frac{1}{2}|\mathbf{x}|} \tag{C.17}$$

in combination with bounded convergence. ■

Prop C.5 (Local Boundedness): For any compact collection of data sets $\tilde{\mathcal{D}} \subseteq \mathcal{D}$,

$$0 < \sup_{D \in \tilde{\mathcal{D}}} \sup_{f^{\mathrm{s}} \in \mathcal{Y}^{\mathcal{X}}} \pi'_f(D)(f^{\mathrm{s}}) < \infty. \tag{C.18}$$

We call this property *local boundedness*.

Proof: Note that

$$M_D := \sup_{f^{\mathrm{s}} \in \mathcal{Y}^{\mathcal{X}}} \pi'_f(D)(f^{\mathrm{s}}) = \frac{(2\pi\sigma^2)^{-\frac{1}{2}|\mathbf{x}|}}{\mathbb{E}_{f^{\mathrm{s}}}[\mathcal{N}(\mathbf{y}; f^{\mathrm{s}}(\mathbf{x}), \sigma^2 \mathbf{I})]}, \tag{C.19}$$

which by an argument similar to the proof of Prop C.4 is continuous in D. Moreover, we have that $0 < M_D < \infty$ for all $D \in \tilde{D}$. Therefore, $0 < \sup_{D \in \tilde{\mathcal{D}}} M_D < \infty$ by continuity of $D \mapsto M_D$ and compactness of $\tilde{\mathcal{D}}$. ■

Def C.5 (Bounded Collection of Data Sets): For $\tilde{\mathcal{D}} \subseteq \mathcal{D}$, define

$$\|\tilde{\mathcal{D}}\|_\infty = \sup_{(\mathbf{x},\mathbf{y})\in\tilde{\mathcal{D}}} \sum_{i=1}^{|\mathbf{x}|} (|y_i| \vee 1).$$

In particular, if $(\mathbf{x},\mathbf{y}) \in \tilde{\mathcal{D}}$, then $|\mathbf{x}| \leq \|\tilde{\mathcal{D}}\|_\infty$ and $\|\mathbf{y}\|_\infty \leq \|\tilde{\mathcal{D}}\|_\infty$. Call a collection of data sets $\tilde{\mathcal{D}} \subseteq \mathcal{D}$ *bounded* if $\|\tilde{\mathcal{D}}\|_\infty < \infty$.

Using this definition of boundedness, we can refine Prop C.5 to obtain a quantitative bound.

Prop C.6 (Local Boundedness (Cont'd)): Assume that

$$V = \sup_{x\in\mathcal{X}} \mathbb{E}[(f^{\mathrm{s}}(x))^2] < \infty. \tag{C.20}$$

Then, for any bounded collection of data sets $\tilde{\mathcal{D}} \subseteq \mathcal{D}$,

$$1 \leq \sup_{D\in\tilde{\mathcal{D}}} \sup_{f^{\mathrm{s}}\in\mathcal{Y}^{\mathcal{X}}} \pi'_f(D)(f^{\mathrm{s}}) \leq 2\exp\left(\frac{2\|\tilde{\mathcal{D}}\|_\infty^2 V}{\sigma^2} + \frac{\|\tilde{\mathcal{D}}\|_\infty^3}{\sigma^2}\right). \tag{C.21}$$

Proof: Start out from (C.19):

$$\frac{1}{M_D} = \mathbb{E}_{f^{\mathrm{s}}}\left[\exp\left(-\frac{\|\mathbf{y} - f^{\mathrm{s}}(\mathbf{x})\|^2}{2\sigma^2}\right)\right] \geq \exp\left(-\frac{\|\mathbf{y}\|^2}{\sigma^2}\right)\mathbb{E}_{f^{\mathrm{s}}}\left[\exp\left(-\frac{\|f^{\mathrm{s}}(\mathbf{x})\|^2}{\sigma^2}\right)\right] \tag{C.22}$$

Therefore,

$$1 \leq M_D \leq \mathbb{E}_{f^{\mathrm{s}}}\left[\exp\left(-\frac{\|f^{\mathrm{s}}(\mathbf{x})\|^2}{\sigma^2}\right)\right]^{-1} \exp\left(\frac{\|\mathbf{y}\|^2}{\sigma^2}\right). \tag{C.23}$$

Now estimate

$$\mathbb{E}_{f^{\mathrm{s}}}\left[\exp\left(-\frac{\|f^{\mathrm{s}}(\mathbf{x})\|^2}{\sigma^2}\right)\right] \geq \mathbb{P}(\|f^{\mathrm{s}}(\mathbf{x})\|_\infty \leq R)\exp\left(-\frac{nR^2}{\sigma^2}\right). \tag{C.24}$$

Choose $R = \sqrt{2nV}$ to obtain

$$\mathbb{P}(\|f^{\mathrm{s}}(\mathbf{x})\|_\infty \leq R) \geq 1 - n\sup_{x\in\mathcal{X}} \mathbb{P}(|f^{\mathrm{s}}(x)| > R) \geq 1 - \frac{nV}{R^2} = \frac{1}{2}. \tag{C.25}$$

With this choice for R, we find

$$1 \leq M_D \leq 2\exp\left(\frac{nR^2}{\sigma^2} + \frac{n\|\mathbf{y}\|_\infty^2}{\sigma^2}\right) = 2\exp\left(\frac{2n^2V}{\sigma^2} + \frac{n\|\mathbf{y}\|_\infty^2}{\sigma^2}\right). \tag{C.26}$$

The result then follows from the observations that $n \leq \|\tilde{\mathcal{D}}\|_\infty$ and $\|\mathbf{y}\|_\infty \leq \|\tilde{\mathcal{D}}\|_\infty$. ■

Def C.6 (Continuous Prediction Map): Call a prediction map $\pi\colon \mathcal{D} \to \mathcal{P}$ *continuous* if $D_i \to D$ implies that $\pi(D_i) \rightharpoonup \pi(D)$. Denote the collection of all continuous prediction maps by $\mathcal{M}$. Write a subscript $_{\mathrm{G}}$ if the prediction maps are also Gaussian: $\mathcal{M}_{\mathrm{G}}$. Write a bar $\overline{}$ if the prediction maps are also noisy: $\overline{\mathcal{M}}$. Call a prediction map $\pi\colon \mathcal{D} \to \mathcal{P}$ *continuous along its finite-dimensional distributions* if $D_i \to D$ implies that $P_{\mathbf{x}}\pi(D_i) \rightharpoonup P_{\mathbf{x}}\pi(D)$ for all $\mathbf{x} \in I$. Write a superscript $^{\text{f.d.d.}}$ if the prediction maps are *only* continuous along their finite-dimensional distributions: $\mathcal{M}^{\text{f.d.d.}}$. Note the following inclusions:

$$\mathcal{M} \subseteq \mathcal{M}^{\text{f.d.d.}}, \quad \mathcal{M}_{\mathrm{G}} \subseteq \mathcal{M}_{\mathrm{G}}^{\text{f.d.d.}}, \quad \overline{\mathcal{M}} \subseteq \overline{\mathcal{M}}^{\text{f.d.d.}}, \quad \overline{\mathcal{M}}_{\mathrm{G}} \subseteq \overline{\mathcal{M}}_{\mathrm{G}}^{\text{f.d.d.}}. \tag{C.27}$$

Prop C.7 (Equality of Continuous Prediction Maps): Let $\pi_1 \in \mathcal{M}^{\text{f.d.d.}}$ and $\pi_2 \in \mathcal{M}^{\text{f.d.d.}}$. Let $\tilde{\mathcal{D}} \subseteq \mathcal{D}$. If $\pi_1 = \pi_2$ are equal on a dense subset of $\tilde{\mathcal{D}}$, then $\pi_1 = \pi_2$ are equal on all of $\tilde{\mathcal{D}}$.

Proof: Let $D \in \mathcal{D}$. Extract $(D_i)_{i\geq 1} \subseteq \tilde{\mathcal{D}}$ convergent to D. Let $\mathbf{x} \in I$. By the assumed continuity, $P_{\mathbf{x}}\pi_1(D_i) \rightharpoonup P_{\mathbf{x}}\pi_1(D)$ and $P_{\mathbf{x}}\pi_2(D_i) \rightharpoonup P_{\mathbf{x}}\pi_2(D)$. Let $L\colon \mathbb{R}^{|\mathbf{x}|} \to \mathbb{R}$ be continuous and bounded. Then

$$P_{\mathbf{x}}\pi_1(D)(L) = \lim_{i\to\infty} P_{\mathbf{x}}\pi_1(D_i)(L) = \lim_{i\to\infty} P_{\mathbf{x}}\pi_2(D_i)(L) = P_{\mathbf{x}}\pi_2(D)(L). \tag{C.28}$$

Since $\mathbf{x}$ and L were arbitrary, $\pi_1(D) = \pi_2(D)$. ■

Prop C.8 (Noisy Posterior Prediction Map is Continuous): Let f be a noisy process and π_f the associated posterior prediction map. Then $\pi_f \in \mathcal{M}$.

Proof: For all $D \in \mathcal{D}$, the noisy part of $\pi_f(D)$ is equal in distribution. Hence, it suffices to show that $D_i \to D$ implies that $f^{\mathrm{s}}_{D_i} \rightharpoonup f^{\mathrm{s}}_{D}$. Let $L\colon C(\mathcal{X}, \mathcal{Y}) \to \mathbb{R}$ be continuous and bounded. Then

$$\mathbb{E}[L(f^{\mathrm{s}}_{D_i})] = \mathbb{E}[\pi'(D_i)(f^{\mathrm{s}})L(f^{\mathrm{s}})] \to \mathbb{E}[\pi'(D)(f^{\mathrm{s}})L(f^{\mathrm{s}})] = \mathbb{E}[L(f^{\mathrm{s}}_{D})] \tag{C.29}$$

by continuity in the data (Prop C.4) and bounded convergence using local boundedness (Prop C.5). ■

Prop C.9 (Noisy Posterior Prediction Map is Bounded): Let f be a noisy process and let $\tilde{\mathcal{D}} \subseteq \mathcal{D}$ be a bounded collection of data sets. Suppose that $\sup_{x\in\mathcal{X}} \mathbb{E}[(f^{\mathrm{s}}(x))^2] < \infty$. Then $\sup_{x\in\mathcal{X},\, D\in\tilde{\mathcal{D}}} \mathbb{E}[(f^{\mathrm{s}}_D(x))^2] < \infty$.

Proof: Follows from

$$\mathbb{E}[(f^{\mathrm{s}}_D(x))^2] \leq \left(\sup_{D\in\tilde{\mathcal{D}}} \sup_{f^{\mathrm{s}}\in\mathcal{Y}^{\mathcal{X}}} \pi'_f(D)(f^{\mathrm{s}})\right)\left(\sup_{x\in\mathcal{X}} \mathbb{E}[(f^{\mathrm{s}}(x))^2]\right) \tag{C.30}$$

in combination with (C.21). ■

C.3. Gaussianised Prediction Maps

Def C.7 (Gaussianised Prediction Map): Given a prediction map $\pi\colon \mathcal{D} \to \mathcal{P}$, the *Gaussianised prediction map* $\mathcal{GP}(\pi)$ is defined by $\mathcal{GP}(\pi)(D) = \mathcal{GP}(\pi(D))$.

The Gaussianisation of a noisy process $f \sim \mu \in \overline{\mathcal{P}}$ is equal to $f^{\mathrm{s}}_{\mathrm{G}} + f^{\mathrm{n}}$ where $f^{\mathrm{s}}_{\mathrm{G}} \sim \mathcal{GP}(\mu^{\mathrm{s}})$. This is perfectly well defined. However, a subtle technical issue is that $f^{\mathrm{s}}_{\mathrm{G}}$ may not be a continuous process, which means that the Gaussianisation of a noisy process is not necessarily a noisy process. To prevent this from happening, we impose regularity conditions on f^{s}.

Prop C.10: Let f be a noisy process and let π_f the associated posterior prediction map. Suppose that there exist $p \geq 2$, $\beta \in (\frac{1}{2}, 1]$, a constant $c > 0$ and a radius $r > 0$ such that

$$\|f^{\mathrm{s}}(x) - f^{\mathrm{s}}(y)\|_{L^p} \leq c|x-y|^{\beta} \quad \text{whenever} \quad |x-y| < r. \tag{C.31}$$

Then, for all $D \in \mathcal{D}$, if $\mathcal{GP}(\pi_f(D))$ exists, it is a noisy process.

Proof: Let $D \in \mathcal{D}$. As explained above, assuming that $\mathcal{GP}(\pi_f(D))$ exists, *i.e.* that $\pi_f(D)$ has a mean function and covariance function, it remains to show that the smooth part of $\mathcal{GP}(\pi_f(D))$ is a continuous process. Let f^{s}_D be the smooth part of $\pi_f(D)$ and let g^{s}_D be the smooth part of $\mathcal{GP}(\pi_f(D))$. Since, by construction of $\mathcal{GP}(\pi_f(D))$, the mean functions and covariance functions of f^{s}_D and g^{s}_D are equal,

$$\mathbb{E}[|g^{\mathrm{s}}_D(x) - g^{\mathrm{s}}_D(y)|^2] = \mathbb{E}[|f^{\mathrm{s}}_D(x) - f^{\mathrm{s}}_D(y)|^2]. \tag{C.32}$$

Therefore, using Jensen's Inequality and concavity of $x \mapsto x^{2/p}$ $(p \geq 2)$,

$$\begin{aligned}
\mathbb{E}[|g^{\mathrm{s}}_D(x) - g^{\mathrm{s}}_D(y)|^2] &\leq \mathbb{E}[|f^{\mathrm{s}}_D(x) - f^{\mathrm{s}}_D(y)|^p]^{2/p} && \text{(C.33)}\\
&= \mathbb{E}[\pi'_f(D)(f^{\mathrm{s}})|f^{\mathrm{s}}(x) - f^{\mathrm{s}}(y)|^p]^{2/p} && \text{(C.34)}\\
&\leq M_D \mathbb{E}[|f^{\mathrm{s}}(x) - f^{\mathrm{s}}(y)|^p]^{2/p} && \text{(C.35)}
\end{aligned}$$

with $M_D = \sup_{f^{\mathrm{s}} \in \mathcal{Y}^{\mathcal{X}}} \pi'_f(D)(f^{\mathrm{s}})$. By (C.19), $0 < M_D < \infty$. We can thus continue our sequence of inequalities:

$$\mathbb{E}[|g^{\mathrm{s}}_D(x) - g^{\mathrm{s}}_D(y)|^2] \leq M_D \|f^{\mathrm{s}}(x) - f^{\mathrm{s}}(y)\|^2_{L^p} \leq M_D c^2 |x-y|^{2\beta} \tag{C.36}$$

whenever $|x-y| < r$. Hence,

$$\|g^{\mathrm{s}}(x) - g^{\mathrm{s}}(y)\|_{L^2} \leq \sqrt{M_D} c |x-y|^{\beta} \quad \text{whenever} \quad |x-y| < r. \tag{C.37}$$

This shows that g^{s} satisfies Kolmogorov's Continuity Criterion and thus admits a continuous version. ■

Throughout, we assume that the condition from Prop C.10 always satisfied. Consequently, the Gaussianisation of any noisy process is always also a noisy process.

Prop C.11 (Gaussianised Noisy Posterior Prediction Map is Cont.): Let f be a noisy process and let π_f the associated posterior prediction map. Suppose that $f^{\mathrm{s}}(x) \in L^2$ for all $x \in \mathcal{X}$. Then $\mathcal{GP}(\pi_f) \in \overline{\mathcal{M}}_{\mathrm{G}}^{\mathrm{f.d.d.}}$.

Proof: Let $D_i \to D$. By the assumptions, $\mathbb{E}[f^{\mathrm{s}}(x)]$ exists. Hence, using continuity in the data (Prop C.4) and dominated convergence in combination with local boundedness (Prop C.5),

$$\mathbb{E}[f^{\mathrm{s}}_{D_i}(x)] = \mathbb{E}[\pi'(D_i)(f^{\mathrm{s}})f^{\mathrm{s}}(x)] \to \mathbb{E}[\pi'(D)(f^{\mathrm{s}})f^{\mathrm{s}}(x)] = \mathbb{E}[f^{\mathrm{s}}_{D}(x)]. \tag{C.38}$$

Similarly, $\mathrm{cov}(f^{\mathrm{s}}_{D_i}(x), f^{\mathrm{s}}_{D_i}(y)) \to \mathrm{cov}(f^{\mathrm{s}}_{D}(x), f^{\mathrm{s}}_{D}(y))$. Therefore, for all $\mathbf{x} \in I$, $P_{\mathbf{x}}\,\mathcal{GP}(\pi_f)(D_i) \rightharpoonup P_{\mathbf{x}}\,\mathcal{GP}(\pi_f)(D)$. ■

Appendix D. The Objective

We before discussing the objective, we first get all issues of measurability out of the way.

Prop D.1: Let $\mu_1, \mu_2 \in \overline{\mathcal{P}}$ and $\mu_{\mathrm{G}} \in \overline{\mathcal{P}}_{\mathrm{G}}$. Fix $n \geq 1$ and consider all $\mathbf{x} \in \mathcal{X}^n$. Then

(i) $\mathbf{x} \mapsto \mathrm{KL}(P_{\mathbf{x}}\mu_1, P_{\mathbf{x}}\mu_{\mathrm{G}})$ is lower semi-continuous, hence measurable;

(ii) $\mathbf{x} \mapsto \mathrm{KL}(P_{\mathbf{x}}\,\mathcal{GP}(\mu_1), P_{\mathbf{x}}\mu_{\mathrm{G}})$ is lower semi-continuous, hence measurable;

(iii) $\mathbf{x} \mapsto \mathrm{KL}(P_{\mathbf{x}}\mu_1, P_{\mathbf{x}}\,\mathcal{GP}(\mu_2))$ is lower semi-continuous, hence measurable.

Proof: Props D.1.(ii) and D.1.(iii) follow from Prop D.1.(i) by the observations that $\mathcal{GP}(\mu_1) \in \overline{\mathcal{P}}_{\mathrm{G}} \subseteq \overline{\mathcal{P}}$ and $\mathcal{GP}(\mu_2) \in \overline{\mathcal{P}}_{\mathrm{G}}$. To prove Prop D.1.(i), let $(\mathbf{x}_i)_{i\geq 1} \subseteq \mathcal{X}^n$ be convergent to $\mathbf{x} \in \mathcal{X}^n$. Then $P_{\mathbf{x}_i}\mu_1 \rightharpoonup P_{\mathbf{x}}\mu_1$ and $P_{\mathbf{x}_i}\mu_{\mathrm{G}} \rightharpoonup P_{\mathbf{x}}\mu_{\mathrm{G}}$ because the smooth parts of μ_1 and μ_{G} are continuous processes. Using that $(\mu, \nu) \mapsto \mathrm{KL}(\mu, \nu)$ is weakly lower semi-continuous (Posner, 1975), we thus find

$$\liminf_{i\to\infty} \mathrm{KL}(P_{\mathbf{x}_i}\mu, P_{\mathbf{x}_i}\nu) \geq \mathrm{KL}(P_{\mathbf{x}}\mu, P_{\mathbf{x}}\nu), \tag{D.1}$$

which shows that $\mathbf{x} \mapsto \mathrm{KL}(P_{\mathbf{x}}\mu, P_{\mathbf{x}}\nu)$ is lower semi-continuous. ■

Prop D.2: Let $\pi_1, \pi_2 \in \overline{\mathcal{M}}^{\text{f.d.d.}}$ and $\pi_{\mathrm{G}} \in \overline{\mathcal{M}}_{\mathrm{G}}^{\text{f.d.d.}}$. Suppose that, for all $D \in \mathcal{D}$, the mean functions and covariance functions $\pi_1(D)$ of $\pi_2(D)$ exist. Fix $n \geq 1$ and let $p(\mathbf{x})$ be a Borel distribution on $\mathcal{X}^n$. Then

(i) $D \mapsto \mathbb{E}_{p(\mathbf{x})}[\mathrm{KL}(P_{\mathbf{x}}\pi_1(D), P_{\mathbf{x}}\pi_{\mathrm{G}}(D))]$ is lower semi-continuous, hence measurable;

(ii) $D \mapsto \mathbb{E}_{p(\mathbf{x})}[\mathrm{KL}(P_{\mathbf{x}}\,\mathcal{GP}(\pi_1)(D), P_{\mathbf{x}}\pi_{\mathrm{G}}(D))]$ is lower semi-continuous, hence measurable;

(iii) $D \mapsto \mathbb{E}_{p(\mathbf{x})}[\mathrm{KL}(P_{\mathbf{x}}\,\mathcal{GP}(\pi_1)(D), P_{\mathbf{x}}\,\mathcal{GP}(\pi_2)(D))]$ is lower semi-continuous, hence measurable.

Proof: By Prop D.1, these expectations are all well defined. Props D.2.(ii) and D.2.(iii) follow from Prop D.2.(i) by the observations that $\mathcal{GP}(\pi_1) \in \overline{\mathcal{M}}_{\mathrm{G}}^{\text{f.d.d.}} \subseteq \overline{\mathcal{M}}^{\text{f.d.d.}}$ and $\mathcal{GP}(\pi_2) \in \overline{\mathcal{M}}^{\text{f.d.d.}}$ (Prop C.11). To prove Prop D.2.(i), let $(D_i)_{i\geq 1} \subseteq \mathcal{D}$ be convergent to $D \in \mathcal{D}$. Since $\pi_1 \in \overline{\mathcal{M}}^{\text{f.d.d.}}$ is continuous, $P_{\mathbf{x}}\pi_f(D_i) \rightharpoonup P_{\mathbf{x}}\pi(D)$ for all $\mathbf{x} \in \mathcal{X}^n$, and the same statement holds for π_{G}. Therefore, using Fatou's Lemma and that $(\mu, \nu) \mapsto \mathrm{KL}(\mu, \nu)$ is weakly lower semi-continuous (Posner, 1975), we thus find

$$\liminf_{i\to\infty} \mathbb{E}_{p(\mathbf{x})}[\mathrm{KL}(P_{\mathbf{x}}\pi_1(D_i), P_{\mathbf{x}}\pi_{\mathrm{G}}(D_i))] \geq \mathbb{E}_{p(\mathbf{x})}[\liminf_{i\to\infty} \mathrm{KL}(P_{\mathbf{x}}\pi_1(D_i), P_{\mathbf{x}}\pi_{\mathrm{G}}(D_i))] \tag{D.2}$$

$$\geq \mathbb{E}_{p(\mathbf{x})}[\mathrm{KL}(P_{\mathbf{x}}\pi_1(D), P_{\mathbf{x}}\pi_{\mathrm{G}}(D))], \tag{D.3}$$

which shows that $D \mapsto \mathbb{E}_{p(\mathbf{x})}[\mathrm{KL}(P_{\mathbf{x}}\pi_f(D), P_{\mathbf{x}}\pi(D))]$ is lower semi-continuous. ■

Prop D.3: Let $\mu \in \overline{\mathcal{P}}$ and $\nu \in \overline{\mathcal{P}}_{\mathrm{G}}$. Fix $n \geq 2$. Assume the following:

(1) The mean functions and covariance functions of μ and ν exist and are uniformly bounded by $M > 0$.

(2) The processes μ and ν are noisy with noise variance greater than $\sigma^2 > 0$.

Then

$$\mathrm{KL}(P_{\mathbf{x}}\mu, P_{\mathbf{x}}\nu) \leq \frac{4n^2(M \vee 1)^2}{\sigma^2}. \tag{D.4}$$

Proof: By the assumption that μ is noisy with noise variance $\sigma^2 > 0$, for all $\mathbf{x} \in \mathcal{X}^n$, the finite-dimensional distribution $P_{\mathbf{x}}\mu$ has the following density with respect to the Lebesgue measure on $\mathbb{R}^n$ (Prop C.2):

$$p_\mu^{\mathbf{x}}(\mathbf{y}) = \mathbb{E}_\mu[\mathcal{N}(\mathbf{y}; f^{\mathrm{s}}(\mathbf{x}), \sigma^2)] \tag{D.5}$$

For all $\mathbf{x} \in I_n$, denote

$$\mathcal{N}(P_{\mathbf{x}}\mu) = \mathcal{N}(\mathbf{m}_\mu^{\mathbf{x}}, \mathbf{K}_\mu^{\mathbf{x}}) \quad \text{and} \quad P_{\mathbf{x}}\nu = \mathcal{N}(\mathbf{m}_\nu^{\mathbf{x}}, \mathbf{K}_\nu^{\mathbf{x}}). \tag{D.6}$$

Start out by expanding the Kullback–Leibler divergence:

$$\begin{aligned} &\mathrm{KL}(P_{\mathbf{x}}\mu, P_{\mathbf{x}}\nu) \\ &\quad = \int p_\mu^{\mathbf{x}}(\mathbf{y}) \left[\log p_\mu^{\mathbf{x}}(\mathbf{y}) + \frac{1}{2}(\log|2\pi\mathbf{K}_\nu^{\mathbf{x}}| + (\mathbf{y} - \mathbf{m}_\nu^{\mathbf{x}})^\mathsf{T}(\mathbf{K}_\nu^{\mathbf{x}})^{-1}(\mathbf{y} - \mathbf{m}_\nu^{\mathbf{x}})) \right] \mathrm{d}\mathbf{y}. \end{aligned} \tag{D.7}$$

Bound

$$p_\mu^{\mathbf{x}}(\mathbf{y}) \leq \sup_{\mathbf{y}'} \mathcal{N}(\mathbf{y}; \mathbf{y}', \sigma^2) \leq (2\pi\sigma^2)^{-\frac{1}{2}n}. \tag{D.8}$$

Therefore, computing the rest of the expectation in closed form,

$$\mathrm{KL}(P_{\mathbf{x}}\mu, P_{\mathbf{x}}\nu) \leq \frac{1}{2}\Bigg(\underbrace{\log \frac{|\mathbf{K}_\nu^{\mathbf{x}}|}{\sigma^{2n}} + \mathrm{tr}((\mathbf{K}_\nu^{\mathbf{x}})^{-1}\mathbf{K}_\mu^{\mathbf{x}})}_{\text{(i)}} + \underbrace{(\mathbf{m}_\mu^{\mathbf{x}} - \mathbf{m}_\nu^{\mathbf{x}})^\mathsf{T}(\mathbf{K}_\nu^{\mathbf{x}})^{-1}(\mathbf{m}_\mu^{\mathbf{x}} - \mathbf{m}_\nu^{\mathbf{x}})}_{\text{(ii)}} \Bigg). \tag{D.9}$$

We separately bound (i) and (ii).

For (i), we use Von Neumann's Trace Inequality: for any two $n \times n$ positive semi-definite matrices $\mathbf{A}$ and $\mathbf{B}$, it holds that

$$\mathrm{tr}(\mathbf{AB}) \leq \sum_{i=1}^n \gamma_i(\mathbf{A})\gamma_i(\mathbf{B}). \tag{D.10}$$

Using this inequality,

$$\text{(i)} \leq \sum_{i=1}^n \left(\log \frac{\gamma_i(\mathbf{K}_\nu^{\mathbf{x}})}{\sigma^2} + \frac{\gamma_i(\mathbf{K}_\mu^{\mathbf{x}})}{\gamma_i(\mathbf{K}_\nu^{\mathbf{x}})} \right) \leq n \left(\log \frac{\gamma_1(\mathbf{K}_\nu^{\mathbf{x}})}{\sigma^2} + \frac{\gamma_1(\mathbf{K}_\mu^{\mathbf{x}})}{\gamma_n(\mathbf{K}_\nu^{\mathbf{x}})} \right). \tag{D.11}$$

Note that, by assumption, $\|\mathbf{m}_\mu^{\mathbf{x}}\|_\infty, \|\mathbf{m}_\nu^{\mathbf{x}}\|_\infty \leq M$ and $\gamma_n(\mathbf{K}_\mu^{\mathbf{x}}), \gamma_n(\mathbf{K}_\nu^{\mathbf{x}}) \geq \sigma^2$. Moreover, it is true that $\gamma_1(\mathbf{K}_\mu^{\mathbf{x}}) \leq \|\mathbf{K}_\mu^{\mathbf{x}}\|$ for any matrix norm $\|\bullet\|$. Taking this norm to be the ∞-norm

$\|\mathbf{K}_\mu^{\mathbf{x}}\|_\infty = \max_{i\in[n]} \sum_{j=1}^n |(\mathbf{K}_\mu^{\mathbf{x}})_{ij}|$, we see that $\gamma_1(\mathbf{K}_\mu^{\mathbf{x}}) \leq nM$; similarly, $\gamma_1(\mathbf{K}_\nu^{\mathbf{x}}) \leq nM$. Plugging in these estimates, we obtain

$$\text{(i)} \leq n\left(\log \frac{nM}{\sigma^2} + \frac{nM}{\sigma^2}\right) \leq \frac{2n^2 M}{\sigma^2}. \tag{D.12}$$

The bound for (ii) is simpler:

$$\text{(ii)} \leq \frac{n}{\gamma_n(\mathbf{K}_\nu^{\mathbf{x}})} \|\mathbf{m}_\mu^{\mathbf{x}} - \mathbf{m}_\nu^{\mathbf{x}}\|_\infty^2 \leq \frac{4nM^2}{\sigma^2}. \tag{D.13}$$

Combining the bounds for (i) and (ii) gives the desired result. ■

In the following, we will repeatedly make use of the following fact. Let $\mu \in \overline{\mathcal{P}}$ have a mean function and covariance function and let $\nu \in \overline{\mathcal{P}}_{\mathrm{G}}$. Then, for all $\mathbf{x} \in I$,

$$\mathrm{G}(P_{\mathbf{x}}\mu, P_{\mathbf{x}}\nu) = \mathrm{KL}(P_{\mathbf{x}}\mu, P_{\mathbf{x}}\nu) - \mathrm{KL}(P_{\mathbf{x}}\mu, P_{\mathbf{x}}\,\mathcal{GP}(\mu)) = \mathrm{KL}(P_{\mathbf{x}}\,\mathcal{GP}(\mu), P_{\mathbf{x}}\nu) \geq 0 \tag{D.14}$$

with equality if and only if $P_{\mathbf{x}}\,\mathcal{GP}(\mu) = P_{\mathbf{x}}\nu$. See App B for more details.

Prop D.4: Assume the assumptions of Prop D.3. Let $p(\mathbf{x})$ be a Borel distribution over $\mathcal{X}^n$ with full support. Then

$$\operatorname*{arg\,min}_{\nu\in\overline{\mathcal{P}}_{\mathrm{G}}} \mathbb{E}_{p(\mathbf{x})}[\mathrm{KL}(P_{\mathbf{x}}\mu, P_{\mathbf{x}}\nu)] = \operatorname*{arg\,min}_{\nu\in\overline{\mathcal{P}}_{\mathrm{G}}} \mathbb{E}_{p(\mathbf{x})}[\mathrm{G}(P_{\mathbf{x}}\mu, P_{\mathbf{x}}\nu)] = \mathcal{GP}(\mu). \tag{D.15}$$

Proof: By Prop D.3, $\mathbf{x} \mapsto \mathrm{KL}(P_{\mathbf{x}}\mu, P_{\mathbf{x}}\,\mathcal{GP}(\mu))$ bounded. Therefore, we can decompose

$$\begin{aligned}
&\mathbb{E}_{p(\mathbf{x})}[\mathrm{KL}(P_{\mathbf{x}}\mu, P_{\mathbf{x}}\nu)] \\
&\quad = \mathbb{E}_{p(\mathbf{x})}[\mathrm{KL}(P_{\mathbf{x}}\mu, P_{\mathbf{x}}\nu) - \mathrm{KL}(P_{\mathbf{x}}\mu, P_{\mathbf{x}}\,\mathcal{GP}(\mu)) + \mathrm{KL}(P_{\mathbf{x}}\mu, P_{\mathbf{x}}\,\mathcal{GP}(\mu))] && \text{(D.16)} \\
&\quad = \underbrace{\mathbb{E}_{p(\mathbf{x})}[\mathrm{G}(P_{\mathbf{x}}\mu, P_{\mathbf{x}}\nu)]}_{\text{(i)}} + \underbrace{\mathbb{E}_{p(\mathbf{x})}[\mathrm{KL}(P_{\mathbf{x}}\mu, P_{\mathbf{x}}\,\mathcal{GP}(\mu))]}_{\text{(ii)}}, && \text{(D.17)}
\end{aligned}$$

using (D.14). Here (i) measures how far $P_{\mathbf{x}}\nu$ is from the best Gaussian approximation of $P_{\mathbf{x}}\mu$ and (ii) measures the unavoidable approximation error due to the restriction to only Gaussian $P_{\mathbf{x}}\nu$. In particular, (i) is zero if and only if $\mathrm{G}(P_{\mathbf{x}}\mu, P_{\mathbf{x}}\nu) = \mathrm{KL}(P_{\mathbf{x}}\,\mathcal{GP}(\mu), P_{\mathbf{x}}\nu) = 0$ for almost all $\mathbf{x} \in \mathcal{X}^n$. Since, $\mathcal{GP}(\mu) \in \overline{\mathcal{P}}_{\mathrm{G}}$ and $\nu \in \overline{\mathcal{P}}_{\mathrm{G}}$, this is true if and only if $\mathcal{GP}(\mu) = \nu$ (Prop C.3), which proves the result. ■

Prop D.5: Let f be a noisy process and let π_f be the associated posterior prediction map. Let $\pi \in \overline{\mathcal{M}}_{\mathrm{G}}^{\text{f.d.d.}}$. Moreover, let $p(\mathbf{x})$ be a Borel distribution with full support over $\mathcal{X}^n$ for a fixed size $n \geq 2$, and let $p(D)$ be a Borel distribution with full support over a collection of data sets $\tilde{\mathcal{D}} \subseteq \mathcal{D}$. Assume the following:

(1) The collection of data sets $\tilde{\mathcal{D}}$ is bounded (Def C.5).

(2) The process f and prediction map π have uniformly bounded second moments:

$$\sup_{x\in\mathcal{X}} \mathbb{E}[f^2(x)] < \infty \quad \text{and} \quad \sup_{D\in\tilde{D}} \sup_{x\in\mathcal{X}} \mathbb{E}_{\pi(D)}[f^2(x)] < \infty.$$

(3) The process f is noisy with noise variance $\sigma^2 > 0$. Also, for all $D \in \tilde{\mathcal{D}}$, the process $\pi(D)$ is noisy with noise variance $\sigma_D^2 > 0$, and $\inf_{D \in \tilde{\mathcal{D}}} \sigma_D^2 > 0$.

Then

$$\mathbb{E}_{p(\mathbf{x})p(D)}[\mathrm{KL}(P_{\mathbf{x}}\pi_f(D), P_{\mathbf{x}}\pi(D))] < \infty. \tag{D.18}$$

Proof: To begin with, using Prop C.8, we confirm that $\pi_f \in \overline{\mathcal{M}} \subseteq \overline{\mathcal{M}}^{\text{f.d.d.}}$. To show (D.18), we show that the supremum over the bounds (D.4) is finite, which amounts to showing that (i) the data sets sizes are bounded, (ii) the collection of mean functions are covariance functions is uniformly bounded, and (iii) the collection of noise variances is bounded away from zero. These follow directly from respectively assumptions (1), (2) in combination with Prop C.9 and (1), and (3). ■

Prop D.6: Assume the assumptions of Prop D.5. Suppose that $\tilde{\mathcal{D}}$ is open. Then

$$\begin{aligned} &\operatorname*{arg\,min}_{\pi \in \overline{\mathcal{M}}_{\mathrm{G}}^{\text{f.d.d.}}} \mathbb{E}_{p(D)p(\mathbf{x})}[\mathrm{KL}(P_{\mathbf{x}}\pi_f(D), P_{\mathbf{x}}\pi(D))] \\ &\quad \overset{\tilde{\mathcal{D}}}{=} \operatorname*{arg\,min}_{\pi \in \overline{\mathcal{M}}_{\mathrm{G}}^{\text{f.d.d.}}} \mathbb{E}_{p(D)p(\mathbf{x})}[\mathrm{G}(P_{\mathbf{x}}\pi_f(D), P_{\mathbf{x}}\pi(D))] \overset{\tilde{\mathcal{D}}}{=} \mathcal{GP}(\pi_f) \end{aligned} \tag{D.19}$$

where the equalities hold for all $D \in \tilde{\mathcal{D}}$.

Proof: To begin with, using assumption (2) and Prop C.11, we confirm that $\mathcal{GP}(\pi_f) \in \overline{\mathcal{M}}_{\mathrm{G}}^{\text{f.d.d.}}$. Then

$$\mathbb{E}_{p(D)p(\mathbf{x})}[\mathrm{KL}(P_{\mathbf{x}}\pi_f(D), P_{\mathbf{x}}\,\mathcal{GP}(\pi_f)(D))] < \infty \tag{D.20}$$

by Prop D.5. Using this, decompose

$$\begin{aligned} &\mathbb{E}_{p(D)p(\mathbf{x})}[\mathrm{KL}(P_{\mathbf{x}}\pi_f(D), P_{\mathbf{x}}\pi(D))] \\ &\quad = \underbrace{\mathbb{E}_{p(D)p(\mathbf{x})}[\mathrm{G}(P_{\mathbf{x}}\pi_f(D), P_{\mathbf{x}}\pi(D))]}_{\text{(i)}} + \underbrace{\mathbb{E}_{p(D)p(\mathbf{x})}[\mathrm{KL}(P_{\mathbf{x}}\pi_f(D), P_{\mathbf{x}}\,\mathcal{GP}(\pi_f)(D))]}_{\text{(ii)}}, \end{aligned} \tag{D.21}$$

using (D.14). Here (i) measures how far $P_{\mathbf{x}}\pi(D)$ is from the best Gaussian approximatio of $P_{\mathbf{x}}\pi_f(D)$ and (ii) measures the unavoidable approximation error due to the restriction to only Gaussian $P_{\mathbf{x}}\pi(D)$. In particular, (i) is zero if and only if $\mathbb{E}_{p(\mathbf{x})}[\mathrm{G}(P_{\mathbf{x}}\pi_f(D), P_{\mathbf{x}}\pi(D))] = 0$ for almost all $D \in \tilde{\mathcal{D}}$. Consequently, by Prop D.4, (i) is zero if and only if $\mathcal{GP}(\pi_f)(D) = \pi(D)$ for almost all $D \in \tilde{\mathcal{D}}$. Using that $\tilde{\mathcal{D}}$ is open and that $p(D)$ has full support, this set of probability one is dense. Therefore, since $\mathcal{GP}(\pi_f) \in \overline{\mathcal{M}}_{\mathrm{G}}^{\text{f.d.d.}}$ and $\pi \in \overline{\mathcal{M}}_{\mathrm{G}}^{\text{f.d.d.}}$, (i) is zero if and only if $\mathcal{GP}(\pi_f)(D) = \pi(D)$ for all $D \in \tilde{\mathcal{D}}$ (Prop C.7), which proves the result. ■

Prop D.7: Assume the assumptions of Prop D.5. Let $(\pi_i)_{i \geq 1} \subseteq \overline{\mathcal{M}}_{\mathrm{G}}^{\text{f.d.d.}}$ be a minimising sequence for the infimum

$$\inf_{\pi \in \overline{\mathcal{M}}_{\mathrm{G}}^{\text{f.d.d.}}} \mathbb{E}_{p(D)p(\mathbf{x})}[\mathrm{KL}(P_{\mathbf{x}}\pi_f(D), P_{\mathbf{x}}\pi(D))]. \tag{D.22}$$

(i) Suppose that, for almost all $D \in \tilde{\mathcal{D}}$, $\pi_i(D)$ has a weak limit $\pi^*(D) \in \overline{\mathcal{P}}_{\mathrm{G}}$. Then $\pi^*(D) = \mathcal{GP}(\pi_f)(D)$ for almost all $D \in \tilde{\mathcal{D}}$.

(ii) Suppose that, for almost all $D \in \tilde{\mathcal{D}}$, $\pi_i(D)$ satisfies the following property: if there exist some $\mu \in \mathcal{P}$ and dense $\tilde{I} \subseteq I$ such that $P_{\mathbf{x}}\pi_i(D) \rightharpoonup P_{\mathbf{x}}\mu$ for all $\mathbf{x} \in \tilde{I}$, then $\pi_i(D) \rightharpoonup \mu$. Then there exists a subsequence $(\pi_{n_i})_{i\geq 1}$ of $(\pi_i)_{i\geq 1}$ such that $\pi_{n_i}(D) \rightharpoonup \mathcal{GP}(\pi_f)(D)$ for almost all $D \in \tilde{\mathcal{D}}$.

Proof: To begin with, by Prop D.5, $\mathbb{E}_{p(D)p(\mathbf{x})}[\mathrm{KL}(P_{\mathbf{x}}\pi_f(D), P_{\mathbf{x}}\,\mathcal{GP}(\pi_f)(D))] < \infty$. Therefore, decompose

$$\begin{aligned}&\mathbb{E}_{p(D)p(\mathbf{x})}[\mathrm{KL}(P_{\mathbf{x}}\pi_f(D), P_{\mathbf{x}}\pi_i(D))]\\ &\quad= \mathbb{E}_{p(D)p(\mathbf{x})}[\mathrm{G}(P_{\mathbf{x}}\pi_f(D), P_{\mathbf{x}}\pi_i(D))] + \mathbb{E}_{p(D)p(\mathbf{x})}[\mathrm{KL}(P_{\mathbf{x}}\pi_f(D), P_{\mathbf{x}}\,\mathcal{GP}(\pi_f)(D))]\end{aligned} \tag{D.23}$$

using (D.14). This shows that

$$\lim_{i\to\infty} \mathbb{E}_{p(D)p(\mathbf{x})}[\mathrm{G}(P_{\mathbf{x}}\pi_f(D), P_{\mathbf{x}}\pi_i(D))] = 0 \tag{D.24}$$

because $\mathbb{E}_{p(D)p(\mathbf{x})}[\mathrm{G}(P_{\mathbf{x}}\pi_f(D), P_{\mathbf{x}}\,\mathcal{GP}(\pi_f)(D))] = 0$ and $\mathcal{GP}(\pi_f) \in \overline{\mathcal{M}}_{\mathrm{G}}^{\text{f.d.d.}}$ ((D.14) and Prop D.5).

Prop D.7.(i): By Fatou's Lemma and the fact that $(\mu, \nu) \mapsto \mathrm{KL}(\mu, \nu)$ is weakly lower semi-continuous (Posner, 1975),

$$\liminf_{i\to\infty} \mathbb{E}_{p(D)p(\mathbf{x})}[\mathrm{KL}(P_{\mathbf{x}}\pi_f(D), P_{\mathbf{x}}\pi_i(D))] \geq \mathbb{E}_{p(D)p(\mathbf{x})}[\mathrm{KL}(P_{\mathbf{x}}\pi_f(D), P_{\mathbf{x}}\pi^*(D))]. \tag{D.25}$$

Therefore, by (D.23),

$$\liminf_{i\to\infty} \mathbb{E}_{p(D)p(\mathbf{x})}[\mathrm{G}(P_{\mathbf{x}}\pi_f(D), P_{\mathbf{x}}\pi_i(D))] \geq \mathbb{E}_{p(D)p(\mathbf{x})}[\mathrm{G}(P_{\mathbf{x}}\pi_f(D), P_{\mathbf{x}}\pi^*(D))]. \tag{D.26}$$

But the left-hand side is zero by (D.24), so the right-hand side is also zero:

$$\mathbb{E}_{p(D)p(\mathbf{x})}[\mathrm{G}(P_{\mathbf{x}}\pi_f(D), P_{\mathbf{x}}\pi^*(D))] = 0, \tag{D.27}$$

which yields that $\mathbb{E}_{p(\mathbf{x})}[\mathrm{G}(P_{\mathbf{x}}\pi_f(D), P_{\mathbf{x}}\pi^*(D))] = 0$ for almost all $D \in \tilde{\mathcal{D}}$. Consequently, using that $\pi^*(D) \in \overline{\mathcal{P}}_{\mathrm{G}}$, it follows that $\pi^*(D) = \mathcal{GP}(\pi_f)(D)$ for almost all $D \in \tilde{\mathcal{D}}$ (Prop D.4).

Prop D.7.(ii): Note that

$$\lim_{i\to\infty} \mathbb{E}_{p(D)p(\mathbf{x})}[\mathrm{G}(P_{\mathbf{x}}\pi_f(D), P_{\mathbf{x}}\pi_i(D))] = 0. \tag{D.28}$$

Therefore, there exists a collection of data sets $A \subseteq \tilde{\mathcal{D}}$ of probability one such that, along a subsequence,

$$\lim_{i\to\infty} \mathbb{E}_{p(\mathbf{x})}[\mathrm{G}(P_{\mathbf{x}}\pi_f(D), P_{\mathbf{x}}\pi_i(D))] = 0 \quad \text{for all } D \in A. \tag{D.29}$$

We show that $\pi_i(D) \rightharpoonup \mathcal{GP}(\pi_f)(D)$ for all $D \in A$. Let $D \in A$. Pass to a further subsequence of $(\pi_i(D))_{i\geq 1}$. It suffices to show that $(\pi_i(D))_{i\geq 1}$ contains a another further subsequence weakly convergent to $\mathcal{GP}(\pi_f)(D)$. Start with the observation that it still holds that

$$\lim_{i\to\infty} \mathbb{E}_{p(\mathbf{x})}[\mathrm{G}(P_{\mathbf{x}}\pi_f(D), P_{\mathbf{x}}\pi_i(D))] = 0. \tag{D.30}$$

Hence, there exists a collection of index sets $B \subseteq \mathcal{X}^n$ of probability one such that, along a further subsequence,

$$\lim_{i \to \infty} \mathrm{G}(P_{\mathbf{x}} \pi_f(D), P_{\mathbf{x}} \pi_i(D)) = 0 \quad \text{for all } \mathbf{x} \in B. \tag{D.31}$$

Consequently, by Pinsker's Inequality and (D.14), $P_{\mathbf{x}} \pi_i(D) \rightharpoonup \mathcal{N}(P_{\mathbf{x}} \pi_f(D)) = P_{\mathbf{x}} \mathcal{GP}(\pi_f)(D)$ for all $\mathbf{x} \in B$. In particular, using that $n \geq 2$, this means that almost all means and covariances converge, which in turn means that $P_{\mathbf{x}} \pi_i(D) \rightharpoonup P_{\mathbf{x}} \mathcal{GP}(\pi_f)(D)$ for all $\mathbf{x} \in \tilde{I}$, for some dense $\tilde{I} \subseteq I$. Therefore, by the assumed property, we conclude that $\pi_i(D) \rightharpoonup \mathcal{GP}(\pi_f)(D)$. ■

Appendix E. The Gaussian Neural Process

We build on the development from Secs 2 and 3, where we defined a Gaussian approximation $\tilde{\pi}\colon \mathcal{D} \to \mathcal{P}_{\mathrm{G}}$ of the posterior prediction map $\pi_f\colon \mathcal{D} \to \mathcal{P}$ corresponding some ground truth stationary stochastic process f. Recall that stationarity of f is equivalent to translation equivariance of π_f (Foong et al., 2020): for all $D \in \mathcal{D}$ and $\tau \in \mathcal{X}$,

$$\mathsf{T}_\tau \pi_f(D) = \pi_f(D + \tau) \tag{E.1}$$

where $\mathsf{T}_\tau f = f(\bullet - \tau)$ is the *shifting operator*, $\mathsf{T}_\tau \pi_f(D)$ is the measure $\pi_f(D)$ pushed through T_τ, and $D + \tau = (\mathbf{x}, \mathbf{y}) + \tau = ((x_1 + \tau, \ldots, x_{|\mathbf{x}|} + \tau), \mathbf{y})$.

Since the approximation $\tilde{\pi}$ is Gaussian, in a way that we now make precise, translation equivariance of $\tilde{\pi}$ is characterised by translation equivariance of the mean functions and kernel functions that $\tilde{\pi}$ maps to. For all $D \in \mathcal{D}$, denote $\tilde{\pi}(D) = \mathcal{GP}(m(D)(\bullet), k(D)(\bullet, \bullet))$. Then $\tilde{\pi}$ is translation equivariant if and only if, for all $D \in \mathcal{D}$ and $\tau \in \mathcal{X}$,

$$m(D + \tau)(\bullet) = m(D)(\bullet - \tau) \quad \text{and} \quad k(D + \tau)(\bullet, \bullet) = k(D)(\bullet - \tau, \bullet - \tau). \tag{E.2}$$

We proceed to find general parametrisations of the mean mapping m and kernel mapping k that then provide a general parametrisation of our approximation $\tilde{\pi}$.

Consider an arbitrary mean mapping $m\colon \mathcal{D} \to C(\mathcal{X}, \mathcal{Y})$ and kernel mapping $k\colon \mathcal{D} \to C^{\text{p.s.d.}}(\mathcal{X}^2, \mathcal{Y})$ where $C^{\text{p.s.d.}}(\mathcal{X}^2, \mathcal{Y})$ is the collection of continuous positive semi-definite functions $\mathcal{X}^2 \to \mathcal{Y}$. Suppose that these mappings satisfy (E.2), *i.e.* they are translation equivariant. Assume that m and k are continuous with respect to the metric on $\mathcal{D}$ (App A) and compact convergence on $C(\mathcal{X}, \mathcal{Y})$ and $C^{\text{p.s.d.}}(\mathcal{X}^2, \mathcal{Y})$.

The goal of this appendix is twofold: establish a universal representation for the kernel mapping k (App E.1) and an implementable neural architecture that can approximate this representation (App E.2). Moreover, in App E.3, we formulate an objective that can be used to train the parameters of this architecture.

E.1. Universal Representation of the Kernel Map

Before we turn our attention to the kernel mapping k, we review how Thm 1 by Gordon et al. (2020) can be used to establish a universal representation of the mean mapping m: for a collection of data sets $\tilde{\mathcal{D}} \subseteq \mathcal{D}$ that is topologically closed, closed under permutations, and closed under translations with finite maximum data set size and multiplicity $K \in \mathbb{N}$ (Def 2 by Gordon et al., 2020)—intuitively, the number of times an observation can occur at the same input is at most K—there exists a Hilbert space $\mathcal{H}$ of functions on $\mathcal{X}$, a continuous stationary kernel $\psi\colon \mathcal{X} \to \mathbb{R}$, a continuous $\phi\colon \mathcal{Y} \to \mathbb{R}^{K+1}$, and a continuous and translation-equivariant $\rho\colon \mathcal{H}' \to C(\mathcal{X}, \mathcal{Y})$ such that, for all $D \in \tilde{\mathcal{D}}$,

$$m(D) = \rho(E(D)) \quad \text{with} \quad E(\mathbf{x}, \mathbf{y}) = \sum_{i=1}^{|\mathbf{x}|} \phi(y_i)\psi(\bullet - x_i), \tag{E.3}$$

where $\mathcal{H}' = E(\tilde{\mathcal{D}}) \subseteq \mathcal{H}$ is a closed subset of $\mathcal{H}$.

We find a similar representation for the kernel mapping k by reducing it to a case where Thm 1 by Gordon et al. (2020) can be applied. Consider a data set $D = (\mathbf{x}, \mathbf{y}) \in \mathcal{D}$. Embed the data set in $\mathcal{D}_2 := \bigcup_{n=0}^{\infty} (\mathcal{X}^2 \times \mathcal{Y})^n$ by duplicating the inputs:

$$D' = (((x_1, x_1), \ldots, (x_{|\mathbf{x}|}, x_{|\mathbf{x}|})), \mathbf{y}) =: (\text{duplicate}(\mathbf{x}), \mathbf{y}). \tag{E.4}$$

Then k satisfies

$$k(D' + (\tau, \tau))(\bullet, \bullet) = k(D')(\bullet - \tau, \bullet - \tau) \quad \text{for all } (\tau, \tau) \in \mathcal{X}^2. \tag{E.5}$$

In other words, k can be viewed as a continuous function $\mathcal{D}_2 \to C(\mathcal{X}^2, \mathcal{Y})$ that is equivariant with respect to *diagonal translations*. If we can continuously extend k to be equivariant with respect to *all* translations, then are in a position to apply Thm 1 by Gordon et al. (2020), now for the input space $\mathcal{X}^2$.

We provide an explicit construction of this desired continuous extension. Set $\mathbf{e}_\parallel = (1, 1)/\sqrt{2} \in \mathcal{X}^2$ and $\mathbf{e}_\perp = (1, -1)/\sqrt{2} \in \mathcal{X}^2$. Then $\mathbf{e}_\parallel$ and $\mathbf{e}_\perp$ form an orthogonal basis for $\mathcal{X}^2$. For $\boldsymbol{\tau} \in \mathcal{X}^2$, let $\mathsf{T}_{\boldsymbol{\tau}} f = f(\bullet - \boldsymbol{\tau})$ be the shifting operator operating on functions on $\mathcal{X}^2$. Lift $k \colon \mathcal{D} \to C(\mathcal{X}, \mathcal{Y})$ to $\overline{k} \colon \mathcal{D}_2 \setminus \{\varnothing\} \to C(\mathcal{X}^2, \mathcal{Y})$ by setting

$$\overline{k}((\mathbf{x}_1, \ldots, \mathbf{x}_n), \mathbf{y}) = \begin{cases} k((x_{11}, \ldots, x_{n1}), \mathbf{y}) & \text{if } x_{i1} = x_{i2} \text{ for all } i \in [n], \\ 0 & \text{otherwise.} \end{cases} \tag{E.6}$$

Note that $D = \varnothing$ is excluded; we will turn to this case after Lem E.1. Then

$$\overline{k}((\mathbf{x}_1 - \boldsymbol{\tau}, \ldots, \mathbf{x}_n - \boldsymbol{\tau}), \mathbf{y}) = \mathsf{T}_{\boldsymbol{\tau}} \overline{k}((\mathbf{x}_1, \ldots, \mathbf{x}_n), \mathbf{y}) \quad \text{for all } \boldsymbol{\tau} = (\tau, \tau) \in \mathcal{X}^2, \tag{E.7}$$

which is the earlier established property that k is equivariant with respect to diagonal translations. Finally, we construct the desired extension $\hat{k} \colon \mathcal{D}_2 \setminus \{\varnothing\} \to C(\mathcal{X}^2, \mathcal{Y})$:

$$\hat{k}((\mathbf{x}_1, \ldots, \mathbf{x}_n), \mathbf{y}) = \mathsf{T}_{\langle \mathbf{e}_\perp, \mathbf{x}_\mathrm{c} \rangle \mathbf{e}_\perp} \overline{k}((\langle \mathbf{e}_\parallel, \mathbf{x}_1 \rangle \mathbf{e}_\parallel, \ldots, \langle \mathbf{e}_\parallel, \mathbf{x}_n \rangle \mathbf{e}_\parallel), \mathbf{y}) \tag{E.8}$$

where $\mathbf{x}_\mathrm{c} = \frac{1}{n} \sum_{i=1}^{n} \mathbf{x}_i$.

Lem E.1: The extended kernel mapping $\hat{k} \colon \mathcal{D}_2 \setminus \{\varnothing\} \to C(\mathcal{X}^2, \mathcal{Y})$

(i) extends k: it agrees with k on the embedding of $\mathcal{D} \setminus \{\varnothing\}$ in $\mathcal{D}_2$;

(ii) is permutation invariant and translation equivariant; and

(iii) is continuous with respect to the metric on $\mathcal{D}$ (App A) and compact convergence on $C(\mathcal{X}^2, \mathcal{Y})$.

Proof: Lem E.1.(i): If all inputs $\mathbf{x} = (x, x) \in \mathcal{X}^2$, then $\langle \mathbf{e}_\parallel, \mathbf{x}_1 \rangle \mathbf{e}_\parallel = \mathbf{x}$ and $\langle \mathbf{e}_\perp, \mathbf{x}_1 \rangle \mathbf{e}_\perp = \mathbf{0}$, so it is clear that $\hat{k}$ then agrees with k.

Lem E.1.(ii): That $\hat{k}$ is permutation invariant is clear. We check translation equivariance. Let $\tau \in \mathcal{X}^2$. Then $\tau = \tau_\parallel + \tau_\perp$ where $\tau_\parallel = \langle \mathbf{e}_\parallel, \tau \rangle \mathbf{e}_\parallel$ and $\tau_\perp = \langle \mathbf{e}_\perp, \tau \rangle \mathbf{e}_\perp$. Therefore,

$$\begin{aligned}
&\hat{k}((\mathbf{x}_1 + \tau, \ldots, \mathbf{x}_n + \tau), \mathbf{y}) && \text{(E.9)}\\
&\quad = \mathsf{T}_{\langle \mathbf{e}_\perp, \mathbf{x}_\mathrm{c} + \tau \rangle \mathbf{e}_\perp} \overline{k}((\langle \mathbf{e}_\parallel, \mathbf{x}_1 + \tau \rangle \mathbf{e}_\parallel, \ldots, \langle \mathbf{e}_\parallel, \mathbf{x}_n + \tau \rangle \mathbf{e}_\parallel), \mathbf{y}) && \text{(E.10)}\\
&\quad = \mathsf{T}_{\langle \mathbf{e}_\perp, \mathbf{x}_\mathrm{c} \rangle \mathbf{e}_\perp + \tau_\perp} \overline{k}((\langle \mathbf{e}_\parallel, \mathbf{x}_1 \rangle \mathbf{e}_\parallel + \tau_\parallel, \ldots, \langle \mathbf{e}_\parallel, \mathbf{x}_n \rangle \mathbf{e}_\parallel + \tau_\parallel), \mathbf{y}) && \text{(E.11)}\\
&\quad = \mathsf{T}_{\langle \mathbf{e}_\perp, \mathbf{x}_\mathrm{c} \rangle \mathbf{e}_\perp + \tau_\perp} \mathsf{T}_{\tau_\parallel} \overline{k}((\langle \mathbf{e}_\parallel, \mathbf{x}_1 \rangle \mathbf{e}_\parallel, \ldots, \langle \mathbf{e}_\parallel, \mathbf{x}_n \rangle \mathbf{e}_\parallel), \mathbf{y}) && \text{(E.12)}\\
&\quad = \mathsf{T}_{\tau} \mathsf{T}_{\langle \mathbf{e}_\perp, \mathbf{x}_\mathrm{c} \rangle \mathbf{e}_\perp} \overline{k}((\langle \mathbf{e}_\parallel, \mathbf{x}_1 \rangle \mathbf{e}_\parallel, \ldots, \langle \mathbf{e}_\parallel, \mathbf{x}_n \rangle \mathbf{e}_\parallel), \mathbf{y}) && \text{(E.13)}\\
&\quad = \mathsf{T}_{\tau} \hat{k}((\mathbf{x}_1, \ldots, \mathbf{x}_n), \mathbf{y}). && \text{(E.14)}
\end{aligned}$$

Since $\tau \in \mathcal{X}^2$ was arbitrary, this shows that $\hat{k}$ is translation equivariant.

Lem E.1.(iii): For $i \in [n]$, let $(\mathbf{x}_i^{(\ell)})_{\ell \geq 1} \subseteq \mathcal{X}^2$ be convergent to $\mathbf{x}_i \in \mathcal{X}^2$, and let $(\mathbf{y}_\ell)_{\ell \geq 1} \subseteq \mathcal{Y}^n$ be convergent to $\mathbf{y} \in \mathcal{Y}^n$. Set $\tau_\ell = \langle \mathbf{e}_\perp, \mathbf{x}_\mathrm{c}^{(\ell)} \rangle \mathbf{e}_\perp$, $\tau = \langle \mathbf{e}_\perp, \mathbf{x}_\mathrm{c} \rangle \mathbf{e}_\perp$,

$$f_\ell = \overline{k}((\langle \mathbf{e}_\parallel, \mathbf{x}_1^{(\ell)} \rangle \mathbf{e}_\parallel, \ldots, \langle \mathbf{e}_\parallel, \mathbf{x}_n^{(\ell)} \rangle \mathbf{e}_\parallel), \mathbf{y}_\ell), \quad f = \overline{k}((\langle \mathbf{e}_\parallel, \mathbf{x}_1 \rangle \mathbf{e}_\parallel, \ldots, \langle \mathbf{e}_\parallel, \mathbf{x}_n \rangle \mathbf{e}_\parallel), \mathbf{y}). \tag{E.15}$$

Then $\tau_\ell \to \tau$ and, by continuity of k, $f_\ell \to f$ compactly. Hence, it remains to show that $\mathsf{T}_{\tau_\ell} f_\ell \to \mathsf{T}_\tau f$ compactly. By convergence of $(\tau_\ell)_{\ell \geq 1}$, assume that the sequence $(\tau_\ell)_{\ell \geq 1}$ and limit τ are contained in $[-R, R]^2$ for some $R > 0$. Let $M > 0$ and consider $\mathbf{x} \in [-M, M]^2$. Set $K = [-(R+M), (M+R)]^2$. Estimate

$$\begin{aligned}
|\mathsf{T}_{\tau_\ell} f_\ell(\mathbf{x}) - \mathsf{T}_\tau f(\mathbf{x})| &\leq |f_\ell(\mathbf{x} - \tau_\ell) - f(\mathbf{x} - \tau_\ell)| + |f(\mathbf{x} - \tau_\ell) - f(\mathbf{x} - \tau)| && \text{(E.16)}\\
&\leq \underbrace{\sup_{\mathbf{z} \in K} |f_\ell(\mathbf{z}) - f(\mathbf{z})|}_{\text{(i)}} + \underbrace{\sup_{\mathbf{z}, \mathbf{z}' \in K^2,\, \|\mathbf{z} - \mathbf{z}'\|_2 \leq \|\tau - \tau_\ell\|_2} |f(\mathbf{z}) - f(\mathbf{z}')|}_{\text{(ii)}}. && \text{(E.17)}
\end{aligned}$$

Here (i) $\to 0$ because $f_\ell \to f$ compactly and (ii) $\to 0$ because f is continuous on K hence uniformly continuous on K (K is compact). We conclude that $\mathsf{T}_{\tau_\ell} f_\ell \to \mathsf{T}_\tau f$ compactly. ■

For $D = \varnothing$, we simply set $\hat{k}(\varnothing) = k(\varnothing)$. Note that there are no issues of continuity of $\hat{k}$ at $\varnothing$, because $\varnothing$ is an isolated point of $\tilde{\mathcal{D}}$ (App A).

Let $\tilde{\mathcal{D}} \subseteq \mathcal{D}$ be collection of data sets that is topologically closed, closed under permutations, and closed under translations with finite maximum data set size and multiplicity $K \in \mathbb{N}$. Let $\tilde{\mathcal{D}}_2$ be $\tilde{\mathcal{D}}$ embedded in $\mathcal{D}_2$ by duplicating the inputs and allowing for a translation:

$$\tilde{\mathcal{D}}_2 = \{(\text{duplicate}(\mathbf{x}), \mathbf{y}) + \tau, : (\mathbf{x}, \mathbf{y}) \in \tilde{\mathcal{D}},\, \tau \in \mathcal{X}^2\}. \tag{E.18}$$

Then also $\tilde{\mathcal{D}}_2$ is topologically closed, closed under permutations, and closed under translations, has finite maximum data set size, and has multiplicity K. Following Thm 1 by Gordon et al. (2020), let the encoding of a data set in $\mathcal{D}_2$ be

$$E_2 \colon \tilde{\mathcal{D}}_2 \to E_2(\tilde{\mathcal{D}}_2), \quad E((\mathbf{x}_1, \ldots, \mathbf{x}_n), \mathbf{y}) = \sum_{i=1}^n \phi(y_i) \psi(\bullet - \mathbf{x}_i), \tag{E.19}$$

where we abuse notation to immediately restrict E_2 to its image. According to Lems 1 to 4 by Gordon et al. (2020), $\mathcal{H}' = E(\tilde{\mathcal{D}}_2) \subseteq \mathcal{H}$ is a closed subset of $\mathcal{H}$, and E is a

translation-equivariant homeomorphism where the inverse recovers the input data set up to a permutation. Set $\rho = \hat{k} \circ E_2^{-1} \colon \mathcal{H}' \to C(\mathcal{X}^2, \mathcal{Y})$. Then, by Lem E.1.(i),

$$\rho(E(\text{duplicate}(\mathbf{x}), \mathbf{y})) = k(D) \text{ for all } D = (\mathbf{x}, \mathbf{y}) \in \tilde{\mathcal{D}}; \tag{E.20}$$

and, by Lem E.1.(iii), ρ is continuous. Moreover, by Lem E.1.(ii), ρ is translation equivariant on $E_2(\tilde{\mathcal{D}}_2 \setminus \{\varnothing\})$. The construction breaks down with translation equivariance of ρ at the zero function $E_2(\varnothing) = 0$. We discuss this issue next.

Suppose that ρ were also translation equivariant at the zero function $E_2(\varnothing) = 0$. Then $\rho(0) = \rho(\mathsf{T}_\tau 0) = \mathsf{T}_\tau \rho(0)$ for all $\tau \in \mathcal{X}^2$, which means that $\rho(E_2(\varnothing))$ must be a constant function. This is an issue, because $k(\varnothing)$ is not a constant function. We fix the issue by avoiding the zero function entirely. In particular, we increase the dimensionality of the embedding E_2 by one by concatenating some fixed continuous function $h \in C(\mathcal{X}^2, \mathcal{Y})$:

$$E_2 \colon \tilde{\mathcal{D}}_2 \to E_2(\tilde{\mathcal{D}}_2), \quad E((\mathbf{x}_1, \ldots, \mathbf{x}_n), \mathbf{y}) = \begin{bmatrix} \sum_{i=1}^n \phi(y_i)\psi(\bullet - \mathbf{x}_i) \\ \mathsf{T}_{\mathbf{x}_\mathrm{c}} h(\bullet) \end{bmatrix} \tag{E.21}$$

where we again abuse notation to immediately restrict E_2 to its image and set $\mathbf{x}_\mathrm{c} = \frac{1}{n}\sum_{i=1}^n \mathbf{x}_i$ if $n > 0$ and $\mathbf{x}_\mathrm{c} = \mathbf{0}$ otherwise. Then clearly $E_2(\tilde{\mathcal{D}}_2) \subseteq \mathcal{H} \times C(\mathcal{X}^2, \mathcal{Y})$ is still a closed subset of $\mathcal{H} \times C(\mathcal{X}^2, \mathcal{Y})$ and clearly E_2 is still a translation-equivariant homeomorphism. Again, set $\rho = \hat{k} \circ E_2^{-1}$. Then again ρ is continuous and agrees with k. The key difference is that $E_2(\varnothing)$ is now not equal to the zero function, so ρ can be extended to also be translation equivariant at $E_2(\varnothing)$: set $\rho(\mathsf{T}_\tau E_2(\varnothing)) := \mathsf{T}_\tau \rho(E(\varnothing)) = \mathsf{T}_\tau k(\varnothing)$ for all $\tau \in \mathcal{X}^2$. We need to make sure that this extension of ρ well defined. For a translation $\tau \in \mathcal{X}^2$, denote $\tau_\parallel = \langle \mathbf{e}_\parallel, \tau \rangle \mathbf{e}_\parallel$ and $\tau_\perp = \langle \mathbf{e}_\perp, \tau \rangle \mathbf{e}_\perp$.

Def E.1: Let $h \in C(\mathcal{X}^2, \mathcal{Y})$. Call h $\mathbf{e}_\perp$*-discriminating* if, for all translations $\tau_1 \in \mathcal{X}^2$ and $\tau_2 \in \mathcal{X}^2$, we have $\mathsf{T}_{\tau_1} h \neq \mathsf{T}_{\tau_2} h$ whenever $(\tau_1)_\perp \neq (\tau_2)_\perp$.

Lem E.2: Suppose that h is $\mathbf{e}_\perp$-discriminating. If $\tau_1 \in \mathcal{X}^2$ and $\tau_2 \in \mathcal{X}^2$ are two translations such that $\mathsf{T}_{\tau_1} E(\varnothing) = \mathsf{T}_{\tau_2} E(\varnothing)$, then $\mathsf{T}_{\tau_1} \rho(E(\varnothing)) = \mathsf{T}_{\tau_2} \rho(E(\varnothing))$. In other words, the extension of ρ is well defined.

Proof: Since $\mathsf{T}_{\tau_1} E(\varnothing) = \mathsf{T}_{\tau_2} E(\varnothing)$, in particular $\mathsf{T}_{\tau_1} h = \mathsf{T}_{\tau_2} h$. Therefore, using that h is $\mathbf{e}_\perp$-discriminating, $(\tau_1)_\perp = (\tau_2)_\perp$. Then

$$\mathsf{T}_{\tau_1} \rho(E(\varnothing)) = \mathsf{T}_{\tau_1} k(\varnothing) \overset{\text{(i)}}{=} \mathsf{T}_{(\tau_1)_\perp} k(\varnothing) \overset{\text{(ii)}}{=} \mathsf{T}_{(\tau_2)_\perp} k(\varnothing) \overset{\text{(i)}}{=} \mathsf{T}_{\tau_2} k(\varnothing) = \mathsf{T}_{\tau_2} \rho(E(\varnothing)), \tag{E.22}$$

using in (i) that $k(\varnothing)$ is invariant to diagonal translations and in (ii) that $(\tau_1)_\perp = (\tau_2)_\perp$. ■

We make one last simplifying assumption: let h be *invariant* to diagonal translations. Then

$$E_2(\text{duplicate}(\mathbf{x}), \mathbf{y}) = \begin{bmatrix} \sum_{i=1}^n \phi(y_i)\psi(\bullet - (x_i, x_i)) \\ h(\bullet) \end{bmatrix} \tag{E.23}$$

where $\{E_2(\text{duplicate}(\mathbf{x}), \mathbf{y}) : (\mathbf{x}, \mathbf{y}) \in \tilde{\mathcal{D}}\} = \mathcal{H}' \times \{h\}$ with $\mathcal{H}' \subseteq \mathcal{H}$ a closed subset of $\mathcal{H}$. We have proved the following theorem.

Thm E.1: Let $k\colon \mathcal{D} \to C^{\text{p.s.d.}}(\mathcal{X}^2, \mathcal{Y})$ be a continuous and translation-equivariant kernel mapping. Let $\tilde{\mathcal{D}} \subseteq \mathcal{D}$ be collection of data sets that is topologically closed, closed under permutations, and closed under translations with finite maximum data set size and multiplicity $K \in \mathbb{N}$. Set $\phi\colon \mathcal{Y} \to \mathbb{R}^{K+1}$, $\phi(y) = (y^0, y^1, \ldots, y^K)$. Choose any $h \in C(\mathcal{X}^2, \mathcal{Y})$ that is $\mathbf{e}_\perp$-discriminating and invariant with respect to diagonal translations. Then there exists a reproducing kernel Hilbert space $\mathcal{H}$ of functions on $\mathcal{X}^2$, a continuous stationary kernel $\psi\colon \mathcal{X} \to \mathbb{R}$, and a continuous and translation-equivariant $\rho\colon \mathcal{H}' \to C(\mathcal{X}^2, \mathcal{Y})$ such that, for all $D \in \tilde{\mathcal{D}}$,

$$k(D) = \rho(E(D)) \quad \text{with} \quad E(\mathbf{x}, \mathbf{y}) = \begin{bmatrix} \sum_{i=1}^{n} \phi(y_i)\psi(\bullet - (x_i, x_i)) \\ h(\bullet) \end{bmatrix} \tag{E.24}$$

where $\mathcal{H}' = E_2(\tilde{\mathcal{D}}_2) \subseteq \mathcal{H} \times C(\mathcal{X}^2, \mathcal{Y})$ is a closed subset of $\mathcal{H} \times C(\mathcal{X}^2, \mathcal{Y})$.

We point out is that ρ is only defined on the closed subset $\mathcal{H}'$. Using a generalisation of Tietze's Extension Theorem by Dugundji (1951), ρ can perhaps be continuously extended to the entirety of $\mathcal{H}$ and an appropriate space containing h, and it appears possible to perform this extension whilst preserving translation equivariance, see *e.g.* the thesis by Feragen (2006). We leave these investigations for future work.

E.2. Implementation of the Kernel Map Representation

In this section, we establish an implementable neural architecture that can approximate the representation in Thm E.1. The key observation is that ρ is a translation-equivariant map between two function spaces. Therefore, if we discretise the functions finely enough, then it appears plausible that the resulting map between discretisations can be approximated by a CNN. We will not make this statement precise; rather, we point the reader to Yarotsky (2018) for the universal approximation properties of CNNs.

We formalise our approximating architecture. We start out by approximating ρ. Let $\mathbf{Z} \in \mathbb{R}^{M \times M}$ be a sufficiently fine grid on $\mathcal{X}^2$. Then, assuming a universal approximation capability of CNNs convenient for our purpose, let CNN be such that

$$\|\rho(e)(\mathbf{Z}) - \text{CNN}(e(\mathbf{Z}))\|_{\text{F}} < \varepsilon \quad \text{for all } e \in E(\tilde{\mathcal{D}}), \tag{E.25}$$

for some chosen level of approximation accuracy $\varepsilon > 0$. Although $\text{CNN}(e(\mathbf{Z}))$ approximates $\rho(e)(\mathbf{Z})$ satisfactorily, it is not guaranteed that $\text{CNN}(e(\mathbf{Z}))$ is a positive semi-definite or even symmetric matrix, which is required for our applications. To fix this, let $\Pi^{\text{p.s.d.}}$ be the operator that takes a matrix to the positive semi-definite matrix closest in Frobenius norm (Higham, 1988). Let $e \in E(\tilde{\mathcal{D}})$. Then

$$\begin{aligned} &\|\rho(e)(\mathbf{Z}) - \Pi^{\text{p.s.d.}} \text{CNN}(e(\mathbf{Z}))\|_{\text{F}} \\ &\quad \leq \|\rho(e)(\mathbf{Z}) - \text{CNN}(e(\mathbf{Z}))\|_{\text{F}} + \|\text{CNN}(e(\mathbf{Z})) - \Pi^{\text{p.s.d.}} \text{CNN}(e(\mathbf{Z}))\|_{\text{F}}. \end{aligned} \tag{E.26}$$

The first term is less than ε by choice of CNN, and it is readily seen that the second term is also less than ε:

$$\|\text{CNN}(e(\mathbf{Z})) - \Pi^{\text{p.s.d.}} \text{CNN}(e(\mathbf{Z}))\|_{\text{F}} \leq \|\text{CNN}(e(\mathbf{Z})) - \rho(e)(\mathbf{Z})\|_{\text{F}} < \varepsilon \tag{E.27}$$

where "$\leq$" follows from the definition of $\Pi^{\text{p.s.d.}}$ and the observation that $\rho(e)(\mathbf{Z})$ is positive semi-definite. Therefore,

$$\|\rho(e)(\mathbf{Z}) - \Pi^{\text{p.s.d.}}\,\text{CNN}(e(\mathbf{Z}))\|_{\text{F}} < 2\varepsilon \quad \text{for all } e \in E(\tilde{\mathcal{D}}), \tag{E.28}$$

which means that $\Pi^{\text{p.s.d.}}\,\text{CNN}$ is also a good approximation of ρ. Crucially, the output $\Pi^{\text{p.s.d.}}\,\text{CNN}$ is always positive semi definite, suited for our applications. To complete the architecture, we assume that the discretisation $\mathbf{Z}$ is sufficiently fine and far-reaching so that we can reasonably construct the covariance between any two points $x_1 \in \mathcal{X}$ and $x_2 \in \mathcal{X}$ of interest simply through interpolation:

$$k(D)(x_1, x_2) \approx \sum_{i=1}^{M}\sum_{j=1}^{M} \hat{\psi}(x_1 - z_1)\hat{\psi}(x_2 - z_2)[\Pi^{\text{p.s.d.}}\,\text{CNN}(E(D)(\mathbf{Z}))]_{ij} \tag{E.29}$$

$$= \langle \hat{\boldsymbol{\psi}}(x_1), \Pi^{\text{p.s.d.}}\,\text{CNN}(E(D)(\mathbf{Z}), h(\mathbf{Z}))\hat{\boldsymbol{\psi}}(x_2)\rangle \tag{E.30}$$

where $Z_{ij} = (z_i, z_j)$, $\hat{\psi}$ is some suitable interpolation kernel, and

$$\hat{\boldsymbol{\psi}}(x) = (\hat{\psi}(x - z_1), \ldots, \hat{\psi}(x - z_M)) \in \mathbb{R}^M \quad \text{for all } x \in \mathcal{X}. \tag{E.31}$$

From (E.30) it is clear that the approximation of $k(D)(x_1, x_2)$ is indeed a positive semi-definite function.

Assume that the multiplicity K of the collection of data sets is one: $K = 1$. Moreover, let $h \in \mathcal{H}$ be such that that $h(\mathbf{Z}) = \mathbf{I}$; indeed, this choice for $h \in \mathcal{H}$ is $\mathbf{e}_\perp$-discriminating and invariant with respect to diagonal translations. Let $D^{(\text{c})} = (\mathbf{x}^{(\text{c})}, \mathbf{y}^{(\text{c})}) \in \mathcal{D}$ be a data set and let $\mathbf{x}^{(\text{t})} \in \mathcal{X}^{|\mathbf{x}^{(\text{t})}|}$ be some target inputs. We then summarise the architecture in the following three-step procedure:

❶ Run the encoder $\mathbf{H} = \text{enc}(D^{(\text{c})}, \mathbf{Z})$, which is defined to produce the following three channels:

$$\underbrace{\mathbf{H}_{::1} = \sum_{i=1}^{|\mathbf{x}|} y_i^{(\text{c})}\psi(\mathbf{Z} - (x_i^{(\text{c})}, x_i^{(\text{c})}))}_{\text{data channel}}, \quad \underbrace{\mathbf{H}_{::2} = \sum_{i=1}^{|\mathbf{x}|} \psi(\mathbf{Z} - (x_i^{(\text{c})}, x_i^{(\text{c})}))}_{\text{density channel}}, \quad \underbrace{\mathbf{H}_{::3} = \mathbf{I}}_{\text{source channel}}. \tag{E.32}$$

❷ Pass the encoding through a CNN, which outputs a single channel, and map it to the closest positive semi-definite matrix:

$$\mathbf{K} = \Pi^{\text{p.s.d.}}\,\text{CNN}(\mathbf{H}). \tag{E.33}$$

❸ Run the decoder to interpolate the covariances to the target inputs:

$$k(D)(\mathbf{x}^{(\text{t})}, \mathbf{x}^{(\text{t})}) \approx \text{dec}(\mathbf{K}, \mathbf{x}^{(\text{t})}) := \psi'(\mathbf{x}^{(\text{t})}, \mathbf{z}^{\mathsf{T}})\mathbf{K}\psi'(\mathbf{z}, \mathbf{x}^{(\text{t})\mathsf{T}}). \tag{E.34}$$

The three channels in ❶ all serve a distinct but crucial service: The *data channel* communicates the values of the observed data to the model. However, as pointed out by Gordon

et al. (2020), the model will not be able to distinguish between a value $y_i = 0$ and no observation. This is what the *density channels* fixes: it communicates to the model where the data is present and where data is missing. Finally, the *source channel* allows the architecture to "start out" with a stationary prior with covariance $\mathbf{I}$, which corresponds to white noise, and then pass it through a CNN to modulate this prior to introduce correlations inferred from the context set. *C.f.*, any Gaussian process can be sampled from by first sampling white noise and then convolving this noise with an appropriate filter; the kernel architecture is a nonlinear generalisation of this procedure.

E.3. Training Objective

The ConvDeepSet for the mean mapping m and the architecture described in App E.2 for the kernel mapping k form the model that we call the Gaussian Neural Process (GNP). The GNP depends on some parameters θ, *e.g.* weights and biases for the CNNs. To train these parameters θ, we maximise the objective (5) from Sec 2:

$$\theta^* = \arg\max_{\theta} \frac{1}{N} \sum_{i=1}^{N} \log \mathcal{N}(\mathbf{y}_i^{(t)} \mid \mathbf{m}_\theta(D_i^{(c)}, \mathbf{x}_i^{(t)}), \mathbf{K}_\theta(D_i^{(c)}, \mathbf{x}_i^{(t)})), \tag{E.35}$$

where $\mathbf{m}_\theta(D_i^{(c)}, \mathbf{x}_i^{(t)})$ and $\mathbf{K}_\theta(D_i^{(c)}, \mathbf{x}_i^{(t)})$ implement the ConvDeepSet architecture and kernel architecture from App E.2 respectively, producing, from the context set $D_i^{(c)}$, a mean and covariance matrix for the target inputs $\mathbf{x}_i^{(t)}$. In practice, we use minibatching in combination with stochastic gradient descent and the adaptive step size method ADAM (Kingma and Ba, 2015).

Appendix F. Experimental Setup

We follow the experimental setup of Foong et al. (2020), with the following exceptions:

(1) We include an additional task Mixture, which samples from EQ, Matern-$\frac{5}{2}$, Noisy Mixture, Weakly Per., or Sawtooth with equal probability and thus consistutes a highly non-Gaussian mixture process.

(2) In line with the theoretical analysis in Apps C and D, we contaminate all data samples with $\mathcal{N}(0, 0.05^2)$-noise.

(3) The margin of the ConvCNP is reduced to 0.1.

The architecture for GNP is analogous to the architecture for ConvCNP, with the following exceptions:

(1) To alleviate memory requirements, the receptive field size is limited to 8.

(2) To alleviate memory requirements, the points per unit is decreased to 20.

Since the tasks are contaminated with noise, we extend the last step ❸ in the kernel the architecture (see App E.2) to also include a term for homogeneous noise, $\mathbf{K}^{(\mathrm{t})} = \mathrm{dec}(\mathbf{K}, \mathbf{x}^{(\mathrm{t})}) + \sigma^2 \mathbf{I}$ where $\sigma^2 > 0$ is a learnable parameter, which comes with the added benefit of stabilising the numerics during training.

Tab F.1 show the parameter count for all models in all tasks in the 1D experiments. The models were trained for roughly five days on a Tesla V100 GPU; Tab F.2 shows the timings of a single epoch for all models in all tasks. Note that, for the computationally more expensive tasks Sawtooth and Mixture, the timing of an epoch for the GNP increases drastically, which is likely due to excessive allocations on the GPU.

Tab 1 shows (where applicable) the performance of the ground-truth Gaussian process (GP), the ground-truth Gaussian process without correlations (GP (diag.)), the GNP, the ConvCNP (Gordon et al., 2020), the ConvNP (Foong et al., 2020), and the ANP (Kim et al., 2019) on all five data sets in (1) an interpolation setup, (2) an interpolation setup where the model was not trained, testing generalisation capability, and (3) and extrapolation setup, also testing generalisation capability. Note that the translation equivariance built into the ConvCNP, ConvNP, and GNP enables these models to to maintain their performance when evaluated outside of the training range.

	EQ	Matérn–$\frac{5}{2}$	Weakly Per.	Sawtooth	Mixture
GNP	101 961	101 961	147 017	360 009	360 009
ConvCNP	42 822	42 822	51 014	100 166	100 166
ConvNP	88 486	88 486	104 870	104 870	203 174
ANP	530 178	530 178	530 178	530 178	530 178

Table F.1: Parameter counts for the GNP, ConvCNP, ConvNP, and ANP in the 1D regression tasks

	EQ	Matérn–$\frac{5}{2}$	Weakly Per.	Sawtooth	Mixture
GNP	145	145	275	1000	1035
ConvCNP	30	30	40	25	60
ConvNP	85	85	110	160	190
ANP	40	40	50	50	75

Table F.2: Rough estimate of the numbers of seconds required for a single epoch of the GNP, ConvCNP, ConvNP, and ANP on a Tesla V100 GPU in the 1D regression tasks

	EQ	Matérn–$\frac{5}{2}$	Weakly Per.	Sawtooth	Mixture
Interpolation inside training range					
GP	$0.70 \pm 4.8\text{E}{-3}$	$0.31 \pm 4.8\text{E}{-3}$	$-0.32 \pm 4.3\text{E}{-3}$	n/a	n/a
GP (diag.)	-0.81 ± 0.01	-0.93 ± 0.01	$-1.18 \pm 7.0\text{E}{-3}$	n/a	n/a
GNP	$0.70 \pm 5.0\text{E}{-3}$	$0.30 \pm 5.0\text{E}{-3}$	$-0.47 \pm 5.0\text{E}{-3}$	0.42 ± 0.01	0.10 ± 0.02
ConvCNP	-0.80 ± 0.01	-0.95 ± 0.01	$-1.20 \pm 7.0\text{E}{-3}$	0.55 ± 0.02	-0.93 ± 0.02
ConvNP	-0.46 ± 0.01	$-0.67 \pm 9.0\text{E}{-3}$	$-1.02 \pm 6.0\text{E}{-3}$	$1.20 \pm 7.0\text{E}{-3}$	-0.50 ± 0.02
ANP	-0.61 ± 0.01	-0.75 ± 0.01	$-1.19 \pm 5.0\text{E}{-3}$	$0.34 \pm 7.0\text{E}{-3}$	-0.69 ± 0.02
Interpolation beyond training range					
GP	$0.70 \pm 4.8\text{E}{-3}$	$0.31 \pm 4.8\text{E}{-3}$	$-0.32 \pm 4.3\text{E}{-3}$	n/a	n/a
GP (diag.)	-0.81 ± 0.01	-0.93 ± 0.01	$-1.18 \pm 7.0\text{E}{-3}$	n/a	n/a
GNP	$0.69 \pm 5.0\text{E}{-3}$	$0.30 \pm 5.0\text{E}{-3}$	$-0.47 \pm 5.0\text{E}{-3}$	0.42 ± 0.01	0.10 ± 0.02
ConvCNP	-0.81 ± 0.01	-0.95 ± 0.01	$-1.20 \pm 7.0\text{E}{-3}$	0.53 ± 0.02	-0.96 ± 0.02
ConvNP	-0.46 ± 0.01	$-0.67 \pm 9.0\text{E}{-3}$	$-1.02 \pm 6.0\text{E}{-3}$	$1.19 \pm 7.0\text{E}{-3}$	-0.53 ± 0.02
ANP	$-1.42 \pm 6.0\text{E}{-3}$	$-1.34 \pm 6.0\text{E}{-3}$	$-1.33 \pm 4.0\text{E}{-3}$	$-0.17 \pm 2.0\text{E}{-3}$	-1.24 ± 0.01
Extrapolation beyond training range					
GP	$0.44 \pm 2.9\text{E}{-3}$	$0.09 \pm 3.1\text{E}{-3}$	$-0.52 \pm 3.4\text{E}{-3}$	n/a	n/a
GP (diag.)	$-1.40 \pm 6.7\text{E}{-3}$	$-1.41 \pm 6.6\text{E}{-3}$	$-1.41 \pm 5.6\text{E}{-3}$	n/a	n/a
GNP	$0.44 \pm 3.0\text{E}{-3}$	$0.08 \pm 3.0\text{E}{-3}$	$-0.62 \pm 4.0\text{E}{-3}$	$0.04 \pm 9.0\text{E}{-3}$	-0.07 ± 0.01
ConvCNP	$-1.41 \pm 7.0\text{E}{-3}$	$-1.42 \pm 6.0\text{E}{-3}$	$-1.41 \pm 6.0\text{E}{-3}$	$0.06 \pm 8.0\text{E}{-3}$	-1.36 ± 0.02
ConvNP	$-1.11 \pm 5.0\text{E}{-3}$	$-1.12 \pm 5.0\text{E}{-3}$	$-1.23 \pm 4.0\text{E}{-3}$	$0.88 \pm 9.0\text{E}{-3}$	-0.93 ± 0.01
ANP	$-1.31 \pm 5.0\text{E}{-3}$	$-1.28 \pm 5.0\text{E}{-3}$	$-1.32 \pm 5.0\text{E}{-3}$	$-0.17 \pm 1.0\text{E}{-3}$	-1.11 ± 0.01

Table F.3: Full results for 1D regression experiments. The numbers are average target point likelihood under the predictive distribution conditioned on the context set. The errors are 95%-confidence intervals. See Foong et al. (2020) for more details.