# Distributional Deep Gaussian Processes

**Sebastian G. Popescu**[1]

[1]Find & Order Labs

## Abstract

Deep Gaussian processes (DGPs) offer a principled Bayesian framework with hierarchical uncertainty propagation, but their reliable propagation of uncertainty and out-of-distribution (OOD) detection performance remains underexplored and often unreliable in safety-critical settings. In this work, we propose a novel kernel operating in both Euclidean and Wasserstein-2 space to better account for the geometry of representation learning spaces, thus circumventing a common pathology called feature collapse, whereby inliers and outliers get mapped to similar spaces. Empirically, our approach consistently improves OOD detection in convolutional image tasks and shows improved performance on tabular datasets.

## 1 INTRODUCTION

Gaussian Processes (GPs) are flexible Bayesian nonparametric models that provide well-calibrated uncertainty estimates and are robust to overfitting [Williams and Rasmussen, 2006], making them particularly suited to data-scarce settings [Timonen et al., 2019, Wang et al., 2020] and safety-critical applications where uncertainty quantification is essential [Chen et al., 2014]. Deep Gaussian Processes (DGPs) [Damianou and Lawrence, 2013] extend GPs into a multi-layer compositional framework, retaining their data-efficiency while substantially increasing representational capacity.

Scalable inference in DGPs has been the primary focus of recent work. Building on sparse variational methods for shallow GPs [Titsias, 2009, Hensman et al., 2013], Salimbeni and Deisenroth [2017] introduced doubly stochastic variational inference (DSVI), which avoids enforcing independence between layers and scales to large datasets. Subsequent work has explored alternative inference schemes, including stochastic gradient Hamiltonian Monte Carlo [Havasi et al., 2018], importance-weighted objectives [Salimbeni et al., 2019], neural-network-parameterised posteriors [Yu et al., 2019] or global inducing points approaches [Ober and Aitchison, 2021b]. Despite this progress, the quality and reliability of uncertainties propagated through the DGP hierarchy has received comparatively little attention.

Ustyuzhaninov et al. [2020] showed that factorised Gaussian variational posteriors cause all but the last GP layer to collapse to near-deterministic transformations in the noise-free regime, calling into question whether DGPs produce meaningful hierarchical uncertainties. Separately, reliably detecting out-of-distribution (OOD) inputs, whether as domain-shifted samples [Lakshminarayanan et al., 2017] or genuine anomalies [Hendrycks and Gimpel, 2016], has emerged as a central challenge for trustworthy ML systems. Malinin and Gales [2018] formalise a useful distinction: *epistemic* uncertainty reflects parameter uncertainty within the data manifold, while *distributional* uncertainty signals that a test point lies outside the training distribution entirely, the critical "unknown-unknown" for OOD detection. Whether DGPs can reliably quantify and propagate distributional uncertainty across layers remains an open question.

Existing work on DGPs for OOD detection is limited. Domingues et al. [2018] use a DGP autoencoder for reconstruction-based anomaly detection, but this sidesteps the question of whether predictive DGPs are inherently distance-aware. We address this directly: we characterise the theoretical conditions under which DGPs preserve distributional uncertainty for OOD inputs across layers, identify the pathological regimes where they fail, and propose a remedy.

Concretely, we introduce a kernel that operates jointly in Euclidean and Wasserstein-2 space, enabling explicit comparison of the hidden-layer Gaussian marginals that DGPs produce. Prior work on Wasserstein-2 kernels [Bachoc et al., 2017, 2018, Thi Thien Trang et al., 2019] has been confined to shallow GPs on distributional inputs; we are the first to extend this to the multi-layer setting. The resulting

*Proceedings of the 42nd Conference on Uncertainty in Artificial Intelligence* (UAI 2026), PMLR 337:5504–5540.

model, Distributional Deep Gaussian Processes (DDGPs), preserves higher distributional uncertainty for OOD inputs at every layer of the hierarchy under conditions we make explicit, and which the GIPL parameterisation is constructed to enforce, thus addressing the failure modes of standard DGPs.

**Contributions.**

- We derive necessary conditions for DGPs to propagate higher total and non-parametric uncertainty for OOD inputs across layers, and show empirically that non-parametric uncertainty collapses in hidden layers under standard inference.
- We identify PCA mean functions as a partial remedy for this collapse, while showing they introduce a new failure mode: erroneously extrapolating low distributional uncertainty beyond the training manifold.
- We propose DDGPs, a novel DGP variant whose hybrid Euclidean & Wasserstein-2 kernel guarantees, under the condition $\Sigma_{Z_{l-1}} \approx V_{in}$, higher non-parametric uncertainty for OOD inputs at each layer; the GIPL parameterisation enforces this condition structurally rather than relying on free-parameter optimisation to discover it.
- We identity DDGPs as more robust to *feature collapse* than standard DGPs.
- We demonstrate empirically that DDGPs achieve superior OOD detection and improved predictive performance across a range of regression and classification benchmarks.

# 2 BACKGROUND

We introduce the theoretical background for this paper.

## 2.1 SPARSE VARIATIONAL GAUSSIAN PROCESSES

Following Hensman et al. [2013], we augment the GP prior with $M$ inducing points $\mathbf{Z} = \{z_m\}_{m=1}^M$ and corresponding inducing variables $\mathbf{u} = f(\mathbf{Z}) \in \mathbb{R}^M$. We approximate the true posterior $p(\mathbf{f}, \mathbf{u} \mid \mathbf{y})$ with a variational distribution $q(\mathbf{f}, \mathbf{u}) = p(\mathbf{f} \mid \mathbf{u})\, q(\mathbf{u})$, where $q(\mathbf{u}) = \mathcal{N}(\mathbf{u}; \mathbf{m}, \mathbf{S})$ with free parameters $\{\mathbf{m}, \mathbf{S}\}$. This yields the *sparse variational GP* (SVGP) evidence lower bound (ELBO):

$$\log p(\mathbf{y}) \geq \mathcal{L}_{\text{SVGP}} \coloneqq \sum_{i=1}^{N} \mathbb{E}_{q(f_i)}[\log p(y_i \mid f_i)] - \mathrm{KL}[q(\mathbf{u}) \,\|\, p(\mathbf{u})]\,, \tag{1}$$

where the marginal $q(f_i) = \mathcal{N}(f_i;\, \tilde{\mu}_i,\, \tilde{\sigma}_i^2)$ is obtained by integrating out $\mathbf{u}$. The predictive mean and variance decompose into a *parametric* component $g(\mathbf{x})$ anchored to the inducing points, and a *non-parametric* residual $h(\mathbf{x})$ drawn from a GP:

$$g(\mathbf{x}) = \mathbf{K}_{fu}\mathbf{K}_{uu}^{-1}\mathbf{m}, \tag{2}$$

$$h(\mathbf{x}) \sim \mathcal{GP}\big(\mathbf{0},\ \mathbf{K}_{ff} - \mathbf{K}_{fu}\mathbf{K}_{uu}^{-1}\mathbf{K}_{uf}\big)\,, \tag{3}$$

so that $\tilde{\boldsymbol{\mu}} = \mathbb{E}[g(\mathbf{x})]$ and $\tilde{\boldsymbol{\Sigma}} = \mathrm{Cov}[g(\mathbf{x})] + \mathrm{Cov}[h(\mathbf{x})]$. Here $g$ is a finite-rank function fully determined by the $M$ inducing variables, while $h$ captures the prior uncertainty not explained by the inducing points, with variance recovering $\sigma^2$ far from their support.

## 2.2 DEEP GAUSSIAN PROCESSES

A Deep GP [Damianou and Lawrence, 2013] is a composition of $L$ GP-distributed functions, $\mathbf{f}_L(\mathbf{x}) = \mathbf{f}_L \circ \cdots \circ \mathbf{f}_1(\mathbf{x})$, with each layer $\mathbf{f}_l \sim \mathcal{GP}(m_l, k_l)$. The joint prior factorises as:

$$p\big(\mathbf{y}, \{\mathbf{f}_l\}_{l=1}^L \mid \mathbf{X}\big) = p(\mathbf{y} \mid \mathbf{f}_L) \prod_{l=1}^{L} p(\mathbf{f}_l \mid \mathbf{f}_{l-1})\,, \tag{4}$$

where $p(\mathbf{f}_l \mid \mathbf{f}_{l-1}) = \mathcal{GP}(m_l(\mathbf{f}_{l-1}),\, k_l(\mathbf{f}_{l-1}, \mathbf{f}_{l-1}))$ and we adopt a squared-exponential kernel with layer-wise hyperparameters $\boldsymbol{\theta}_l = \{\sigma_l^2, \ell_{l,1}^2, \ldots, \ell_{l,D_l}^2\}$. Exact inference is intractable due to the non-Gaussian propagation of uncertainty across layers.

Following the SVGP construction in §2.1, we augment each layer with $M_l$ inducing points $(\mathbf{Z}_{l-1}, \mathbf{U}_l)$ and introduce a mean-field variational posterior:

$$q\big(\{\mathbf{U}_l\}_{l=1}^L\big) = \prod_{l=1}^{L} \mathcal{N}(\mathbf{U}_l;\, \mathbf{m}_l,\, \mathbf{S}_l)\,. \tag{5}$$

Marginalising the inducing variables yields the DGP ELBO:

$$\log p(\mathbf{y}) \geq \mathcal{L}_{\text{DGP}} \coloneqq \mathbb{E}_{\prod_l q(\mathbf{f}_l \mid \mathbf{f}_{l-1})}[\log p(\mathbf{y} \mid \mathbf{f}_L)] - \sum_{l=1}^{L} \mathrm{KL}[q(\mathbf{U}_l) \,\|\, p(\mathbf{U}_l)]\,, \tag{6}$$

where each inter-layer marginal $q(\mathbf{f}_l \mid \mathbf{f}_{l-1}) = \mathcal{N}(\mathbf{f}_l;\, \tilde{\boldsymbol{\mu}}_l,\, \tilde{\boldsymbol{\Sigma}}_l)$ follows the SVGP predictive equations (2)–(3) with mean-corrected inducing targets:

$$\tilde{\boldsymbol{\mu}}_l = m_l(\mathbf{f}_{l-1}) + \mathbf{K}_{fu}\mathbf{K}_{uu}^{-1}\left[\mathbf{m}_l - m_l(\mathbf{Z}_{l-1})\right], \tag{7}$$

$$\tilde{\boldsymbol{\Sigma}}_l = \mathbf{K}_{ff} - \mathbf{K}_{fu}\mathbf{K}_{uu}^{-1}\left(\mathbf{K}_{uu} - \mathbf{S}_l\right)\mathbf{K}_{uu}^{-1}\mathbf{K}_{uf}. \tag{8}$$

The expectation in $\mathcal{L}_{\text{DGP}}$ is intractable in closed form and is estimated via doubly stochastic variational inference [Salimbeni and Deisenroth, 2017].

## 2.3 WASSERSTEIN-2 DISTANCE KERNELS ON GAUSSIAN MEASURES

The *Wasserstein-2 space* $\mathcal{W}_2(\mathbb{R}^d)$ is the set of probability measures on $\mathbb{R}^d$ with finite second moment, metrised by the

optimal transport cost:

$$W_2(\mu,\nu) = \left( \inf_{\pi \in \Pi(\mu,\nu)} \int \|\mathbf{x}-\mathbf{y}\|^2 \, \mathrm{d}\pi(\mathbf{x},\mathbf{y}) \right)^{1/2}, \quad (9)$$

where $\Pi(\mu,\nu)$ denotes the set of couplings with marginals $\mu$ and $\nu$ [Villani, 2008]. Bachoc et al. [2017] showed that the RBF kernel on $\mathcal{W}_2$ is positive definite:

**Theorem 2.1** (Bachoc et al. 2017, Thm. IV.1)**.** *For any $\mu,\nu \in \mathcal{W}_2(\mathbb{R})$, the kernel $k^{W_2}(\mu,\nu) = \sigma^2 \exp\!\left(-W_2^2(\mu,\nu)/\ell^2\right)$ is positive definite.*

A proof is provided in Appendix A.3. Since products of positive definite kernels are positive definite, the ARD extension to $D$-dimensional distributional inputs follows immediately:

$$k^{W_2}\left(\{\mu_d\}_{d=1}^D, \{\nu_d\}_{d=1}^D\right) = \sigma^2 \exp\left( - \sum_{d=1}^{D} \frac{W_2^2(\mu_d,\nu_d)}{\ell_d^2} \right). \quad (10)$$

**Wasserstein-2 distance between Gaussians.** Since Gaussian measures have finite second moments, $W_2$ is well-defined on them. For $\mu = \mathcal{N}(\mathbf{m}_1,\mathbf{\Sigma}_1)$ and $\nu = \mathcal{N}(\mathbf{m}_2,\mathbf{\Sigma}_2)$, the closed form is $W_2^2(\mu,\nu) = \|\mathbf{m}_1 - \mathbf{m}_2\|_2^2 + \mathrm{Tr}\left[\mathbf{\Sigma}_1 + \mathbf{\Sigma}_2 - 2\left(\mathbf{\Sigma}_1^{1/2}\mathbf{\Sigma}_2\mathbf{\Sigma}_1^{1/2}\right)^{1/2}\right]$ [Dowson and Landau, 1982]. In the univariate case this reduces to $W_2^2(\mu,\nu) = (m_1-m_2)^2 + (\sqrt{\sigma_1^2}-\sqrt{\sigma_2^2})^2$, which is the form used throughout this paper.

## 3 PROPAGATION OF VARIANCE IN DGP

This section derives the conditions a DGP must satisfy for OOD inputs to be propagated forward with elevated predictive variance. The SVGP decomposition $f = g + h$ admits two distinct regimes depending on how the inducing points are arranged. We study them as two case studies. **Case A (regression regime, Prop. 3.1):** inducing points are pooled near the origin (typical for centred regression data), so the non-parametric component $h$ collapses on OOD inputs and total variance is driven entirely by the parametric component $g$. We derive the condition on $\partial\tilde{\mu}_2/\partial f_1$ under which $g$ propagates higher variance to OOD inputs. **Case B (classification regime, Prop. 3.2):** inducing points form well-separated clusters (typical when classes are separable at a hidden layer), so the parametric component vanishes for OOD inputs that fall between clusters. The only signal is non-parametric, and we derive the condition on the length-scale $\ell_2^2$ under which $h$ propagates correctly. The two cases together cover the qualitative regimes a deep GP encounters in practice.

We derive theoretical conditions under which a zero-mean DGP preserves higher predictive uncertainty for out-of-distribution (OOD) inputs relative to in-distribution ones

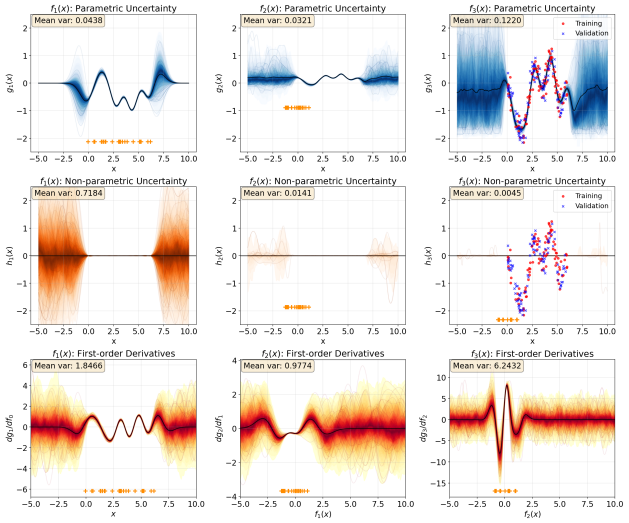



Figure 1: **Collapse of non-parametric variance in DGP.** Layer-wise decomposition of uncertainty into parametric and non-parametric for a zero-mean DGP.

across all layers, and characterise which uncertainty component (parametric ($g$) or non-parametric ($h$) ) drives this behaviour through the hierarchy. We consider two practically motivated regimes. In the first (**Case A**), the inducing locations $\{\mathbf{Z}_l\}_{l=0}^L$ are uniformly distributed on $[-3\sigma_{l-1}, 3\sigma_{l-1}]$, ensuring that the support of the previous layer's output lies within the inducing neighbourhood.

**Proposition 3.1.** *Consider an OOD input $\mathbf{x}_{\text{out}}$ such that $\tilde{\boldsymbol{\mu}}_1(\mathbf{x}_{\text{out}}) = \mathbf{0}$ and $\tilde{\mathbf{\Sigma}}_1(\mathbf{x}_{\text{out}}) = \sigma_1^2\mathbf{I}$ (prior variance recovered), and an in-distribution input $\mathbf{x}_{\text{in}}$ with $\tilde{\boldsymbol{\mu}}_1(\mathbf{x}_{\text{in}}) = \mathbf{m}$ and $\tilde{\mathbf{\Sigma}}_1(\mathbf{x}_{\text{in}}) = V_{\text{in}} \le \sigma_1^2$. If the layer-2 inducing points are uniformly distributed on $[-3\sigma_1, 3\sigma_1]$, then $\tilde{\mathbf{\Sigma}}_2(\mathbf{x}_{\text{out}}) \ge \tilde{\mathbf{\Sigma}}_2(\mathbf{x}_{\text{in}})$ whenever:*

$$\frac{\left(\dfrac{\partial\,\tilde{\boldsymbol{\mu}}_2}{\partial \mathbf{f}_1}\right)^2\Bigg|_{\mathbf{f}_1=\mathbf{m}}}{\left(\dfrac{\partial\,\tilde{\boldsymbol{\mu}}_2}{\partial \mathbf{f}_1}\right)^2\Bigg|_{\mathbf{f}_1=\mathbf{0}}} \le \frac{\sigma_1^2}{V_{\text{in}}}. \quad (11)$$

*Proof.* See Appendix E.1. □

**Remark 3.1.** *The condition (11) is most easily satisfied when $\tilde{\boldsymbol{\mu}}_2$ has maximum slope at $\mathbf{f}_1 = \mathbf{0}$ (the average image of OOD inputs under the first layer) since the RHS satisfies $1 \le \frac{\sigma_1^2}{V_{\text{in}}}$ by construction. Figure 1 confirms this empirically: total uncertainty is correctly amplified across layers for OOD inputs precisely because $|\partial\tilde{\boldsymbol{\mu}}_2/\partial\mathbf{f}_1|$ peaks in the vicinity of the origin.*

Figure 1 reveals that the non-parametric component $h$ collapses to zero in all hidden layers, rendering the DGP effectively parametric beyond the first layer. The reason behind

this pathological behaviour is that DGP without a mean function will invariably suffer from *feature collapse*. We will touch on this subject later in the paper.

We also consider a second scenario (**Case B**) that arises naturally in binary classification when the two classes are well-separated at a given layer. Specifically, we place inducing points in two separated clusters, $\mathbf{Z}_{l,1} \subset (-\infty, -c]$ and $\mathbf{Z}_{l,2} \subset [c, \infty)$ for some $c > 0$. Recalling the SVGP decomposition $f(\cdot) = \underbrace{\mathbf{K}_{fu}\mathbf{K}_{uu}^{-1}\mathbf{u}}_{g(\cdot)} + \underbrace{\mathcal{N}(\mathbf{0},\, \mathbf{K}_{ff} - \mathbf{K}_{fu}\mathbf{K}_{uu}^{-1}\mathbf{K}_{uf})}_{h(\cdot)}$, we now derive conditions under which $h$ correctly propagates uncertainty for OOD inputs across all layers.

**Proposition 3.2.** *Suppose $g_l(\mathbf{x}_{\text{out}}) \approx \mathbf{0}$ at every layer, so that OOD inputs are governed solely by the non-parametric component $h$. Consider two inducing points $\mathbf{Z}_1 = \{-c, c\}$ with $c \geq 0$ and an OOD input whose first-layer output satisfies $h_1(\mathbf{x}_{\text{out}}) \sim \mathcal{N}(0, \sigma_1^2)$. Then $\mathbb{V}[h_2(\mathbf{x}_{\text{out}})]$ is non-monotone: it attains a unique minimum at*

$$\sigma_1^{2*} = \frac{1}{2}\left(\frac{c^2}{w^*} - \ell_2^2\right) > 0, \tag{12}$$

*where $w^* \in (0, \frac{1}{2})$ is the unique solution to $(1 - 2w)\, e^{-w} = \exp\left(-\frac{3c^2}{\ell_2^2}\right)$, and the threshold satisfies $\sigma_1^{2*} \geq \max\big(0, (2c^2 - \ell_2^2)/2\big)$. In particular, $\ell_2^2 \geq 2c^2$ is a necessary condition for rapid onset of uncertainty propagation.*

*Proof.* See Appendix E.2. □

**Remark 3.2.** *For $\sigma_1^2 < \sigma_1^{2*}$, $\mathbb{V}[h_2(\mathbf{x}_{\text{out}})]$ decreases, whereas for $\sigma_1^2 > \sigma_1^{2*}$, $\mathbb{V}[h_2(\mathbf{x}_{\text{out}})]$ is pulled towards the prior $\sigma_2^2$. The threshold $\sigma_1^{2*}$ marks the transition between these two regimes. The condition $\ell_2^2 \geq 2c^2$ ensures this transition occurs at a moderate level of input uncertainty, so that OOD inputs (not just extreme outliers) trigger increasing non-parametric variance.*

# 4 DISTRIBUTIONAL DEEP GAUSSIAN PROCESSES

Section 3 showed that standard DGPs propagate OOD uncertainty correctly only under fragile hyperparameter conditions: in the regression regime (**Case A**) the parametric component must satisfy a derivative condition that is not enforced during training, while in the classification regime (**Case B**) the non-parametric component requires a lengthscale that the ELBO has no incentive to choose. In this section we construct a model that has uncertainty propagation *by design* rather than by chance.

**Desiderata for kernels in DGPs.** The conditions derived above are not guaranteed to hold in practice. To actively enforce reliable OOD behaviour, we propose two desiderata for kernels at each layer.

**I *Distance-awareness.*** We adopt the standard notion that an outlier is a point residing in a low-probability region under the data distribution $\mathcal{P}$, i.e. $\mathcal{A} = \{x \in \mathcal{X} \mid p(x) \leq \xi\}$ for threshold $\xi \geq 0$ [Ruff et al., 2021]. A model should reflect this through its predictive distribution:

**Definition 4.1** (Distance-awareness [Liu et al., 2020a])**.** *The predictive distribution $p(y^* \mid \mathbf{x}^*)$ is* distance-aware *if it admits a summary statistic*

$$u(\mathbf{x}^*) = v\big[\mathbb{E}_{\mathbf{x}\sim\mathcal{X}_{\text{in}}}\big[\|\mathbf{x}^* - \mathbf{x}\|^2\big]\big]\,,$$

*where $v$ is monotonically increasing with distance from the training manifold $\mathcal{X}_{\text{in}}$.*

Since SVGPs use translation-invariant kernels, the non-parametric component $h$ naturally provides such a statistic: its variance increases as $\mathbf{x}^*$ moves away from the inducing support, making it a principled measure of *distributional uncertainty*. Conversely, the variance of $g$ captures *epistemic uncertainty*, uncertainty within the data manifold. We quantify distributional uncertainty via the differential entropy of $h$:

$$\mathcal{H}[h(\mathbf{x})] = \frac{1}{2}\log\det\big(\mathbf{K}_{ff} - \mathbf{K}_{fu}\mathbf{K}_{uu}^{-1}\mathbf{K}_{uf}\big) + \text{const.} \tag{13}$$

This serves as our OOD score throughout the paper. We desire each DGP layer to preserve this distance-awareness property of its shallow SVGP counterpart.

**II *Smoothness* in hidden layers.** Smoothness in intermediate representations is known to promote generalisation, robustness, and reliable uncertainty estimates [Rosca et al., 2020]. Following Liu et al. [2020a], we require DGP kernels to be *bi-Lipschitz* across layers: points close in one layer should remain close in the next, guarding against *feature collapse* [van Amersfoort et al., 2021] and ensuring that the distance structure, on which distance-awareness relies, is preserved throughout the hierarchy.

## 4.1 HYBRID KERNEL

**Distributional Deep Gaussian Processes (DDGPs).** The Wasserstein-2 kernel (10) is attractive for OOD detection because it is sensitive to differences in variance: two Gaussians with different spreads are far apart in $\mathcal{W}_2$, so a GP built on this kernel is naturally *distance-aware*. However, naively stacking a DistGP on top of a GP layer breaks Desideratum II: since all hidden-layer marginals share the same variance $\sigma_1^2$ in prior space, the Wasserstein-2 kernel evaluates to a constant, collapsing inter-point correlations and causing prior samples to degenerate.

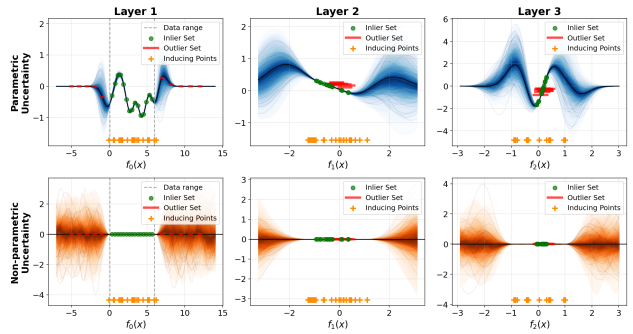



Figure 2: **Feature-collapse in DGP.** Outliers get mapped in hidden layers within the same space as inliers.

The fix is to couple two complementary pathways at each hidden layer:

- **Stochastic** ($\mathbf{f}_l^{\text{sth}}$): a Monte Carlo sample from the layer-$l$ GP marginal, which carries Euclidean geometry and preserves inter-point correlations in the prior, thus satisfying Desideratum II.
- **Deterministic** ($\mathbf{f}_l^{\text{det}}$): the full Gaussian marginal, passed as a distribution into the Wasserstein-2 kernel, which encodes whether a point is OOD via its posterior variance, thus satisfying Desideratum I.

We combine these via a *hybrid kernel* (inspired by Varifold theory; see Appendix A.2 for an introduction and connections between the two): $k^H(\{x_i, \mu_i\}, \{x_j, \mu_j\}) = k^{\text{SE}}(x_i, x_j) \cdot \exp\Big(-\sum_{d=1}^{D} \frac{W_2^2(\mu_{i,d}, \mu_{j,d})}{\ell_d^2}\Big)$, where $\mu_{i,d} = \mathcal{N}(m_d, \sigma_d^2)$ is the $d$-th marginal of the hidden-layer GP and $x_i$ is a sample from it. The resulting two-layer DDGP generative process is:

$$\mathbf{f}_1 \sim \mathcal{GP}\big(\mathbf{0},\, k^{\text{SE}}(\mathbf{x}, \mathbf{x})\big)\,, \tag{14}$$

$$\mathbf{f}_1^{\text{sth}} = \boldsymbol{\mu}(\mathbf{f}_1) + \boldsymbol{\sigma}(\mathbf{f}_1) \odot \boldsymbol{\epsilon}, \quad \boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}), \tag{15}$$

$$\mathbf{f}_1^{\text{det}} = \mathcal{N}\big(\boldsymbol{\mu}(\mathbf{f}_1),\, \operatorname{diag} \boldsymbol{\sigma}^2(\mathbf{f}_1)\big)\,, \tag{16}$$

$$\mathbf{f}_2 \sim \mathcal{GP}\big(\mathbf{0},\, k^H\big(\{\mathbf{f}_1^{\text{sth}}, \mathbf{f}_1^{\text{det}}\}, \cdot\big)\big)\,. \tag{17}$$

**Remark 4.1.** *DDGPs coincide with DGPs in prior space since the $\mathcal{W}_2$ term equals 1 when all marginals share the same prior variance $\sigma_1^2$, so correlations are determined entirely by $k^{\text{SE}}$. The two frameworks diverge in posterior space, where heterogeneous predictive variances activate the Wasserstein-2 pathway, endowing DDGPs with layer-wise distance-awareness absent in standard DGPs.*

**Hybrid kernels and OOD detection.** A key failure mode of standard DGPs is that sampling from $q(\mathbf{f}_1)$ can map an OOD point $\mathbf{x}_{\text{ood}}$ and an in-distribution point $\mathbf{x}_{\text{in}}$ to equally distant locations from the layer-2 inducing points $\mathbf{Z}_1$ despite their very different posterior variances in $\mathbf{f}_1$ (see Figure 2). When this occurs, the non-parametric variance $h_2$ is identical for both points, and the OOD signal is lost (violating Desideratum I).

The hybrid kernel resolves this via its $\mathcal{W}_2$ component. Because $k^{\text{H}}$ compares pairs (point, measure), the inducing inputs of layer 2 are themselves Gaussian measures with learnable mean and variance (i.e., *distributional* inducing locations), formalised in §4.2.

If the *distributional* inducing locations $\mathbf{Z}_1$ are calibrated to match the posterior variance of in-distribution points in $\mathbf{f}_1$, then $W_2(\mathbf{f}_1^{\text{det}}(\mathbf{x}_{\text{ood}}), \mathbf{z})$ is strictly larger than $W_2(\mathbf{f}_1^{\text{det}}(\mathbf{x}_{\text{in}}), \mathbf{z})$, since $\mathbf{x}_{\text{ood}}$ carries higher posterior variance. This inflates the non-parametric variance $h_2(\mathbf{x}_{\text{ood}})$ relative to $h_2(\mathbf{x}_{\text{in}})$, restoring Desideratum I at layer 2. The mechanism is illustrated in Figure 9 in Appendix H alongside a more in-depth discussion of DDGPs. In Appendix D we provide an intuitive explanation for why distributional variance does not collapse for DDGPs.

### 4.2 EVIDENCE LOWER BOUND

DDGPs differ from DGPs in layers $l \geq 2$: the hybrid kernel $k^H$ requires both a *sample* from $q(\mathbf{f}_{l-1})$ (Euclidean pathway) and the full *marginal distribution* $q(\mathbf{f}_{l-1})$ (Wasserstein-2 pathway). The augmented joint prior factorises as:

$$p\big(\mathbf{y}, \{\mathbf{f}_l, \mathbf{U}_l\}_{l=1}^L\big) = p(\mathbf{y} \mid \mathbf{f}_L)\, \underbrace{p(\mathbf{f}_1 \mid \mathbf{U}_1; \mathbf{X})\, p(\mathbf{U}_1)}_{\text{Euclidean (layer 1)}} \underbrace{\prod_{l=2}^{L} p(\mathbf{f}_l \mid \mathbf{f}_{l-1}, \mathbf{U}_l; \mathbf{Z}_{l-1})\, p(\mathbf{U}_l)}_{\text{Euclidean + Wasserstein-2 (layers}\geq 2)}\,. \tag{18}$$

We adopt the mean-field variational posterior $q\big(\{\mathbf{f}_l, \mathbf{U}_l\}_{l=1}^L\big) = p(\mathbf{f}_L \mid \mathbf{U}_L; \mathbf{Z}_{L-1}) \prod_{l=1}^L q(\mathbf{U}_l)$, with $q(\mathbf{U}_l) = \mathcal{N}(\mathbf{U}_l; \mathbf{m}_l, \mathbf{S}_l)$. For $l \geq 2$, marginalising $\mathbf{U}_l$ yields the inter-layer conditional $q(\mathbf{f}_l \mid \mathbf{f}_{l-1}) = \mathcal{N}\Big(\mathbf{f}_l;\ \mathbf{K}_{fu}^H (\mathbf{K}_{uu}^H)^{-1} \mathbf{m}_l,\ \mathbf{K}_{ff}^H - \mathbf{K}_{fu}^H (\mathbf{K}_{uu}^H)^{-1} \big(\mathbf{K}_{uu}^H - \mathbf{S}_l\big) (\mathbf{K}_{uu}^H)^{-1} \mathbf{K}_{uf}^H\Big)$, where the hybrid kernel matrices are evaluated using samples from $q(\mathbf{f}_{l-1})$ in the Euclidean component and the full marginal $q(\mathbf{f}_{l-1})$ in the $\mathcal{W}_2$ component, with distributional inducing locations $\mathbf{Z}_{l-1} \sim \mathcal{N}(\mathbf{m}_{Z_{l-1}}, \boldsymbol{\Sigma}_{Z_{l-1}})$. Following the same derivation as for the DGP ELBO (6), we obtain:

$$\log p(\mathbf{y}) \geq \mathcal{L}_{\text{DDGP}} \coloneqq \mathbb{E}_{\prod_l q(\mathbf{f}_l \mid \mathbf{f}_{l-1})}[\log p(\mathbf{y} \mid \mathbf{f}_L)] - \operatorname{KL}[q(\mathbf{U}_1) \,\|\, p(\mathbf{U}_1)] - \sum_{l=2}^{L} \mathbb{E}_{\mathbf{Z}_{l-1}}[\operatorname{KL}[q(\mathbf{U}_l) \,\|\, p(\mathbf{U}_l)]]\,. \tag{19}$$

The expectation over $\mathbf{Z}_{l-1}$ in the KL terms arises because the distributional inducing locations are themselves random, distinguishing $\mathcal{L}_{\text{DDGP}}$ from $\mathcal{L}_{\text{DGP}}$. For a deeper dive into the conceptual differences between DGPs and DDGPs we refer the reader to Appendix H.

### 4.3 PROPAGATION OF VARIANCE IN DDGP

We now characterise when DDGPs reliably preserve higher uncertainty for OOD inputs across layers.

**Proposition 4.1.** *Under the same layer-1 setup as Proposition 3.1, assume distributional inducing locations with means $\mathbf{m}_{Z_1}$ uniformly spread on $[-3\sigma_1, 3\sigma_1]$ and variances $\mathbf{\Sigma}_{Z_1} \leq \sigma_1^2$. Then $\tilde{\mathbf{\Sigma}}_2(\mathbf{x}_{\text{ood}}) \geq \tilde{\mathbf{\Sigma}}_2(\mathbf{x}_{\text{in}})$ if:*

$$\mathbf{\Sigma}_{Z_1} \approx V_{\text{in}} \ll \sigma_1^2. \tag{20}$$

*Proof.* See Appendix E.3. □

**Remark 4.2.** *Condition* (20) *has a natural interpretation: the distributional inducing locations should match the low posterior variance of in-distribution points, so that the $\mathcal{W}_2$ component of the hybrid kernel registers a large distance to OOD inputs, which carry prior variance $\sigma_1^2$. This is self-reinforcing: optimisation drives inducing points towards the in-distribution manifold, naturally shrinking $\mathbf{\Sigma}_{Z_{l-1}}$ to the in-distribution posterior variance at every layer. DDGPs therefore satisfy Desideratum I more robustly than DGPs, which rely on the brittle derivative condition* (11).

### 4.4 GLOBAL INDUCING-POINT LOCATIONS

Naively, the distributional inducing locations $\mathbf{Z}_{l-1} \sim \mathcal{N}(\mathbf{m}_{Z_{l-1}}, \mathbf{\Sigma}_{Z_{l-1}})$ are free variational parameters at each layer $l \geq 2$, factorised across $l$. This raises two concerns specific to DDGPs. First, our guarantee in Prop. 4.1 requires $\mathbf{\Sigma}_{Z_{l-1}} \approx V_{\text{in}}$, but free parameters give optimisation no reason to converge there. Second, layer-wise factorisation collapses the compositional uncertainty [Ustyuzhaninov et al., 2020] that motivates a deep model in the first place. We therefore replace these free parameters with the global-inducing scheme of Ober and Aitchison [2021b]: a single set of $M$ inducing inputs $\mathbf{Z}_0 \in \mathbb{R}^{M \times D}$ is learned at the input layer and propagated through the hierarchy, so that for $l \geq 2$:

$$\begin{aligned} q(\mathbf{Z}_l \mid \mathbf{Z}_{l-1}) &= \int p(\mathbf{Z}_l \mid \mathbf{U}_l, \mathbf{Z}_{l-1})\, q(\mathbf{U}_l)\, \mathrm{d}\mathbf{U}_l \\ &= \mathcal{N}(\mathbf{Z}_l;\, \boldsymbol{\mu}_{Z_l}(\mathbf{Z}_{l-1}),\, \boldsymbol{\sigma}^2_{Z_l}(\mathbf{Z}_{l-1})), \end{aligned} \tag{21}$$

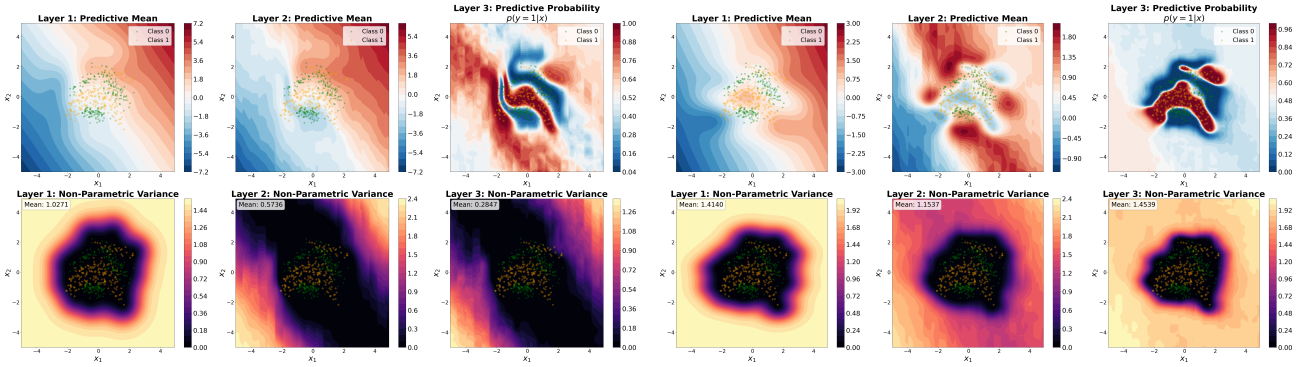
where $(\boldsymbol{\mu}_{Z_l}, \boldsymbol{\sigma}^2_{Z_l})$ are the layer-$l$ predictive moments evaluated at $\mathbf{Z}_{l-1}$. Two properties make this the natural pairing for the hybrid kernel. **(i) Self-calibration.** The propagated variance $\boldsymbol{\sigma}^2_{Z_l}$ tracks the in-distribution posterior variance at every layer by construction, so the condition required by Prop. 4.1 is enforced rather than hoped for. **(ii) Cross-layer correlation.** Although $q(\{\mathbf{U}_l\}) = \prod_l q(\mathbf{U}_l)$ factorises over inducing *outputs*, the kernels $K^H(\mathbf{Z}_{l-1}, \cdot)$ depend on $\mathbf{U}_{1:l-1}$ through propagation, so the induced posterior over functions $q(\{\mathcal{F}_l\})$ is not factorised. We refer to this variant as GIPL-DDGP and use it throughout the experiments alongisde a GIPL variant of DGPs; Appendix G gives a full derivation, contrasts the construction with Ustyuzhaninov et al. [2020], and details the modified ELBO.

## 5 EXPERIMENTS

**Effect of dimensionality reduction on layer-wise uncertainty.** Figure 3 examines the behaviour of each model when the PCA mean function encodes the 2D banana dataset down to a single hidden dimension. For the DGP, this creates a fundamental ambiguity: OOD points that project near an in-distribution location along the principal axis are indistinguishable from genuine inliers under the SE kernel, which only has access to the 1D mean representation. The non-parametric variance maps in panel (a) make this visible; uncertainty propagates in stripe-like structures along the PCA direction, leaving large OOD regions with spuriously low uncertainty and corrupting the output probability surface. The DDGP resolves this because the Wasserstein-2 component of the hybrid kernel operates on the *full posterior distribution*, not just the projected mean. Two points sharing the same 1D projection but differing in posterior variance (one confidently in-distribution, one reverting to the prior) receive very different kernel similarities, and the non-parametric variance maps in panel (b) remain tightly localised around the training manifold at every layer.

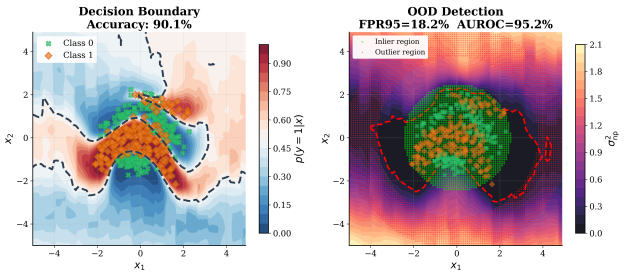
**Uncertainty localisation without dimensionality reduction.** Figure 4 removes the confound of the previous experiment by keeping all hidden dimensions at 2, so the PCA mean function does not suffer from information loss. Both models achieve comparable accuracy (DGP: $90.1\%$; DDGP: $90.5\%$), confirming that any difference in uncertainty structure is attributable to the kernel rather than discriminative capacity. Despite the absence of a bottleneck, the DGP's non-parametric uncertainty remains diffuse and poorly localised: it fails to contract around the training manifold, leaving broad OOD regions with spuriously low uncertainty and yielding FPR@95 $= 18.2\%$, AUROC $= 95.2\%$. This confirms that the SE kernel's limitation is intrinsic; without access to distributional information, it cannot reliably distinguish an in-distribution point from an OOD point at the same Euclidean distance from the inducing locations. The DDGP's Wasserstein-2 term corrects this directly: the non-parametric uncertainty collapses tightly onto the data manifold, the OOD boundary based on FPR@95 is sharp and well-defined, and detection performance improves dramatically to FPR@95 $= 1.4\%$, AUROC $= 99.\%$.

We evaluate OOD detection on two benchmarks: **MNIST** (inlier) using a 3-layer Deep Convolutional GP (DConvGP) [Blomqvist et al., 2018] and its distributional variant (DistDConvGP), alongside a shallow ConvGP [Van der Wilk et al., 2017] baseline; and **CIFAR-10** (inlier) using analogous DGP and DDGP architectures with a shallow GP baseline.

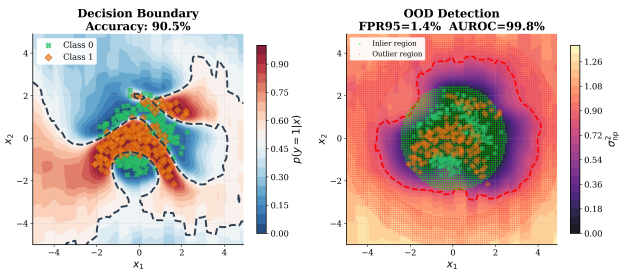



(a) DGP (b) DDGP

Figure 3: **Layer-wise uncertainty localisation on the banana dataset.** Posterior decision boundary and non-parametric uncertainty $\sigma_{\text{np}}$ for a 3-layer DGP (a) and DDGP (b) with PCA-mean functions and 1 dim. for hidden layers.




(a) DGP




(b) DDGP

Figure 4: **Uncertainty localisation on the banana dataset.** Posterior decision boundary, non-parametric uncertainty for a 3-layer DGP (a) and DDGP (b) with PCA-mean functions and 2 dim. for hidden layers.

To benchmark against strong deterministic detectors, we additionally train *parameter-matched* Deep Ensembles (DE) [Lakshminarayanan et al., 2017] and a Spectral-normalised Neural Gaussian Process (SNGP) [Liu et al., 2020a], so that every method operates at a comparable parameter budget. For each benchmark we consider three OOD categories: Near-OOD (semantically similar), Far-OOD (semantically dissimilar) and Covariate Shift. Performance is measured by AUROC ↑ and FPR@95 ↓ under four uncertainty scores spanning two families: the *distributional* scores native to our models, *predictive entropy* (Pred. Ent.) and *distributional entropy* (Dist. Ent.), and the post-hoc feature/logit scores *energy* [Liu et al., 2020b] and *Mahalanobis* [Lee et al., 2018], which apply to every method. To facilitate model comparison across datasets, we report rank $mean \pm std$ aggregated over all $dataset \times metric$ combinations per score.

The parameter-matched baselines set a demanding bar on the post-hoc scores: DE is the single strongest detector under predictive entropy, energy and Mahalanobis on *both* benchmarks (e.g. ranks 2.0–2.1 under Pred. Ent. and energy), with SNGP consistently mid-pack. However, DistDConvGPs remain competitive with parameter-matched DE and SNGP on the shared post-hoc scores (3L-D ranks 2.6–3.2 under energy).

Under *distributional entropy*, the distributional variants are decisive, on MNIST the shallow GP leads ($1.0_{\pm 0.0}$) with 3L-D second ( 2.0), and on CIFAR-10 the distributional models take the top two ranks (2L-D 2.3, 3L-D 2.4) (see Table 1). Critically, the distributional variant outranks its standard counterpart under this score on both benchmarks, confirming that the Wasserstein-2 kernel pathway is the primary driver of OOD signal from the distributional viewpoint.

Under *predictive entropy*, the ranking inverts entirely, with depth becoming the decisive factor: on MNIST, 3L and 3L-D lead ($2.5_{\pm 1.0}$ and $3.2_{\pm 1.1}$) while ConvGP collapses to last ($6.9_{\pm 0.3}$); on CIFAR-10, 3L-D dominates ($2.9_{\pm 1.1}$). Shallow models thus rely entirely on their distributional entropy for OOD signal and provide no useful predictive entropy, whereas deep models benefit from both pathways. Taken together, DistDConvGPs are the only model class that consistently performs competitively under *both* scores, albeit not on par with DE. For the OOD results on the MNIST setting, we refer the reader to Appendix I.

**UCI regression results.** We additionally compare against deep kernel processes (DKP) [Aitchison et al., 2021], a Wishart-based hierarchical model that propagates uncer-

Table 1: OOD detection on CIFAR-10 (inlier). For each OOD score we report AUROC (A↑) and FPR95 (F↓) on three families of distribution shift: near-OOD, far-OOD and covariate shift. Values show mean over 10 random seeds. Std is not shown for improved visualization. Within each block, every column is rank-shaded across models (**best** darkest, **worst** lightest). *mean rank* is the per-model rank averaged over all AUROC/FPR95 evaluations in the block; lower is better, the best is in bold.

| CIFAR-10 (inlier) | | Near-OOD | | | | | | Far-OOD | | | | | | | | Cov. Shift | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | *CIFAR-100* | | *STL-10* | | *Tiny ImageNet* | | *Places365* | | *Rnd. Noise* | | *SVHN* | | *Textures* | | *CIFAR-10.1* | | *CIFAR-10-C* | | *CIFAR-10 (g)* | |
| **Score** | **Model** | A↑ | F↓ | A↑ | F↓ | A↑ | F↓ | A↑ | F↓ | A↑ | F↓ | A↑ | F↓ | A↑ | F↓ | A↑ | F↓ | A↑ | F↓ | A↑ | F↓ |
| Pred. Ent. | **2L** | .659 | .803 | .532 | .935 | .631 | .900 | .612 | .891 | .992 | .031 | .653 | .739 | .583 | .963 | .577 | .887 | .609 | .894 | .641 | .881 |
| | **2L-D** | .652 | .822 | .502 | .947 | .578 | .948 | .578 | .916 | .853 | .268 | .651 | .701 | .505 | .994 | .578 | .897 | .370 | .964 | .654 | .855 |
| | **3L** | .663 | .814 | .529 | .924 | .631 | .887 | .625 | .868 | 1.000 | .000 | .727 | .658 | .619 | .923 | .574 | .896 | .614 | .879 | .637 | .882 |
| | **3L-D** | .684 | .779 | .547 | .918 | .647 | .852 | .654 | .842 | **1.000** | **.000** | .716 | .654 | .581 | .930 | .587 | .890 | **.839** | **.541** | .634 | .877 |
| | **GP** | .612 | .862 | .495 | .960 | .570 | .933 | .576 | .909 | .555 | .604 | .596 | .783 | .483 | .996 | .568 | .914 | .522 | .938 | **.674** | **.830** |
| | **DE** | **.836** | **.466** | **.666** | **.860** | **.823** | **.529** | **.829** | **.506** | .584 | .636 | **.847** | **.398** | **.798** | **.582** | **.623** | **.879** | .718 | .805 | .589 | .892 |
| | **SNGP** | .779 | .594 | .634 | .898 | .777 | .618 | .777 | .625 | .504 | .808 | .778 | .523 | .744 | .661 | .611 | .906 | .679 | .807 | .578 | .925 |
| | *mean rank* ↓ | **2L**: $4.4_{\pm 1.0}$ | | **2L-D**: $5.4_{\pm 1.5}$ | | **3L**: $3.9_{\pm 0.9}$ | | **3L-D**: $2.9_{\pm 1.1}$ | | **GP**: $6.0_{\pm 1.8}$ | | **DE**: $\mathbf{2.0_{\pm 1.9}}$ | | **SNGP**: $3.3_{\pm 2.1}$ | | | | | |
| Dist. Ent. | **2L** | .503 | .957 | .643 | .912 | .653 | .894 | .648 | .890 | .991 | .011 | .193 | .996 | .486 | .996 | .485 | .948 | .817 | .373 | .380 | .986 |
| | **2L-D** | .527 | .961 | **.706** | **.819** | **.743** | **.777** | **.719** | .817 | **.999** | **.001** | .122 | 1.000 | .625 | .988 | .501 | .951 | **.995** | **.010** | .351 | .988 |
| | **3L** | .474 | .964 | .630 | .915 | .642 | .899 | .654 | .882 | .989 | .013 | .210 | 1.000 | .468 | .997 | .477 | .951 | .766 | .502 | **.384** | .984 |
| | **3L-D** | .498 | **.952** | .664 | .880 | .691 | .833 | .694 | **.811** | .971 | .030 | **.358** | **.965** | .592 | .966 | .476 | **.941** | .965 | .079 | .365 | **.980** |
| | **GP** | **.561** | .952 | .644 | .899 | .673 | .857 | .634 | .869 | .995 | .005 | .320 | .984 | **.661** | **.946** | **.503** | .953 | .869 | .233 | .310 | .981 |
| | *mean rank* ↓ | **2L**: $3.5_{\pm 0.7}$ | | **2L-D**: $\mathbf{2.3_{\pm 1.5}}$ | | **3L**: $4.2_{\pm 1.0}$ | | **3L-D**: $2.4_{\pm 1.4}$ | | **GP**: $2.6_{\pm 1.2}$ | | **DE**: – | | **SNGP**: – | | | | | |
| Energy | **2L** | .654 | .810 | .526 | .969 | .651 | .883 | .654 | .878 | .965 | .070 | .577 | .864 | .559 | .970 | .581 | .906 | .658 | .860 | .672 | .869 |
| | **2L-D** | .646 | .828 | .509 | .977 | .618 | .922 | .634 | .895 | .719 | .372 | .462 | .826 | .467 | .989 | .586 | .909 | .414 | .957 | .670 | .851 |
| | **3L** | .664 | .803 | .532 | .947 | .653 | .868 | .669 | .852 | .964 | .041 | .624 | .780 | .585 | .928 | .580 | .897 | .661 | .829 | .656 | .887 |
| | **3L-D** | .686 | .774 | .541 | .937 | .665 | .833 | .696 | .808 | **.969** | **.032** | .638 | .749 | .567 | .949 | .594 | .895 | **.866** | **.386** | .679 | .864 |
| | **GP** | .604 | .886 | .488 | .979 | .586 | .927 | .605 | .905 | .546 | .566 | .503 | .909 | .445 | .999 | .573 | .905 | .580 | .906 | **.710** | **.813** |
| | **DE** | **.857** | **.452** | **.663** | **.871** | **.851** | **.510** | **.862** | **.496** | .558 | .639 | **.853** | **.365** | **.786** | **.618** | **.629** | **.887** | .717 | .783 | .607 | .885 |
| | **SNGP** | .801 | .579 | .631 | .901 | .800 | .614 | .801 | .642 | .532 | .760 | .808 | .477 | .771 | .638 | .613 | .892 | .688 | .783 | .585 | .924 |
| | *mean rank* ↓ | **2L**: $4.7_{\pm 1.0}$ | | **2L-D**: $5.5_{\pm 1.3}$ | | **3L**: $4.0_{\pm 0.9}$ | | **3L-D**: $2.6_{\pm 0.9}$ | | **GP**: $6.0_{\pm 1.8}$ | | **DE**: $\mathbf{2.0_{\pm 1.8}}$ | | **SNGP**: $3.1_{\pm 2.0}$ | | | | | |
| Maha. | **2L** | .527 | .961 | .700 | .867 | .712 | .861 | .701 | .877 | **1.000** | **.000** | .068 | 1.000 | .485 | 1.000 | .514 | .945 | .984 | .044 | .380 | .986 |
| | **2L-D** | .527 | .962 | .717 | .840 | .736 | .826 | .720 | .860 | **1.000** | **.000** | .058 | 1.000 | .525 | .998 | .515 | .948 | .992 | .020 | .375 | .987 |
| | **3L** | .496 | .970 | .682 | .886 | .689 | .882 | .688 | .880 | 1.000 | **.000** | .077 | 1.000 | .440 | 1.000 | .495 | .947 | .948 | .133 | .379 | .988 |
| | **3L-D** | .510 | .965 | .693 | .867 | .711 | .846 | .704 | .851 | **1.000** | **.000** | .130 | .999 | .498 | .997 | .490 | .953 | .982 | .045 | .356 | .981 |
| | **GP** | .543 | .959 | **.893** | **.424** | .768 | .724 | **.880** | .553 | **1.000** | **.000** | .016 | 1.000 | .784 | .955 | .527 | .948 | **1.000** | **.000** | .172 | .997 |
| | **DE** | **.833** | **.526** | .683 | .829 | **.848** | **.511** | .875 | **.452** | .913 | .267 | **.962** | **.239** | **.927** | **.358** | **.604** | **.888** | .759 | .725 | **.576** | .924 |
| | **SNGP** | .755 | .670 | .620 | .854 | .751 | .689 | .799 | .612 | .525 | .891 | .890 | .449 | .783 | .696 | .585 | .896 | .629 | .877 | .575 | **.918** |
| | *mean rank* ↓ | **2L**: $4.2_{\pm 1.6}$ | | **2L-D**: $3.9_{\pm 1.5}$ | | **3L**: $5.7_{\pm 1.6}$ | | **3L-D**: $4.5_{\pm 1.7}$ | | **GP**: $2.9_{\pm 2.0}$ | | **DE**: $\mathbf{2.4_{\pm 2.0}}$ | | **SNGP**: $3.5_{\pm 2.1}$ | | | | | |

tainty through positive-definite Gram matrices rather than through function values, included here as a further GP-family baseline at depths 2 and 3 (DKP-2, DKP-3). Table 2 reports results across 10 UCI datasets under four metrics. On the three primary accuracy metrics (RMSE, MAE, and NLPD), DDGP-2 consistently ranks best (mean ranks 1.6, 1.5, 2.1), with DDGP-3 second on RMSE and MAE (2.8, 2.6), demonstrating that the inherent deterministic pathway yields systematic gains over the sole stochastic pathway specific to the DGP counterparts. The deep kernel process baselines rank last on these accuracy metrics (DKP-2 and DKP-3 mean ranks 5.1–5.6), falling behind even the standard DGP variants and confirming that the Wishart-based construction does not translate into competitive tabular predictive performance under the same sparse-inference budget. Coverage at the 95% nominal level is broadly similar across all models (mean ranks 2.8–4.5), confirming that none of the architectures achieves a decisive advantage on interval coverage alone, with the shallower SVGP and DGP-2 marginally closest to nominal. Taken together, DDGPs offer the best accuracy & calibration profile when the primary objective is predictive performance, while shallow GPs remain competitive when interval reliability is prioritised.

## 6 DISCUSSION

**Uncertainty propagation in DGPs.** Total uncertainty is correctly propagated in DGPs only when the predictive mean's first-order derivative attains its maximum at the image of OOD inputs in hidden layers, a condition that is brittle in practice. With more inducing points, OOD inputs increasingly land near inducing locations, suppressing $h(\cdot)$ and reducing the model to a purely parametric composition

Table 2: UCI regression benchmark (10 datasets, 10 splits). Metrics: RMSE, MAE, NLPD (lower↓ is better); 95% Coverage (closer to nominal 0.95 is best, denoted →). Cell shading indicates per-column rank: **darkest** = best, lightest = worst. Values show mean ± std over 10 random splits. Bottom rows report rank mean ± std across all datasets (lower is better).

| Metric | Model | *Boston* | *Concrete* | *Energy* | *Kin8nm* | *Naval* | *Power* | *Protein* | *Wine-R* | *Wine-W* | *Yacht* |
|---|---|---|---|---|---|---|---|---|---|---|---|
| RMSE↓ | **SVGP** | 3.04±0.44 | 5.79±0.54 | 0.665±0.088 | 0.101±0.003 | 0.000±0.000 | 4.13±0.11 | 4.74±0.05 | 0.616±0.024 | 0.708±0.018 | 0.368±0.104 |
| | **DGP-2** | 3.50±0.86 | 6.02±1.12 | 0.536±0.063 | 0.091±0.002 | 0.000±0.000 | 4.06±0.12 | 4.48±0.05 | 0.626±0.021 | 0.709±0.022 | 1.07±0.96 |
| | **DGP-3** | 4.00±0.79 | 6.11±0.95 | 0.533±0.078 | 0.092±0.002 | 0.000±0.000 | 4.07±0.12 | 4.47±0.05 | 0.627±0.018 | 0.709±0.023 | 1.72±1.27 |
| | **DDGP-2** | 3.08±0.91 | 4.97±0.59 | 0.507±0.061 | 0.083±0.002 | 0.000±0.000 | 3.97±0.12 | 4.22±0.05 | 0.625±0.028 | 0.697±0.017 | 0.871±0.840 |
| | **DDGP-3** | 3.66±0.95 | 5.13±0.70 | 0.510±0.045 | 0.076±0.002 | 0.000±0.000 | 3.98±0.12 | 4.08±0.05 | 0.669±0.032 | 0.699±0.021 | 1.63±0.80 |
| | **DKP-2** | 3.73±0.88 | 6.36±1.16 | 0.568±0.071 | 0.096±0.003 | 0.000±0.000 | 4.22±0.14 | 4.70±0.06 | 0.648±0.023 | 0.728±0.024 | 1.15±0.91 |
| | **DKP-3** | 4.28±0.85 | 6.54±1.02 | 0.501±0.069 | 0.098±0.003 | 0.000±0.000 | 4.19±0.13 | 4.59±0.06 | 0.642±0.021 | 0.722±0.023 | 1.88±1.18 |
| | *mean rank* ↓ | **SVGP**: $3.6_{\pm 2.5}$ | **DGP-2**: $3.3_{\pm 1.0}$ | **DGP-3**: $4.1_{\pm 1.4}$ | **DDGP-2**: $1.6_{\pm 0.5}$ | **DDGP-3**: $2.8_{\pm 1.9}$ | **DKP-2**: $5.3_{\pm 1.7}$ | **DKP-3**: $5.1_{\pm 2.2}$ | | | |
| MAE↓ | **SVGP** | 2.06±0.28 | 4.31±0.34 | 0.498±0.062 | 0.078±0.003 | 0.000±0.000 | 3.15±0.03 | 3.83±0.04 | 0.480±0.023 | 0.551±0.013 | 0.222±0.049 |
| | **DGP-2** | 2.24±0.36 | 4.35±0.54 | 0.400±0.048 | 0.070±0.001 | 0.000±0.000 | 3.08±0.06 | 3.48±0.04 | 0.487±0.021 | 0.557±0.014 | 0.550±0.239 |
| | **DGP-3** | 2.48±0.27 | 4.38±0.53 | 0.404±0.066 | 0.071±0.002 | 0.000±0.000 | 3.09±0.05 | 3.46±0.05 | 0.485±0.019 | 0.554±0.015 | 0.820±0.419 |
| | **DDGP-2** | 2.04±0.37 | 3.57±0.39 | 0.371±0.039 | 0.064±0.002 | 0.000±0.000 | 2.98±0.06 | 3.21±0.04 | 0.479±0.023 | 0.545±0.013 | 0.437±0.213 |
| | **DDGP-3** | 2.43±0.50 | 3.60±0.42 | 0.363±0.029 | 0.059±0.002 | 0.000±0.000 | 2.99±0.06 | 3.02±0.04 | 0.489±0.023 | 0.541±0.019 | 0.909±0.441 |
| | **DKP-2** | 2.36±0.39 | 4.60±0.57 | 0.423±0.054 | 0.073±0.002 | 0.000±0.000 | 3.20±0.07 | 3.65±0.05 | 0.505±0.022 | 0.570±0.015 | 0.598±0.252 |
| | **DKP-3** | 2.63±0.31 | 4.73±0.56 | 0.386±0.058 | 0.074±0.002 | 0.000±0.000 | 3.18±0.06 | 3.58±0.05 | 0.500±0.020 | 0.565±0.015 | 0.872±0.392 |
| | *mean rank* ↓ | **SVGP**: $3.8_{\pm 2.4}$ | **DGP-2**: $3.4_{\pm 1.0}$ | **DGP-3**: $4.0_{\pm 1.3}$ | **DDGP-2**: $1.5_{\pm 0.5}$ | **DDGP-3**: $2.6_{\pm 2.1}$ | **DKP-2**: $5.3_{\pm 1.8}$ | **DKP-3**: $5.3_{\pm 1.8}$ | | | |
| NLPD↓ | **SVGP** | 0.267±0.115 | 0.352±0.088 | −1.29±0.07 | 0.497±0.019 | −2.31±0.02 | 0.002±0.026 | 1.16±0.01 | 1.15±0.04 | 1.19±0.03 | −1.89±0.05 |
| | **DGP-2** | 0.328±0.267 | 0.362±0.155 | −1.48±0.10 | 0.387±0.011 | −2.19±0.05 | −0.014±0.028 | 1.11±0.01 | 1.16±0.04 | 1.19±0.03 | −1.32±0.24 |
| | **DGP-3** | 0.380±0.249 | 0.373±0.140 | −1.44±0.13 | 0.403±0.014 | −2.20±0.05 | −0.012±0.028 | 1.10±0.01 | 1.16±0.03 | 1.19±0.03 | −1.10±0.23 |
| | **DDGP-2** | 0.376±0.410 | 0.184±0.128 | −1.59±0.07 | 0.268±0.030 | −2.27±0.07 | −0.036±0.030 | 1.05±0.01 | 1.20±0.08 | 1.18±0.03 | −1.66±0.18 |
| | **DDGP-3** | 0.914±0.595 | 0.308±0.202 | −1.59±0.06 | 0.170±0.029 | −2.09±0.22 | −0.036±0.030 | 1.01±0.01 | 1.51±0.28 | 1.20±0.04 | −0.853±0.532 |
| | **DKP-2** | 0.365±0.285 | 0.391±0.162 | −1.39±0.10 | 0.420±0.015 | −2.12±0.06 | 0.006±0.031 | 1.15±0.02 | 1.20±0.05 | 1.22±0.03 | −1.25±0.28 |
| | **DKP-3** | 0.431±0.271 | 0.414±0.153 | −1.51±0.09 | 0.376±0.016 | −2.18±0.06 | −0.020±0.031 | 1.08±0.02 | 1.18±0.04 | 1.18±0.03 | −1.18±0.27 |
| | *mean rank* ↓ | **SVGP**: $3.7_{\pm 2.6}$ | **DGP-2**: $3.5_{\pm 0.9}$ | **DGP-3**: $4.3_{\pm 1.2}$ | **DDGP-2**: $2.1_{\pm 1.3}$ | **DDGP-3**: $4.0_{\pm 2.8}$ | **DKP-2**: $5.6_{\pm 1.2}$ | **DKP-3**: $4.0_{\pm 1.7}$ | | | |
| Cov.95 → | **SVGP** | 0.961±0.030 | 0.947±0.027 | 0.970±0.017 | 0.980±0.004 | 1.00±0.00 | 0.962±0.004 | 0.960±0.003 | 0.943±0.012 | 0.945±0.007 | 0.994±0.013 |
| | **DGP-2** | 0.941±0.032 | 0.949±0.024 | 0.984±0.013 | 0.973±0.005 | 1.00±0.00 | 0.962±0.004 | 0.955±0.004 | 0.935±0.014 | 0.942±0.010 | 0.997±0.010 |
| | **DGP-3** | 0.935±0.029 | 0.932±0.032 | 0.991±0.010 | 0.975±0.005 | 1.00±0.00 | 0.965±0.005 | 0.954±0.003 | 0.939±0.019 | 0.944±0.007 | 0.987±0.021 |
| | **DDGP-2** | 0.871±0.053 | 0.906±0.032 | 0.964±0.025 | 0.966±0.006 | 1.00±0.00 | 0.960±0.006 | 0.950±0.003 | 0.894±0.024 | 0.933±0.009 | 0.990±0.021 |
| | **DDGP-3** | 0.786±0.059 | 0.870±0.034 | 0.970±0.013 | 0.960±0.005 | 1.000±0.001 | 0.961±0.006 | 0.941±0.004 | 0.809±0.059 | 0.924±0.011 | 0.965±0.057 |
| | **DKP-2** | 0.928±0.033 | 0.940±0.027 | 0.986±0.014 | 0.975±0.005 | 1.00±0.00 | 0.964±0.005 | 0.957±0.004 | 0.927±0.015 | 0.939±0.011 | 0.998±0.009 |
| | **DKP-3** | 0.918±0.031 | 0.921±0.034 | 0.993±0.009 | 0.977±0.005 | 1.00±0.00 | 0.967±0.005 | 0.956±0.004 | 0.932±0.019 | 0.941±0.009 | 0.985±0.022 |
| | *mean rank* ↓ | **SVGP**: $3.1_{\pm 2.3}$ | **DGP-2**: $2.8_{\pm 1.5}$ | **DGP-3**: $3.3_{\pm 1.6}$ | **DDGP-2**: $3.4_{\pm 2.3}$ | **DDGP-3**: $4.1_{\pm 2.7}$ | **DKP-2**: $4.4_{\pm 1.5}$ | **DKP-3**: $4.5_{\pm 1.9}$ | | | |

of $g(\cdot)$ components.

**PCA mean functions and hybrid kernels.** PCA-based mean functions as introduced in Salimbeni and Deisenroth [2017] partially alleviate distributional uncertainty collapse, but introduce a new failure mode: they extrapolate low distributional uncertainty in orthogonal directions to the leading eigenvectors, misidentifying those regions as in-distribution (Figure 3). Hybrid kernels resolve this directly since even when the mean is dominated by PCA projections, elevated posterior variance inflates the $\mathcal{W}_2$ distance, correctly flagging OOD inputs with higher distributional uncertainty. Our theoretical guarantees for DDGPs require inducing location variances to match in-distribution posterior variance, a condition that is self-enforcing under optimisation.

Taken together, Figures 3 and 4 provide complementary evidence for the theoretical claims of Section 4.3: DDGPs satisfy the OOD desiderata not by compressing representations, but by embedding a distributional notion of proximity into the kernel itself, which propagates coherently through depth and produces well-calibrated, spatially localised uncertainty at every layer of the hierarchy.

**Feature collapse and high-dimensional pathologies.** DDGPs also provide theoretical guarantees against *feature collapse* and smoother inter-layer representations (Appendix F), which we argue underlies their strong OOD detection performance. We further identify a previously unreported pathology in wide DGPs: hidden-layer samples collapse onto a thin hypersphere, demanding a disproportionate number of inducing points for adequate coverage. Hybrid kernels mitigate this pathology (Appendix C).

**Conclusion.** DDGPs consistently outperform standard DGPs on OOD detection across toy and real-world benchmarks, with improvements on in-distribution predictive performance. Promising directions for future work include applications to safety-critical settings such as biomedical image segmentation [Czolbe et al., 2021] and semantic segmentation [Franchi et al., 2020], where principled distributional uncertainty is of paramount importance.

## References


Laurence Aitchison, Adam Yang, and Sebastian W Ober. Deep kernel processes. In *International Conference on Machine Learning*, pages 130–140. PMLR, 2021.

William K. Allard. On the first variation of a varifold. *Annals of Mathematics*, 95(3):417–491, 1972. doi: 10.2307/1970859.

Frederick J. Almgren. The theory of varifolds: A variational calculus in the large for the $k$-dimensional area integrand. Mimeographed notes, Institute for Advanced Study, Princeton, 1965.

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

Francois Bachoc, Alexandra Suvorikova, David Ginsbourger, Jean-Michel Loubes, and Vladimir Spokoiny. Gaussian processes with multidimensional distribution inputs via optimal transport and hilbertian embedding. *arXiv preprint arXiv:1805.00753*, 2018.

Christian Berg, Jens Peter Reus Christensen, and Paul Ressel. *Harmonic Analysis on Semigroups: Theory of Positive Definite and Related Functions*, volume 100 of *Graduate Texts in Mathematics*. Springer, New York, 1984. doi: 10.1007/978-1-4612-1128-0.

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

Avrim Blum, John Hopcroft, and Ravindran Kannan. *Foundations of data science*. Cambridge University Press, 2020.

Nicolas Charon and Alain Trouvé. The varifold representation of nonoriented shapes for diffeomorphic registration. *SIAM journal on Imaging Sciences*, 6(4):2547–2580, 2013.

Tongtong Chen, Bin Dai, Ruili Wang, and Daxue Liu. Gaussian-process-based real-time ground segmentation for autonomous land vehicles. *Journal of Intelligent & Robotic Systems*, 76(3):563–582, 2014.

Michael Cowling, Uffe Haagerup, and Roger Howe. Almost $L^2$ matrix coefficients. *Journal für die reine und angewandte Mathematik*, 387:97–110, 1988. The completely monotone result used here is classical; see also Donoghue (1974) *Monotone Matrix Functions and Analytic Continuation*.

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

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

Rémi Domingues, Pietro Michiardi, Jihane Zouaoui, and Maurizio Filippone. Deep gaussian process autoencoders for novelty detection. *Machine Learning*, 107(8):1363–1383, 2018.

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

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

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

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

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

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

James Hensman, Nicolo Fusi, and Neil D Lawrence. Gaussian processes for big data. *arXiv preprint arXiv:1309.6835*, 2013.

José Miguel Hernández-Lobato and Ryan Adams. Probabilistic backpropagation for scalable learning of bayesian neural networks. In *International conference on machine learning*, pages 1861–1869. PMLR, 2015.

Diederik Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Soheil Kolouri, Yang Zou, and Gustavo K Rohde. Sliced wasserstein kernels for probability distributions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5258–5267, 2016.

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

Kimin Lee, Kibok Lee, Honglak Lee, and Jinwoo Shin. A simple unified framework for detecting out-of-distribution samples and adversarial attacks. *Advances in neural information processing systems*, 31, 2018.

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

Weitang Liu, Xiaoyun Wang, John Owens, and Yixuan Li. Energy-based out-of-distribution detection. *Advances in neural information processing systems*, 33:21464–21475, 2020b.

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

Sebastian Ober and Laurence Aitchison. A variational approximate posterior for the deep wishart process. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 6567–6579. Curran Associates, Inc., 2021a. URL https://proceedings.neurips.cc/paper_files/paper/2021/file/33ceb07bf4eeb3da587e268d663aba1a-Paper.pdf.

Sebastian W Ober and Laurence Aitchison. Global inducing point variational posteriors for bayesian neural networks and deep gaussian processes. In *International Conference on Machine Learning*, pages 8248–8259. PMLR, 2021b.

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

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

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

Hugh Salimbeni, Vincent Dutordoir, James Hensman, and Marc Peter Deisenroth. Deep gaussian processes with importance-weighted variational inference. *arXiv preprint arXiv:1905.05435*, 2019.

Isaac J. Schoenberg. Metric spaces and completely monotone functions. *Annals of Mathematics*, 39(4):811–841, 1938. doi: 10.2307/1968466.

Leon Simon. *Lectures on Geometric Measure Theory*, volume 3 of *Proceedings of the Centre for Mathematical Analysis*. Australian National University, Canberra, 1984. ISBN 0-86784-429-9.

Bui Thi Thien Trang, Jean-Michel Loubes, Laurent Risser, and Patricia Balaresque. Distribution regression model with a reproducing kernel hilbert space approach. *Communications in Statistics-Theory and Methods*, pages 1–23, 2019.

Juho Timonen, Henrik Mannerström, Aki Vehtari, and Harri Lähdesmäki. lgpr: An interpretable nonparametric method for inferring covariate effects from longitudinal data. *arXiv preprint arXiv:1912.03549*, 2019.

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

Ivan Ustyuzhaninov, Ieva Kazlauskaite, Markus Kaiser, Erik Bodin, Neill Campbell, and Carl Henrik Ek. Compositional uncertainty in deep gaussian processes. In Jonas Peters and David Sontag, editors, *Proceedings of the 36th Conference on Uncertainty in Artificial Intelligence (UAI)*, volume 124 of *Proceedings of Machine Learning Research*, pages 480–489. PMLR, 03–06 Aug 2020. URL https://proceedings.mlr.press/v124/ustyuzhaninov20a.html.

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

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

Cédric Villani. *Optimal transport: old and new*, volume 338. Springer Science & Business Media, 2008.

Cédric Villani. *Optimal Transport: Old and New*, volume 338 of *Grundlehren der mathematischen Wissenschaften*. Springer, Berlin, 2009. doi: 10.1007/978-3-540-71050-9.

Dennis Wang, James Hensman, Ginte Kutkaite, Tzen S Toh, Ana Galhoz, Jonathan R Dry, Julio Saez-Rodriguez, Mathew J Garnett, Michael P Menden, Frank Dondelinger, et al. A statistical framework for assessing pharmacological responses and biomarkers using uncertainty estimates. *Elife*, 9:e60352, 2020.

Christopher KI Williams and Carl Edward Rasmussen. *Gaussian processes for machine learning*, volume 2. MIT press Cambridge, MA, 2006.

Haibin Yu, Yizhou Chen, Zhongxiang Dai, Kian Hsiang Low, and Patrick Jaillet. Implicit posterior variational inference for deep gaussian processes. *arXiv preprint arXiv:1910.11998*, 2019.

# Distributional Deep Gaussian Processes (Supplementary Material)

**Sebastian G. Popescu**[1]

[1]Find & Order Labs

## A ADDITIONAL BACKGROUND MATERIAL

### A.1 GAUSSIAN PROCESS PREDICTION WITH UNCERTAIN INPUTS

For SVGP we established that the predictive distribution of a GP at a *noise-free* test point is Gaussian. We now consider the practically important setting where the test input is itself uncertain, modelled as a random variable. This situation arises naturally in DGPs, where the output of one layer serves as the (stochastic) input to the next: computing the inter-layer predictive distribution requires integrating out the uncertain intermediate representation. The treatment here follows Girard [2004].

**Problem formulation.** Suppose the test input is distributed as $x^* \sim \mathcal{N}(\boldsymbol{\mu}_{x^*}, \boldsymbol{\Sigma}_{x^*})$. The marginal predictive distribution is

$$p(y^* \mid \mathcal{D},\, \boldsymbol{\mu}_{x^*}, \boldsymbol{\Sigma}_{x^*}) = \int p(y^* \mid \mathcal{D}, x^*)\; \mathcal{N}(x^* \mid \boldsymbol{\mu}_{x^*}, \boldsymbol{\Sigma}_{x^*})\; \mathrm{d}x^*, \tag{22}$$

where the inner term is the standard GP predictive density at a fixed $x^*$,

$$p(y^* \mid \mathcal{D}, x^*) = \mathcal{N}\big(y^* \mid \mu(x^*),\, \sigma^2(x^*)\big)\,, \qquad \mu(x^*) = \mathbf{k}_{*f}\mathbf{K}_{ff}^{-1}\mathbf{y}, \quad \sigma^2(x^*) = k_{**} - \mathbf{k}_{*f}\mathbf{K}_{ff}^{-1}\mathbf{k}_{f*}. \tag{23}$$

Here we adopt the shorthand $k_{**} = K(x^*, x^*)$, $\mathbf{k}_{*f} = [K(x^*, x_i)]_{i=1}^N$, and $K_{ij} = K(x_i, x_j) + \delta_{ij}\sigma^2_{\text{noise}}$. Because $\mu(x^*)$ and $\sigma^2(x^*)$ are non-linear functions of $x^*$, the integral in (22) is analytically intractable for general kernels. We therefore seek a tractable Gaussian approximation.

#### A.1.1 Gaussian Approximation via the Law of Total Variance

We approximate (22) by a Gaussian whose first two moments match those of the true marginal:

$$p(y^* \mid \mathcal{D}, \boldsymbol{\mu}_{x^*}, \boldsymbol{\Sigma}_{x^*}) \;\approx\; \mathcal{N}(y^* \mid m(\boldsymbol{\mu}_{x^*}, \boldsymbol{\Sigma}_{x^*}),\, v(\boldsymbol{\mu}_{x^*}, \boldsymbol{\Sigma}_{x^*}))\,. \tag{24}$$

Applying the law of total expectation and the law of total variance to (22) yields the exact moment expressions:

$$m(\boldsymbol{\mu}_{x^*}, \boldsymbol{\Sigma}_{x^*}) = \mathbb{E}_{x^*}[\mu(x^*)]\,, \tag{25}$$

$$v(\boldsymbol{\mu}_{x^*}, \boldsymbol{\Sigma}_{x^*}) = \underbrace{\mathbb{E}_{x^*}\big[\sigma^2(x^*)\big]}_{\text{average noise}} + \underbrace{\mathbb{V}_{x^*}[\mu(x^*)]}_{\text{mean variability}}, \tag{26}$$

where $\mathbb{V}_{x^*}[\,\cdot\,]$ denotes variance with respect to the input distribution. The decomposition in (26) has a natural interpretation: the total predictive uncertainty is the sum of the *aleatoric* uncertainty averaged over the input distribution, plus the additional *epistemic* uncertainty arising from not knowing $x^*$ exactly. In the limit $\boldsymbol{\Sigma}_{x^*} \to \mathbf{0}$ both extra terms vanish and (25)–(26) recover the standard noise-free GP predictive moments.

The remaining challenge is computing the expectations in (25) and (26), since $\mu(x^*)$ and $\sigma^2(x^*)$ are non-linear in $x^*$. We now derive tractable approximations via Taylor expansion.

### A.1.2 Approximate Moments via Taylor Expansion

**Univariate case.** Let $x$ be a scalar random variable with $\mathbb{E}[x] = \mu_x$ and $\mathbb{V}[x] = \sigma_x^2$. For a smooth function $f$ and sufficiently small $\sigma_x$, a second-order Taylor expansion about $\mu_x$ gives

$$f(x) \;\approx\; f(\mu_x) + f'(\mu_x)(x - \mu_x) + \frac{1}{2} f''(\mu_x)(x - \mu_x)^2. \tag{27}$$

Taking expectations term by term, and using $\mathbb{E}[x - \mu_x] = 0$ and $\mathbb{E}[(x - \mu_x)^2] = \sigma_x^2$:

$$\mathbb{E}[f(x)] \approx f(\mu_x) + \frac{1}{2}\sigma_x^2 f''(\mu_x), \tag{28}$$

$$\mathbb{V}[f(x)] \approx \sigma_x^2 \left[f'(\mu_x)\right]^2. \tag{29}$$

Equation (29) follows from $\mathbb{V}[f(x)] = \mathbb{E}[f(x)^2] - \mathbb{E}[f(x)]^2$ by expanding to first order in $\sigma_x^2$: $\mathbb{E}[f(x)^2] \approx f(\mu_x)^2 + \sigma_x^2[f'(\mu_x)]^2 + \sigma_x^2 f(\mu_x) f''(\mu_x)$ and $\mathbb{E}[f(x)]^2 \approx f(\mu_x)^2 + \sigma_x^2 f(\mu_x) f''(\mu_x)$, so the $f(\mu_x)f''(\mu_x)$ terms cancel exactly, leaving $\mathbb{V}[f(x)] \approx \sigma_x^2[f'(\mu_x)]^2$.

**Multivariate generalisation.** For a vector $\mathbf{x} \sim \mathcal{N}(\boldsymbol{\mu}_x, \boldsymbol{\Sigma}_x)$ the second-order expansion of $f(\mathbf{x})$ about $\boldsymbol{\mu}_x$ is

$$f(\mathbf{x}) \approx f(\boldsymbol{\mu}_x) + \nabla f(\boldsymbol{\mu}_x)^\top (\mathbf{x} - \boldsymbol{\mu}_x) + \frac{1}{2}(\mathbf{x} - \boldsymbol{\mu}_x)^\top \mathbf{H}_f(\boldsymbol{\mu}_x)(\mathbf{x} - \boldsymbol{\mu}_x), \tag{30}$$

where $\mathbf{H}_f(\boldsymbol{\mu}_x)$ is the Hessian of $f$ evaluated at $\boldsymbol{\mu}_x$. The corresponding moment approximations are:

$$\mathbb{E}_{\mathbf{x}}[f(\mathbf{x})] \approx f(\boldsymbol{\mu}_x) + \frac{1}{2}\,\mathrm{Tr}[\mathbf{H}_f(\boldsymbol{\mu}_x)\,\boldsymbol{\Sigma}_x], \tag{31}$$

$$\mathbb{V}_{\mathbf{x}}[f(\mathbf{x})] \approx \nabla f(\boldsymbol{\mu}_x)^\top \boldsymbol{\Sigma}_x\, \nabla f(\boldsymbol{\mu}_x). \tag{32}$$

**Application to GP prediction.** Applying (31)–(32) to the exact moment expressions (25)–(26) gives the following closed-form approximations:

$$m(\boldsymbol{\mu}_{x^*}, \boldsymbol{\Sigma}_{x^*}) \approx \mu(\boldsymbol{\mu}_{x^*}) + \frac{1}{2}\,\mathrm{Tr}[\mathbf{H}_\mu(\boldsymbol{\mu}_{x^*})\,\boldsymbol{\Sigma}_{x^*}]\,, \tag{33}$$

$$v(\boldsymbol{\mu}_{x^*}, \boldsymbol{\Sigma}_{x^*}) \approx \sigma^2(\boldsymbol{\mu}_{x^*}) + \frac{1}{2}\,\mathrm{Tr}[\mathbf{H}_{\sigma^2}(\boldsymbol{\mu}_{x^*})\,\boldsymbol{\Sigma}_{x^*}] + \nabla\mu(\boldsymbol{\mu}_{x^*})^\top \boldsymbol{\Sigma}_{x^*}\, \nabla\mu(\boldsymbol{\mu}_{x^*}), \tag{34}$$

where $\mathbf{H}_\mu$ and $\mathbf{H}_{\sigma^2}$ denote the Hessians of the GP predictive mean and variance with respect to $x^*$, evaluated at the input mean $\boldsymbol{\mu}_{x^*}$.

**Remark A.1.** *Equation* (33) *reveals that input uncertainty inflates the predictive mean by a curvature-weighted correction: if the GP mean function is concave (convex) at $\boldsymbol{\mu}_{x^*}$, the correction is negative (positive), reflecting Jensen's inequality. Equation* (34) *shows that the total predictive variance is the sum of (i) the noise-free predictive variance at the input mean, (ii) a second-order correction from the curvature of $\sigma^2$, and (iii) the variance of the GP mean induced by input uncertainty, which grows with both the gradient magnitude and the input variance $\boldsymbol{\Sigma}_{x^*}$. The final term is the dominant contribution when the GP mean is steep and input uncertainty is non-negligible.*

**Remark A.2.** *In the DSVI framework for DGPs [Salimbeni and Deisenroth, 2017], the single-sample estimator $T = 1$ is used at each layer to propagate uncertainty, effectively replacing the integral in* (22) *with a point estimate. The Taylor approximation* (33)–(34) *provides a more principled alternative by propagating the full Gaussian uncertainty analytically, at the cost of computing first and second derivatives of the GP predictive functions—quantities available in closed form for SE kernels [Girard, 2004].*

## A.2 VARIFOLD THEORY AND CONNECTIONS TO HYBRID KERNELS

**Background.** Varifolds are measure-theoretic generalisations of differentiable manifolds, introduced by Almgren [1965] and developed into a rigorous framework by Allard [1972]. Formally, a *$k$-varifold* on an open set $\Omega \subset \mathbb{R}^n$ is a Radon

measure on the Grassmannian bundle $\Omega \times G(n,k)$, where $G(n,k)$ denotes the space of $k$-dimensional linear subspaces of $\mathbb{R}^n$. A varifold $V$ is *rectifiable* if it can be written as:

$$V = \theta(x)\, \mathcal{H}^k \lfloor_M \otimes \delta_{T_x M}, \tag{35}$$

where $M \subset \mathbb{R}^n$ is a $k$-rectifiable set, $T_x M \in G(n,k)$ is the approximate tangent plane at $x$, $\mathcal{H}^k$ is the $k$-dimensional Hausdorff measure, and $\theta : M \to \mathbb{R}^+$ is an integer-valued density function [Allard, 1972, Simon, 1984]. Intuitively, a rectifiable varifold simultaneously encodes the *position* of a surface (via its support) and its *local geometry* (via the tangent plane field), without requiring a global orientation.

**Varifolds for shape comparison.** Charon and Trouvé [2013] adapted this framework to computational anatomy, enabling metric comparisons between unoriented surfaces in $\mathbb{R}^3$. Given a rectifiable surface $M \subset \mathbb{R}^d$, one associates the varifold $[M] \in W'$ defined by its action on test functions $\omega \in W$ as:

$$[M](\omega) = \int_M \omega(x,\, \tau_x M)\, \mathrm{d}\mathcal{H}^k(x), \tag{36}$$

where $\tau_x M \in \mathbb{S}^{d-1}$ is the unit normal at $x$. The space $W$ is the RKHS associated to the tensor-product kernel $k_e \otimes k_t$, where $k_e : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ is a position kernel and $k_t : \mathbb{S}^{d-1} \times \mathbb{S}^{d-1} \to \mathbb{R}$ is a tangent kernel [Charon and Trouvé, 2013]:

**Proposition A.1** (Charon and Trouvé 2013, Prop. 4.1)**.** *Let $k_e$ be a continuous, positive-definite kernel on $\mathbb{R}^d$ that vanishes at infinity, and $k_t$ a continuous, positive-definite kernel on $\mathbb{S}^{d-1}$. Then the RKHS $W$ of $k_e \otimes k_t$ embeds continuously into $C_0(\mathbb{R}^d \times \mathbb{S}^{d-1})$, and the inner product between two surface varifolds $[S]$ and $[T]$ is:*

$$\langle [S],\, [T] \rangle_W = \int_S \int_T k_e(x,y)\, k_t(\tau_x S,\, \tau_y T)\, \mathrm{d}\mathcal{H}^k(x)\, \mathrm{d}\mathcal{H}^k(y). \tag{37}$$

Equation (37) jointly measures *spatial proximity* via $k_e$ and *geometric similarity* of tangent planes via $k_t$. Points close in position but with misaligned normals—e.g. across a sharp fold—are correctly penalised by the tangent term.

**Connections to the hybrid kernel.** The hybrid kernel (4.1) is directly inspired by (37). Consider drawing $S$ samples from the layer-$(l-1)$ posterior $q(\mathbf{f}_{l-1})$, yielding points $\{x_i\}_{i=1}^S$ and associated Gaussian marginals $\{\mu_i\}_{i=1}^S$. The multi-sample hybrid kernel evaluates as:

$$k^H(\{x_i,\mu_i\},\{x_j,\mu_j\}) = \sum_{i=1}^S \sum_{j=1}^S k^{\mathrm{SE}}(x_i,x_j) \cdot \exp\left( -\sum_{d=1}^D \frac{W_2^2(\mu_{i,d},\, \mu_{j,d})}{\ell_d^2} \right). \tag{38}$$

The structural analogy with (37) is transparent:

| **Varifold kernel** | | **Hybrid kernel** |
|---|---|---|
| Position kernel $k_e(x,y)$ | $\longleftrightarrow$ | $k^{\mathrm{SE}}(x_i,x_j)$ |
| Tangent kernel $k_t(\tau_x S, \tau_y T)$ | $\longleftrightarrow$ | $\exp(-W_2^2(\mu_{i,d},\mu_{j,d})/\ell_d^2)$ |
| Tangent plane $\tau_x M \in \mathbb{S}^{d-1}$ | $\longleftrightarrow$ | Gaussian marginal $\mu_i \in \mathcal{W}_2(\mathbb{R})$ |
| Surface integral $\int_S \int_T$ | $\longleftrightarrow$ | Sample sum $\sum_i \sum_j$ |

In the varifold setting, $k_t$ compares the *orientation* of surface patches; in DDGPs, the $\mathcal{W}_2$ term compares the *distributional geometry* of hidden-layer GP marginals. Just as a varifold assigns geometric meaning to each surface point through its tangent plane, a DDGP assigns distributional meaning to each hidden-layer sample through its posterior Gaussian—encoding whether that point lies inside or outside the training distribution. The hybrid kernel is thus a distributional analogue of the varifold kernel, operating on the *Grassmannian of Gaussian measures* rather than the Grassmannian of tangent planes.

### A.3 PROOF OF THEOREM 2.1

We provide a self-contained proof that the hybrid kernel's $\mathcal{W}_2$ component is positive definite, following Thi Thien Trang et al. [2019] and Bachoc et al. [2017]. The argument proceeds in two steps: we first show that $W_2^{2H}$ is a negative definite kernel (Proposition A.2), and then apply Schoenberg's theorem (Theorem A.1) to lift this to positive definiteness.

**Setup.** Let $\Omega \subset \mathbb{R}$ be a compact subset and let $\mathcal{W}_2(\Omega)$ denote the space of probability measures on $\Omega$ with finite second moment, endowed with the quadratic Wasserstein distance

$$W_2(\mu,\nu) = \left( \inf_{\pi \in \Pi(\mu,\nu)} \int |x-y|^2 \, \mathrm{d}\pi(x,y) \right)^{1/2}, \tag{39}$$

where $\Pi(\mu,\nu)$ is the set of couplings of $\mu$ and $\nu$. For $\mu \in \mathcal{W}_2(\Omega)$, denote by $F_\mu^{-1}$ the quantile function $F_\mu^{-1}(t) = \inf\{u \in \Omega : F_\mu(u) \geq t\}$. Since $\Omega$ is compact, for any $\mu, \nu \in \mathcal{W}_2(\Omega)$ the Wasserstein distance admits the closed-form expression [Villani, 2009, Theorem 2.18]:

$$W_2^2(\mu,\nu) = \mathbb{E}\Big[ \big( F_\mu^{-1}(V) - F_\nu^{-1}(V) \big)^2 \Big], \tag{40}$$

where $V$ is a uniform random variable on $[a,b] \subseteq [0,1]$, and $(F_\mu^{-1}(V),\, F_\nu^{-1}(V))$ is the optimal coupling.

**Theorem A.1** (Schoenberg Schoenberg, 1938)**.** *Let $F : \mathbb{R}^+ \to \mathbb{R}^+$ be a completely monotone function, and let $\mathcal{K}$ be a negative definite kernel. Then the kernel $(x,y) \mapsto F(\mathcal{K}(x,y))$ is positive definite.*

Recall that $F$ is *completely monotone* if $(-1)^n F^{(n)}(t) \geq 0$ for all $t > 0$ and $n \geq 0$; in particular, $F(x) = e^{-\lambda x}$ for $\lambda > 0$ is completely monotone [Cowling et al., 1988].

**Proposition A.2** (Thi Thien Trang et al., 2019, Prop. 2.3)**.** *The function $W_2^{2H}$ is a negative definite kernel on $\mathcal{W}_2(\Omega)$ if and only if $0 < H \leq 1$.*

*Proof.* We follow Thi Thien Trang et al. [2019] and extend the result of Kolouri et al. [2016] from absolutely continuous distributions to all distributions in $\mathcal{W}_2(\Omega)$.

Let $\mu_1, \ldots, \mu_n \in \mathcal{W}_2(\Omega)$ and $c_1, \ldots, c_n \in \mathbb{R}$ with $\sum_{i=1}^n c_i = 0$. Using the quantile representation (40),

$$\begin{aligned} \sum_{i,j=1}^n c_i c_j \, W_2^2(\mu_i,\mu_j) &= \sum_{i,j=1}^n c_i c_j \, \mathbb{E}\left[ \left( F_{\mu_i}^{-1}(V) - F_{\mu_j}^{-1}(V) \right)^2 \right] \\ &= \sum_{i,j=1}^n c_i c_j \, \mathbb{E}\big[ F_{\mu_i}^{-1}(V)^2 \big] + \sum_{i,j=1}^n c_i c_j \, \mathbb{E}\Big[ F_{\mu_j}^{-1}(V)^2 \Big] - 2 \sum_{i,j=1}^n c_i c_j \, \mathbb{E}\Big[ F_{\mu_i}^{-1}(V) \, F_{\mu_j}^{-1}(V) \Big]. \end{aligned} \tag{41}$$

Since $\sum_{i=1}^n c_i = 0$, the first two sums vanish: $\sum_{i,j} c_i c_j = (\sum_i c_i)(\sum_j c_j) = 0$. Hence,

$$\sum_{i,j=1}^n c_i c_j \, W_2^2(\mu_i,\mu_j) = -2 \sum_{i,j=1}^n c_i c_j \, \mathbb{E}\Big[ F_{\mu_i}^{-1}(V) \, F_{\mu_j}^{-1}(V) \Big] = -2 \, \mathbb{E}\left[ \left( \sum_{i=1}^n c_i \, F_{\mu_i}^{-1}(V) \right)^2 \right] \leq 0. \tag{42}$$

This establishes that $W_2^2$ is a negative definite kernel ($H = 1$). For general $0 < H \leq 1$, the function $t \mapsto t^H$ is a Bernstein function, and the composition of a Bernstein function with a negative definite kernel is again negative definite [Berg et al., 1984, Corollary 3.2]. The converse ($H > 1$ fails) follows from the fact that $t \mapsto t^H$ is not a Bernstein function for $H > 1$. □

We are now in a position to prove the main theorem.

**Theorem A.2** (Thi Thien Trang et al., 2019, Thm. 2.1)**.** *Let $\Theta := (\gamma, H, \ell)$ with $\gamma \neq 0$, $\ell > 0$, and $0 < H \leq 1$. The kernel $k_\Theta : \mathcal{W}_2(\Omega) \times \mathcal{W}_2(\Omega) \to \mathbb{R}$ defined by*

$$k_\Theta(\mu,\nu) := \gamma^2 \exp\left( -\frac{W_2^{2H}(\mu,\nu)}{\ell} \right) \tag{43}$$

*is a positive definite kernel.*

*Proof.* By Proposition A.2, $W_2^{2H}(\mu,\nu)$ is a negative definite kernel for $0 < H \leq 1$. Define $F : \mathbb{R}^+ \to \mathbb{R}^+$ by $F(x) = \gamma^2 e^{-x/\ell}$. Since $\ell > 0$ and $\gamma^2 > 0$, the function $F$ is completely monotone (as the composition of a positive constant with a decaying exponential). Applying Schoenberg's Theorem (Theorem A.1) with $\mathcal{K} = W_2^{2H}$ and this choice of $F$ yields that $k_\Theta(\mu,\nu) = F(W_2^{2H}(\mu,\nu))$ is a positive definite kernel. □

**Remark A.3.** *Theorem A.2 justifies the use of the $\mathcal{W}_2$ component in our hybrid kernel* (4.1)*. In practice we set $H = 1$, which corresponds to the standard squared Wasserstein distance and yields the tightest negative definiteness constant. The scale parameter $\ell > 0$ plays the role of a lengthscale and is learned during training via gradient-based optimisation of the variational ELBO.*

# B EXPERIMENTS DETAILS

## B.1 DATASETS

We evaluate on two benchmark suites.

**Regression.** We use 10 datasets from the UCI repository: Boston Housing ($N$=506, $D$=13), Concrete ($N$=1030, $D$=8), Energy Efficiency ($N$=768, $D$=8), Kin8nm ($N$=8192, $D$=8), Naval Propulsion ($N$=11934, $D$=16), Power Plant ($N$=9568, $D$=4), Protein Structure ($N$=45730, $D$=9), Wine Red ($N$=1599, $D$=11), Wine White ($N$=4898, $D$=11), and Yacht Hydrodynamics ($N$=308, $D$=6). Following the protocol of Hernández-Lobato and Adams [2015], each dataset is split into 90% training and 10% test partitions over 20 independent random splits (base seed 42), with inputs and targets normalised to zero mean and unit variance using training-set statistics.

**OOD detection.** For the MNIST inlier setting, we evaluate against 11 OOD datasets spanning three categories: *Near-OOD* (Fashion-MNIST, MNIST-C, KMNIST), *Far-OOD* (CIFAR-10 grayscale, NotMNIST, Omniglot, Random Noise, SVHN grayscale), and *Covariate Shift* (MNIST-M, Scaled/Translated MNIST, USPS). For the CIFAR-10 inlier setting, we evaluate against 10 OOD datasets: *Near-OOD* (CIFAR-100, STL-10, Tiny ImageNet), *Far-OOD* (Places365, Random Noise, SVHN, Textures), and *Covariate Shift* (CIFAR-10.1, CIFAR-10-C, CIFAR-10 grayscale). OOD detection performance is measured by AUROC↑ and FPR@95↓, computed by comparing uncertainty scores on the held-out test set against each OOD dataset.

## B.2 MODELS

We compare five model families across all experiments. A single-layer sparse variational GP (**SVGP**) serves as the shallow baseline; on the image tasks this shallow layer is the convolutional GP of Van der Wilk et al. [2017], whose inducing points live in patch space and whose output is the weighted response over image patches. Deep GPs with 2 and 3 hidden layers (**DGP-2**, **DGP-3**) and their distributional counterparts (**DDGP-2**, **DDGP-3**) form the core comparison; the deep image models stack convolutional GP layers (patch size 5, stride 2 in the first layer and 1 thereafter, 5 response channels per hidden layer) in the deep-convolutional-GP construction of Blomqvist et al. [2018]. All models use RBF (squared-exponential) kernels with a Gaussian likelihood for the tabular regression tasks (learned observation-noise variance initialised at $10^{-2}$) and a robust-max/softmax likelihood for image classification. Training uses doubly-stochastic variational inference [Salimbeni and Deisenroth, 2017]: the ELBO is estimated by propagating samples through the layer composition, with 1 Monte-Carlo sample per data point for the tabular models and 5 for the image models.

All deep models use $M$=100 inducing points per layer for tabular experiments and $M$=384 (MNIST) or $M$=500 (CIFAR-10) for image experiments. Hidden dimensionality is set to $\min(D, 2)$ for tabular tasks and 5 units per hidden layer for image tasks. A whitened variational parameterisation is used throughout, with the variational covariance associated to inducing points in inner layers initialised to identity scaled by $10^{-5}$ and PCA linear mean functions at every layer.

**Deterministic baselines (image experiments).** For the OOD image benchmarks we additionally compare against two strong deterministic detectors, each *parameter-matched* to the 3-layer DConvGP so that every row of the OOD tables operates at a comparable per-model budget. Both share a WideResNet backbone (residual blocks with GroupNorm, 8 groups): **WRN-16-3** for MNIST (stage channels $[48, 96, 192]$, 2 blocks per stage, $\sim$1.55M parameters) and **WRN-16-4** for CIFAR-10 (stage channels $[64, 128, 256]$, 2 blocks per stage, $\sim$2.76M parameters), matching the 3-layer DConvGP at $\sim$1.59M (MNIST) and $\sim$2.68M (CIFAR-10) almost exactly. These are downsized from the canonical WRN-28-10 specifically to enforce per-model parity with the GP rows rather than to maximise raw accuracy.

**Deep Ensembles (DE)** [Lakshminarayanan et al., 2017] use $M$=10 members, each an independently initialised and separately trained WideResNet of the backbone above; members are optimised sequentially so only one member resides in GPU memory at a time. Predictions are the mean of the per-member softmax logits and the epistemic uncertainty is their across-member variance. Note the total ensemble budget is therefore $M\times$ a single member ($\sim$15.5M on MNIST, $\sim$27.5M on CIFAR-10); parity is enforced *per member* against the 3-layer DConvGP.

**SNGP** [Liu et al., 2020a] is a single deterministic model using the *same* WideResNet backbone but with (i) spectral normalisation applied to every hidden convolution (spectral-norm multiplier $6.0$, one power-iteration step per forward), rendering the feature extractor approximately bi-Lipschitz, and (ii) the linear classifier replaced by a distance-aware random-Fourier-feature Gaussian-process head with a Laplace-approximated posterior. The GP head uses orthogonal random features ($D_{\mathrm{RFF}}$=128 for MNIST, 1024 for CIFAR-10), kernel init stddev $0.05$, ridge penalty $\tau$=1, and mean-field logit calibration with $\lambda$=10 (MNIST) / $\lambda$=20 (CIFAR-10); this keeps the total trainable count at $\sim$1.55M / $\sim$2.76M, matching both the 3-layer DConvGP and a single DE member. Epistemic uncertainty is read off the GP head's predictive variance.

Crucially, neither DE nor SNGP admits a *distributional* (Wasserstein-2) uncertainty in the sense of our distributional inducing points: both expose only a predictive/epistemic uncertainty. We therefore report the *distributional entropy* score exclusively for the GP-family models, which is why the DE and SNGP rows are left blank (–) under that score in Tables 4.

### B.3 KERNEL CONFIGURATION

**Euclidean kernels.** All Squared Exponential (SE) kernels employ automatic relevance determination (ARD), with both lengthscale and signal variance initialised to $1.351$ per input dimension.

**Hybrid kernels.** DDGP models use the SE kernel in the first layer and the hybrid $\mathcal{W}_2$-SE kernel in all subsequent layers. Lengthscale and signal variance are initialised identically to the Euclidean case ($1.351$); the distributional inducing point variances are initialised to $0.01$.

### B.4 INDUCING POINTS

**Standard Inducing Points** We initialize the inducing point locations to the k-means of the training data. In the case of image data, inducing points are initialized by running k-means on image patches from the training set.

**Distributional inducing points** In the tabular setting, the GIPL variant maintains a single set of $M$ learnable base inducing locations $\mathbf{Z}_0 \in \mathbb{R}^{M\times D}$ at the model level, initialized via K-means clustering on the training inputs. Layer 0 is a standard sparse GP layer with a squared exponential kernel; at each forward pass, $\mathbf{Z}_0$ is propagated through this layer's predictive posterior to obtain distributional inducing representations $(\boldsymbol{\mu}_{Z_{(1)}}, \boldsymbol{\sigma}^2_{Z_{(1)}})$ for the next layer, alongside a reparameterized sample $\mathbf{Z}_1$. For layers $\ell \geq 1$, each GIPL distributional layer uses a Wasserstein-2 hybrid kernel that accepts these propagated moments; the layer computes its own conditional at the current inducing locations to produce $(\mathbf{Z}_{\ell+1}, \boldsymbol{\mu}_{Z_{\ell+1}}, \boldsymbol{\sigma}^2_{Z_{\ell+1}})$ for subsequent layers. Crucially, no per-layer inducing locations are stored beyond $\mathbf{Z}_0$. All downstream distributional inducing representations are derived on-the-fly via this dual propagation of samples and moments. The variational parameters $\mathbf{q}_\mu$ are initialized to zero, while the Cholesky factors are set to $\alpha\mathbf{I}$ with $\alpha = 10^{-5}$ for inner layers (yielding near-deterministic posteriors initially) and $\mathbf{I}$ for the final layer. When whitening is enabled (default), the KL divergence simplifies to $\mathrm{KL}[q(\mathbf{U})\|\mathcal{N}(\mathbf{0}, \mathbf{I})]$, avoiding costly kernel evaluations at the distributional inducing locations during training.

In the case of image data, we maintain a set of $M_z$ learnable inducing images $\mathbf{Z}_0 \in \mathbb{R}^{M_z\times H\times W\times C}$ at the model level, initialized as a random subset of training images. Layer 0 is a standard convolutional GP layer with inducing patch locations initialized via K-means on patches extracted from the training data. At inference, the inducing images $\mathbf{Z}_0$ are propagated through layer 0 to produce distributional representations $(\boldsymbol{\mu}_{Z_1}, \boldsymbol{\sigma}^2_{Z_1})$ via the layer's predictive posterior, which will serve as the distributional inducing locations (means and variances) for the next layer. This amortization across layers means that all inducing locations share the same $M_z$ base images, with the number of inducing points per layer being $M_z \times P_\ell$ where $P_\ell$ is the number of patches at layer $\ell$.

### B.5 TRAINING

**Tabular and UCI experiments.** All models are trained for 20,000 iterations using the Adam optimiser [Kingma and Ba, 2014] with learning rate $10^{-3}$ ($\beta_1$=0.9, $\beta_2$=0.999) and a cosine annealing schedule with 500 linear warm-up steps. Mini-batches of size 512 are used; no gradient clipping is applied. For regression, a Gaussian likelihood with initial noise variance $\sigma^2$=0.01 is used. For classification, the likelihood is selected by number of classes: Bernoulli for binary tasks; Softmax with 100 Monte Carlo integration points for multi-class. The doubly-stochastic variational ELBO is estimated with a single Monte Carlo sample ($S$=1) per iteration. At test time, predictive distributions are obtained from 100 Monte Carlo forward passes, batched in groups of 512.

**Image experiments.** Convolutional DGP and DDGP models are trained for 50,000 iterations with mini-batches of size 32, using the full MNIST or CIFAR-10 training set as appropriate. All other optimiser settings are identical to the tabular configuration. Results are reported for architectures with 1 and 2 hidden layers.

**Deterministic baselines.** The parameter-matched Deep Ensemble (DE) and SNGP baselines follow the *identical* optimisation schedule: 50,000 iterations, mini-batches of size 32, and the same Adam optimiser with cosine decay and 500-step linear warmup used for the GP models. For DE, each of the $M{=}10$ WideResNet members is trained *independently and sequentially*, with per-member initialisation and batch-ordering seeds derived from the base seed so that both the weights and the SGD trajectory decorrelate across members; only one member occupies GPU memory at a time. For SNGP, the spectral-normalised backbone and the random-feature GP head are trained jointly for the same 50,000 iterations; the head's Laplace precision matrix is reset at the start of every "epoch", and after the optimiser has converged a single final pass over the full training set accumulates the posterior precision $\Sigma^{-1}$ used at test time.

# C ISOTROPIC KERNEL PROPERTIES IN (DISTRIBUTIONAL) DEEP GAUSSIAN PROCESSES

**Point-mass inducing points fail in high dimensions.** We present a theoretical analysis of DGP behaviour in high-dimensional hidden spaces, noting that the argument extends directly to DeepConvGP architectures. The central observation is that concentration-of-measure phenomena cause Euclidean-distance-based kernels to become uninformative as layer width grows, an issue that distributional inducing points circumvent by design.

We begin with the classical result that underpins the argument.

**Definition C.1** (Isotropic Gaussian)**.** *A $d$-dimensional isotropic Gaussian with mean $\boldsymbol{\mu} \in \mathbb{R}^d$ and scalar variance $\sigma^2 > 0$ is the distribution $\mathcal{N}(\boldsymbol{\mu},\, \sigma^2 \mathbf{I}_d)$, with density*

$$p(\mathbf{x}) = \frac{1}{(2\pi\sigma^2)^{d/2}} \exp\left[-\frac{\|\mathbf{x}-\boldsymbol{\mu}\|^2}{2\sigma^2}\right]. \tag{44}$$

**Theorem C.1** (Gaussian Annulus Theorem, Blum et al. 2020, Ch. 2)**.** *Let $\mathbf{x} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_d)$. For any $\beta > 0$,*

$$\Pr\Big(\,|\,\|\mathbf{x}\| - \sqrt{d}\,| \geq \beta\Big) \;\leq\; 2\exp\big(-c\beta^2\big)\,, \tag{45}$$

*where $c > 0$ is a universal constant. Consequently, with high probability all samples lie within a thin shell of radius $\sqrt{d}$ and width $O(1)$ centred at the origin.*

**Why point-mass inducing points fail.** Consider the $l$-th hidden layer of a DGP with approximate posterior $q(\mathbf{F}_l) = \mathcal{N}(\tilde{\mathbf{U}}_l, \tilde{\boldsymbol{\Sigma}}_l)$, where $d_l$ denotes the layer width. Suppose the predictive covariance is approximately isotropic, $\tilde{\boldsymbol{\Sigma}}_l \approx \tilde{\sigma}_l^2 \mathbf{I}_{d_l}$, a reasonable assumption under mean-field approximations or when the layer has not collapsed. Then the centred samples $\mathbf{f}_l - \tilde{\mathbf{U}}_l \sim \mathcal{N}(\mathbf{0}, \tilde{\sigma}_l^2 \mathbf{I}_{d_l})$, and by Theorem C.1 applied with the rescaling $\mathbf{x} = (\mathbf{f}_l - \tilde{\mathbf{U}}_l)/\tilde{\sigma}_l$, these samples concentrate on a hyperspherical shell of radius $\tilde{\sigma}_l\sqrt{d_l}$ centred at $\tilde{\mathbf{U}}_l$, with deviations from this radius decaying exponentially in $\beta^2$.

Now let $\mathbf{Z}_l = \{\mathbf{z}_l^{(m)}\}_{m=1}^M \subset \mathbb{R}^{d_l}$ be point-mass inducing locations. The kernel between a hidden-layer sample $\mathbf{f}_l^*$ and an inducing point $\mathbf{z}_l^{(m)}$ depends on the Euclidean distance $\|\mathbf{f}_l^* - \mathbf{z}_l^{(m)}\|$. For any two points drawn uniformly from a sphere of radius $r = \tilde{\sigma}_l\sqrt{d_l}$, their expected squared distance is $2r^2 = 2\tilde{\sigma}_l^2 d_l$, with fluctuations of order $O(\sqrt{d_l})$. Consequently, as $d_l$ grows, all pairs of points on the shell become approximately equidistant: the ratio of standard deviation to mean distance tends to zero. Under a stationary kernel $k(\mathbf{f}_l^*, \mathbf{z}_l^{(m)}) = k_0(\|\mathbf{f}_l^* - \mathbf{z}_l^{(m)}\|^2)$, this implies that all kernel evaluations $k(\mathbf{f}_l^*, \mathbf{z}_l^{(m)})$ converge to the same constant for every test point $\mathbf{f}_l^*$ on the shell, rendering the inducing point locations informationally equivalent and the sparse GP approximation degenerate.

**Remark C.1.** *This pathology is specific to isotropic kernels evaluated with Euclidean distances. The $M$ inducing locations must effectively cover the surface of a $d_l$-sphere, whose area grows as $d_l^{d_l/2}$ exponentially in the layer width. No fixed number $M$ of inducing points can provide adequate coverage as $d_l$ scales.*

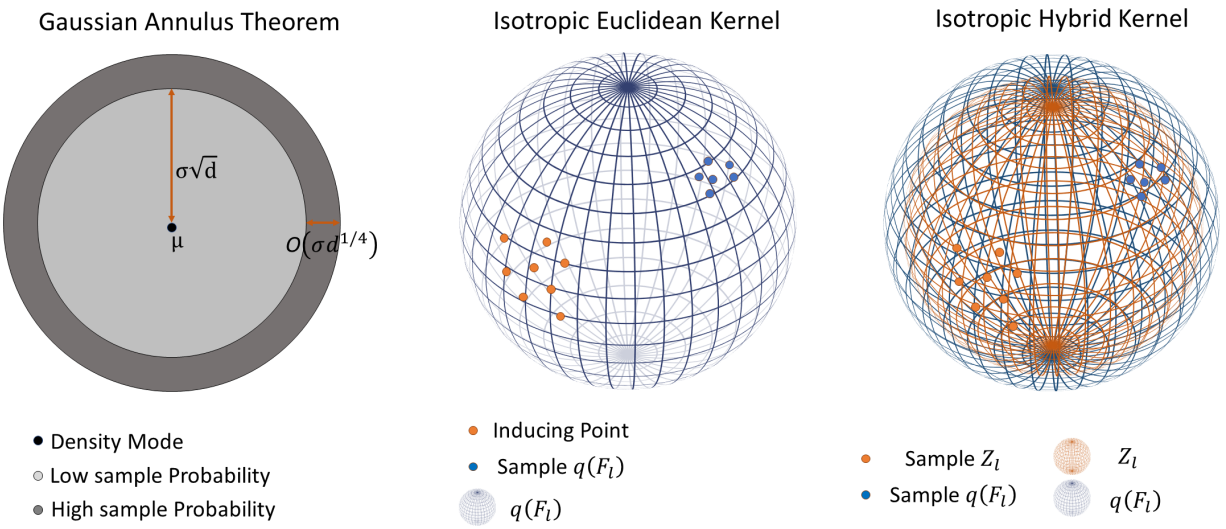



Figure 5: **Concentration of measure and inducing-point coverage in (D)DGPs.** *Left:* Theorem C.1 illustrated: samples from $\mathcal{N}(\mathbf{0}, \mathbf{I}_d)$ concentrate on a thin hyperspherical shell of radius $\sqrt{d}$. *Middle:* In a standard DGP, hidden-layer samples $\mathbf{f}_l$ concentrate on a shell of radius $\sqrt{d_l}\,\tilde{\sigma}_l$; point-mass inducing locations must cover the entire surface to remain informative, an increasingly intractable task as $d_l$ grows. *Right:* In a DDGP, distributional inducing points $\mathcal{N}(\mathbf{z}_l, \mathbf{\Sigma}_{z_l})$ encode covariance structure; the $\mathcal{W}_2$ component of the hybrid kernel measures distributional discrepancy, which remains informative regardless of the shell radius.

**How distributional inducing points mitigate this.** In a DDGP, the inducing locations at layer $l$ are themselves Gaussians, $\mathbf{Z}_l = \{\mathcal{N}(\mathbf{z}_l^{(m)}, \mathbf{\Sigma}_{\mathbf{z}_l}^{(m)})\}_{m=1}^{M}$, and the hybrid kernel (4.1) compares a test point against these inducing distributions via the $\mathcal{W}_2$ metric. The predictive distribution at any hidden-layer test point $\mathbf{f}_l^*$ is itself a Gaussian $q(\mathbf{f}_l^*) = \mathcal{N}(\tilde{\boldsymbol{\mu}}_l, \tilde{\mathbf{\Sigma}}_l)$, and the Bures–Wasserstein distance between $q(\mathbf{f}_l^*)$ and an inducing distribution $\mathcal{N}(\mathbf{z}_l^{(m)}, \mathbf{\Sigma}_{\mathbf{z}_l}^{(m)})$ decomposes as

$$\mathcal{W}_2^2\left(q(\mathbf{f}_l^*), \mathcal{N}\left(\mathbf{z}_l^{(m)}, \mathbf{\Sigma}_{\mathbf{z}_l}^{(m)}\right)\right) = \underbrace{\|\tilde{\boldsymbol{\mu}}_l - \mathbf{z}_l^{(m)}\|^2}_{\text{mean term}} + \underbrace{\mathrm{Tr}\left[\tilde{\mathbf{\Sigma}}_l + \mathbf{\Sigma}_{\mathbf{z}_l}^{(m)} - 2\left(\tilde{\mathbf{\Sigma}}_l^{1/2}\mathbf{\Sigma}_{\mathbf{z}_l}^{(m)}\tilde{\mathbf{\Sigma}}_l^{1/2}\right)^{1/2}\right]}_{\text{covariance term}}. \tag{46}$$

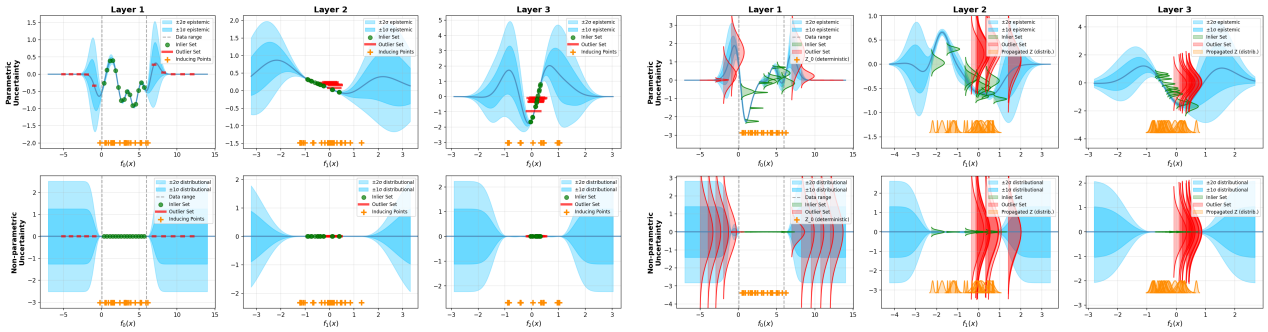
Both terms escape the concentration pathology, but for distinct reasons, and together they provide complementary discriminative signal.

The mean term $\|\tilde{\boldsymbol{\mu}}_l - \mathbf{z}_l^{(m)}\|^2$ compares two *deterministic* quantities: the predictive mean $\tilde{\boldsymbol{\mu}}_l$, which is a fixed function of the input $\mathbf{x}$, and the inducing location $\mathbf{z}_l^{(m)}$. Crucially, no sampling from a high-dimensional Gaussian is involved here: the concentration-of-measure argument that afflicts the SE kernel applies to *samples* $\mathbf{f}_l^* \sim q(\mathbf{f}_l^*)$, not to predictive means. The mean term therefore provides reliable distance information—for instance, an OOD input whose predictive mean lies far from all inducing locations will be correctly penalised—without inheriting the equidistance pathology.

The covariance term provides complementary, and in many cases decisive, additional signal. Two inputs can share similar predictive means yet differ substantially in their predictive covariance: in-distribution inputs propagate through well-supported kernel paths and acquire covariances closely aligned with the inducing distributions, whereas OOD inputs accumulate additional posterior variance along under-supported directions. The covariance term in (46) captures this discrepancy via the trace of the matrix square root, measuring the geometric difference in the *shape* of the predictive distribution rather than just its location. Taken together, the $\mathcal{W}_2$ metric operates entirely on the *distributional* output of each layer rather than on raw samples, sidestepping the concentration-of-measure problem at both the mean and covariance level, and providing robust discriminability that does not require tiling the hyperspherical surface with inducing points.

## D NON-PARAMETRIC VARIANCE PROPAGATION IN (D)DGPS

To understand the mechanism underlying this pathology, we examine a 3-layer DGP and DDGP trained on a toy regression task (Figure 6). Consider the outlier situated at $x \approx -7.5$, well outside the training support.




(a) DGP (b) DDGP

Figure 6: **Layer-wise uncertainty propagation for a 3-layer DGP (a) and DDGP (b) on a toy regression task, with PCA mean functions and 2-dimensional hidden layers.** Each panel shows three columns corresponding to Layers 1–3, with parametric uncertainty $\sigma_{\text{para}}$ (top row) and non-parametric (distributional) uncertainty $\sigma_{\text{np}}$ (bottom row). The training data occupies the range indicated by the dashed vertical line; the outlier input lies at $x \approx -7.5$. In each layer, the red distributions depict the predictive posterior moments of outlier points, and green points depict those of inlier points.

**DGP failure mode.** In Layer 1 of the DGP (Figure 6a), the predictive posteriors of outlier points (red) and inlier points (green) are clearly separated under the non-parametric uncertainty: the outlier representations are far from the inducing locations and the SE kernel correctly assigns them elevated uncertainty. However, this separation does not survive propagation to Layer 2. The mean function maps the outlier's Layer 1 representation into a region of hidden space that happens to coincide with inducing point locations of the next GP, causing the SE kernel to treat it as in-distribution. The predictive posteriors of outlier and inlier points consequently collapse onto the same uncertainty level, and the OOD signal is lost. By Layer 3, the red and green predictive posteriors are indistinguishable, confirming that depth alone provides no remedy. The fundamental issue is that the SE kernel operates only on *point locations* in hidden space: it is entirely blind to whether a representation arrived there via an in-distribution or out-of-distribution input path.

**DDGP resolution.** The DDGP avoids this collapse by replacing point inducing locations with *distributional* inducing points whose full posterior Gaussians are propagated forward at every layer boundary. In Layer 1 (Figure 6b), the $\mathcal{W}_2$ kernel already separates outlier from inlier predictive posteriors by comparing the posterior Gaussian of each input against those of the inducing points. This separation is geometrically transparent in the figure: the distributional inducing points (orange) carry tightly concentrated posterior variances that closely match the predictive moments of inlier inputs, reflecting the local geometry of the training manifold. Outlier inputs, by contrast, produce posterior Gaussians with markedly larger variance (visible as the wide red distributions in each layer of panel (b)), which are metrically distant from the compact orange inducing distributions under $\mathcal{W}_2$. The kernel therefore assigns them low similarity to any inducing point, and non-parametric uncertainty remains high. Crucially, the distributional inducing points are then propagated forward ("Propagated $Z$ (distrib.)" in Layers 2 and 3), carrying the accumulated distributional information into subsequent layers. As a result, even if the *mean* of an outlier's hidden representation coincides with that of an inlier, its *posterior variance* will differ markedly from the compact orange inducing distributions, and the $\mathcal{W}_2$ kernel penalises this discrepancy at every layer. The separation between wide red outlier posteriors and tightly concentrated orange inducing distributions therefore persists visibly across all three layers, yielding a coherent and depth-stable OOD signal. This is precisely the guarantee of Proposition 4.1: distributional inducing points embed proximity in distribution space directly into the kernel, preventing the unsafe collapse of non-parametric uncertainty under arbitrary layer-wise mappings.

## E PROPAGATION OF UNCERTAINTY IN (D)DGPS

The proofs in this section rely on the Gaussian approximation framework for GPs with uncertain inputs developed in Appendix A.1; the reader is encouraged to review that material before proceeding. Throughout we adopt the shorthand $\tilde{U}_l(\cdot)$ and $\tilde{\Sigma}_l(\cdot)$ for the posterior predictive mean and variance of the $l$-th layer SVGP, and denote by $\sigma_l^2$ the prior signal variance of that layer.

### E.1 PROOF OF PROPOSITION 3.1

*Proof.* The approximate posterior of the $l$-th DGP layer is

$$q(F_l \mid F_{l-1}) = \int p(F_l \mid U_l, F_{l-1})\, q(U_l)\, \mathrm{d}U_l = \mathcal{N}\Big(F_l \mid \tilde{U}_l(F_{l-1}),\, \tilde{\Sigma}_l(F_{l-1})\Big)\,. \tag{47}$$

For a two-layer DGP the marginal distribution of $F_2$ at input $x$ is

$$q(F_2)(x) = \int p(F_2 \mid F_1)\, q(F_1(x))\, \mathrm{d}F_1, \qquad q(F_1(x)) = \mathcal{N}\Big(F_1 \mid \tilde{U}_1(x),\, \tilde{\Sigma}_1(x)\Big)\,. \tag{48}$$

This integral has the same form as the uncertain-input GP prediction studied in Appendix A.1: $F_1$ plays the role of the stochastic test input, with mean $\tilde{U}_1(x)$ and variance $\tilde{\Sigma}_1(x)$. Applying the Taylor approximation of Section A.1.2 (equations (33)–(34)) gives the following approximate moments of $q(F_2)(x)$. For the mean we use a first-order Taylor expansion for the approximation,

$$m(F_2(x)) = \tilde{U}_2\Big(\tilde{U}_1(x)\Big)\,, \tag{49}$$

and for the variance we use the second-order expression:

$$v(F_2(x)) = \tilde{\Sigma}_2\Big(\tilde{U}_1(x)\Big) + \tilde{\Sigma}_1(x)\left[\frac{1}{2}\frac{\partial^2 \tilde{\Sigma}_2(F_1)}{\partial F_1^2}\Big|_{F_1=\tilde{U}_1(x)} + \left(\frac{\partial \tilde{U}_2(F_1)}{\partial F_1}\right)^2\Big|_{F_1=\tilde{U}_1(x)}\right]. \tag{50}$$

Following **?**, who showed empirically that $\partial^2\tilde{\Sigma}_2/\partial F_1^2 \to 0$ as the number of inducing points $M \to \infty$ (the second derivative follows a sinusoidal path whose amplitude decays to zero with $M$), we drop that term, yielding

$$v(F_2(x)) \approx \tilde{\Sigma}_2\Big(\tilde{U}_1(x)\Big) + \tilde{\Sigma}_1(x)\left(\frac{\partial \tilde{U}_2(F_1)}{\partial F_1}\right)^2\Big|_{F_1=\tilde{U}_1(x)}. \tag{51}$$

We now evaluate (51) at an out-of-distribution point $x_{\mathrm{out}}$ and a boundary in-distribution point $x_{\mathrm{in}}$. Under the proposition's assumptions:

- $\tilde{U}_1(x_{\mathrm{out}}) = 0$ and $\tilde{\Sigma}_1(x_{\mathrm{out}}) = \sigma_1^2$ (the posterior has not contracted from the prior);
- $\tilde{U}_1(x_{\mathrm{in}}) = M_{\mathrm{in}}$ and $\tilde{\Sigma}_1(x_{\mathrm{in}}) = V_{\mathrm{in}}$ with $V_{\mathrm{in}} \le \sigma_1^2$.

Substituting into (51):

$$v(F_2(x_{\mathrm{out}})) = \tilde{\Sigma}_2(0) + \sigma_1^2\left(\frac{\partial \tilde{U}_2(F_1)}{\partial F_1}\right)^2\Big|_{F_1=0}, \tag{52}$$

$$v(F_2(x_{\mathrm{in}})) = \tilde{\Sigma}_2(M_{\mathrm{in}}) + V_{\mathrm{in}}\left(\frac{\partial \tilde{U}_2(F_1)}{\partial F_1}\right)^2\Big|_{F_1=M_{\mathrm{in}}}. \tag{53}$$

We require $v(F_2(x_{\mathrm{in}})) \le v(F_2(x_{\mathrm{out}}))$. Under the assumption that the second-layer inducing point locations $Z_1^{(m)}$ are placed equidistantly on $[-3\sigma_1, 3\sigma_1]$ and their posterior variances are equal, the posterior variance $\tilde{\Sigma}_2(\cdot)$ is approximately constant near the origin, giving $\tilde{\Sigma}_2(M_{\mathrm{in}}) \approx \tilde{\Sigma}_2(0)$. Cancelling this common term, the desired inequality holds if and only if

$$\frac{\left(\dfrac{\partial \tilde{U}_2(F_1)}{\partial F_1}\right)^2\Big|_{F_1=M_{\mathrm{in}}}}{\left(\dfrac{\partial \tilde{U}_2(F_1)}{\partial F_1}\right)^2\Big|_{F_1=0}} \;\le\; \frac{\sigma_1^2}{V_{\mathrm{in}}}, \tag{54}$$

which is the statement of Proposition 3.1. □

### E.2 PROOF OF PROPOSITION 3.2

*Proof.* We specialise to a two-layer model in which the first layer has not observed any data (prior regime), so that the output of layer 1 is distributed as $H_1 \mid H_0 \sim \mathcal{N}(0, \sigma_1^2)$, and compute the propagated variance at layer 2 analytically for an SE kernel with two inducing points.

**Setup.** Consider a zero-mean second-layer GP with SE kernel

$$K(x, x') = \sigma_2^2 \exp\left(-\frac{(x - x')^2}{2l^2}\right)$$

and two inducing points at $z_1 = +c$ and $z_2 = -c$. We wish to compute the predictive variance of $H_2$ when its input is $H_1 \sim \mathcal{N}(0, \sigma_1^2)$. Since the prior mean is zero, $\tilde{U}_2(H_1) = 0$ identically, and the Gaussian-approximation variance reduces to the average predictive variance:

$$v(0, \sigma_1^2) = \mathbb{E}_{H_1}\left[\tilde{\Sigma}_2(H_1)\right] = \mathbb{E}_{H_1}[k(H_1, H_1)] - \mathrm{Tr}\left[K_{uu}^{-1}\, \Psi_2(\sigma_1^2)\right], \tag{55}$$

where $(\Psi_2)_{ij} = \mathbb{E}_{H_1}[K(H_1, z_i)K(H_1, z_j)]$ and $\mathbb{E}[K(H_1, H_1)] = \sigma_2^2$.

**Computing $K_{uu}$ and $\Psi_2$.** The inducing-point covariance matrix and its inverse are

$$K_{uu} = \sigma_2^2 \begin{pmatrix} 1 & e^{-2c^2/l^2} \\ e^{-2c^2/l^2} & 1 \end{pmatrix}, \qquad K_{uu}^{-1} = \frac{1}{\sigma_2^2\left(1 - e^{-4c^2/l^2}\right)} \begin{pmatrix} 1 & -e^{-2c^2/l^2} \\ -e^{-2c^2/l^2} & 1 \end{pmatrix}. \tag{56}$$

The entries of $\Psi_2$ are computed by integrating products of SE kernel evaluations against $\mathcal{N}(H_1 \mid 0, \sigma_1^2)$. Since $K(x, z_i)^2 = \sigma_2^4 \exp(-(x - z_i)^2/l^2)$ and $K(x, c)K(x, -c) = \sigma_2^4 \exp(-(x^2 + c^2)/l^2)$, we apply the standard Gaussian product identity

$$\int \exp\left(-A(x - b)^2\right) \mathcal{N}(x \mid 0, \sigma^2)\, \mathrm{d}x = \frac{1}{\sqrt{1 + 2A\sigma^2}} \exp\left(-\frac{Ab^2}{1 + 2A\sigma^2}\right)$$

with $(A, b) = (1/l^2,\, c)$ for the diagonal entries and $(A, b) = (1/l^2,\, 0)$ for the off-diagonal entries, to obtain:

$$(\Psi_2)_{11} = (\Psi_2)_{22} = \frac{\sigma_2^4}{\sqrt{1 + 2\sigma_1^2/l^2}} \exp\left(-\frac{c^2}{l^2 + 2\sigma_1^2}\right), \tag{57}$$

$$(\Psi_2)_{12} = (\Psi_2)_{21} = \frac{\sigma_2^4}{\sqrt{1 + 2\sigma_1^2/l^2}} \exp\left(-\frac{c^2}{l^2}\right). \tag{58}$$

**Closed-form variance.** Taking the trace of $K_{uu}^{-1}\Psi_2$ and substituting into (55) (the $\sigma_2^4$ in $\Psi_2$ and $\sigma_2^{-2}$ in $K_{uu}^{-1}$ combine to give an overall $\sigma_2^2$ factor):

$$v(0, \sigma_1^2) = \sigma_2^2 \left[1 - \frac{2\left(e^{-c^2/(l^2 + 2\sigma_1^2)} - e^{-3c^2/l^2}\right)}{\left(1 - e^{-4c^2/l^2}\right)\sqrt{1 + 2\sigma_1^2/l^2}}\right]. \tag{59}$$

**Boundary behaviour.** At $\sigma_1^2 = 0$:

$$v(0, 0) = \sigma_2^2 \left[1 - \frac{2e^{-c^2/l^2}}{1 + e^{-2c^2/l^2}}\right] < \sigma_2^2. \tag{60}$$

As $\sigma_1^2 \to \infty$, both exponential terms in the numerator of (59) are dominated by the growing denominator $\sqrt{1 + 2\sigma_1^2/l^2}$, so $v(0, \sigma_1^2) \to \sigma_2^2$. Intuitively, when the input distribution is extremely diffuse, the Nyström correction collapses to zero and the variance recovers to the prior level.

**Derivative and non-monotonicity.** Let $s := \sigma_1^2$, $K_c := \sigma_2^2(1 - e^{-4c^2/l^2}) > 0$, $E_0 := e^{-3c^2/l^2}$, and define

$$f(s) = e^{-c^2/(l^2+2s)} - E_0, \qquad h(s) = \sqrt{1 + 2s/l^2},$$

so that $v(s) = \sigma_2^2 - 2f(s)/(K_c\, h(s))$. Differentiating by the quotient rule, with $f'(s) = \dfrac{2c^2}{(l^2+2s)^2} e^{-c^2/(l^2+2s)}$ and $h'(s)/h(s) = 1/(l^2+2s)$, gives

$$\frac{dv}{ds} = \frac{2}{K_c\, h(s)} \left[ \frac{e^{-c^2/(l^2+2s)} \big[(l^2+2s) - 2c^2\big]}{(l^2+2s)^2} - \frac{e^{-3c^2/l^2}}{l^2+2s} \right]. \tag{61}$$

Since $K_c > 0$ and $h(s) > 0$, the sign of $dv/ds$ is determined solely by the bracket.

**Step 1: $dv/ds < 0$ at $s = 0$ for all $c, l > 0$.** Evaluating the bracket in (61) at $s = 0$ (so $l^2 + 2s = l^2$):

$$\left[ \frac{e^{-c^2/(l^2+2s)} \big[(l^2+2s) - 2c^2\big]}{(l^2+2s)^2} - \frac{e^{-3c^2/l^2}}{l^2+2s} \right] \Big|_{s=0} = \frac{e^{-c^2/l^2}(l^2 - 2c^2)}{l^4} - \frac{e^{-3c^2/l^2}}{l^2}. \tag{62}$$

We factor out $e^{-c^2/l^2}/l^2 > 0$, thus:

$$\begin{aligned} \frac{e^{-c^2/l^2}(l^2 - 2c^2)}{l^4} - \frac{e^{-3c^2/l^2}}{l^2} &= \frac{e^{-c^2/l^2}}{l^2} \left[ \frac{l^2 - 2c^2}{l^2} - e^{-2c^2/l^2} \right] \\ &= \frac{e^{-c^2/l^2}}{l^2} \left[ (1 - 2t) - e^{-2t} \right], \qquad t := \frac{c^2}{l^2} > 0. \end{aligned} \tag{63}$$

Since the prefactor $e^{-c^2/l^2}/l^2 > 0$, the sign is determined by

$$\varphi(t) := 1 - 2t - e^{-2t}. \tag{64}$$

We show $\varphi(t) < 0$ for all $t > 0$. At $t = 0$, we have $\varphi(0) = 1 - 0 - 1 = 0$. Hence, if we differentiate we obtain

$$\varphi'(t) = -2 + 2e^{-2t} = 2\big(e^{-2t} - 1\big). \tag{65}$$

Since $e^{-2t} < 1$ for all $t > 0$, we have $\varphi'(t) < 0$ for all $t > 0$. Therefore $\varphi$ is strictly decreasing from $\varphi(0) = 0$, which gives

$$\varphi(t) < 0 \quad \forall\, t > 0, \tag{66}$$

and hence $dv/ds|_{s=0} < 0$ for all $c, l > 0$.

**Step 2: $dv/ds > 0$ for all sufficiently large $s$.** Let $u := l^2 + 2s$, so $u \to \infty$ as $s \to \infty$. Writing the bracket with the common denominator $u^2$:

$$\frac{1}{u^2} \left[ e^{-c^2/u}(u - 2c^2) - u\, e^{-3c^2/l^2} \right] = \frac{u}{u^2} \left[ e^{-c^2/u} \left( 1 - \frac{2c^2}{u} \right) - e^{-3c^2/l^2} \right]. \tag{67}$$

As $u \to \infty$, we expand $e^{-c^2/u} = 1 - c^2/u + O(u^{-2})$, thus:

$$e^{-c^2/u} \left( 1 - \frac{2c^2}{u} \right) = \left( 1 - \frac{c^2}{u} + O(u^{-2}) \right) \left( 1 - \frac{2c^2}{u} \right) = 1 - \frac{3c^2}{u} + O(u^{-2}). \tag{68}$$

Therefore the bracket satisfies:

$$\frac{1}{u} \left[ 1 - \frac{3c^2}{u} - e^{-3c^2/l^2} + O(u^{-2}) \right] \xrightarrow{u \to \infty} \frac{1 - e^{-3c^2/l^2}}{u} > 0, \tag{69}$$

since $e^{-3c^2/l^2} < 1$ for all $c, l > 0$. Hence $dv/ds > 0$ for all sufficiently large $s$.

**Step 3: Unique threshold $\sigma_1^{2*}$.** Since $dv/ds < 0$ at $s = 0$ (Step 1), $dv/ds > 0$ for large $s$ (Step 2), and $dv/ds$ is continuous, it must change sign at least once. We show the sign change is unique by reducing the zero-crossing condition to a one-dimensional equation with a unique solution.

Setting the bracket in (61) to zero with $u := l^2 + 2s$:

$$\frac{e^{-c^2/u}(u - 2c^2)}{u^2} = \frac{e^{-3c^2/l^2}}{u}. \tag{70}$$

Multiplying both sides by $u > 0$:

$$e^{-c^2/u} \cdot \frac{u - 2c^2}{u} = e^{-3c^2/l^2}. \tag{71}$$

Now substitute $w := c^2/u$, so $u = c^2/w$ and:

$$\frac{u - 2c^2}{u} = 1 - \frac{2c^2}{u} = 1 - 2w. \tag{72}$$

Equation (71) becomes:

$$(1 - 2w)\, e^{-w} = e^{-3c^2/l^2}. \tag{73}$$

Define $\psi(w) := (1 - 2w)e^{-w}$. Its derivative is:

$$\psi'(w) = -2e^{-w} + (1 - 2w)(-e^{-w}) = -(3 - 2w)e^{-w}. \tag{74}$$

For $w \in (0, \frac{1}{2})$, we have $3 - 2w \in (2, 3) > 0$, so $\psi'(w) < 0$: the function $\psi$ is *strictly decreasing* on $(0, \frac{1}{2})$. At the endpoints:

$$\psi(0) = 1, \qquad \psi\left(\frac{1}{2}\right) = 0. \tag{75}$$

The right-hand side $e^{-3c^2/l^2} \in (0, 1)$ for all $c, l > 0$. By the intermediate value theorem and strict monotonicity of $\psi$, equation (73) has a *unique* solution $w^* \in (0, \frac{1}{2})$.

The corresponding threshold in the original variable is:

$$u^* = \frac{c^2}{w^*}, \qquad \sigma_1^{2*} = \frac{u^* - l^2}{2} = \frac{1}{2}\left(\frac{c^2}{w^*} - l^2\right). \tag{76}$$

To confirm $\sigma_1^{2*} > 0$, we need $w^* < c^2/l^2 =: t$. Evaluating $\psi$ at $w = t$:

$$\psi(t) - e^{-3t} = (1 - 2t)e^{-t} - e^{-3t} = e^{-t}\left[(1 - 2t) - e^{-2t}\right] = e^{-t}\,\varphi(t) < 0, \tag{77}$$

where the last inequality uses Step 1. Hence $\psi(t) < e^{-3t}$, and since $\psi$ is strictly decreasing, the solution $w^*$ satisfies $w^* < t$, confirming $\sigma_1^{2*} > 0$.

Finally, since $dv/ds$ is continuous and $\psi$ has a unique zero at $w^*$ (equivalently, a unique $s^*$), the sign of $dv/ds$ changes from negative to positive exactly once. Therefore $v(s)$ is strictly decreasing on $[0, \sigma_1^{2*})$ and strictly increasing on $(\sigma_1^{2*}, \infty)$.

**Step 4: Lower bound on $\sigma_1^{2*}$.** Recall the derivative (61) has the form

$$\frac{dv}{ds} \propto \underbrace{\frac{e^{-c^2/u}(u - 2c^2)}{u^2}}_{\text{Term A}} - \underbrace{\frac{e^{-3c^2/l^2}}{u}}_{\text{Term B}}, \qquad u := l^2 + 2s. \tag{78}$$

Term B is always strictly positive. For $dv/ds > 0$, Term A must not only be positive but must *exceed* Term B. A necessary first requirement is therefore that Term A is positive at all:

$$\text{Term A} > 0 \iff u - 2c^2 > 0 \iff l^2 + 2\sigma_1^2 > 2c^2. \tag{79}$$

The geometric meaning is transparent: $u = l^2 + 2\sigma_1^2$ is the effective input scale seen by the second-layer kernel, and $2c^2$ is twice the squared separation between the inducing points. The condition $u > 2c^2$ says the input spread must be *wide enough*

*to reach beyond the inducing point cluster*. Until that happens, Term A is negative and both terms pull $dv/ds$ downward, so the variance cannot yet increase.

Solving for $\sigma_1^2$:

$$\sigma_1^2 > \frac{2c^2 - l^2}{2}, \tag{80}$$

which is only a binding constraint when $l^2 < 2c^2$ (i.e. when the lengthscale is shorter than the inter-inducing-point distance $c\sqrt{2}$). If $l^2 \geq 2c^2$, then $u = l^2 + 2\sigma_1^2 \geq l^2 \geq 2c^2$ already at $\sigma_1^2 = 0$, so Term A is non-negative from the outset and the lower bound is trivially zero. In either case:

$$\sigma_1^{2*} \;\geq\; \max\!\left(0,\; \frac{2c^2 - l^2}{2}\right). \tag{81}$$

Intuitively, a large lengthscale $l$ relative to $c$ means the kernel is already "looking across" the full inducing point spread even at small input uncertainty, so the transition to increasing variance happens sooner. Conversely, a short lengthscale relative to $c$ means large input uncertainty is needed before the distribution spreads far enough to escape the inducing cluster, delaying the onset of uncertainty propagation.

**Conclusion.** Combining Steps 1–4, $v(0, \sigma_1^2)$ is strictly decreasing on $[0, \sigma_1^{2*})$, attains its unique minimum at $\sigma_1^{2*}$, and is strictly increasing on $(\sigma_1^{2*}, \infty)$ toward the prior level $\sigma_2^2$, establishing the three-case characterisation in the proposition. □

### E.3 PROOF OF PROPOSITION 4.1

*Proof.* We revisit the desired second-layer inequality from Proposition 3.1:

$$\tilde{\Sigma}_2(M_{\text{in}}) + V_{\text{in}} \left( \frac{\partial \tilde{U}_2(F_1)}{\partial F_1} \right)^2 \Bigg|_{F_1 = M_{\text{in}}} \;\leq\; \tilde{\Sigma}_2(0) + \sigma_1^2 \left( \frac{\partial \tilde{U}_2(F_1)}{\partial F_1} \right)^2 \Bigg|_{F_1 = 0}. \tag{82}$$

In the standard DGP the posterior variance $\tilde{\Sigma}_2(\cdot)$ uses a scalar kernel, so $\tilde{\Sigma}_2(M_{\text{in}}) \approx \tilde{\Sigma}_2(0)$ under the equidistant-inducing-point assumption, and the gradient ratio condition (54) is the binding constraint.

In the DDGP, the second-layer kernel is the hybrid kernel $K^H$ that incorporates distributional information via the Wasserstein-2 distance. We therefore examine how the non-parametric variance terms $\tilde{\Sigma}_2(M_{\text{in}})$ and $\tilde{\Sigma}_2(0)$ differ under $K^H$.

To expose the mechanism clearly, consider a single inducing point $\mathcal{Z}_m = \mathcal{N}(Z_m, Z_{\text{var}})$ located midway between 0 and $M_{\text{in}}$, with $Z_{\text{var}} \ll \sigma_1^2$. The non-parametric variance at $F_1$ is $\tilde{\Sigma}_2(F_1) = \sigma_2^2 - \left[K^H(F_1, Z_m)\right]^2 / K^H(Z_m, Z_m)$. The condition $\tilde{\Sigma}_2(M_{\text{in}}) \leq \tilde{\Sigma}_2(0)$ then becomes

$$K^H\big(\mathcal{N}(0, \sigma_1^2),\, \mathcal{N}(Z_m, Z_{\text{var}})\big) \;\leq\; K^H(\mathcal{N}(M_{\text{in}}, V_{\text{in}}),\, \mathcal{N}(Z_m, Z_{\text{var}}))\,. \tag{83}$$

Using the DDGP hybrid kernel $K^H(\mu_i, \mu_j) = \sigma_2^2 \exp(-W_2^2(\mu_i, \mu_j)/l^2) \exp(-(\bar{x}_i - \bar{x}_j)^2/l^2)$, where $\bar{x}$ denotes the mean of a Gaussian argument, inequality (83) after taking logarithms is equivalent to:

$$W_2^2\big(\mathcal{N}(0, \sigma_1^2),\, \mathcal{N}(Z_m, Z_{\text{var}})\big) + (0 - Z_m)^2 \;\geq\; W_2^2(\mathcal{N}(M_{\text{in}}, V_{\text{in}}),\, \mathcal{N}(Z_m, Z_{\text{var}})) + (M_{\text{in}} - Z_m)^2. \tag{84}$$

We now expand the $W_2^2$ terms using the closed-form formula for Gaussians:

$$W_2^2(\mathcal{N}(\mu_1, \Sigma_1),\, \mathcal{N}(\mu_2, \Sigma_2)) = (\mu_1 - \mu_2)^2 + \Sigma_1 + \Sigma_2 - 2\sqrt{\Sigma_1 \Sigma_2}. \tag{85}$$

For the left-hand side of (84):

$$W_2^2(\mathcal{N}(0, \sigma_1^2),\, \mathcal{N}(Z_m, Z_{\text{var}})) + Z_m^2 = 2Z_m^2 + \sigma_1^2 + Z_{\text{var}} - 2\sqrt{\sigma_1^2 Z_{\text{var}}}. \tag{86}$$

For the right-hand side of (84), recalling $Z_m = M_{\text{in}}/2$:

$$\begin{aligned} &W_2^2(\mathcal{N}(M_{\text{in}}, V_{\text{in}}),\, \mathcal{N}(Z_m, Z_{\text{var}})) + (M_{\text{in}} - Z_m)^2 \\ &\quad = (M_{\text{in}} - Z_m)^2 + V_{\text{in}} + Z_{\text{var}} - 2\sqrt{V_{\text{in}} Z_{\text{var}}} + (M_{\text{in}} - Z_m)^2 \\ &\quad = 2Z_m^2 + V_{\text{in}} + Z_{\text{var}} - 2\sqrt{V_{\text{in}} Z_{\text{var}}}. \end{aligned} \tag{87}$$

Subtracting the right from the left, inequality (84) holds if and only if

$$\sigma_1^2 - 2\sqrt{\sigma_1^2 Z_{\text{var}}} \;\geq\; V_{\text{in}} - 2\sqrt{V_{\text{in}} Z_{\text{var}}}, \tag{88}$$

or equivalently $f(\sigma_1^2) \geq f(V_{\text{in}})$ where $f(t) = t - 2\sqrt{t\, Z_{\text{var}}}$. Since $f'(t) = 1 - Z_{\text{var}}^{1/2}/t^{1/2}$ and the condition $Z_{\text{var}} \ll \sigma_1^2$ implies $f'(\sigma_1^2) > 0$ and $f'(V_{\text{in}}) > 0$ whenever $V_{\text{in}} > Z_{\text{var}}$, $f$ is increasing on $[Z_{\text{var}}, \infty)$. Hence $f(\sigma_1^2) \geq f(V_{\text{in}})$ holds whenever $\sigma_1^2 \geq V_{\text{in}}$, which is guaranteed by assumption. More precisely, the condition $\sigma_1^2 \gg Z_{\text{var}}$ and $V_{\text{in}} \approx Z_{\text{var}}$ ensures the strictest inequality, since then $f(\sigma_1^2) \gg f(V_{\text{in}}) \approx 0$.

Thus inequality (82) holds: $\tilde{\Sigma}_2(M_{\text{in}}) < \tilde{\Sigma}_2(0)$ for DDGP, providing a strictly tighter second-layer posterior at in-distribution inputs than at OOD inputs, in contrast to the standard DGP where both terms cancel. □

# F *FEATURE-COLLAPSE* IN (D)DGPS

Dunlop et al. [2018] provide a detailed analysis of the function-space properties of DGPs, which DDGPs match in prior space (see Section 4.1). Here we derive the conditions under which, in the *posterior*, the hidden-layer representations of two in-distribution points collapse to the same value, a phenomenon we call *feature-collapse*. Proposition F.1 establishes these conditions for standard DGPs, and Proposition F.2 shows that DDGPs are strictly more resistant to feature-collapse due to the additional distributional discrepancy encoded in the hybrid kernel.

**Setup and notation.** Let $\mathcal{D} \subset \mathbb{R}^{d_0}$ be the input domain and $\mathcal{D}_l \subset \mathbb{R}^{d_l}$ the $l$-th hidden-layer feature space. Each layer defines a function $f_l : \mathcal{D}_{l-1} \to \mathcal{D}_l$, and the $n$-layer DGP is the composition $f_n \circ \cdots \circ f_1$. We denote by $Z_l = \{z_l^{(m)}\}_{m=1}^M$ the inducing point locations at layer $l$, by $U_l$ the corresponding inducing outputs, and write $[K_{uu}]_{mm'} = K_l(z_l^{(m)}, z_l^{(m')})$. Throughout we assume the inducing sets are sufficiently rich that $K_{ff} \approx K_{fu}K_{uu}^{-1}K_{uf}$ (tight Nyström approximation), so the posterior conditional variance is negligible at in-distribution points and $f_l(x) \mid U_l, f_{l-1}$ is effectively deterministic with mean $K_{xu}K_{uu}^{-1}U_l$. We define the per-layer contraction coefficient

$$C_l \;\coloneqq\; \frac{\sigma_l^4}{4l_l^4} \sum_{d=1}^{d_l} \|K_{uu}^{-1} U_l^d\|_2^2, \tag{89}$$

and write $\delta_l \coloneqq \mathbb{E}[\|f_l(x) - f_l(x^*)\|_2^2 |\ \{U_j\}_{j=1}^l]$.

## F.1 FEATURE-COLLAPSE IN DGP

**Proposition F.1 (Feature-collapse condition, DGP).** *Assume $f_0$ is bounded on bounded sets almost surely. Let $x, x^* \in \mathcal{D}_{\text{in}}$ be two in-distribution inputs, and suppose the Nyström approximation $Q_{ff} \approx K_{ff}$ holds at each layer. If at each layer $l \in \{1, \ldots, n\}$ the per-layer contraction coefficient satisfies $C_l \cdot \delta_{l-1} < 1$, then*

$$P\Big(\|f_n(x) - f_n(x^*)\|_2 \to 0 \;\Big|\; \{U_l\}_{l=1}^n\Big) = 1. \tag{90}$$

*Proof.* We begin by noticing that:

**Single-layer contraction bound.** Under the Nyström approximation, the conditional expectation at layer $n$ is:

$$\begin{aligned}\delta_n &= \mathbb{E}\Big[\|f_n(x) - f_n(x^*)\|_2^2 \;\Big|\; f_{n-1}, U_n\Big] \\ &= \sum_{d=1}^{d_n} |(K_{xu} - K_{x^*u})K_{uu}^{-1}U_n^d|^2,\end{aligned} \tag{91}$$

where we write $K_{xu} = [K_n(f_{n-1}(x), z_n^{(m)})]_m$ and similarly for $K_{x^*u}$. Applying the Cauchy–Schwarz inequality to each summand:

$$\delta_n \leq \sum_{d=1}^{d_n} \|K_{xu} - K_{x^*u}\|_2^2 \, \|K_{uu}^{-1}U_n^d\|_2^2. \tag{92}$$

Now let $h = f_{n-1}(x) - f_{n-1}(x^*)$ be the feature difference at layer $n-1$, and assume $f_{n-1}(x^*)$ is close to the nearest inducing point $z_n^{(m^*)}$, so $\|f_{n-1}(x^*) - z_n^{(m^*)}\|_2^2 \approx 0$. Under the SE kernel $K_n(a,b) = \sigma_n^2 \exp(-\|a-b\|^2/(2l_n^2))$:

$$\|K_{xu} - K_{x^*u}\|_2^2 \le \sigma_n^4 \left| \exp\left(-\frac{\|h\|_2^2}{2l_n^2}\right) - 1 \right|^2 \le \frac{\sigma_n^4}{4l_n^4} \|h\|_2^4, \tag{93}$$

where the last step uses the inequality $1 - e^{-t} \le t$ for $t \ge 0$, applied with $t = \|h\|_2^2/(2l_n^2)$. Substituting (93) into (92) and using the definition (89):

$$\delta_n \le C_n \cdot \|f_{n-1}(x) - f_{n-1}(x^*)\|_2^4 = C_n \cdot \delta_{n-1}^2. \tag{94}$$

**Recursive unrolling.** The recursion (94) is *quadratic*: each layer squares the previous discrepancy and multiplies by $C_l$. Unrolling:

$$\delta_n \le C_n\, \delta_{n-1}^2 \le C_n\, C_{n-1}^2\, \delta_{n-2}^4 \le \cdots \le \prod_{l=1}^{n} C_l^{\,2^{n-l}} \cdot \delta_0^{\,2^n}, \tag{95}$$

where $\delta_0 = \|f_0(x) - f_0(x^*)\|_2^2$ is bounded by assumption.

**Almost-sure convergence via Borel–Cantelli.** Let $C = \max_l C_l$. If $C\,\delta_0 < 1$ then

$$\delta_n \le (C\,\delta_0)^{2^n}/C \;\longrightarrow\; 0 \quad \text{super-exponentially fast.} \tag{96}$$

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

$$P\Big(\|f_n(x) - f_n(x^*)\|_2 \ge \epsilon \;\Big|\; \{U_l\}_{l=1}^n\Big) \le \frac{\delta_n}{\epsilon^2} \le \frac{(C\,\delta_0)^{2^n}}{C\,\epsilon^2}. \tag{97}$$

Since $C\,\delta_0 < 1$, the series $\sum_{n=1}^{\infty} (C\,\delta_0)^{2^n}$ converges (its terms decay faster than any geometric series). By the first Borel–Cantelli lemma:

$$P\Big(\|f_n(x) - f_n(x^*)\|_2 \ge \epsilon \text{ i.o.} \;\Big|\; \{U_l\}_{l=1}^n\Big) = 0 \quad \forall\, \epsilon > 0. \tag{98}$$

Intersecting over a countable sequence $\epsilon_k = 1/k \to 0$:

$$\begin{aligned} P\Big(\|f_n(x) - f_n(x^*)\|_2 \to 0 \;\Big|\; \{U_l\}_{l=1}^n\Big) &= P\left(\bigcap_{k=1}^{\infty} \bigcup_{N=1}^{\infty} \bigcap_{n=N}^{\infty} \left\{\|f_n(x) - f_n(x^*)\|_2 \le \frac{1}{k}\right\}\right) \\ &= 1 - P\left(\bigcup_{k=1}^{\infty} \bigcap_{N=1}^{\infty} \bigcup_{n=N}^{\infty} \left\{\|f_n(x) - f_n(x^*)\|_2 \ge \frac{1}{k}\right\}\right) \\ &\ge 1 - \sum_{k=1}^{\infty} P\left(\|f_n - f_n^*\|_2 \ge \frac{1}{k} \text{ i.o.}\right) = 1. \end{aligned}$$

□

**Remark F.1** (Avoiding feature-collapse in DGPs)**.** *Feature-collapse is driven by the per-layer contraction coefficient* $C_l = \frac{\sigma_l^4}{4l_l^4} \sum_d \|K_{uu}^{-1} U_l^d\|_2^2$. *To* resist *collapse one should maximise* $C_l$*, which is achieved by increasing the signal variance* $\sigma_l^2$ *or the norm of the projected inducing outputs* $\|K_{uu}^{-1} U_l\|_2^2$ *(both control the amplitude of sampled functions), or by* decreasing *the lengthscale* $l_l^2$*. Intuitively, a shorter lengthscale makes the GP mean function more sensitive to position: two nearby points that straddle an inducing location are pushed toward different mean values, preventing their representations from coinciding. Conversely, a very large lengthscale flattens the mean function, assigning nearly identical values to all in-distribution inputs and thereby directly inducing collapse.*

### F.2 FEATURE-COLLAPSE IN DDGP

**Proposition F.2 (Feature-collapse condition, DDGP).** *Under the same assumptions as Proposition F.1, but with the DDGP hybrid kernel* $K^H$ *replacing the SE kernel at layers* $l \ge 2$*, define the* distributional *per-layer contraction coefficient*

$$C_l^H \;\coloneqq\; \frac{\sigma_l^4}{4l_l^4} \sum_{d=1}^{d_l} \|K_{uu}^{-1} U_l^d\|_2^2 \left[1 + \frac{\|\Delta m_l\|_2^2 + \|\Delta v_l\|_2^2}{\|f_l(x) - f_l(x^*)\|_2^2}\right]^2, \tag{99}$$

*where $\Delta m_l = m(f_l(x)) - m(f_l(x^*))$ and $\Delta v_l = v(f_l(x)) - v(f_l(x^*))$ are the differences in the posterior predictive mean and variance of layer $l$ evaluated at $x$ and $x^*$. If $C_l^H \cdot \delta_{l-1} < 1$ at each layer, then*

$$P\Big(\|f_n(x) - f_n(x^*)\|_2 \to 0 \;\Big|\; \{U_l\}_{l=1}^n\Big) = 1. \tag{100}$$

*Proof.* We begin by noticing that:

**Single-layer bound under the hybrid kernel.** The posterior joint distribution at layer $n$ takes the same form as in Proposition F.1, with the SE kernel replaced by $K^H$. The starting bound (92) therefore reads:

$$\delta_n \le \sum_{d=1}^{d_n} \|K_{xu}^H - K_{x^*u}^H\|_2^2 \, \|K_{uu}^{-1} U_n^d\|_2^2, \tag{101}$$

where $[K_{xu}^H]_m = K^H(\mu(f_{n-1}(x)),\, \mu(z_n^{(m)}))$ and $\mu(\cdot)$ denotes the posterior predictive distribution at that location. The hybrid kernel evaluates as

$$K^H(\mu(a), \mu(b)) = \sigma_n^2 \exp\!\left(-\frac{(a-b)^2 + W_2^2(\mu(a), \mu(b))}{2l_n^2}\right). \tag{102}$$

**Bounding the kernel difference.** Let $h = f_{n-1}(x) - f_{n-1}(x^*)$ as before, and assume $f_{n-1}(x^*) \approx z_n^{(m^*)}$ so that $\mu(f_{n-1}(x^*)) \approx \mu(z_n^{(m^*)})$. For a distributional perturbation of size $h$, write the joint increment as

$$\rho^2(h) \;\coloneqq\; \|h\|_2^2 + W_2^2(\mu(f_{n-1}(x)),\, \mu(f_{n-1}(x^*)))\,, \tag{103}$$

so that $K^H(\mu(f_{n-1}(x)), \mu(z_n^{(m^*)})) = \sigma_n^2 \exp(-\rho^2(h)/(2l_n^2))$. Using the closed-form $W_2^2$ distance for Gaussians (see Appendix **??**),

$$W_2^2(\mu(f_{n-1}(x)),\, \mu(f_{n-1}(x^*))) = \|\Delta m_{n-1}\|_2^2 + \|\Delta v_{n-1}\|_2^2, \tag{104}$$

where $\Delta m_{n-1}$ and $\Delta v_{n-1}$ are the mean and variance differences of the layer-$(n-1)$ posteriors at $x$ and $x^*$, respectively. Hence $\rho^2(h) = \|h\|_2^2 + \|\Delta m_{n-1}\|_2^2 + \|\Delta v_{n-1}\|_2^2$. Applying $1 - e^{-t} \le t$ for $t \ge 0$:

$$\begin{aligned}
\|K_{xu}^H - K_{x^*u}^H\|_2^2 &\le \sigma_n^4 \left(1 - \exp\!\left(-\frac{\rho^2(h)}{2l_n^2}\right)\right)^2 \\
&\le \frac{\sigma_n^4}{4l_n^4}\, \rho^4(h) \\
&= \frac{\sigma_n^4}{4l_n^4} \left(\|h\|_2^2 + \|\Delta m_{n-1}\|_2^2 + \|\Delta v_{n-1}\|_2^2\right)^2.
\end{aligned} \tag{105}$$

**Factoring the bound.** Let $c = \|\Delta m_{n-1}\|_2^2 + \|\Delta v_{n-1}\|_2^2 \ge 0$. We factor:

$$\left(\|h\|_2^2 + c\right)^2 = \|h\|_2^4 + 2c\|h\|_2^2 + c^2 = \|h\|_2^2 \left(\|h\|_2^2 + 2c + \frac{c^2}{\|h\|_2^2}\right) = \|h\|_2^2 \left(\|h\|_2 + \frac{c}{\|h\|_2}\right)^2. \tag{106}$$

Substituting (106) into (105) and then into (101):

$$\delta_n \le \frac{\sigma_n^4}{4l_n^4} \sum_{d=1}^{d_n} \|K_{uu}^{-1} U_n^d\|_2^2 \cdot \|h\|_2^2 \left(\|h\|_2 + \frac{\|\Delta m_{n-1}\|_2^2 + \|\Delta v_{n-1}\|_2^2}{\|h\|_2}\right)^2. \tag{107}$$

Recognising $\|h\|_2^2 = \delta_{n-1}$ and writing the bracket in terms of $\delta_{n-1}$:

$$\delta_n \le C_n^H \cdot \delta_{n-1}^2, \tag{108}$$

where $C_n^H$ is exactly the distributional contraction coefficient defined in (99).

**Almost-sure convergence.** The recursion (108) is of the same form as (94), so the Borel–Cantelli argument from Steps 2–3 of Proposition F.1 applies verbatim with $C_l$ replaced by $C_l^H$, completing the proof. □

**Remark F.2** (DDGPs are strictly more resistant to feature-collapse)**.** *Comparing* (89) *and* (99)*, we have*

$$C_l^H = C_l \cdot \left[1 + \frac{\|\Delta m_l\|_2^2 + \|\Delta v_l\|_2^2}{\|f_l(x) - f_l(x^*)\|_2^2}\right]^2 \geq C_l, \tag{109}$$

*with equality only when the two posterior distributions at $x$ and $x^*$ are identical. Since the hybrid kernel augments the Euclidean feature distance with a Wasserstein-2 distributional discrepancy,* any *difference in the posterior moments of the two inputs tightens the collapse condition, making feature-collapse strictly harder to achieve in DDGPs than in DGPs with the same kernel hyperparameters.*

*Furthermore, since the distributional component depends on the posterior predictive mean and variance—both computed via SE kernels—the correction terms $\Delta m_l$ and $\Delta v_l$ inherit the smoothness of the SE kernel. This endows DDGPs with smoother hidden-layer representations compared to DGPs, directly satisfying Desideratum II.*

**Empirical validation.** The analytical distinction in Remark F.2 is confirmed empirically on the "banana" dataset (Figure 7), where we track feature correlations between two focus points across the full input space. DDGPs exhibit markedly smoother and more localised correlations at hidden layer 3, with a visible improvement already at layer 2, consistent with the strictly larger collapse-resistance coefficient $C_l^H > C_l$.

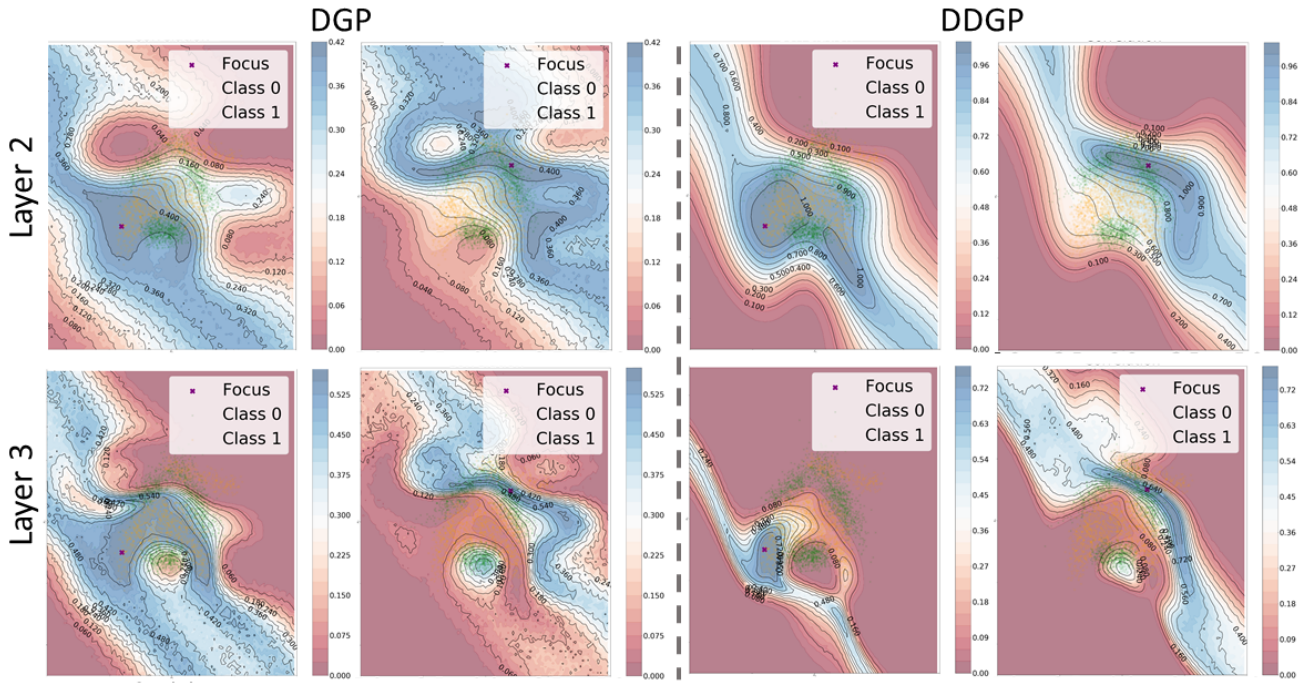



Figure 7: **Smoothness comparison between DGPs and DDGPs.** Pairwise correlation of two focus points (purple dots) with the full input space at each hidden layer, for DGPs (left) and DDGPs (right) with PCA mean function trained on the "banana" dataset. DDGPs exhibit smoother, more localised correlations, consistent with the strictly larger per-layer collapse-resistance coefficient $C_l^H \geq C_l$ established in Proposition F.2.

# G GIPL-DDGP: GLOBAL INDUCING POINT LOCATIONS.

## G.1 FAILURE MODE OF THE STANDARD PARAMETERISATION

The standard DDGP attaches a separate set of distributional inducing locations $\mathbf{Z}_{l-1} \sim \mathcal{N}(\mathbf{m}_{Z_{l-1}}, \mathbf{\Sigma}_{Z_{l-1}})$ at every layer $l \geq 2$, with $(\mathbf{m}_{Z_{l-1}}, \mathbf{\Sigma}_{Z_{l-1}})$ optimised as free variational parameters. This construction is convenient but has one failure mode specific to DDGP.

**Calibration mismatch.** Our OOD guarantee at layer $l$ (Prop. 4.1) requires $\mathbf{\Sigma}_{Z_{l-1}} \approx V_{\text{in}} \ll \sigma^2_{l-1}$, where $V_{\text{in}}$ is the typical posterior variance for in-distribution inputs at layer $l-1$. With $\mathbf{\Sigma}_{Z_{l-1}}$ optimised freely, nothing in the ELBO actively pushes it to track $V_{\text{in}}$ across training; the condition is asymptotically self-enforcing only in the sense that ELBO maxima will eventually settle near it, but only weakly.

## G.2 CONSTRUCTION

We adopt the global-inducing scheme of Ober and Aitchison [2021b], originally introduced for BNNs and standard DGPs and subsequently used in deep Wishart processes [Ober and Aitchison, 2021a]. A single set of inducing inputs $\mathbf{Z}_0 \in \mathbb{R}^{M \times D}$ is learned at the input layer. Layer 1 uses an Euclidean kernel on point inputs $\mathbf{X}$ with $\mathbf{Z}_0$ as inducing inputs, identically to standard sparse GPs. For $l \geq 2$, the inducing locations are no longer free parameters but are obtained by marginalising the inducing outputs $\mathbf{U}_l$ from the variational posterior at layer $l$ and evaluating the resulting predictive distribution at $\mathbf{Z}_{l-1}$:

$$q(\mathbf{Z}_l \mid \mathbf{Z}_{l-1}) = \int p(\mathbf{Z}_l \mid \mathbf{U}_l, \mathbf{Z}_{l-1})\, q(\mathbf{U}_l)\, \mathrm{d}\mathbf{U}_l = \mathcal{N}(\mathbf{Z}_l;\, \boldsymbol{\mu}_{Z_l},\, \boldsymbol{\sigma}^2_{Z_l}), \tag{110}$$

where the moments are the standard SVGP predictive moments of Hensman et al. [2013] applied at $\mathbf{Z}_{l-1}$ rather than at the data:

$$\boldsymbol{\mu}_{Z_l} = K^H_{\mathbf{Z}_l\mathbf{U}_l}(K^H_{\mathbf{U}_l\mathbf{U}_l})^{-1}\mathbf{m}_l, \tag{111}$$

$$\boldsymbol{\sigma}^2_{Z_l} = \mathrm{diag}(K^H_{\mathbf{Z}_l\mathbf{Z}_l} - K^H_{\mathbf{Z}_l\mathbf{U}_l}(K^H_{\mathbf{U}_l\mathbf{U}_l})^{-1}(K^H_{\mathbf{U}_l\mathbf{U}_l} - \mathbf{S}_l)(K^H_{\mathbf{U}_l\mathbf{U}_l})^{-1}K^H_{\mathbf{U}_l\mathbf{Z}_l}), \tag{112}$$

with $K^H$ the hybrid kernel, $\mathbf{m}_l$ and $\mathbf{S}_l$ the mean and covariance of $q(\mathbf{U}_l) = \mathcal{N}(\mathbf{m}_l, \mathbf{S}_l)$, and inducing inputs at layer $l$ given by samples from $q(\mathbf{Z}_{l-1})$.

The propagated inducing locations are *distributional by construction*: their uncertainty arises from the accumulated variance through the composition $f_l \circ \cdots \circ f_1$, exactly the quantity the hybrid kernel needs to evaluate its $W_2$ pathway. There is no need for separate variational parameters $(\mathbf{m}_{Z_l}, \mathbf{\Sigma}_{Z_l})$ because the propagation itself produces these moments.

## G.3 WHY THIS IS THE NATURAL CHOICE FOR DDGP

**(i) Self-enforcing calibration.** The condition $\boldsymbol{\sigma}^2_{Z_{l-1}} \approx V_{\text{in}}$ that Prop. 4.1 requires is now automatic: $\boldsymbol{\sigma}^2_{Z_{l-1}}$ *is* the in-distribution posterior variance, by definition of the propagation (110). Free-parameter DDGP relies on the optimiser to discover this condition; GIPL-DDGP encodes it structurally.

**(ii) Compatibility with the hybrid kernel.** The hybrid kernel $k^H$ takes two arguments per inducing point: a sample (Euclidean pathway) and a Gaussian marginal ($W_2$ pathway). Equation (110) produces both in a single propagation step, with the moments needed for the $W_2$ pathway being $(\boldsymbol{\mu}_{Z_{l-1}}, \boldsymbol{\sigma}^2_{Z_{l-1}})$ and the sample for the Euclidean pathway being a draw from $\mathcal{N}(\boldsymbol{\mu}_{Z_{l-1}}, \boldsymbol{\sigma}^2_{Z_{l-1}})$. Free-parameter DDGP must instead train two distinct objects (the mean and variance of $\mathbf{Z}_{l-1}$) that play asymmetric roles in the two pathways, with no mechanism to keep them mutually consistent.

**(iii) Cross-layer correlation in the function-space posterior.** Although $q(\{\mathbf{U}_l\}) = \prod_l q(\mathbf{U}_l)$ factorises across layers in the inducing *outputs*, the function-space posterior $q(\{\mathcal{F}_l\})$ does not, because at layer $l$ the kernel matrices $K^H_{\cdot\mathbf{Z}_{l-1}}$ depend on a sample of $\mathbf{Z}_{l-1}$, which in turn depends on $\mathbf{U}_{1:l-1}$. This matches the property that Ober and Aitchison [2021b] prove for global inducing points in BNNs and DGPs: posterior dependence between non-adjacent layers, not just between immediate neighbours.

### G.4 RELATION TO PRIOR WORK

The closest precedent in the DGP literature is the "inducing points as inducing locations" construction of Ustyuzhaninov et al. [2020], which similarly propagates a single set of inducing inputs through the layers (their §4.2). Two distinctions matter for DDGP. First, Ustyuzhaninov et al. [2020] keep $q(\{\mathbf{f}_\ell^z\})$ factorised across $\ell$, which gives *marginal* dependence between adjacent layers but renders functions in non-adjacent layers independent (see Appendix G of Ober and Aitchison [2021b] for a precise statement). The Ober–Aitchison construction we adopt induces dependence across all layers because the covariance $\mathbf{\Sigma}_\ell^u = (K^{-1}(\mathbf{U}_{\ell-1}) + \mathbf{\Lambda}_\ell)^{-1}$ of $q(\mathbf{U}_\ell \mid \mathbf{U}_{\ell-1})$ in their formulation, and equivalently the SVGP variance in ours, are functions of the previous layer's sample. Second, Ustyuzhaninov et al. [2020] did not consider distributional kernels: their $\mathbf{f}_\ell^z$ enters only via point evaluations, whereas the hybrid kernel in DDGP requires the full marginal $(\boldsymbol{\mu}_{Z_l}, \boldsymbol{\sigma}^2_{Z_l})$. Equation (110) happens to deliver both, which is why GIPL is a particularly clean fit for DDGP rather than being merely a swap of one inducing-point parameterisation for another.

A more recent application of the global-inducing scheme is to deep Wishart processes [Ober and Aitchison, 2021a], which provides further evidence that propagated-inducing-point posteriors are well suited to kernel-based deep hierarchies. Our GIPL-DDGP is the analogous extension to the $W_2$-aware hybrid kernel.

### G.5 MODIFIED ELBO

The ELBO retains the structure of Eq. (19) in the main paper, with one modification: the KL term at layer $l \geq 2$ is now an expectation over samples of $\mathbf{Z}_{l-1}$, since the prior $p(\mathbf{U}_l) = \mathcal{N}(\mathbf{0}, K^H_{\mathbf{U}_l\mathbf{U}_l}(\mathbf{Z}_{l-1}))$ depends on a propagated quantity. Following the Monte-Carlo treatment of the data-dependent KLs in Ober and Aitchison [2021b], we estimate this expectation with a single sample from $q(\mathbf{Z}_{l-1})$, reusing the sample drawn for the Euclidean pathway of the hybrid kernel at no additional cost. The resulting ELBO is unbiased and a strict lower bound on $\log p(\mathbf{y})$.

**Computational cost.** Per-layer cost is unchanged from standard DDGP: $O(M^3 + NM^2)$ for the GP operations, plus $O(M^2)$ extra for evaluating the moments at $\mathbf{Z}_{l-1}$. The main saving is in parameter count: GIPL-DDGP has $O(MD)$ inducing-related parameters in total, against $O(LMD)$ for the standard DDGP, a factor of $L$ reduction.

**Variant: GIPL-DGP.** The same construction applies to standard DGPs by setting the $W_2$ pathway weight to zero, recovering a pure-Euclidean global-inducing DGP. We use this variant as a baseline in the experiments to isolate the contribution of the hybrid kernel from that of GIPL.

## H FROM DEEP GPS TO DISTRIBUTIONAL DEEP GPS

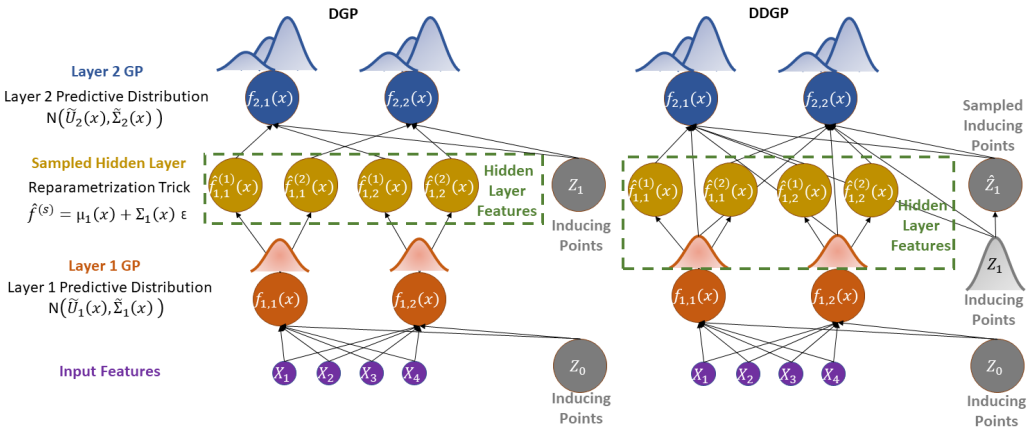



Figure 8: **Conceptual differences between DGPs and DDGPs.** DGPs operate in Euclidean space of sampled hidden layer features. DDGPs also operate on Wasserstein space of underlying hidden layer GPs, with inducing locations following a Gaussian distribution.

Figure 8 illustrates the key architectural difference between a standard Deep GP (DGP) and a Distributional Deep GP

(DDGP) on a two-layer example with four input features and two hidden units per layer.

**Standard DGP.** In a DGP, each layer $\ell$ consists of $D_\ell$ independent GP outputs $f_{\ell,d}$, each conditioned on a shared set of *point-valued* inducing locations $Z_\ell \subset \mathbb{R}^{D_{\ell-1}}$. The layer-1 GPs receive the raw input features $\{X_1, \ldots, X_4\}$ and produce a predictive distribution $\mathcal{N}(\tilde{U}_1(x), \tilde{\Sigma}_1(x))$. Because exact propagation of this distribution to layer 2 is intractable, a single sample is drawn via the reparametrisation trick,

$$\hat{f}^{(s)} = \mu_1(x) + \Sigma_1(x)^{1/2}\,\varepsilon, \quad \varepsilon \sim \mathcal{N}(0, I), \tag{113}$$

yielding sampled hidden-layer features $\{\hat{f}^{(1)}_{1,1}, \hat{f}^{(2)}_{1,1}, \hat{f}^{(1)}_{1,2}, \hat{f}^{(2)}_{1,2}\}$ (gold nodes, dashed box in Figure 8). These point samples are then passed as inputs to the layer-2 GPs, which again condition on point-valued inducing locations $Z_1$. Crucially, the inducing points at every layer are fixed deterministic locations in Euclidean space; the kernel therefore measures similarity between inputs and inducing points purely through Euclidean distance, with no awareness of the uncertainty carried by the hidden-layer representation.

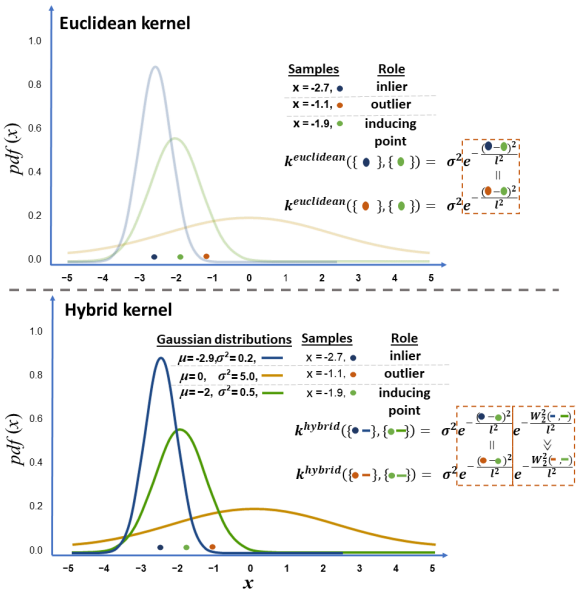



Figure 9: **Conceptual difference between Euclidean and hybrid kernel.** Explicit difference between moments in Wasserstein-2 distance ensures *distance-awareness*.

**Distributional DGP.** The DDGP modifies this architecture in one targeted but consequential way: the inducing points at layers $\ell \geq 1$ are replaced by *distributional inducing points* $\hat{Z}_\ell = \{\mathcal{N}(Z_m, Z_m^{\text{var}})\}_{m=1}^{M}$, each of which is a Gaussian distribution rather than a fixed location (grey shaded nodes in Figure 8, right). The layer-1 computation proceeds identically to the DGP: the same reparametrisation trick is applied and identical sampled hidden features are produced. The difference emerges at the kernel level in layer 2, where the standard SE kernel is replaced by a *hybrid kernel* that is a product of two components. To evaluate it, one draws a point sample $\hat{f}^{(s)} \sim \mathcal{N}(\tilde{\mu}_1(x), \tilde{\Sigma}_1(x))$ from the layer-1 predictive distribution, and independently draws a point sample $\hat{z}_m \sim \mathcal{N}(Z_m, Z_m^{\text{var}})$ from the distributional inducing point. The hybrid kernel is then:

$$k^{\mathcal{H}}\left(\hat{f}^{(s)}, \hat{Z}_m\right) = \sigma^2 \exp\left(-\frac{\mathcal{W}_2^2\left(\mathcal{N}(\hat{f}^{(s)}, \tilde{\Sigma}_1(x)), \mathcal{N}(Z_m, Z_m^{\text{var}})\right)}{2l^2}\right) \cdot \exp\left(-\frac{\|\hat{f}^{(s)} - \hat{z}_m\|^2}{2l^2}\right). \tag{114}$$

The first factor is the $\mathcal{W}_2$-based term: it encodes the *distributional discrepancy* between the hidden-layer predictive distribution and the inducing distribution, capturing both mean and variance differences. The second factor is the standard SE term, now evaluated between the two point samples $\hat{f}^{(s)}$ and $\hat{z}_m$. A test point whose hidden-layer posterior has high residual variance will be assigned low similarity by the first factor regardless of where its mean lands, driving the layer-2 posterior back towards its prior. This distributional awareness is precisely the OOD behaviour absent in standard DGPs, where a high-uncertainty hidden representation is indistinguishable from a confident one once collapsed to a point sample.

**Summary of differences.** Table 3 summarises the two architectures. The input features, GP parameterisation, variational family, and sampling procedure are identical; the sole modification is the promotion of inducing locations at hidden layers from fixed points to Gaussian distributions, and the corresponding substitution of the SE kernel with the hybrid $\mathcal{W}_2$-SE kernel. This minimal change has no effect on the model's in-distribution predictive capacity but endows the DDGP with a principled mechanism for uncertainty propagation across layers that respects the distributional geometry of the hidden representations.

Table 3: Architectural comparison of DGP and DDGP.

| **Component** | **DGP** | **DDGP** |
|---|---|---|
| Input inducing points ($Z_0$) | Point-valued $\in \mathbb{R}^D$ | Point-valued $\in \mathbb{R}^D$ |
| Hidden inducing points ($Z_\ell$, $\ell \geq 1$) | Point-valued $\in \mathbb{R}^{D_\ell}$ | Distributional $\mathcal{N}(Z_m, Z_m^{\text{var}})$ |
| Layer-1 kernel | SE (ARD) | SE (ARD) |
| Layer-$\ell$ kernel ($\ell \geq 2$) | SE (ARD) | Hybrid $\mathcal{W}_2$-SE |
| Hidden sampling | Reparametrisation | Reparametrisation |
| OOD sensitivity | Euclidean distance only | Euclidean + variance discrepancy |

# I ADDITIONAL EXPERIMENTS

## I.1 ADDITIONAL OOD EXPERIMENTS

Due to lack of space in the main paper, we include the table corresponding to the MNIST scenario here.

Table 4: OOD detection on MNIST (inlier). For each OOD score we report AUROC (A↑) and FPR95 (F↓) on three families of distribution shift: near-OOD, far-OOD and covariate shift. Values show mean over 10 random seeds. Std is not shown for improved visualization. Within each block, every column is rank-shaded across models (**best** darkest, **worst** lightest). *mean rank* is the per-model rank averaged over all AUROC/FPR95 evaluations in the block; lower is better, the best is in bold.

| **MNIST (inlier)** | | **Near-OOD** | | | | **Far-OOD** | | | | | | | | | | | | **Cov. Shift** | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | *Fashion-MNIST* | | *MNIST-C* | | *CIFAR-10 (gray)* | | *KMNIST* | | *NotMNIST* | | *Omniglot* | | *Rnd. Noise* | | *SVHN (g)* | | *MNIST-M* | | *Scaled/Translated* | | *USPS* | |
| **Score** | **Model** | A↑ | F↓ | A↑ | F↓ | A↑ | F↓ | A↑ | F↓ | A↑ | F↓ | A↑ | F↓ | A↑ | F↓ | A↑ | F↓ | A↑ | F↓ | A↑ | F↓ | A↑ | F↓ |
| Pred. Ent. | **2L** | .474 | .828 | .661 | .669 | .423 | .789 | .587 | .723 | .681 | .620 | .676 | .620 | .090 | .975 | .416 | .796 | .393 | .810 | .754 | .505 | .615 | .612 |
| | **2L-D** | .412 | .959 | .713 | .625 | .507 | .776 | .638 | .714 | .596 | .847 | .745 | .567 | .116 | .964 | .503 | .851 | .472 | .786 | .781 | .504 | .602 | .701 |
| | **3L** | .950 | .135 | **.924** | **.227** | .956 | .109 | .915 | .263 | .913 | .276 | .917 | .242 | .834 | .299 | .962 | .083 | **.950** | **.117** | **.934** | **.200** | .925 | .255 |
| | **3L-D** | .897 | .267 | .871 | .361 | .877 | .257 | .877 | .342 | .916 | .273 | .827 | .406 | .611 | .600 | .848 | .297 | .864 | .286 | .931 | .231 | **.947** | **.159** |
| | **GP** | .343 | .960 | .613 | .885 | .210 | .971 | .501 | .930 | .584 | .946 | .680 | .750 | .076 | .978 | .224 | .975 | .199 | .982 | .599 | .915 | .395 | .976 |
| | **DE** | **.990** | **.029** | .744 | .807 | **.995** | **.015** | **.988** | **.033** | **.992** | **.023** | **.983** | **.062** | .990 | .025 | **.989** | **.037** | .932 | .454 | .875 | .512 | .802 | .776 |
| | **SNGP** | .989 | .034 | .638 | .912 | .993 | .025 | .986 | .042 | .989 | .037 | .968 | .171 | **.993** | **.024** | .978 | .067 | .887 | .736 | .789 | .815 | .825 | .810 |
| | *mean rank* ↓ | | | **2L**: $5.4_{\pm 0.9}$ | | **2L-D**: $5.0_{\pm 0.9}$ | | **3L**: $2.5_{\pm 1.0}$ | | **3L-D**: $3.2_{\pm 1.1}$ | | **GP**: $6.9_{\pm 0.3}$ | | **DE**: $\mathbf{2.1_{\pm 1.4}}$ | | **SNGP**: $3.0_{\pm 1.7}$ | | | | | | | |
| Dist. Ent. | **2L** | .020 | 1.000 | .006 | 1.000 | .006 | 1.000 | .010 | 1.000 | .110 | 1.000 | .027 | 1.000 | .002 | 1.000 | .026 | 1.000 | .002 | 1.000 | .025 | 1.000 | .003 | 1.000 |
| | **2L-D** | .209 | .991 | .094 | .989 | .303 | .845 | .106 | .994 | .286 | .988 | .225 | .906 | .472 | .603 | .350 | .796 | .233 | .979 | .189 | .988 | .179 | .971 |
| | **3L** | .612 | .873 | .447 | .878 | .738 | .561 | .389 | .925 | .307 | .963 | .509 | .871 | .950 | .140 | .726 | .608 | .728 | .704 | .365 | .947 | .484 | .805 |
| | **3L-D** | .724 | .724 | .847 | .375 | .876 | .336 | .753 | .550 | .616 | .829 | .856 | .341 | .996 | .018 | .865 | .473 | .853 | .441 | .717 | .590 | .732 | .540 |
| | **GP** | **.981** | **.086** | **.909** | **.147** | **.999** | **.003** | **.927** | **.139** | **.951** | **.156** | **.901** | **.148** | **1.000** | **.000** | **.999** | **.007** | **.998** | **.009** | **.897** | **.170** | **.988** | **.041** |
| | *mean rank* ↓ | | | **2L**: $5.0_{\pm 0.0}$ | | **2L-D**: $4.0_{\pm 0.0}$ | | **3L**: $3.0_{\pm 0.0}$ | | **3L-D**: $2.0_{\pm 0.0}$ | | **GP**: $\mathbf{1.0_{\pm 0.0}}$ | | **DE**: – | | **SNGP**: – | | | | | | | |
| Energy | **2L** | .509 | .810 | .720 | .576 | .421 | .759 | .663 | .610 | .672 | .619 | .728 | .532 | .106 | .958 | .413 | .784 | .388 | .790 | .816 | .381 | .656 | .550 |
| | **2L-D** | .552 | .809 | .773 | .531 | .513 | .704 | .740 | .558 | .736 | .626 | .807 | .436 | .084 | .967 | .517 | .741 | .475 | .733 | .868 | .311 | .737 | .501 |
| | **3L** | .985 | .040 | **.962** | **.149** | .983 | .039 | .976 | .091 | .988 | .042 | .969 | .099 | .908 | .183 | .985 | .030 | **.977** | **.051** | **.984** | **.053** | **.983** | .060 |
| | **3L-D** | .960 | .122 | .916 | .281 | .951 | .120 | .951 | .174 | .973 | .102 | .915 | .241 | .735 | .420 | .945 | .132 | .937 | .158 | .975 | .079 | .981 | **.058** |
| | **GP** | .099 | .993 | .191 | .990 | .035 | .998 | .206 | .992 | .340 | .976 | .154 | .964 | .011 | .998 | .027 | .998 | .050 | .999 | .188 | .991 | .168 | .993 |
| | **DE** | **.996** | **.014** | .733 | .803 | **.999** | **.005** | **.994** | **.021** | **.997** | **.013** | **.987** | **.044** | **.997** | **.011** | **.994** | **.021** | .936 | .455 | .881 | .529 | .826 | .765 |
| | **SNGP** | .986 | .050 | .623 | .918 | .992 | .023 | .982 | .076 | .989 | .039 | .959 | .224 | .993 | .023 | .974 | .146 | .878 | .809 | .767 | .813 | .794 | .813 |
| | *mean rank* ↓ | | | **2L**: $5.5_{\pm 0.7}$ | | **2L-D**: $4.7_{\pm 0.9}$ | | **3L**: $2.1_{\pm 0.8}$ | | **3L-D**: $3.2_{\pm 1.0}$ | | **GP**: $7.0_{\pm 0.0}$ | | **DE**: $\mathbf{2.0_{\pm 1.5}}$ | | **SNGP**: $3.5_{\pm 1.6}$ | | | | | | | |
| Maha. | **2L** | .027 | 1.000 | .013 | 1.000 | .142 | 1.000 | .058 | 1.000 | .311 | .973 | .013 | 1.000 | .315 | .800 | .098 | 1.000 | .103 | 1.000 | .015 | 1.000 | .020 | 1.000 |
| | **2L-D** | .014 | 1.000 | .009 | 1.000 | .066 | 1.000 | .034 | 1.000 | .178 | .997 | .006 | 1.000 | .263 | .835 | .043 | 1.000 | .052 | 1.000 | .005 | 1.000 | .006 | 1.000 |
| | **3L** | .000 | 1.000 | .000 | 1.000 | .000 | 1.000 | .000 | 1.000 | .002 | 1.000 | .000 | 1.000 | .001 | 1.000 | .000 | 1.000 | .000 | 1.000 | .000 | 1.000 | .000 | 1.000 |
| | **3L-D** | .000 | 1.000 | .000 | 1.000 | .000 | 1.000 | .001 | 1.000 | .002 | 1.000 | .000 | 1.000 | .006 | 1.000 | .000 | 1.000 | .001 | 1.000 | .000 | 1.000 | .000 | 1.000 |
| | **GP** | .953 | .281 | .088 | 1.000 | 1.000 | **.000** | .938 | .996 | **1.000** | **.000** | .080 | 1.000 | 1.000 | **.000** | **1.000** | **.000** | **1.000** | **.000** | .352 | 1.000 | **1.000** | **.001** |
| | **DE** | **1.000** | **.002** | **.884** | **.421** | **1.000** | **.000** | **.998** | **.009** | .999 | .002 | **.993** | **.031** | **1.000** | **.000** | .999 | .004 | .975 | .156 | .922 | .295 | .976 | .103 |
| | **SNGP** | .999 | .004 | .769 | .673 | .999 | .003 | .996 | .018 | .999 | .006 | .988 | .056 | .999 | .003 | .995 | .023 | .949 | .324 | **.933** | **.251** | .976 | .083 |
| | *mean rank* ↓ | | | **2L**: $3.9_{\pm 0.3}$ | | **2L-D**: $4.5_{\pm 0.7}$ | | **3L**: $5.5_{\pm 1.5}$ | | **3L-D**: $5.1_{\pm 1.2}$ | | **GP**: $2.0_{\pm 1.0}$ | | **DE**: $\mathbf{1.5_{\pm 0.7}}$ | | **SNGP**: $2.4_{\pm 0.6}$ | | | | | | | |

### I.2 ABLATION EXPERIMENTS: DISENTANGLING THE $W_2$ KERNEL FROM GLOBAL INDUCING POINTS

In this subsection we analyse the separate contributions of the two ingredients of our proposed model: the distributional $\mathrm{SE} \times W_2$ kernel and the global inducing-point approach. We therefore run a $2 \times 2$ ablation that crosses the two factors : kernel $\in \{\text{SE only}, \mathrm{SE} \times W_2\}$ and inducing $\in \{\text{local } Z_\ell, \text{ GIPL } Z_l\}$. All models are 3-layer; we use 5 cross-validation splits and 20,000 optimisation iterations per cell. Datasets span small (BOSTON, $N$=506), mid (ENERGY, $N$=768) and large (PROTEIN, $N$=45,730) regimes.

Table 5: Ablation grid: RMSE (lower is better), mean $\pm$ std over 5 splits.

| Dataset | Kernel | Local $Z_\ell$ | GIPL $Z_\ell$ |
|---|---|---|---|
| BOSTON | SE only | $4.01 \pm 0.728$ | $3.93 \pm 0.699$ |
| | SE $\times W_2$ | $3.82 \pm 0.620$ | $3.65 \pm 1.01$ |
| ENERGY | SE only | $0.52 \pm 0.037$ | $0.488 \pm 0.036$ |
| | SE $\times W_2$ | $0.62 \pm 0.139$ | $0.50 \pm 0.053$ |
| PROTEIN | SE only | $4.62 \pm 0.035$ | $4.48 \pm 0.052$ |
| | SE $\times W_2$ | $4.24 \pm 0.024$ | $4.078 \pm 0.027$ |

Table 6: Ablation grid: NLPD (lower is better), mean $\pm$ std over 5 splits.

| Dataset | Kernel | Local $Z_\ell$ | GIPL $Z_\ell$ |
|---|---|---|---|
| BOSTON | SE only | $0.41 \pm 0.186$ | $0.309 \pm 0.172$ |
| | SE $\times W_2$ | $1.87 \pm 1.333$ | $1.133 \pm 0.726$ |
| ENERGY | SE only | $-1.29 \pm 0.078$ | $-1.512 \pm 0.078$ |
| | SE $\times W_2$ | $-1.430 \pm 0.046$ | $-1.617 \pm 0.084$ |
| PROTEIN | SE only | $1.23 \pm 0.007$ | $1.105 \pm 0.011$ |
| | SE $\times W_2$ | $1.10 \pm 0.006$ | $1.011 \pm 0.006$ |

The ablation shows that the patterns observed in Table 2 are also present in the version of these models without GIPL, with the latter adding an almost uniform improvement in any setting of roughly $5 - 10\%$.

### I.3 GRADIENT-VARIANCE DIAGNOSTIC

In this subsection we are interested in finding out if the $W_2$/distributional pathway affects the variance of the stochastic ELBO-gradient estimator during training (i.e. whether it stabilises or destabilises optimisation relative to the standard DGP pathway) and how the global inducing-point layer (GIPL) interacts with this. We log the full-model gradient norm $\|\nabla \mathcal{L}\|_2$ at every optimisation step and report both the raw trace and its rolling variance over a 100-step window. We run 3 seeds per model and summarise stability by the post-warmup coefficient of variation $\mathrm{CV}(\|\nabla \mathcal{L}\|_2) = \mathrm{std}/\mathrm{mean}$ (lower is smoother). We cover both the tabular regime (3-layer DGP/DDGP and their GIPL variants on PROTEIN, $N$=45,730) and the convolutional regime (2- and 3-layer DeepConvGP vs. its distributional counterpart on MNIST).

**Gradient-variance results.** The diagnostic shows that the $W_2$ pathway does *not* destabilise optimisation, and for the deepest model it markedly improves gradient-noise stability. On MNIST, the 3-layer distributional model attains the lowest post-warmup gradient-norm variability of any model ($\mathrm{CV} = 0.254$) with a variance that *decreases* throughout training, while the deterministic 3-layer DeepConvGP is the least stable ($\mathrm{CV} = 0.536$) with variance that grows by roughly two orders of magnitude (Fig. 11). At 2 layers the two variants are comparable ($\mathrm{CV} = 0.371$ vs. $0.340$), indicating the stabilising effect of the $W_2$ pathway emerges with depth. In the tabular regime, the distributional kernel likewise yields a smoother estimator on the large PROTEIN dataset ($\mathrm{CV} = 0.217$ for DDGP-3 vs. $0.289$ for DGP-3; Fig. 10). Adding GIPL raises gradient variance modestly and, as expected, increases per-step cost (roughly $30\%$ more wall-clock on PROTEIN: 10.7 vs. $8.2$ ms/step for the distributional models), but all variants converge without instability. Overall, the diagnostic supports the claim that the accuracy gains of the distributional pathway are not obtained at the expense of optimisation stability.

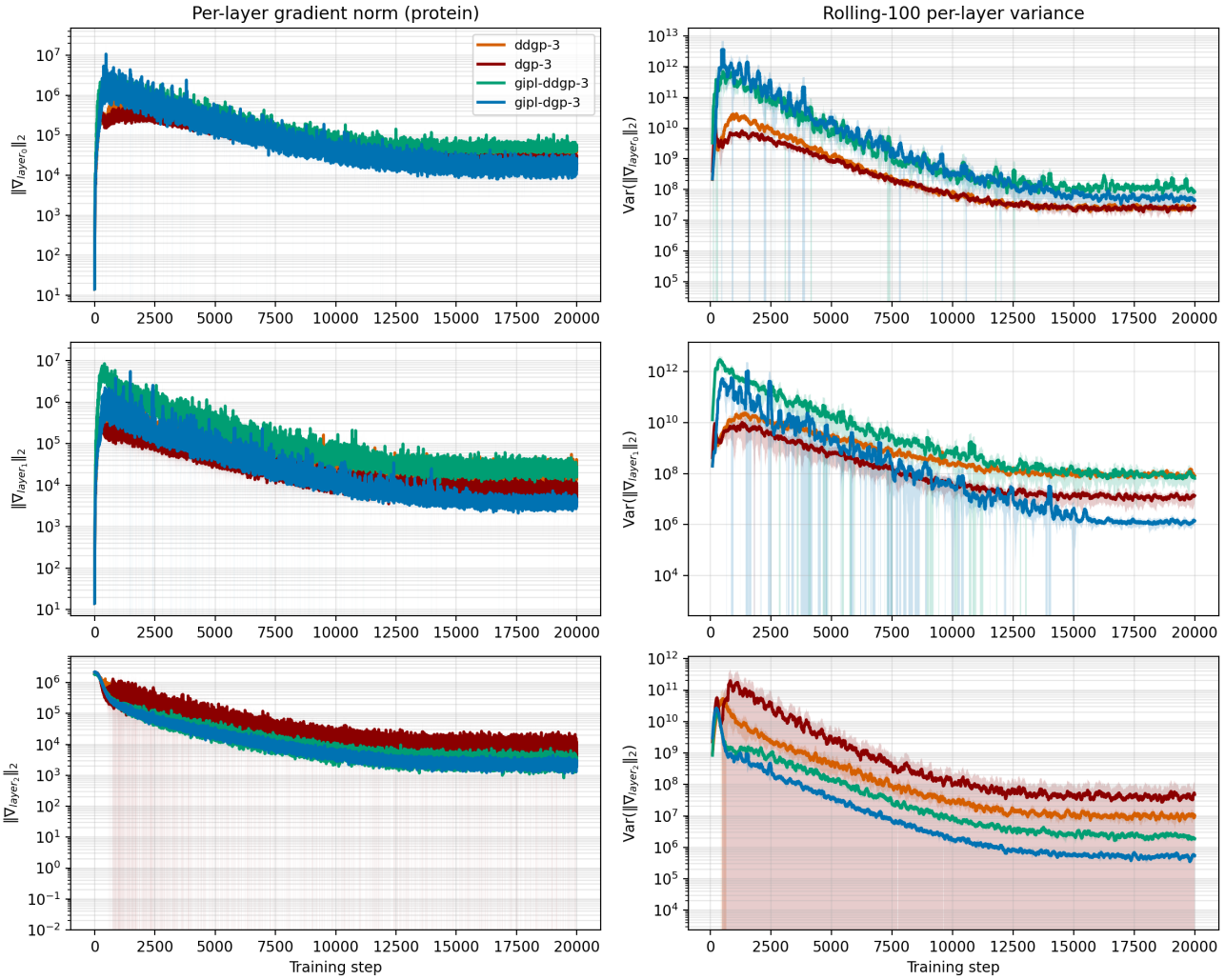



Figure 10: Gradient-norm diagnostic on PROTEIN (3-layer models, mean ± std over 3 seeds). *Left:* per-step gradient norm $\|\nabla\mathcal{L}\|_2$ (log scale). *Right:* rolling-100 variance. All models converge to a comparable low-variance regime; the local-inducing DGP/DDGP settle at the lowest norm and variance, while the GIPL variants carry somewhat higher gradient variance throughout training.

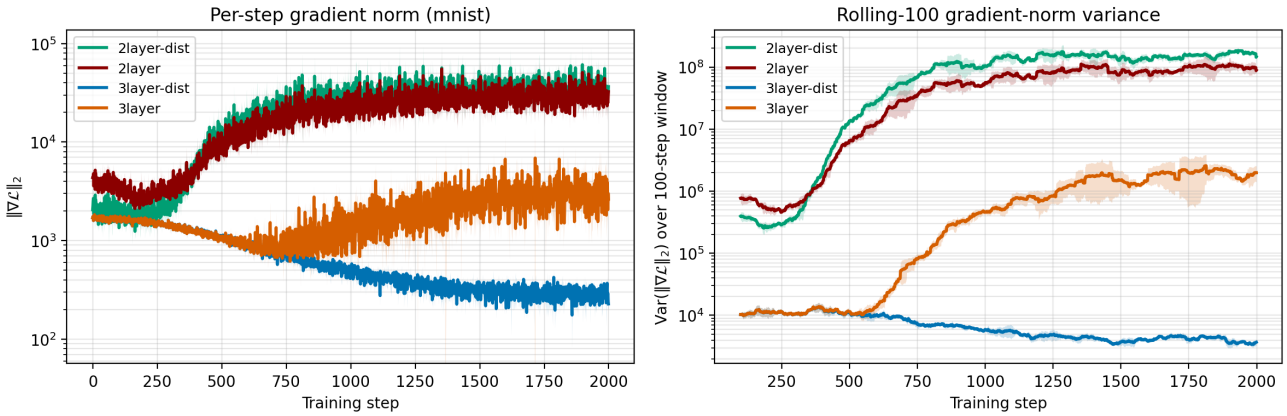



Figure 11: Gradient-norm diagnostic on MNIST convolutional models (mean ± std over 3 seeds). The distributional 3-layer model (`3layer-dist`) has monotonically *decreasing* gradient-norm variance and is the smoothest of all models, whereas its deterministic counterpart (`3layer`) shows variance growing by two orders of magnitude over training. The two 2-layer models behave comparably.

### I.4 COMPUTE AND MEMORY OVERHEAD

Using the same instrumented training runs as the gradient-variance diagnostic (App. I.3), we measured per-step wall-time, total wall-clock, and peak GPU memory across 3 seeds in two regimes: (i) BOSTON UCI regression at depth 3 (5000 iterations, full-batch), and (ii) MNIST classification at depths 2 and 3 (2000 iterations, batch 32). Table 7 reports the aggregate (mean ± std over seeds); peak memory is the per-run JAX BFC peak-allocation delta measured in an isolated process.

Table 7: Compute and peak-memory cost, mean ± std over 3 seeds. Peak memory is the JAX peak-allocation delta from baseline (per-seed variation was below 0.01 MiB, hence no visible spread). Percentages in the text are computed relative to the corresponding non-distributional / non-GIPL baseline.

| Setting | Model | ms/step | Wall (s) | Peak mem (MiB) |
|---|---|---|---|---|
| BOSTON depth 3 5000 iter | DGP-3 | $7.94 \pm 0.03$ | $39.7 \pm 0.2$ | 18.5 |
| | DDGP-3 | $8.70 \pm 0.07$ | $43.5 \pm 0.3$ | 18.5 |
| | GIPL-DGP-3 | $10.18 \pm 0.05$ | $50.9 \pm 0.3$ | 18.5 |
| | GIPL-DDGP-3 | $10.99 \pm 0.08$ | $54.9 \pm 0.4$ | 18.5 |
| MNIST depths 2/3 2000 iter | DConvGP-2 | $61.18 \pm 0.12$ | $122.4 \pm 0.2$ | 795.6 |
| | DistDConvGP-2 | $68.51 \pm 0.10$ | $137.0 \pm 0.2$ | 838.6 |
| | DConvGP-3 | $144.81 \pm 0.20$ | $289.6 \pm 0.4$ | 1103.9 |
| | DistDConvGP-3 | $155.86 \pm 0.11$ | $311.7 \pm 0.2$ | 1148.2 |

**Compute and memory results.** *Tabular (BOSTON, depth 3).* At a fixed inducing strategy the distributional pathway is cheap: DGP-3 runs at 7.94 ms/step (39.7 s total) versus DDGP-3 at 8.70 ms/step (43.5 s), a $+9.6\%$ wall-time overhead, and peak GPU memory is indistinguishable across all four variants ($\approx 18.5$ MiB). The dominant cost is instead the inducing strategy: GIPL adds $\sim 28\%$ wall-time over the standard baseline (50.9 s for GIPL-DGP-3, 54.9 s for GIPL-DDGP-3), with the $W_2$ pathway again adding $\sim 8\%$ on top of GIPL. *Convolutional (MNIST, depths 2/3).* The same picture holds: DistDConvGP-2 is $+12.0\%$ wall-time and $+5.4\%$ peak memory over DConvGP-2, and DistDConvGP-3 is $+7.6\%$ and $+4.0\%$ over DConvGP-3. Depth dominates total cost: going from 2 to 3 layers is $\approx 2.4\times$ wall-time and $\approx 1.4\times$ peak memory, roughly an order of magnitude larger than the $W_2$-pathway overhead. Across both regimes the distributional pathway adds at most $\sim 12\%$ wall-clock and at most $\sim 5.4\%$ peak GPU memory over the corresponding non-distributional baseline, with no observable destabilisation of the optimiser (App. I.3).