Structure and Distribution Metric for Quantifying the Quality of Uncertainty: Assessing Gaussian Processes, Deep Neural Nets, and Deep Neural Operators for Regression

Ethan Pickering * **Themistoklis P. Sapsis**
Mechanical Engineering
Massachusetts Institute of Technology
Cambridge, MA 02139

Abstract

We propose two bounded comparison metrics that may be implemented to arbitrary dimension in regression tasks. One quantifies the *structure* of uncertainty and the other quantifies the *distribution* of uncertainty. The structure metric assesses the similarity in shape and location of uncertainty with the true error, while the distribution metric quantifies the supported magnitudes between the two. We apply these metrics to Gaussian Processes (GPs), Ensemble Deep Neural Nets (DNNs), and Ensemble Deep Neural Operators (DNOs) on high-dimensional and nonlinear test cases. We find that comparing a model's uncertainty estimates with the model's squared error provides a compelling ground truth assessment. We also observe that both DNNs and DNOs, especially when compared to GPs, provide encouraging metric values in high dimensions with either sparse or plentiful data.

1 Introduction

Deep neural networks (DNN/NN) have become increasing popular as a surrogate model of choice. This is largely due to their flexibility, propensity for large-data training, and their predictive performance on unseen data. However, the former property, flexibility, comes at a cost. DNNs lack closed analytical forms, rigorous proofs, or unique solutions. Most importantly, there exists no general approach for quantifying the uncertainty (or uncertainty quantification, UQ) for any DNN model. This leads to a hesitancy, reluctance, or a general lack of trust one may have in DNNs, particularly when safety- and mission-critical real world applications are the subject of training and prediction. Finding a general approach for uncertainty quantification for DNNs will greatly remove such concerns and allow modelers to present new data to alleviate large model uncertainties.

However, there are other, and perhaps more fundamental reasons for a lack of DNN uncertainty quantification. What uncertainty quantification ground truth should a technique be compared to and through what measure should we even compare uncertainty (Gawlikowski et al., 2021)? Quite often, measures are empirical, qualitative, and low-dimensional. For example, in 1-dimensional problems the "eyeball" norm, or intuition, is commonly used to demonstrate the potential of various methods (Lakshminarayanan et al., 2016; Yao et al., 2019). This measure is clearly not robust nor scalable to high dimensions, and biased by human perceptions of what is a superior uncertainty.

Our objective of this study is to propose a predictive-variance and squared-error comparison metric that efficiently quantifies the quality of uncertainty quantification in arbitrarily large dimensions. We propose two scalar metrics for assessing the similarity between uncertainty fields *and* their relationship to the true error. This requires decoupling the structure, where uncertainty and error lie in the parameter space, and distribution of values, i.e. the magnitudes, of uncertainty and error. The former replaces the "eyeball" norm when visualization is impossible, while the latter provides confidence that the magnitudes of predictive variance are reasonable. These metrics allow for us to systemically ask the question: For any given model, what is the model's "quality of uncertainty"?

This work is also motivated from a Bayesian Experimental Design (BED) and Bayesian Optimization (BO) viewpoint, where measures of the predictive variance are the key ingredient for informing the acquisition of new training data. Thus, the metrics posed aim to answer whether the predictive variances found through various modeling strategies are sufficient for use in BED or BO. Here, we specifically consider Gaussian process (GP) regression, deep neural networks (DNNs/NNs) and deep neural operators (DN0s) for this purpose. Although there is no "perfect" quantification of uncertainty, uncertainty quantification found by Gaussian Processes are universally trusted and often cited

*Corresponding author: pickering@mit.edu

as the gold-standard and compose the backbone of BED and BO. Despite this, we emphasize that GPs do not generalize well to large-parameter spaces and high dimensions, motivating our curiosity in the structure of uncertainty in DNNs and DNOs.

2 Uncertainty Metrics

Several quality of uncertainty metrics exist in the literature, but often these metrics are misleading (Yao et al., 2019). Examples include high test log likelihood (LogLL), RMSE, prediction interval coverage probability (PICP) (Pearce et al., 2018), and mean prediction interval width (MPIW) (Su et al., 2018). For high LogLL, the goal is find as much diversity as possible in regions that data has yet to be observed. However, the metric does not possess meaningful bounds and can range from -100 (bad model) to 2 (good model). This is useful for model selection, but not for approximating the true posterior. RMSE only provides performance and greater confidence in the model, not a true uncertainty measure. Similarly, MPIW determines the average width of the 2.5% to 97.5% percentile interval with the goal of minimizing MPIW. This metric directly competes with the concept of model diversity and identifying regions of model concern, especially for high-dimensional problems.

Commonly used for ensemble methods, PICP provides a more suitable measure for the structure/quality of uncertainty. PICP directly measures the probability of test data lying within the 2.5% to 97.5% percentile interval, where the ideal value is 95%. Although PICP provides a probability for capturing test data, it does not provide a measure of how the models do so. PICP does not measure the relative structure of the underlying uncertainty, nor does it weight regions with intriguing uncertainty or test error. As we will show, a large, constant, and "boring" predictive variance over a high dimensional space consistently provide large scores, despite providing limited information about the underlying regression problem. The metrics we pose are not fooled by uninteresting uncertainty quantification.

2.1 Structure Metric: R

The structure metric we implement is no more than the correlation coefficient between the squared error and the predictive variance, but as we will show, it brings far more information than traditionally used metrics, such as PICP or LogLL. Our reasoning for directly comparing squared error and the predictive variance is motivated by BED and BO. For cases where the purpose of uncertainty quantification is to assist in providing models that perform well, in that they generalize to unseen data with small error, we argue the ideal metric be one that measures the ability of uncertainty to identify generalization error. Specifically, our question is whether a model strategy, equipped with a special set of kernels or functions, accurately reflects its perceived errors via its predictive variance. The correlation coefficient does just that, it measures the degree to which the *spatial structure* of the error and the predictive variance are similar.

To compute the correlation metric, both the squared error, $\epsilon^2(\boldsymbol{\theta})$, where $\boldsymbol{\theta}$ is an $n-$dimensional random variable, and $\sigma^2(\boldsymbol{\theta})$ are evaluated at $\boldsymbol{\theta}_{\text{test}}$ points and represented as a 1-dimensional vector. As is standard for calculating the correlation, we center each vector by its mean and normalize it to unity such that:

$$\begin{aligned} \boldsymbol{\sigma}^{2T}\boldsymbol{\sigma}^2 &= 1 \\ \boldsymbol{\epsilon}^{2T}\boldsymbol{\epsilon}^2 &= 1. \end{aligned} \tag{1}$$

We may then take the inner product of the two normalized vectors to assess their similarity or agreement. Due to the normalization the projection leads to the correlation coefficient, R, ranging from -1 to 1,

$$R = \boldsymbol{\sigma}^{2T}\boldsymbol{\epsilon}^2. \tag{2}$$

The correlation coefficient brings an unusual set of bounds. A value of 1 indicates that the fields are identical to a scalar multiple, 0 indicates no agreement, or orthogonality between the two methods, while -1 presents vectors that are identically inverse. For comparing σ^2 and ϵ^2, all 0 and negative values are effectively useless. Negative values are extremely rare and problematic when observed, as we are comparing positive valued fields anchored by identical training datasets. However, the ability of the correlation coefficient to significantly *penalize* inverse behavior between the variance and error is the *attractive feature*. This is specifically what keeps this measure from being tricked by large, constant, and boring predictive variances found by GPs later. As a consequence of the penalization, values should be interpreted as reporting that **at least** an R fraction of the predictive variance regions/mass is in agreement with the squared error.

These bounded values provide a clear and interpretable metric that is defined only by the number of query points in an arbitrarily large parameter space (i.e. θ) and is not hampered by large dimensions.

2.2 Distribution Metric: NDIP

With structure considered, we are also interested in the similarity of supported magnitudes of the predictive variance and pose a metric for assessing the quality of the distribution of uncertainty. Although structure is critical for BED or BO, an understanding of the amplitude of the predictive variance is important for model confidence. In order to define the distribution metric, we remove the notion of structure and look solely at the distribution of predictive variances and squared errors.

Again, we begin with a test vector of predictive variances and squared errors, but instead of centering and normalizing, we fit the values, using a kernel density estimator, to a model agnostic distribution $p_{\sigma^2}(\sigma^2)$. We then discretize and normalize this distribution such that

$$\begin{aligned} \boldsymbol{p}_{\sigma^2}{}^T \boldsymbol{p}_{\sigma^2} &= 1 \\ \boldsymbol{p}_{\epsilon^2}{}^T \boldsymbol{p}_{\epsilon^2} &= 1. \end{aligned} \tag{3}$$

Taking the inner product of this normalized distribution gives the Normalized Distribution Inner Product (NDIP)

$$\text{NDIP} = \boldsymbol{p}_{\sigma^2}{}^T \boldsymbol{p}_{\epsilon^2}. \tag{4}$$

Just as the correlation coefficient, a value of 1 presents two identical distributions, while a value of 0 gives orthogonality (as all value are positive for probability distributions, negative values are not possible).

2.3 Surrogate Models

2.3.1 Gaussian Process Regression

For low-dimensional problems, Gaussian process (GP) regression (Rasmussen, 2003) is seen as the "gold standard" for Bayesian design and uncertainty quantification. A Gaussian process $\bar{f}(\boldsymbol{\theta})$, where $\boldsymbol{\theta}$ is a random variable, is completely specified by its mean function $m(\boldsymbol{\theta})$ and covariance function $k(\boldsymbol{\theta}, \boldsymbol{\theta}')$. For a dataset $\mathcal{D}$ of input-output pairs $(\{\boldsymbol{\Theta}, \mathbf{y}\})$ and a Gaussian process with constant mean $m_{\mathbf{0}}$, the random process $\bar{f}(\boldsymbol{\theta})$ conditioned on $\mathcal{D}$ follows a normal distribution with posterior mean and variance

$$\mu(\boldsymbol{\theta}) = m_0 + k(\boldsymbol{\theta}, \boldsymbol{\Theta})\mathbf{K}^{-1}(\mathbf{y} - m_0) \tag{5}$$

$$\sigma^2(\boldsymbol{\theta}) = k(\boldsymbol{\theta}, \boldsymbol{\theta}) - k(\boldsymbol{\theta}, \boldsymbol{\Theta})\mathbf{K}^{-1}k(\boldsymbol{\Theta}, \boldsymbol{\theta}) \tag{6}$$

respectively, where $\mathbf{K} = k(\boldsymbol{\Theta}, \boldsymbol{\Theta}) + \sigma_\epsilon^2\mathbf{I}$. Equation (5) can be used to predict the value of the surrogate model at any point $\boldsymbol{\theta}$, and (6) to quantify uncertainty in prediction at that point (Rasmussen, 2003). Here, the kernel is chosen as a radial-basis-function (RBF) kernel with automatic relevance determination (ARD),

$$k(\boldsymbol{\theta}, \boldsymbol{\theta}') = \sigma_f^2 \exp\left[-(\boldsymbol{\theta} - \boldsymbol{\theta}')^\top \mathbf{L}^{-1}(\boldsymbol{\theta} - \boldsymbol{\theta}')/2\right], \tag{7}$$

where $\boldsymbol{L}$ is a diagonal matrix containing the lengthscales for each dimension and the GP hyperparameters appearing in the covariance function (σ_f^2 and $\boldsymbol{L}$ in (6) are trained by maximum likelihood estimation).

2.3.2 Deep Neural Networks and Operators

Here we implement the architecture proposed by Lu et al. (2021) for approximating nonlinear operators: *DeepONet*, only covering the basic details here. DeepONet seeks approximations of nonlinear operators by constructing two deep neural networks, one representing the input function at a fixed number of sensors and another for encoding the "locations" of evaluation of the output function. The first neural network, termed the "branch", takes input functions, u, observed at discrete sensors, $x_i, i = 1...m$. The second neural network, the "trunk", encodes inherent qualities of the operator, denoted as z. Together, these networks seek to approximate the nonlinear operation upon u and z as $G(u)(z) = y$, where y denotes the scalar output from the u, z input pair. Therefore, our set of input-output pairs when discussing DeepONet are, $\{[\mathbf{u}, \mathbf{z}], G(\mathbf{u})(\mathbf{z})\}$.

Although NNs are attractive for approximating nonlinear regression tasks, their complexity rids them of analytical expressions for uncertainty. There are several techniques for quantifying uncertainty in neural networks, however we only consider ensemble methods here (see Gawlikowski et al. (2021) and Psaros et al. (2022) for comprehensive reviews of methods for uncertainty quantification in NNs).

Ensemble approaches have been used quite extensively throughout the literature (Hansen and Salamon, 1990; Lakshminarayanan et al., 2016) and despite their improved results for identifying the underlying tasks at hand (Gustafsson et al., 2020), their utility for quantifying uncertainty in a model remains a topic of debate. The are several approaches for creating ensembles. These include random weight initialization (Lakshminarayanan et al., 2016), different network architectures (including activation functions), data shuffling, data augmentation, bagging, bootstrapping, and snapshot ensembles (Loshchilov and Hutter, 2016; Huang et al., 2017; Smith, 2015) among others. Here we employ random weight initialization, as it has been found to perform similarly or better as BNN approaches (Monte Carlo Dropout and Probablistic Backpropagation) for evaluation accuracy and out-of-distribution detection for both classification and regression tasks (Lakshminarayanan et al., 2016). Although much of the literature is skeptical of the generality of ensembles to provide uncertainty estimates, recent viewpoints, specifically Wilson and Izmailov (2020), have argued that DNN ensembles provide a very good approximation of the posterior.

We train $N = 10$ randomly weight-initialized NN models, each denoted as $\tilde{G}_n$, that find the associated solution field Y for inputs u and z. This allows us to then determine the point-wise variance of the models as

$$\sigma^2(u, z) = \frac{1}{(N-1)} \sum_{n=1}^{N} (\tilde{G}_n(u)(z) - \overline{\tilde{G}(u)(z)})^2 \tag{8}$$

where $\overline{\tilde{G}(u)(z)}$ is the mean solution of the model ensemble and N are the total number of models retained from the initialized weight models. Finally, we must adjust the above representation to match the description for GPs and to permit a systematic scaling in dimension. The input parameters, $\boldsymbol{\theta}$, represent the union of two set of parameters, the stochastic parameters $\boldsymbol{\theta}_u$ and the operation parameters

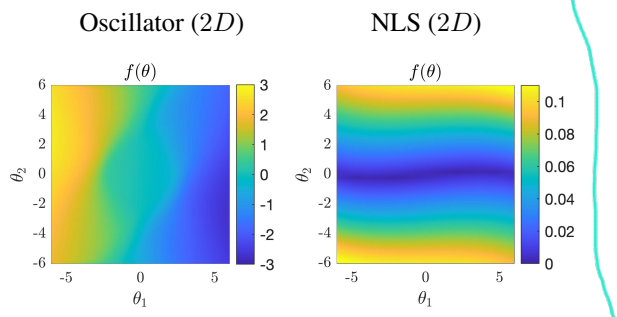

Figure 1: The objective functions, as shown by the contour values, for the 2-dimensional stochastic forcing applied to the SO (left) and the stochastic initial condition to the NLS (right).

$\boldsymbol{\theta}_z$. The parameters $\boldsymbol{\theta}_u$ typically represent coefficients to a set of functions that represent a decomposition of a random function $u = \boldsymbol{\theta}_u \boldsymbol{\Phi}(x_1, ...x_m)$, while $\boldsymbol{\theta}_z = z$ represent non-functional parameters. Thus, the DNN/DNO description for UQ may be recast as:

$$\begin{aligned}\sigma^2(\boldsymbol{\theta}) &= \sigma(\boldsymbol{\theta}_u \cup \boldsymbol{\theta}_z) \\ &= \frac{1}{(N-1)} \sum_{n=1}^{N} (\tilde{G}_n(\boldsymbol{\theta}_u \boldsymbol{\Phi}(x_1, ...x_m))(\boldsymbol{\theta}_z) \\ &- \overline{\tilde{G}(\boldsymbol{\theta}_u \boldsymbol{\Phi}(x_1, ...x_m))(\boldsymbol{\theta}_z)})^2.\end{aligned} \tag{9}$$

For this study, a modest 10 randomly initialized ensemble members are used.

3 Results

Here we demonstrate the metric on two applications, a stochastic oscillator (SO) (Mohamad and Sapsis, 2018; Blanchard and Sapsis, 2020) and a version of the nonlinear Schrödinger equation (NLS) (Majda et al., 1997). Details for each set of equations and the appropriate output definitions are given in Appendix A.

We are specifically interested in how uncertainty is quantified for low- to high-dimensional stochastic processes and with regard to sparsely and densely populated training sets. The dimension of the stochastic excitation (SO, $u(t)$) or initial condition (NLS, $u(x)$), is defined by a finite number of random variables using the Karhunen-Loève expansion,

$$u \approx \boldsymbol{\theta}_u \boldsymbol{\Phi}, \tag{10}$$

where $\boldsymbol{\theta}_u \in \mathbb{R}^m$ is a vector of random variables and $\boldsymbol{\Phi}$ are the eigenvectors of an associated correlation matrix. This definition allows a systematic increase in the input space, on a [-6,6] domain, to arbitrarily large dimensions.

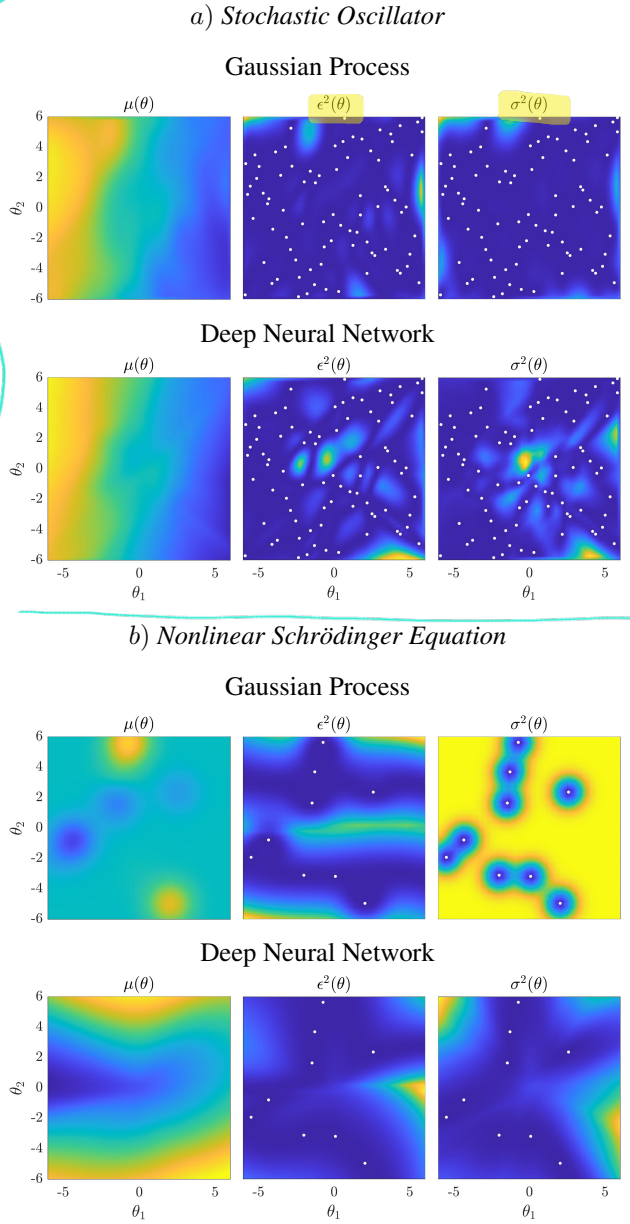

Figure 2: $2D$ GP and NN regressions, squared errors, and predictive variances for SO $a)$ with dense training data ($\boldsymbol{\theta}_n = 81$) and NLS $b)$ with sparse training data ($\boldsymbol{\theta}_n = 9$).

Table 1: Metric values for the visualized squared errors and predictive variances in figure 2. "Better" scores are bold for each case.

SO	R	NDIP	PICP	LogLL	MIWP	MSE
GP	**0.76**	**0.99**	**0.55**	**-1.65**	0.12	0.009
NN	**0.76**	0.98	0.46	-3.24	**0.07**	**0.004**
NLS						
GP	0.45	0.42	**0.88**	**2.31**	0.046	0.19
NN	**0.69**	**0.93**	0.14	-1.09	**0.007**	**0.06**

3.1 Gaussian Processes and Deep Neural Networks

3.2 $2D$ Example

For the 2D case, we provide visual examples of our approach. Figure 1 presents the objective output, or regression task, for the 2-dimensional stochastic representations for both the SO and NLS. Considering that only stochastic dimensions, $\boldsymbol{\theta}_u$, are considered, the DeepONet architecture reduces to a standard DNN (or simply NN).

Figure 2 a) provides the estimated maps, the squared error, and the predictive uncertainties for SO (81 training points), NLS (9 training points) using both GPs and DNNs, while table 1 provides the scores from our two metrics and the four metrics featured in Yao et al. (2019) (defined in Appendix B). Immediately from the figure and the reported R value, we observe that both the GP and NN provide predictive variances that agree similarly in structure with the squared error. The similarity is so close that their R value is 0.76 for both models, despite significant differences between the GP and NN predictive variances. Even with the visual and R value agreement, the traditionally trusted PICP and LogLL favor the GP model, while the PICP for both GP and NN is quite low, and would be considered unacceptable. For this case, PICP fails to capture the relationship between true error and predictive variance.

The visualization and metrics of the sparsely trained NLS example tell an alternate story with a similar conclusion: The standard metrics cannot identify agreement in structure between the predictive variance and squared error. Here the R value for the NN is 0.69 and visual inspection confirms a similar structure, but the PICP is only 0.14 and the LogLL underwhelms the GP case by a factor of e^3. The GP, however, possess a reasonable R at 0.45, but a substantially large PICP of 0.88. Clearly, the GP model provides a predictive uncertainty that is less informative to the underlying error than the NN. PICP is susceptible to constant uncertainty. This will be exacerbated in higher dimensions.

3.3 Scaling: Dimensions and Data

We now look to explore the metrics as we increase dimension and vary the training data from sparse to dense. Here we only compare MSE, PCIP, R, and NDIP, as LogLL and MPIW provided consistently monotonic values with increasing data (LogLL increasing, MPIW decreasing), and limited insight for varied dimensionality.

For the SO case, we provide the median results of 25 independent experiments from 1-5D and training points ranging from $n_{\text{train}} = (D+1)^x$ where D is the dimension and $x = [1, 2, 3, 4]$ with $x = 1$ relating to a sparse distribution and $x = 4$ being a dense distribution (e.g. $D = 2$ gives $n_{\text{train}} = [3, 9, 27, 81]$.). The metrics are evaluated over 10^3 test points for $1D$, 10^4 for $2D$ and 10^5 for $3-5D$.

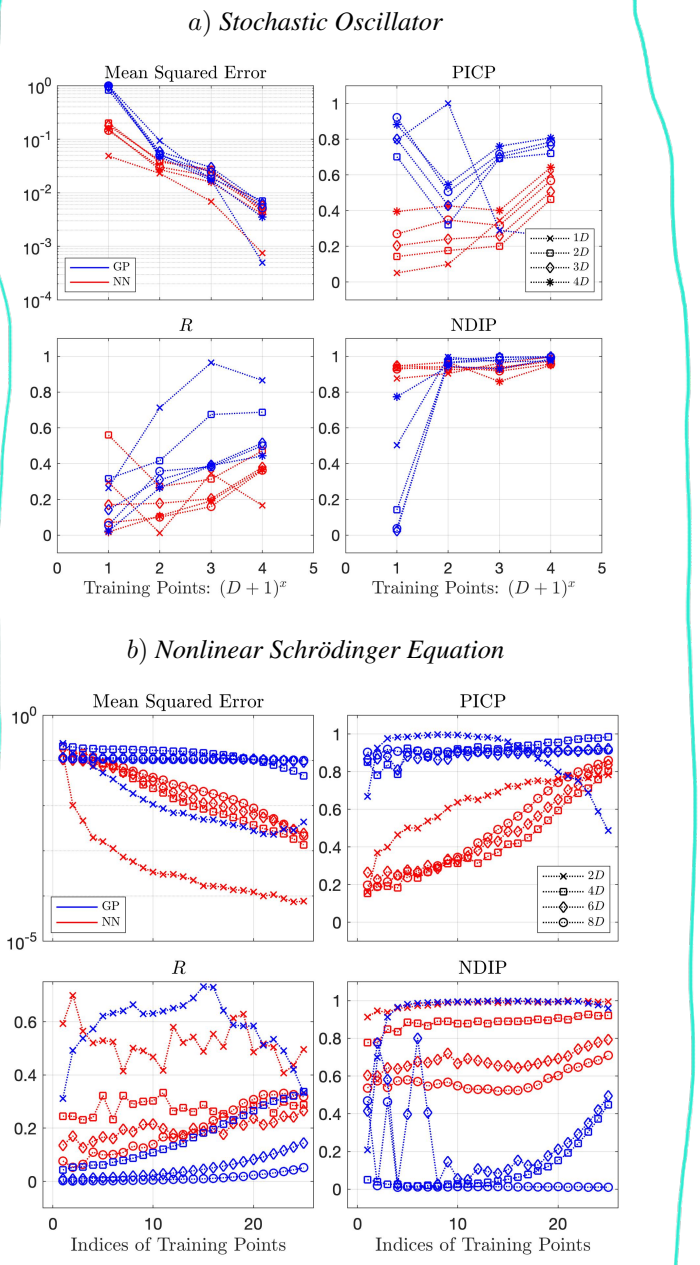

Figure 3: Median values of the MSE, PICP, R and NDIP from sparse to dense training points for many dimensions. a) the stochastic oscillator (25 experiments) and b) the Nonlinear Schrödinger equation (10 experiments).

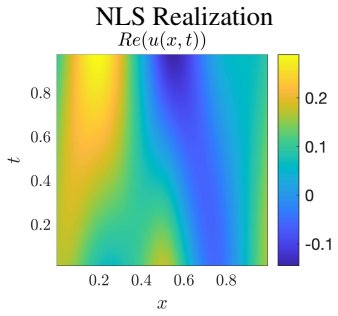

Figure 4: One realization (i.e. $\boldsymbol{\theta}$) of the NLS solution.

All training and test points are computed using Latin hypercube sampling (LHS). Figure 3 a) presents the median metrics related to both models (GP as blue and NN as red) and all five dimensions. We can see that the R values agree with the visual evidence presented earlier and PICP tracks similarly to R for this example. Juxtaposing all subplots it is not clear which modeling approach is "better". NNs provide better MSE, GPs provide slightly better structural metric values over all cases (especially low-D), and NNs perform better in both metrics for sparse data. Based on these results, if one is happy with one model's uncertainty quantification, there is no justifiable reason to not be just as happy with the other.

Turning to the more complex NLS example in 3 b), we test 10 independent experiments over 25 log-spaced training points, $2D$: 3-300, $4D$:5-1000, $6D$:7-2500, $8D$:9-5000, where the x-axis denotes the indice of the interval. The MSE shows that for $4D$ and higher, the GPs are breaking down due to the complexity of the map and high-dimensional regression. Despite this, the PICP for GPs at all dimensions and training sizes, except $2D$ at large n_{train}, is nearly 1. The R metric is not fooled by the poor GP regression and gives values of nearly zero for $> 2D$, until approximately $> 10^3$ points are provided. This is a significant amount of training data for GPs. Additionally, the NDIP metric begins to take action. The low NDIP scores for GPs at $> 2D$ stress that the predictive variance does not present the rich distribution of underlying error values. The NNs, however, provide relatively impressive scores for R, NDIP, and MSE, for all dimensions. Only PICP reports poor NN uncertainty quantification.

3.4 Deep Neural Operator: DeepONet

We now extend our analysis for deep neural operators for the NLS case. Instead of only regressing on a set of random initial conditions from $2-8D$, we add both the spatial and temporal dimensions, x, t. This is a particularly challenging regression. Considering the difficulty of GPs to parameterize the non-operator case, without x, t, we only consider the uncertainty of an ensemble of DeepONets.

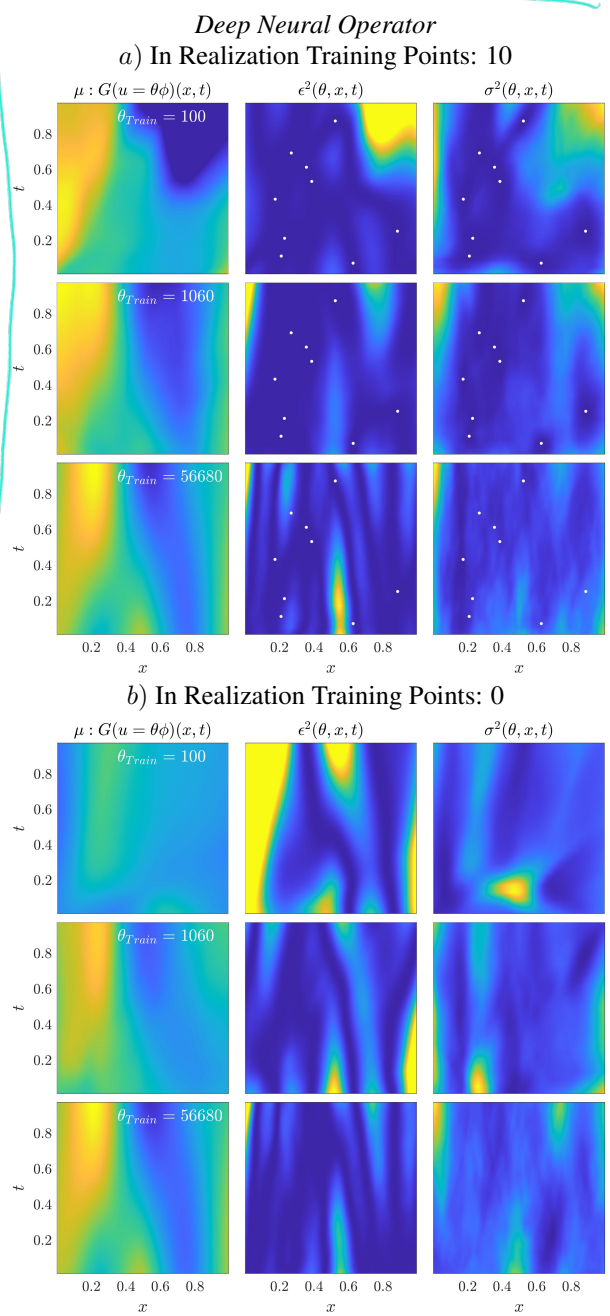

Figure 5: Mean regression, squared error, and predictive variance over x, t for one realization of $\boldsymbol{\theta}$ with 100, 1060, and 56680 training samples, top to bottom, respectively. a) provides training that included 10 points from the visualized realization, while b) does not. R from top to bottom: a) 0.72, 0.61, 0.42; b) -0.18, 0.24, 0.57.

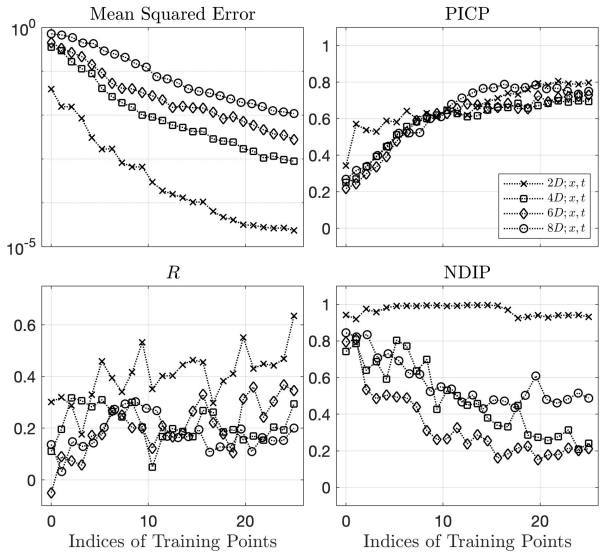

Figure 6: Median values of the MSE, PICP, R and NDIP from sparse to dense training points for the operator case.

Figure 4 visualizes one realization of an $8D$ initial condition propagated in time and space, while 5 provides two sets of training snapshots for the same realization. Figure 5 a) provides the mean DNO regressions when 10 randomly chosen points from this realization are include in the training set, denoted as white points. Regressions are given for $\theta_{\text{train}} = [100, 1060, 56650]$, where $\theta_{u,\text{train}} = \theta_{\text{train}}/10$ as each function set of parameters includes 10 randomly chosen x, t points. We observe that with little training data, the squared error and predictive variance are small around the training points, but as training increases the individual influence of these points diminishes. Interestingly, similar errors and variances appear for 5 b) when training increases, despite the absence of training points from this realization.

With a visualization of the task at hand, we apply the metrics over the various dimensions and with increasing data size. Here the training data sizes are log-spaced over an interval with 25 indices, $2D; x, t$: 50-10000, $4D; x, t$:70-25000, $6D; x, t$:90-50000, $8D; x, t$:110-75000. Figure 6 provides the median metric values for 10 independent experiments. Despite the dimensionality and the large data sizes considered, both the R and NDIP values are promising. For the R value it is particularly intriguing that large data sizes produce substantial agreement with the underlying error. This implies that even at these scales, BED and BO with DNOs is likely fruitful. The NDIP metric does show reduction with increased data size at large dimension, however, the values appear to converge. This suggests the variance and error distributions are not substantially disconnected. Although the PCIP value increases monotonically with data size, our previous observations leave little room for prescribing meaning to these trends.

4 Conclusions

The proposed correlation metric computed over predictive uncertainty and squared error provides a representative measure a model's underlying predictive deficiencies. The metric consistently quantifies the similarity in structure between uncertainty and error, while other metrics do not. Metrics, such as PICP, appear to provide useful information at low dimensions, but when complexity and dimensionality are increased these metrics are easily fooled (e.g. figure 3 b)). The correlation metric provides a means for quantifying the relationship of the topology between predictive variance and error. Such a measure is critical for instilling confidence that the surrogate model of choice is performing well. This is especially important for Bayesian experimental design and optimization that rely on intriguing predictive variance estimates for efficient data acquisition.

Comparing the uncertainties emitted by GPs, DNNs, and DNOs, or other models, unjustly biases what a "good" uncertainty is for a given model. Often, models such as GPs or Hamiltonian Monte Carlo (Neal et al., 2011) are lauded for ideal uncertainty quantification, but this is biased to conceptions of how uncertainty is perceived in 1 or 2 dimensions (and these models do not scale well). As the correlation metric shows, uncertainty quantification can take many forms with similar quantification of the underlying error as shown in figure 6. Any one model is not superior to another through direct comparison of predictive variance. Instead, models must be directly assessed against their own inherent deficiencies. If a model consistently reflects correlation of error with predictive variance, then it honestly informs the user of its deficiencies.

Shallow ensembles of DNNs and DNOs provide encouraging predictive variance structure and distribution with respect to the true error. Here we employ an ensemble approach consisting of only 10 members, yet the test accuracy and structure of the predictive variance are often superior in MSE and correlation metric than GPs. This is especially true for problems of higher complexity (e.g. NLS and figure 3 b)) and greater dimensionality. Considering computational cost for training N members is a critical disadvantage for ensemble NNs, these results are promising for those aiming to apply ensemble NNs or DNOs to BED or BO.

A Applications

A.1 Stochastic Oscillator

Investigated previously by Mohamad and Sapsis (2018) and Blanchard and Sapsis (2020), the stochastic oscillator

is described as

$$\frac{d^2s}{dt^2} + \delta\frac{ds}{dt} + F(s) = u(t), \tag{11}$$

where $s(t) \in \mathbb{R}$ is the state variable, $u(t)$ is a stationary stochastic process with correlation function $\sigma_u^2 \exp\left[-\tau^2/(2\ell_u^2)\right]$, and F is a nonlinear restoring force defined by

$$F(u) = \begin{cases} \alpha s, & \text{for } 0 \le |s| \le s_1 \\ \alpha s_1, & \text{for } s_1 \le |s| \le s_2 \\ \alpha s_1 + \beta\left(s - s_2\right)^3, & \text{for } s_2 \le |s| \end{cases} \tag{12}$$

The remaining parameters take the values, $\delta = 1.5, \alpha = 1, \beta = 0.1, s_1 = 0.5, s_2 = 1.5, \sigma_\xi^2 = 0.1, \ell_\xi = 4$, and $T = 25$. The specific output of interest, shown in figure 2 is the mean value of $u(t)$ over the interval $[0, T]$:

$$f(\boldsymbol{\theta}) = \frac{1}{T}\int_0^T s(t;\boldsymbol{\theta})\mathrm{d}t. \tag{13}$$

A.2 Nonlinear Schrödinger Equation

We implement a version of the nonlinear Schrödinger (NLS) equation, supplemented with a dissipation term for stability proposed by Majda, McLaughlin, and Tabak (Majda et al., 1997) for studying 1D wave turbulence. It is a one-dimensional, dispersive nonlinear prototype model with intermittent events described by

$$iu_t = \left|\partial_x\right|^\alpha u + \lambda\left|\partial_x\right|^{-\beta/4}\left(\left|\left|\partial_x\right|^{-\beta/4}u\right|^2\left|\partial_x\right|^{-\beta/4}u\right) + iDu, \tag{14}$$

where u is a complex scalar, exponents α and β are chosen model parameters, and D is a selective Laplacian. See (Majda et al., 1997; Cai et al., 1999) for details on the rich dynamics and (Zakharov et al., 2001, 2004; Pushkarev and Zakharov, 2013) for its application in understanding extreme rogue waves. We refer the reader to Cousins and Sapsis (2014) for details in computing this version of the NLS, but note the chosen parameters for that description here: $\alpha = 1/2$, $\beta = 0$, $\lambda = -0.5$, $k^* = 20$, $f(k) = 0$, $dt = 0.001$, and a grid that is periodic between 0-1 with $N_x = 512$ grid points.

To propose a stochastic and complex initial condition, $u(x, t = 0)$, we use the complex-valued kernel

$$k(x, x') = \sigma_u^2 e^{i(x-x')}e^{-\frac{(x-x')^2}{\ell_u}}, \tag{15}$$

with $\sigma_u^2 = 1$ and $\ell_u = 0.35$. The objective function we define for the NNs is

$$f(\boldsymbol{\theta}) = ||Re(u(x, t = T;\boldsymbol{\theta}))||_\infty, \tag{16}$$

where $T = 20$, while the objective function for DNOs is simply $Re(u(x, t)$.

B Metrics

All typically used metrics are discussed here where y_n refer to test samples.

Average marginal log-likelihood (LogLL): Maximize

$$\frac{1}{N}\sum_{n=1}^{N}\log(p(y_n|\theta_n)) \tag{17}$$

Normalized Mean Squared-Error (MSE): Minimize

$$\frac{1}{N}\sum_{n=1}^{N}(\mu(\theta_n) - y_n)^2 \tag{18}$$

Prediction Interval Coverage Probability (PICP): Values close to 0.95

$$\frac{1}{N}\sum_{n=1}^{N}\mathbb{1}_{y_n \le \tilde{y}_n^{97.5}} \cdot \mathbb{1}_{y_n \ge \tilde{y}_n^{2.5}} \tag{19}$$

Mean Prediction Interval Width (MPIW): Minimize

$$\frac{1}{N}\sum_{n=1}^{N}\left(\tilde{y}_n^{97.5} - \tilde{y}_n^{2.5}\right) \tag{20}$$

where 97.5 and 2.5 refer to percentiles of the posterior distribution, $\tilde{y}_n$.

References

Blanchard, A. and Sapsis, T. P. (2020). Output-weighted optimal sampling for bayesian experimental design and uncertainty quantification. *arXiv e-prints*, pages arXiv–2006.

Cai, D., Majda, A. J., McLaughlin, D. W., and Tabak, E. G. (1999). Spectral bifurcations in dispersive wave turbulence. *Proceedings of the National Academy of Sciences*, 96(25):14216–14221.

Cousins, W. and Sapsis, T. P. (2014). Quantification and prediction of extreme events in a one-dimensional nonlinear dispersive wave model. *Physica D: Nonlinear Phenomena*, 280:48–58.

Gawlikowski, J., Tassi, C. R. N., Ali, M., Lee, J., Humt, M., Feng, J., Kruspe, A., Triebel, R., Jung, P., Roscher, R., et al. (2021). A survey of uncertainty in deep neural networks. *arXiv preprint arXiv:2107.03342*.

Gustafsson, F. K., Danelljan, M., and Schon, T. B. (2020). Evaluating scalable bayesian deep learning methods for robust computer vision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 318–319.

Hansen, L. K. and Salamon, P. (1990). Neural network ensembles. *IEEE transactions on pattern analysis and machine intelligence*, 12(10):993–1001.

Huang, G., Li, Y., Pleiss, G., Liu, Z., Hopcroft, J. E., and Weinberger, K. Q. (2017). Snapshot ensembles: Train 1, get m for free. *arXiv preprint arXiv:1704.00109*.

Lakshminarayanan, B., Pritzel, A., and Blundell, C. (2016). Simple and scalable predictive uncertainty estimation using deep ensembles. *arXiv preprint arXiv:1612.01474*.

Loshchilov, I. and Hutter, F. (2016). Sgdr: Stochastic gradient descent with warm restarts. *arXiv preprint arXiv:1608.03983*.

Lu, L., Jin, P., Pang, G., Zhang, Z., and Karniadakis, G. E. (2021). Learning nonlinear operators via deeponet based on the universal approximation theorem of operators. *Nature Machine Intelligence*, 3(3):218–229.

Majda, A. J., McLaughlin, D. W., and Tabak, E. G. (1997). A one-dimensional model for dispersive wave turbulence. *Journal of Nonlinear Science*, 7(1):9–44.

Mohamad, M. A. and Sapsis, T. P. (2018). Sequential sampling strategy for extreme event statistics in nonlinear dynamical systems. *Proceedings of the National Academy of Sciences*, 115(44):11138–11143.

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

Pearce, T., Brintrup, A., Zaki, M., and Neely, A. (2018). High-quality prediction intervals for deep learning: A distribution-free, ensembled approach. In *International conference on machine learning*, pages 4075–4084. PMLR.

Psaros, A. F., Meng, X., Zou, Z., Guo, L., and Karniadakis, G. E. (2022). Uncertainty quantification in scientific machine learning: Methods, metrics, and comparisons. *arXiv preprint arXiv:2201.07766*.

Pushkarev, A. and Zakharov, V. E. (2013). Quasibreathers in the MMT model. *Physica D: Nonlinear Phenomena*, 248:55–61.

Rasmussen, C. E. (2003). Gaussian processes in machine learning. In *Summer school on machine learning*, pages 63–71. Springer.

Smith, L. N. (2015). No more pesky learning rate guessing games. *CoRR, abs/1506.01186*, 5.

Su, D., Ting, Y. Y., and Ansel, J. (2018). Tight prediction intervals using expanded interval minimization. *arXiv preprint arXiv:1806.11222*.

Wilson, A. G. and Izmailov, P. (2020). Bayesian deep learning and a probabilistic perspective of generalization. *arXiv preprint arXiv:2002.08791*.

Yao, J., Pan, W., Ghosh, S., and Doshi-Velez, F. (2019). Quality of uncertainty quantification for bayesian neural network inference. *arXiv preprint arXiv:1906.09686*.

Zakharov, V. E., Dias, F., and Pushkarev, A. (2004). One-dimensional wave turbulence. *Physics Reports*, 398(1):1–65.

Zakharov, V. E., Guyenne, P., Pushkarev, A. N., and Dias, F. (2001). Wave turbulence in one-dimensional models. *Physica D: Nonlinear Phenomena*, 152:573–619.