Beyond Marginal Uncertainty: How Accurately can Bayesian Regression Models Estimate Posterior Predictive Correlations?

Abstract

While uncertainty estimation is a well-studied topic in deep learning, most such work focuses on marginal uncertainty estimates, i.e. the predictive mean and variance at individual input locations. But it is often more useful to estimate predictive correlations between the function values at different input locations. In this paper, we consider the problem of benchmarking how accurately Bayesian models can estimate predictive correlations. We first consider a downstream task which depends on posterior predictive correlations: transductive active learning (TAL). We find that TAL makes better use of models’ uncertainty estimates than ordinary active learning, and recommend this as a benchmark for evaluating Bayesian models. Since TAL is too expensive and indirect to guide development of algorithms, we introduce two metrics which more directly evaluate the predictive correlations and which can be computed efficiently: meta-correlations (i.e. the correlations between the models’ correlation estimates and the true values), and cross-normalized likelihoods (XLL). We validate these metrics by demonstrating their consistency with TAL performance and obtain insights about the relative performance of current Bayesian neural net and Gaussian process models.

1 Introduction

Uncertainty estimation is a key desideratum for modern deep learning systems, and is essential for guiding exploration and making robust decisions. Most works on uncertainty estimation have focused on marginal uncertainty, i.e. the posterior predictive variance \( \text{Var}(f(x)|D_{tr}) \) of a function \( f \) at a single location \( x \) conditioned on the training data \( D_{tr} \). Marginal uncertainty is used in many active learning and Bayesian optimization algorithms [14, 10, 3, 6], as high uncertainty at a point indicates it is favorable to query it.

However, some algorithms can explore even more efficiently by exploiting posterior predictive correlations (PPCs) between function values at multiple locations. Mathematically, we are interested in \( \rho(f(x), f(x')|D_{tr}), \) where \( \rho \) denotes the Pearson correlation. Algorithms that exploit PPCs include transductive experiment design [36], where the goal is to acquire information relevant to a specific prediction \( f(x_*) \); in this case, it is advantageous to query locations whose values have a high PPC with \( f(x_*) \). Similarly, in cost-sensitive Bayesian optimization, it would be very useful to make cheap queries in order to indirectly learn about more expensive (and likely higher-performing) regions [11, 33, 7]. Arguably, any task requiring exploration directed towards a specific goal should benefit from accurately modeling dependencies between function values that are (cheaply) observable and also relevant to the goal.

Despite its importance, the problem of estimating PPCs has received little attention in machine learning, and applications of PPCs currently rely on a handful of models for which they can be tractably computed, such as Gaussian processes. We believe this neglect is due to PPC estimates being much harder to evaluate than marginal uncertainty estimates. For evaluating marginal uncertainty, predictive log-likelihoods and reliability plots [10] are widely accepted tools. However, no analogous metrics exist for evaluating PPCs. In this paper, we introduce and validate three metrics for evaluating PPCs under different considerations and use them to benchmark various Bayesian models.

Conceptually, the most straightforward way to evaluate PPC estimators is by comparing their estimates to the true PPC values. We formalize this by measuring metacorrelations, i.e. the correlations between a model’s
Posterior Predictive Correlations

PPC estimates and the true values. Unfortunately, metacorrelations have limited applicability because the true PPC values are only defined for synthetic data drawn from a known distribution. However, they are very useful as a tool for validating other metrics which can be computed on real data.

If no single metric is completely satisfactory, it is often useful to measure both performance on a downstream task (which is practically relevant but indirect and expensive), as well as cheaper metrics which measure PPCs more directly. For the downstream task, we consider Transductive active learning (TAL), which chooses training data interactively to quickly improve the performance at known testing locations. Since the test locations are known, a good strategy is to query points that have high PPCs with test locations [21]. Hence, better PPC estimates ought to lead to better TAL performance. We validate the TAL task itself as a benchmark for PPC, and find that the TAL performances closely align to metacorrelations on synthetic datasets where both can be evaluated. We also find that TAL algorithms make better use of uncertainty than ordinary active learning algorithms.

Because TAL requires training each model multiple times and averaging over many training runs, it is too expensive to use as a routine metric to guide algorithm development. Hence, we need a proxy which is cheaper to evaluate. We observe that predictive joint log-likelihoods are insufficient as a metric because they depend too heavily on the marginal uncertainty. To disentangle the effect of predictive correlations, we introduce cross-normalized log-likelihood (XLL), which measures the predictive joint likelihoods of a model after its marginal predictions have been transformed to match a reference model. We find that XLL closely aligns with both metacorrelations and TAL performance, validating its usefulness as a routine metric for improving PPC estimation.

Using the three metrics we developed, we evaluate a variety of Gaussian process and Bayesian neural net models in terms of PPC estimation. We first observe that these three metrics align well with each other on synthetic data. Further, under these metrics, we consistently find that different models differ significantly in the accuracy of their PPC estimations. Our code for the experiments and the benchmarks is made available at:  

https://github.com/ssydasheng/predictive-correlation-benchmark.

2 Setup

Notations. Let \( \mathbf{x} \in \mathbb{R}^d \) be the input vector and \( f \) be the function. We focus on the single-output regression problem with Gaussian observation noises [8, 14, 20], i.e., the target \( y \) is a noisy observation of the function value, \( y = f(\mathbf{x}) + \epsilon, \epsilon \sim \mathcal{N}(0, \sigma^2_\epsilon(\mathbf{x})) \). We use \( \mathcal{D}_t \) and \( \mathcal{D}_v \) to denote the training set and the test set, respectively. For active learning problems, we also use \( \mathcal{D}_p \) to denote the pool set. Given the training set, for \( n \) locations \( \mathbf{X} \in \mathbb{R}^{n \times d} \), the predictive distribution for \( f(\mathbf{X})|\mathcal{D}_p \) is represented as a multivariate Gaussian \( \mathcal{N}(\mu(\mathbf{X}), \Sigma(\mathbf{X}, \mathbf{X})) \). Here \( \mu(\mathbf{X}) = [\mu_1, ..., \mu_n]^\top \in \mathbb{R}^n \) is the predictive mean and \( \Sigma(\mathbf{X}, \mathbf{X}) = [\Sigma(x, x')]_{x, x' \in \mathbf{X}} \) is the predictive covariance. Then the predictive variance \( \sigma^2_\mathbf{x} = \Sigma(\mathbf{x}, \mathbf{x}) \) and the predictive correlation \( \rho(\mathbf{x}, \mathbf{x}') = \Sigma(\mathbf{x}, \mathbf{x}')/(\sigma_\mathbf{x}\sigma_{\mathbf{x}'}), \) (1)

Models. Our experiments involve the Gaussian Processes (GP) [26] as well as sparse variational Gaussian Processes (SVGP) [13]. Given a kernel function \( k(\cdot, \cdot) \), a GP defines a prior \( p(f) \) over the functions. Specifically, for any finite set \( \{x_1, ..., x_n\} \), the function values \( f = (f(x_1), f(x_2), ..., f(x_n)) \) have a multivariate Gaussian distribution \( f \sim \mathcal{N}(0, \mathbf{K}(\mathbf{X}, \mathbf{X})) \), where \( \mathbf{K}(\mathbf{X}, \mathbf{X}) = [k(x_i, x_j)]_{i,j=1}^n \) is the kernel matrix. Moreover, if the likelihood is Gaussian, the posterior \( p(f|\mathcal{D}_t) \) can be computed in closed form.

We also include several models for Bayesian neural networks (BNNs), including Hamiltonian Monte Carlo (HMC)[24], Bayes-by-Backprop (BBB)[2], Noisy natural gradients (NNG)[37], Functional BNNs (FBNN)[32], Dropout BNNs [8], and Deep Ensemble [20]. Typically, a BNN defines a prior \( p(w) \) over the weights \( w \) of a neural network, and conducts inference for the posterior distribution \( p(w|\mathcal{D}_t) \). Furthermore, as pointed out by Neal [23], when the hidden sizes of BNNs grow to infinity, the BNNs prior becomes equivalent to a Gaussian Process prior.

Covariance Computation. The posterior predictive correlations (PPCs) can be obtained from the predictive covariance \( \Sigma(\mathbf{X}, \mathbf{X}) \) by \( \rho(\mathbf{x}, \mathbf{x}') = \Sigma(\mathbf{x}, \mathbf{x}')/(\sigma_\mathbf{x}\sigma_{\mathbf{x}'}) \). The covariance matrices of GP and FBNN are explicit. For example, for a GP with the kernel \( K \), the predictive covariance given training inputs \( \mathbf{X}_{tr} \) is,

\[ K(\mathbf{X}, \mathbf{X}) - K(\mathbf{X}, \mathbf{X}_{tr}) (K(\mathbf{X}_{tr}, \mathbf{X}_{tr}) + \sigma^2_\epsilon \mathbb{I})^{-1} K(\mathbf{X}_{tr}, \mathbf{X}) \]

Whereas for other models like BNNs, we need to draw samples to estimate the covariance matrix. For two points \( \mathbf{x}, \mathbf{x}' \), we empirically estimate the covariance by:

\[ \hat{\Sigma}(\mathbf{x}, \mathbf{x}') = \frac{1}{m} \sum_{i=1}^m (f_i(\mathbf{x}) - \hat{\mu}_\mathbf{x}) (f_i(\mathbf{x}') - \hat{\mu}_{\mathbf{x}'}), \]

where \( f_i \sim p(f|\mathcal{D}_v) \) are random function samples from the posterior, and \( \hat{\mu}_\mathbf{x} := \frac{1}{m} \sum_{i=1}^m f_i(\mathbf{x}) \).

1 Following the original papers of each model, the observation variance \( \sigma^2_\epsilon(\mathbf{x}) \) is constant for all models other than Deep Ensemble [20]. If Deep Ensemble uses a constant \( \sigma^2_\epsilon(\mathbf{x}) \), severe overfitting occurs in our experiments.
3 Benchmarking PPC Estimators

In this section, we describe our methodology for evaluating PPC estimators. We first introduce metacorrelations as a gold standard when the true PPC values can be obtained. Then, for practical settings, we introduce transductive active learning (TAL) as a downstream task for evaluating PPCs. Lastly, since TAL is expensive to run, we introduce Cross-Normalized Log Likelihood (XLL) as a more efficient proxy for TAL.

3.1 If we have an Oracle Model: metacorrelations

The most conceptually straightforward way to evaluate PPC estimators is by comparing their estimates to the ground truth PPC values. When the data generating distribution is known and has a simple form (e.g., synthetic data generated from a GP), we can compute ground truth PPCs; we refer to this as the oracle model. We can then benchmark models based on how closely their PPCs match those of the oracle. We quantify this by computing the Pearson correlation of the PPCs of the candidate model with the PPCs of the oracle model; we refer to these correlations of correlations as metacorrelations. Higher metacorrelations imply better PPC estimates.

3.2 Evaluating PPC Estimators on Transductive Active Learning

Active Learning improves sample efficiency by allowing the learning algorithm to choose training data interactively. In each iteration, we use the selection model to compute the acquisition function for choosing points, and we report the test performance of the prediction model trained on the accumulated training data. A diagram visualizing the active learning process is in Figure 1. For practical applications, the selection model is likely to be the same as the prediction model. However, we are interested in active learning as a downstream task for evaluating uncertainty estimates. Therefore, we fix a single prediction model across all conditions, and vary only the selection model, as this is the only part of the active learning algorithm that uses the uncertainty estimates.

The Total Information Gain (TIG) [21] is one acquisition function that has been used to benchmark predictive marginal uncertainty [4, 16, 15, 37, 9, 1]. TIG computes the informativeness of each point by measuring the mutual information between it and the model parameters \(w\): \(\text{TIG}(x) := I(y_x; w|D_{tr})\), where \(I\) represents mutual information. When the observation noise is assumed to be homoscedastic, choosing points by TIG is equivalent to choosing the points with the largest predictive variances. However, as pointed out by MacKay [21], TIG is problematic since it favors points a at the edges of the input space, hence leading to useless or even harmful queries. Moreover, as TIG relies only on marginal uncertainty, it cannot be used for benchmarking PPCs.

Instead we consider the Transductive Active Learning (TAL) setting for benchmarking PPCs. In TAL, a region of interest, such as the locations of test points or the distribution of the locations, is assumed to be known in advance. MacKay [21] presents the Marginal Information Gain (MIG) acquisition function, which measures the information gained about the region of interest by querying a new point. Supposing \(x_u\) is the point of interest\(^2\), \(\text{MIG}(x; x_u) := I(y_{x}; f(x_u)|D_{tr})\). We prefer MIG to TIG both because it’s been found to avoid TIG’s outlier problem [21] and because it makes direct use of PPCs.

In practice, selecting one point at a time, and retraining the models in between, is unreasonably expensive, so instead we would like to select batches of points. Naïvely selecting the set of points with highest scores often results in inefficiently selecting a cluster of nearby points which provide redundant information. To encourage diversity in a batch, we introduce the Batch Marginal Information Gain (BatchMIG) acquisition function, inspired by the BatchBALD algorithm [18] in active learning. For a batch \(x_{1:q}\),

\[
\text{BatchMIG}(x_{1:q}; x_u) := I(y_{x_{1:q}}; f(x_u)|D_{tr}).
\]

BatchMIG quantifies the amount of information carried by the selected batch. Though selecting the optimal batch \(x^*_{1:q}\) for BatchMIG is intractable, we adopt a greedy algorithm [18] to approximate it. We note that

\(^2\)For a test set \(\{x^*_1\}\), of interest, we adopt the mean marginal information gain [21].
To illustrate the differences between acquisition functions, we compare TIG, MIG and BatchMIG through a toy example shown in Figure 2. We observe that TIG chooses points nearby the boundary and MIG chooses redundant points close to each other, while BatchMIG chooses diverse points close to the test set. In consequence, the BatchMIG predictions after query match the ground truth better at the test locations.

3.3 Efficient Metrics beyond TAL

TAL is a useful downstream task for evaluating the PPCs, but running TAL requires training the model multiple times in succession, which is computationally expensive. Furthermore, as the PCCs are only part of a larger pipeline, it is rather indirect as a means of evaluating PPC estimators. Both factors limit its routine use as a guide for algorithmic development. In this section we introduce more direct and efficient metrics for PPC evaluation.

3.3.1 Joint Log-Likelihoods?

The log marginal likelihood $\sum_{i=1}^{n} \log p(y_i|x_i)$ is a widely used metric for evaluating predictive marginal uncertainty. By analogy, it would be natural to evaluate PPCs using the joint log-likelihood for a batch of points, $\log p(y_1, ..., y_b|x_1, ..., x_b)$. However, this is unsatisfying for two reasons:

Impact of predictive marginals. We’ve found the joint log-likelihood scores to be determined almost entirely by the marginal log-likelihood scores, with only a small dependence on the PPCs. Hence, in practice, they provide little new information beyond marginal log-likelihoods, as shown by Figure 9 in the appendix.

Uncorrelated random batches. The points in a random batch $\{(x_i, y_i)\}_{i=1}^{b}$ are almost uncorrelated because they usually scatter far away from each other.

For both reasons, joint log-likelihoods for random batches do not indicate the quality of PPCs.

3.3.2 Cross-Normalized Log Likelihood

As discussed in Section 3.3.1, joint log-likelihoods are appealing because they directly measure uncertainty and are efficient to evaluate, but they have the flaw that the scores are dominated by the marginal predictive distributions. To disentangle the effect of predictive marginals, we propose to substitute each model’s predictive means and variances with those of a reference model, whose predictive means and variances are believed to be reasonably good. Consequently, the joint log-likelihoods depend only on the PPCs but not on the predictive marginals. We refer to this criterion as cross-normalized log-likelihood (XLL). To compute the XLL of a batch of $b$ points $\{X, y\}$, we define

$$XLL(y|X, M, M_{ref}) = \log \mathcal{N}(y|\mu_{ref}, \text{diag}(\sigma_{ref})C_M \text{diag}(\sigma_{ref})).$$

Here, $M$ and $M_{ref}$ denote the candidate and reference model, $\mu_{ref}$ and $\sigma_{ref}$ are the predictive mean and variance given by $M_{ref}$, and $C_M$ is the predictive correlation matrix computed by $M$. Additionally, we can also rank each candidate model by the XLL under the same reference model, and we refer to the resulting criterion as cross-normalized log-likelihood rank (XLLR). Furthermore, to mitigate the problem that most pairs of randomly chosen locations have small correlations between the function values, we use the reference model to select top-correlated points (sorted by the absolute value of correlations) to form batches for evaluating the joint log-likelihoods.

Choosing the reference model. Intuitively, the ideal reference model would be the oracle, i.e. the true data generating distribution. That way, the XLL fully represents how the PPCs $C_M$ match the ground truth correlations $C_{gt}$. Although the predictive marginals of the oracle model are unknown in real-world problems, we can show that, as long as the reference marginals are nearly optimal, the XLL still reflects the quality of the PPCs. Informally, let $\text{LogDet}(C_{gt}, C_M)$ be the LogDet divergence, then

$$\text{LogDet}(C_{gt}, C_M) = -EX, y \text{XLL} + O\left(\frac{b^{3/2}}{\lambda} \sqrt{\xi}\right) + c,$$

where $c$ is a constant, $b$ is the batch size (set to 5 in our experiments). Here, $\xi$ denotes the KL-divergence between the reference marginal $(\mu_{ref}, \sigma_{ref})$ and the true marginal $(\mu_{gt}, \sigma_{gt})$. $\lambda$ denotes the smallest eigenvalue of $C_M$, which is usually not much smaller than 1 due to

![Figure 2: A visual comparison between BatchMIG, MIG and TIG. The red curve is the prediction after query, blue curve is the ground-truth, the • on top are the testing points, the ● are the training points, and the ● are the queried points of the corresponding acquisition function.](image)
to the observation noise. This results indicates that, with a nearly-optimal reference marginal, a larger XLL implies a smaller LogDet divergence. A more formal statement is given as Theorem 1 in the appendix.

To avoid favoring any particular model in our comparisons, we propose to iterate through every candidate model to serve once as the reference model. The reported XLL and XLLR values are averaged over all choices of the reference model. Still, one would like to validate that the results are not overly sensitive to the choice of reference model. In Figure 6, we observe that the XLLR values are consistent between choices of reference model (see Figure 6). We also observe that XLL and XLLR align well with the TAL performances as well as the oracle-based metacorrelations. Pseudocode for computing XLL and XLLR can be found in Algorithm 2 in the appendix.

4 Related Works

Benchmarking Uncertainty. There have been numerous attempts to reliably evaluate Bayesian models. The UCI regression benchmarks [15] are used extensively for evaluating Bayesian neural networks. Calibration metrics [10, 19] are used for testing whether the predictions are over- or under-confident. Snoek et al. [29] studied how the predictive accuracy and calibration are affected by a dataset shift. Researchers have also related the performance of various downstream tasks to the handling of uncertainty. Riquelme et al. [27] developed a contextual bandit benchmark which uses marginal uncertainty to balance exploitation and exploration. Diabetic retinopathy diagnosis [6] was also used for comparing uncertainty estimates. However, all of these benchmarks are on evaluating marginal uncertainty estimations.

Algorithms exploiting PPCs. Albeit less attention has been paid to PPCs, there are still several algorithms that exploit PPCs in their designs. In transductive active learning [21, 36], mutual information gain (MIG) improves the data efficiency by gathering points that have high PPCs with the test set. Similarly in Bayesian optimization, entropy search approaches [11, 15, 35] make the query to acquire the most information about the optimum of the underlying function. Knowledge gradient [7] makes the query so that the expected next step performance is maximized. Furthermore, in cost-sensitive Bayesian optimization where different queries might incur different costs, obtaining cheap queries that can indirectly acquire information about more expensive ones [11, 33]. Nevertheless, they only concern the usage of PPCs, but leave the question of which model is better at predicting PPCs unanswered.

5 Experiments

In this section, we first introduce the details of the setup in our experiments. Then, we conduct experiments on synthetic datasets to validate TAL and XLL(R) as metrics for evaluating PPCs. We show that both metrics correspond closely with our ground-truth metric, metacorrelations. Finally, we use our TAL and XLL(R) benchmarks to address the main question of this paper: how accurately can Bayesian regression models estimate PPCs in real-world datasets?

5.1 Experimental Setup

Synthetic Datasets. Some of our experiments were done on synthetic data drawn from a Gaussian process. Our motivation for this was twofold. Firstly, having access to the true data generating distribution allows us to compute metacorrelations with the oracle model. Secondly, the prior distributions for all models could be chosen to exactly or approximately match the true distribution.

We generated synthetic datasets using a Gaussian process whose kernel was obtained from the limiting distribution of infinitely wide Bayesian ReLU networks with one hidden layer [23]. Hence, the priors for the finite BNN weights could be chosen to approximately match this distribution. To generate a $d$-dimensional dataset, we sampled $5d, 500, 200$ points from the standard Normal distribution, as the training, test and pool sets, respectively. Then we sampled a random function $f$ from the oracle Gaussian process. The corresponding observed function value at $x$ is then $y = f(x) + \epsilon, \epsilon \sim \mathcal{N}(0, 0.01)$. For all models, we used 1,000 epochs for training and the true observation variance$^3$. All results were averaged over 50 datasets which are randomly sampled in this manner.

UCI Regression Datasets. We also conducted experiments using eight UCI regression datasets which have served as standard benchmarks for BNNs [15, 8, 37, 20, 32]. This includes five small datasets (Boston, Concrete, Energy, Wine and Yacht) and three large datasets (Kin8nm, Naval and Power_Plant). For all experiments, we used 20% of the entire dataset as the initial training set $D_{tr}, 20\%$ as the test set $D_{te}$, and the remaining 60% as the pool set $D_{pl}$ in active learning. In each active learning iteration, we selected a number of points from the pool set corresponding to 1% of the original dataset, so that the final prediction was made with 30% of the data (see Algorithm 1 in the Appendix). All experiments were run ten times with varied random seeds.

$^3$For the synthetic setting, each network in the Deep Ensemble only predicts the mean.
An Oracle Prediction Model for AL. The active learning performance is affected by two factors: the prediction model and the selection model. To disentangle their effects, we used the same ‘Oracle’ prediction model when varying the selection models. For the synthetic setting, we adopt the oracle GP model as the prediction model. For real-world data, the data generating distribution is unknown, so instead we fit the hyperparameters of a flexible GP kernel structure using the union of the training and pool set (rather than the much smaller initial training set). While the resulting distribution is probably not an exact match to the data generating distribution, it is likely superior to what the other methods can realistically obtain using the much smaller training set. For the flexible GP prior, we used the neural kernel network (NKN) kernel [31], a differentiable architecture representing sums and products of simpler kernels.

5.2 Is TAL a Suitable Downstream Task?

Active learning using TIG has been used as a benchmark for uncertainty in BNNs, but the validity of this benchmark is debatable, as improved uncertainty modeling has not been shown to consistently benefit active learning performance [15]. We now aim to determine if transductive active learning is a good downstream task for evaluating PPCs.

Is transductive information useful for active learning? We firstly conducted experiments on the synthetic datasets and UCI regression datasets to determine if transductive information is useful for active learning. Specifically, for the synthetic setting, we compared the active learning acquisition functions with all posterior predictive distributions obtained from the true GP prior that used to generate the data. For the realistic setting, we used the predictive distributions given by the ‘Oracle’ NKN model for computing these criteria. The results on synthetic datasets are reported in Figure 3 comparing TIG, MIG, BatchMIG and random selection. We first observe that the transductive criteria can in general achieve much better sample efficiency than the non-transductive ones. Furthermore, BatchMIG outperforms MIG, due to the increased diversity of points within the query batch. Both observations also hold on the UCI datasets; results can be found in Figure 10 (deferred to the Appendix to save space). Hence, it appears that transductive information can be exploited for sample efficiency.

Do more accurate PPCs enable better TAL performance? To investigate this question, we used the synthetic data so that ground-truth correlations and variances (from the Oracle) were available. We conducted active learning experiments using various models for query selection, each with multiple choices of hyperparameters. For each model, we evaluate the test log-likelihoods after one iteration. From Figure 4, we find that the BatchMIG performance is well aligned with the quality of PPCs, as measured by the metacorrelations. Hence, TAL is likely to be a good downstream task for evaluating PPC estimates. In contrast, the TIG performance appears to be unrelated to the quality of predictive variances, as expected. The contrast between (BatchMIG, Correlation) and (TIG, Variance) highlights the usefulness of TAL for benchmarking PPC estimators.

So far, we have demonstrated that (1) we can enjoy significantly better sample efficiency in active learning by incorporating the transductive information; and (2) the TAL performance is clearly and positively correlated with the quality of the PPCs. These evidence support that TAL is a suitable benchmark for evaluating PPCs.

Figure 3: Results on synthetic datasets using BatchMIG, MIG, TIG and random selection with the oracle model.

Figure 4: Left: Transductive active learning. BatchMIG performance vs. the quality of predictive correlations (metacorrelations). Right: Active learning. TIG performance vs. the quality of predictive variances. The BatchMIG performance and the metacorrelations are strongly correlated with a coefficient 0.762.
Table 1: The average XLLR for each model on UCI datasets. We use color red to highlight the best ones (within one standard error), and color blue for the worst ones (within one standard error).

| Dataset / Method | (SV)GP-RBF | BBB | NNG | HMC | FBNN | Dropout | Ensemble |
|------------------|------------|-----|-----|-----|------|---------|----------|
| Boston           | 2.53 (0.20) | 4.24 (0.18) | 2.57 (0.19) | 0.93 (0.13) | 3.20 (0.21) | 5.31 (0.13) | 2.21 (0.19) |
| Concrete         | 2.13 (0.14) | 4.50 (0.14) | 3.19 (0.17) | 1.49 (0.18) | 2.49 (0.14) | 5.81 (0.09) | 1.40 (0.21) |
| Energy           | 1.93 (0.17) | 4.07 (0.19) | 4.71 (0.15) | 1.59 (0.22) | 2.00 (0.17) | 4.57 (0.15) | 2.13 (0.21) |
| Wine             | 1.89 (0.18) | 4.61 (0.13) | 3.00 (0.16) | 2.14 (0.15) | 1.14 (0.21) | 5.59 (0.12) | 2.63 (0.18) |
| Yacht            | 1.90 (0.19) | 3.74 (0.14) | 3.49 (0.17) | 2.63 (0.27) | 1.76 (0.16) | 5.71 (0.11) | 1.77 (0.15) |
| Kin8nm           | 1.24 (0.11) | 4.37 (0.07) | 4.27 (0.16) | 1.44 (0.15) | 1.81 (0.15) | 5.14 (0.25) | 2.71 (0.18) |
| Naval            | 3.33 (0.14) | 5.54 (0.12) | 4.76 (0.08) | 2.40 (0.18) | 2.31 (0.07) | 1.43 (0.28) | 1.23 (0.12) |
| Power_plant      | 1.19 (0.11) | 3.81 (0.12) | 5.04 (0.15) | 1.93 (0.15) | 3.54 (0.11) | 4.33 (0.29) | 1.16 (0.15) |

Mean ↓ 2.02 4.36 3.88 1.82 2.28 4.73 1.91

Figure 5: Metacorrelations and BatchMIG performance versus XLL and XLLR.

5.3 Are XLL and XLLR Reliable?

We have discussed three evaluation metrics with different strengths and weaknesses: TAL, which is meaningful but expensive to evaluate; metacorrelation, which is fast and directly measures correlations but requires an oracle model; and XLL/XLLR, which are fast and do not require an oracle, but are less obviously related to the quality of PPCs. How well do they match each other? We measure all three metrics on the synthetic datasets. The results are shown in Figure 5. We observe that XLL/XLLR align well with both metacorrelations and the BatchMIG performance in TAL. This indicates that XLL/XLLR could be good proxy metrics for evaluating PPCs when TAL experiments are too expensive and the oracle model is not available.

In Figure 6, we plot the XLL and XLLR for each individual reference model. The x-axis indicates the reference model.

Figure 6: XLL and XLLR comparisons for each individual reference model. The x-axis indicates the reference model.

5.4 PPC Estimation for Real-World Datasets

The previous sections validated TAL and XLL/XLLR as metrics for the quality of PPCs. We now turn to our central question: how well can existing Bayesian models estimate PPCs? For this section, we consider both XLL/XLLR and TAL on UCI datasets. We first talk about the empirical results on both benchmarks, and then discuss how they connect to each other.

XLL and XLLR. Since the XLL cannot be directly averaged across different datasets, we instead average the ranks (i.e. XLLR values). This is analogous to the approach taken in the Bayesian bandits benchmark paper [27]. We present the empirical results of XLLR computed for different models in Table 1; results for XLL are given in Appendix E.3. Firstly, we observe that HMC, GP-RBF, FBNN and Ensemble usually outperform BBB, NNG and Dropout. Moreover, GP-RBF and FBNN both conduct function-space inference under the same prior, thus they perform similarly with FBNN being slightly worse. Besides, in contrast to HMC, BBB and NNG cannot estimate the correlations well, which highlights the importance of accurate posterior inference in BNNs. Finally, we found that Dropout estimates the PPCs the worst overall, even though we observed that Dropout performs well in log marginal likelihoods.

Transductive Active Learning. Figure 7a and 7b shows the TAL performance where various algorithms are used for query selection but predictions are made using the HMC and NKN model respectively. (We
omitted naval, since active learning appears to provide no benefit; see Figure 10 in the appendix). Clearly, the model used for query selection has a large effect even when the prediction model is fixed. In general, even though the prediction models are different, the best-performing models were typically the HMC BNN, FBNN, GP-RBF and Ensemble. The variational BNNs and dropout performed less well, consistent with the results on the synthetic datasets (see Figure 12 in the Appendix).

Using just a single model for prediction gives an incomplete picture of the query selection performance of different methods, as some methods might happen to be better matches to the NKN model or the HMC. Table 4 (in the Appendix) shows the results of mixing and matching a wider variety of prediction and selection models. In general, we observe that regardless of which model is used for prediction, the best results are obtained when queries are selected using the most accurate models, rather than the same models used for prediction. We believe the TAL experiment indicates that high-quality posterior distributions are useful for data selection, above and beyond the benefits from making better predictions from a fixed training set.

**Discussion.** Overall, the XLL(R) and TAL metrics present a consistent picture for evaluating PPCs. First, on both benchmarks, HMC, GP-RBF, FBNN and Ensemble are in general better than BBB, NNG and Dropout. Second, for the boston, concrete and energy datasets, HMC performed the best according to both metrics. Lastly, for the wine and yacht datasets, FBNN outperforms GP-RBF, reflected again by the XLL(R) and TAL performance.

### 6 Conclusion

In this work, we developed three evaluation metrics for evaluating PPC estimators: metacorrelations, transductive active learning, and cross-normalized log likelihoods. We used synthetic data to validate that the three metrics are suitable for evaluating PPC estimators. Finally, we introduced benchmarks on real-world data to examine how accurately Bayesian models can estimate PPCs.
Acknowledgements

We thank Saminul Haque, Guodong Zhang, Wenyuan Zeng and Cong Han Lim for their insightful comments and discussions on this project. We also thank the Vector Institute for providing the scientific computing resources. This research project has been supported by LG Electronics. SS was supported by the Connaught Fellowship. RG acknowledges support from the CIFAR Canadian AI Chairs program.

References

[1] William H Beluch, Tim Genewein, Andreas Nürnberger, and Jan M Köhler. The power of ensembles for active learning in image classification. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 9368–9377, 2018.

[2] Charles Blundell, Julien Cornebise, Koray Kavukcuoglu, and Daan Wierstra. Weight uncertainty in neural network. In International Conference on Machine Learning, pages 1613–1622, 2015.

[3] Adam D Cobb, Stephen J Roberts, and Yarin Gal. Loss-calibrated approximate inference in bayesian neural networks. arXiv preprint arXiv:1805.03901, 2018.

[4] David A Cohn, Zoubin Ghahramani, and Michael I Jordan. Active learning with statistical models. Journal of artificial intelligence research, 4:129–145, 1996.

[5] Bernard Delyon, Marc Lavielle, and Eric Moulines. Convergence of a stochastic approximation version of the em algorithm. Annals of statistics, pages 94–128, 1999.

[6] Angelos Filos, Sebastian Farquhar, Aidan N. Gomez, Tim G. J. Rudner, Zachary Kenton, Lewis Smith, Milad Alizadeh, Arnoud de Kroon, and Yarin Gal. Benchmarking bayesian deep learning with diabetic retinopathy diagnosis. https://github.com/DAOML/bdl-benchmarks, 2019.

[7] Peter Frazier, Warren Powell, and Savas Dayanik. The knowledge-gradient policy for correlated normal beliefs. INFORMS journal on Computing, 21(4):599–613, 2009.

[8] Yarin Gal and Zoubin Ghahramani. Dropout as a Bayesian approximation: Representing model uncertainty in deep learning. In International Conference on Machine Learning, pages 1050–1059, 2016.

[9] Yarin Gal, Riashat Islam, and Zoubin Ghahramani. Deep bayesian active learning with image data. In International Conference on Machine Learning, pages 1183–1192, 2017.

[10] Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. On calibration of modern neural networks. In International Conference on Machine Learning, pages 1321–1330, 2017.

[11] Philipp Hennig and Christian J Schuler. Entropy search for information-efficient global optimization. Journal of Machine Learning Research, 13(Jun):1809–1837, 2012.

[12] James Hensman, Nicolo Fusi, and Neil D Lawrence. Gaussian processes for big data. In Uncertainty in Artificial Intelligence, page 282, 2013.

[13] James Hensman, Alexander Matthews, and Zoubin Ghahramani. Scalable variational Gaussian process classification. In Artificial Intelligence and Statistics, pages 351–360, 2015.

[14] 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, 2015.

[15] José Miguel Hernández-Lobato, Matthew W Hoffman, and Zoubin Ghahramani. Predictive entropy search for efficient global optimization of black-box functions. In Advances in Neural Information Processing Systems, pages 918–926, 2014.

[16] Neil Houlsby, Ferenc Huszár, Zoubin Ghahramani, and Mátyás Lengyel. Bayesian active learning for classification and preference learning. arXiv preprint arXiv:1112.5745, 2011.

[17] Diederik P Kingma and Max Welling. Auto-encoding variational Bayes. arXiv preprint arXiv:1312.6114, 2013.

[18] Andreas Kirsch, Joost van Amersfoort, and Yarin Gal. Batchbald: Efficient and diverse batch acquisition for deep bayesian active learning, 2019.

[19] Volodymyr Kuleshov, Nathan Fenner, and Stefano Ermon. Accurate uncertainties for deep learning using calibrated regression. In International Conference on Machine Learning, pages 2796–2804, 2018.

[20] Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. In Advances in Neural Information Processing Systems, pages 6402–6413, 2017.

[21] David JC MacKay. Information-based objective functions for active data selection. Neural computation, 4(4):590–604, 1992.
[22] James Martens and Roger Grosse. Optimizing neural networks with kronecker-factored approximate curvature. In *International conference on machine learning*, pages 2408–2417, 2015.

[23] Radford M Neal. *Bayesian Learning for Neural Networks*. PhD thesis, University of Toronto, 1995.

[24] Radford M Neal et al. MCMC using Hamiltonian dynamics. *Handbook of Markov Chain Monte Carlo*, 2(11), 2011.

[25] Ali Rahimi and Benjamin Recht. Random features for large-scale kernel machines. In *Advances in Neural Information Processing Systems*, pages 1177–1184, 2008.

[26] Carl Edward Rasmussen and Christopher KI Williams. *Gaussian Processes for Machine Learning*. MIT Press, 2006.

[27] Carlos Riquelme, George Tucker, and Jasper Snoek. Deep Bayesian bandits showdown: An empirical comparison of Bayesian deep networks for Thompson sampling. In *International Conference on Learning Representations*, 2018.

[28] Jiaxin Shi, Mohammad Emteiyaz Khan, and Jun Zhu. Scalable training of inference networks for gaussian-process models. *arXiv preprint arXiv:1905.10969*, 2019.

[29] Jasper Snoek, Yaniv Ovadia, Emily Fertig, Balaji Lakshminarayanan, Sebastian Nowozin, D Sculley, Joshua Dillon, Jie Ren, and Zachary Nado. Can you trust your model’s uncertainty? evaluating predictive uncertainty under dataset shift. In *Advances in Neural Information Processing Systems*, pages 13969–13980, 2019.

[30] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: a simple way to prevent neural networks from overfitting. *The journal of machine learning research*, 15(1):1929–1958, 2014.

[31] Shengyang Sun, Guodong Zhang, Chaoqi Wang, Wenyuan Zeng, Jiaman Li, and Roger Grosse. Differentiable compositional kernel learning for Gaussian processes. In *International Conference on Machine Learning*, pages 4828–4837, 2018.

[32] Shengyang Sun, Guodong Zhang, Jiaxin Shi, and Roger Grosse. Functional variational bayesian neural networks. *arXiv preprint arXiv:1903.05779*, 2019.

[33] Kevin Swersky, Jasper Snoek, and Ryan P Adams. Multi-task bayesian optimization. In *Advances in neural information processing systems*, pages 2004–2012, 2013.

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

[35] Zi Wang and Stefanie Jegelka. Max-value entropy search for efficient Bayesian optimization. In *International Conference on Machine Learning*, pages 3627–3635, 2017.

[36] Kai Yu, Jinbo Bi, and Volker Tresp. Active learning via transductive experimental design. In *Proceedings of the 23rd international conference on Machine learning*, pages 1081–1088. ACM, 2006.

[37] Guodong Zhang, Shengyang Sun, David Duvenaud, and Roger Grosse. Noisy natural gradient as variational inference. In *International Conference on Machine Learning*, pages 5852–5861, 2018.
A Notation

| Notation | Description |
|----------|-------------|
| $x$      | the input vector |
| $\{x\}$ | a set of input vectors |
| $f$      | the underlying function |
| $y$      | the observed value of the function at a given location $x$ |
| $\sigma_{e_x}$ | the standard variance of the observation noise at $x$ |
| $\sigma_e$ | the standard variance of a homoscedastic observation noise |
| $\sigma_{\epsilon_x}$ | a vector, whose element is corresponding to the standard variance of the observation noise at each $x \in \{x\}$ |
| $\eta$  | the prior variance for BNNs |

B The Pseudocodes

**Algorithm 1** A procedure of (Transductive) Active Learning. We use red and blue to show the difference between active learning and TAL. TIG and MIG can be replaced by any other acquisition functions.

Require: Selection Model: $M^s$; Prediction Model: $M^p$.

Require: Datasets: $D_{tr} = \{X_{tr}, y_{tr}\}$, $D_{te} = \{X_{te}, y_{te}\}$, $D_{pl} = \{X_{pl}, y_{pl}\}$.

Require: Total active learning iterations: $T$; #Queried samples per iteration: $m$.

1: $\mathcal{R} = \emptyset$.
2: for $t = 1$ to $T$ do
3: Train $M^p, M^s$ on $D_{tr}$ until convergence.
4: Test $M^s$ over $D_{te}$ and put the result to $\mathcal{R}$.
5: InfoG = TIG($X_{pl}, M^p$) or InfoG = MIG($X_{pl}, X_{te}, M^p$).
6: Sort InfoG in descending order and retrieve top $m$ samples from $D_{pl}$ as $D_{qc}$.
7: $D_{tr} \leftarrow D_{tr} \cup D_{qc}$; $D_{pl} \leftarrow D_{pl} \setminus D_{qc}$.
8: $t \leftarrow t + 1$.
9: end for
10: return $\mathcal{R}, M^p, D_{tr}$.

C Information Gains for Active Learning

In this section, we first introduce three types of information gains, Total Information Gain (TIG), Marginal Information Gain (MIG), Batched Marginal Information Gain (BatchMIG), and present their analytical forms for Gaussian predictive distribution. Then, we provide a greedy approximation for computing the optimal batch corresponding to BatchMIG.

C.1 Three Types of Information Gains

We firstly specify the analytic expressions for computing the information gain acquisition functions:

**Total Information Gain (TIG)**, measures the mutual information between the queried point $x$ and the model parameters $w$,

$$\text{TIG}(x) := I(y_x; w|D_{tr}) = \frac{1}{2} \log \left( 1 + \frac{\sigma_x^2}{\sigma_{\epsilon_x}^2} \right),$$  \hspace{1cm} (4)

**Marginal Information Gain (MIG)**, measures the mutual information between the queried point $x$ and a point $x_u$ of interest,

$$\text{MIG}(x; x_u) := I(y_x; f(x_u)|D_{tr}) = -\frac{1}{2} \log \left( 1 - \frac{\text{Cov}(x, x_u)^2}{\sigma_x^2 \sigma_{x_u}^2} \frac{\sigma_x^2}{\sigma_{\epsilon_x}^2} \sigma_{x_u}^2 \right),$$  \hspace{1cm} (5)
Algorithm 2 Computing XLL and XLLR.

Require: Model Predictions \( \{(\mu_i, \Sigma_i)\}_{i=1}^m \); Test set \( D_{te} \); Batch size \( b \)

1: for \( j = 1 \) to \( m \) do
   \( \triangleright \) Reference Model
2:     \( \triangleright \) Normalize Predictive Marginals
3:     for \( i = 1 \) to \( m \) do
4:         \( D_0^i = \sqrt{\text{diag}(\Sigma_j)} / \sqrt{\text{diag}(\Sigma_i)} \).
5:         \( \bar{\mu}_i = \mu_j, \quad \bar{\Sigma}_i = D_0^i \Sigma_i D_0^i \).
6:     end for
7: \( T' = \{ \} \).
8: for \( (x, y) \in D_{te} \) do
9:   \( \triangleright \) Build Test Batches
10:     Top correlated points \( B_x := \{(x_k, y_k)\}_{k=1}^b \); Add \( B_x \) to \( T' \).
11: end for
12: for \( i = 1 \) to \( m \) do
13:     \( \triangleright \) Compute Log Joints
14:     \( \text{lld}_i^j = \frac{1}{|T'|} \sum_{B \in T'} \log \mathcal{N}(B|\bar{\mu}_i, \bar{\Sigma}_i) \).
15: end for
16: \{rank\}_i^m from sorting \( \{\text{lld}_i^j\}_i^m \).
17: end for
18: return \( \{\text{lld}_i\}_i^m \) and \( \{\text{rank}\}_i^m \).

Batched Marginal Information Gain (BatchMIG), measures the mutual information between a batch of queried points \( x_{1:q} \) and the point \( x_u \) of interest,

\[
\text{BatchMIG}(x_{1:q}; x_u) := \mathbb{I}(y_{x_{1:q}}; f(x_u)|D_{tr})
= -\frac{1}{2} \log \left( 1 - \frac{\text{Cov}(x_{1:q}, x_u) \text{T} \left( \text{Cov}(x_{1:q}, x_{1:q}) + \sigma_{\text{err}} \right)^{-1} \text{Cov}(x_{1:q}, x_u)}{\sigma_{x_u}^2} \right), \tag{6}
\]

Again for MIG and BatchMIG, assuming that we are interested at a set of points \( \{x_u^i\}_{i=1}^f \), as recommended in MacKay [21], we adopt the mean marginal information gains: \( \frac{1}{f} \sum_{i=1}^f \text{MIG}(x; x_u^i) \) and \( \frac{1}{f} \sum_{i=1}^f \text{BatchMIG}(x; x_u^i) \).
A Greedy Approximation of the Optimal Batch

Algorithm 3 BatchBald [18]: a greedy approximation of the optimal batch.

Require: Model $\mathcal{M}$, Points of interest $\mathcal{I}$, Query Batch Size $q$

Require: The information gain acquisition function IG.

1: $A \leftarrow \emptyset$
2: for $i = 1$ to $q$ do
3: \hspace{1em} $x^* \in \arg\min_x IG(x, \mathcal{M}, \mathcal{I})$
4: \hspace{1em} $A = A \cup \{x^*\}$
5: end for
6: return $A$.

In practice we will usually query a batch of points at each iteration for efficiency. For TIG and MIG, selecting a batch corresponds to selecting the points with highest information gains, correspondingly. For BatchMIG, although extending the information gain acquisition functions from the single-point scenario to the batch scenario is straightforward, solving for the optimal batch requires a combinatorial explosion of computations. Following [18], we adopt a greedy approximation of the optimal batch, which is specified in Alg 3.

D Experimental Details

D.1 Models and Algorithms

GP/SVGP. A Gaussian process (GP) defines a distribution $p(f)$ over functions $\mathcal{X} \rightarrow \mathcal{R}$ for some domain $\mathcal{X}$. For any finite set $\{x_1, \ldots, x_n\} \subset \mathcal{X}$, the function values $f = (f(x_1), f(x_2), \ldots, f(x_n))$ have a multivariate Gaussian distribution $f \sim \mathcal{N}\left(\mu, \mathbf{K}_{XX}\right)$, where $\mu := (\mu(x_1), \ldots, \mu(x_n))^\top$ using the mean function $\mu(\cdot)$, and $\mathbf{K}_{XX}$ denotes the matrix $[k(x_i, x_j)]_{i,j}$ using the kernel function $k(\cdot, \cdot)$. For Gaussian likelihoods with variance $\epsilon^2$, we can make predictions $p(y_*|x_*, D_{tr})$ in closed form:

$$p(y_*|x_*, D_{tr}) = \mathcal{N}(\mathbf{K}_{*X} (\mathbf{K}_{XX} + \epsilon^2 \mathbf{I})^{-1} y, \Sigma), \quad \Sigma := \mathbf{K}_{**} - \mathbf{K}_{*X} (\mathbf{K}_{XX} + \epsilon^2 \mathbf{I})^{-1} \mathbf{K}_{X*} + \epsilon^2.$$  

GP performs exact posterior inference from the prior and training data. But its computational cost scales cubically with the training size. Thus, for large scale problems, the stochastic variational Gaussian Processes (SVGP) [34, 12] are usually adopted.

Bayesian neural networks (BNNs). Given a $L$-layer neural network, the weights $\mathbf{w} = \{\mathbf{w}_l\}_{l=1}^L$ is the collection of $V_l \times (V_{l-1} + 1)$ weight matrices in each layer, where $V_l$ is the number of units in the $l$-th layer and the $+1$ accounts for the biases. Assuming the pre-activations and activations in the $l$-th layer are $\mathbf{z}_l$ and $\mathbf{a}_l$, we have $\mathbf{a}_l = \mathbf{w}_l \mathbf{z}_{l-1}/\sqrt{V_{l-1} + 1}$, where $\sqrt{V_{l-1} + 1}$ keeps the scales of $\mathbf{a}_l$ independent of the number of input neurons. BNNs specify a prior distribution $p(\mathbf{w})$ over the weights and perform posterior inference for uncertainty estimation. In our experiments, we use $p(\mathbf{w}) = \mathcal{N}(\mathbf{0}, \eta^2)$, where the scalar $\eta$ is the prior variance.

HMC BNNs. The posterior of weights $p(\mathbf{w}|D)$ is intractable in BNNs. Hamiltonian Monte Carlo (HMC) [24] is one Markov Chain Monte Carlo (MCMC) method that uses the gradient of the target distribution to give better proposals than a random walk MCMC, which makes it more tractable in higher dimensional parameter spaces, such as in neural networks. However, Neal et al. [24] points out that jointly sampling weights and the prior variance $\eta$ makes it difficult to mix, then they suggest to perform Gibbs sampling for prior variances combined with HMC for weights. Instead, we adopt Monte-Carlo EM [5] for optimizing hyperparameters: the prior variance $\eta$ and the observation variance $\epsilon^2$. Specifically, we run HMC for the weights and directly optimize hyperparameters using maximum-likelihood under the current weights.

Bayes By Backprop (BBB). Variational methods provide another way to resolve the intractable posterior problem. It fits an approximate posterior $q(\mathbf{w})$ to maximize the evidence lower bound:

$$L_q = \mathbb{E}_q[\log p(D_{tr}|\mathbf{w})] - KL(q(\mathbf{w})||p(\mathbf{w})).$$

Specifically, Bayes-by-backprop (BBB) [2] uses a fully factorized Gaussian $\mathcal{N}(\mathbf{\mu}, \mathbf{\sigma}^2)$ to approximate the posterior. Using the reparameterization trick [17], the gradients towards $(\mathbf{\mu}, \mathbf{\sigma}^2)$ can be computed by backpropagation, and then be used for updates. We fix the prior variance $\eta = 1$ because we found it is not beneficial when optimized.
Noisy Natural Gradients (NNG). Among the variational Bayesian families, noisy natural gradient (NNG) [37] is an efficient method to fit multivariate Gaussian posteriors by adding adaptive weight noise to ordinary natural gradient updates. Assuming $q(w) = \mathcal{N}(\mu, \Sigma)$ and $Dw = \nabla_w \log p(y|x, w)$, then the update rules are:

$$F \leftarrow (1 - \beta)F + \beta \left( DwDw^T \right) \quad \text{and} \quad \mu \leftarrow \mu + \alpha \left( F + \frac{1}{N\eta}I \right)^{-1} \left( Dw - \frac{1}{N\eta}w \right).$$  \hspace{1cm} (7)

In particular, when using a Kronecker-factored approximation [22] for the fisher matrix $F$, NNG is equivalent to imposing a matrix-variate Gaussian distribution for the variational posterior $q(w)$. Similarly to BBB, we fix the prior variance $\eta = 1$ because we found it is not beneficial when optimized.

Functional BNNs (FBNN) The weights usually have a complicated relationship with the corresponding function, making it difficult to specify a meaningful prior over weights. Functional BNNs [32] proposes to directly apply priors over the function and perform functional variational inference. For estimating the functional KL divergence $KL(q(f)||p(f))$, they use mini-batch marginal KL divergences, resulting in,

$$\mathcal{L}_q = \frac{1}{|D_s|} \sum_{(x,y)} E_q[\log p(y|f(x))] - \lambda KL \left( q \left( \left[ f_{D_s} \right] \right) || p \left( \left[ f_{M} \right] \right) \right),$$

where $f_{D_s}$ and $f_{M}$ are the function values on locations within a random mini-batch and $M$ random locations from a heuristic distribution, respectively. Across all experiments, we set $\lambda = 1$ and we use a GP-RBF prior. Following Shi et al. [28], we use a RBF random feature network [25] with one hidden layer for the posterior $q(f)$, where the first layer is deterministically trainable and the second layer is a Bayesian linear regression layer applied on the features from the first layer. In consequence, the posterior distribution $q$ becomes Gaussian with explicit means and covariances. We set the heuristic distribution for $f_{M}$ as Gaussian distributions centered at random training locations with variances being $s^2/2$, where $s$ is the lengthscale of the prior kernel.

Dropout BNNs. Dropout [30] is a technique employed in training neural networks wherein the output units at each layer has a probability of $p$ to be discarded independently. Typically, the trained network is made deterministic by scaling the output of each layer by $p$ rather than randomly zeroing it out. For obtaining uncertainty, Gal and Ghahramani [8] proposes to keep dropout stochastic in test time and use many forward passes to compute the means and variances. It is also shown that dropout can be seen as implicitly optimizing a variational objective. Deep Ensemble. Deep Ensemble [20] consists of training multiple networks concurrently towards the same objective. Then the predictions of all networks can be aggregated to compute the predictive mean and variance. One distinction to note is that the networks used in the ensemble output not only predicted mean $\mu_i(x)$, but also predicted aleatoric variance $\sigma_i^2(x)$. Note that $\sigma_i^2(x)$ represents aleatoric variance, and $\mu_i(x)$ can be seen as a random function sample. Therefore, this also yields the epistemic uncertainty through the variance of $\{\mu_i(x)\}_{i=1}^m$. 

Posterior Predictive Correlations
D.2 Hyperparameters

We use the standard regression task for tuning hyperparameters with respect to each method and each dataset. Specifically, we split the dataset into train (60%), valid (20%) and test (20%). Across 10 different runs, we use the same validation set but split train/test randomly. Finally the averaged validation log likelihood will be used for selecting the hyperparameters. A list of details about hyperparameters is shown in Table 3.

With the tuned hyperparameters, we conduct transductive active learning and compute the XN-LLDR metrics. To avoid that the test set being used for tuning hyper-parameters, we make sure the randomly selected test set is disjoint with the validation set for hyperparameter tuning.

| Methods | Hyperparameters to tune |
|---------|-------------------------|
| BBB     | lr: [0.001, 0.003, 0.01], hidden units: [50, 400], #eval_cov_samples: [100, 700, 5000] |
| NNG     | lr: [0.001, 0.003, 0.01], hidden units: [50, 400] |
| HMC     | lr: [0.001, 0.003, 0.01], hidden units: [50, 400] |
| FBNN    | lr: [0.001, 0.003, 0.01], number of random measurement points [5, 20, 100], hidden units: [50, 400] |
| Dropout | lr: [0.001, 0.003, 0.01], hidden units: [50, 400], Dropout Rate: [0.0025, 0.01, 0.05], Observation variance: [0.005, 0.025, 0.125] |
| Ensemble| lr: [0.001, 0.003, 0.01], hidden units: [50, 400] |

| Methods | Other Settings |
|---------|----------------|
| (SV)GP  | Optimizer=Adam, lr=0.003, epochs=10,000, batch_size=min(5,000, #training data), length_scale are initialized with k-means on training data, ARD=True, min_obsvar=1e-5 (except for Wine dataset, we use min_obsvar = 1e-8); For large datasets, we adopt SVGP with 1,000 inducing points; For (SV)GP-NKN, we adopt the same NKN as in Sun et al. [31] and epochs=5,000. |
| BBB     | Optimizer=Adam, epochs=10,000, batch_size=100, #training_particles=10, #test_particles=5,000. |
| NNG     | Optimizer=NG-KFAC(damping=1e-5, ema_cov_decay=0.999), epochs=10,000, lr decay by a factor 0.1 every 5000 epochs, #training_particles=10, #test_particles=5,000, #eval_cov_samples=5000. |
| HMC     | #chains = 10, burnin=5,000 for small datasets and 15,000 for larger ones, step_size starts at 0.01 but is adapted according to the acceptance rate, #leap_frog_steps=5; We select one particle every 100 samples after burnin until we collected 100 samples in each chain, which results at 1,000 samples for testing and computing the covariance. We use Adam Optimizer for optimizing the prior hyperparameters $\eta, \xi$ every 10 HMC steps. |
| FBNN    | Optimizer=Adam, epochs=10,000, batch_size=#training data for small datasets and 900 for larger datasets in order to match the computation complexity of SVGP. The network has 400 hidden units with cosine activations. |
| Dropout | Optimizer=Adam, epochs=10,000, batch_size=100. We use 5,000 samples for test and computing the covariance. L2 regularization with $10^{-4} \ast (1 - \text{dropout}_\text{rate})/(2 \ast N \ast \xi)$. |
| Ensemble| Optimizer=Adam, epochs=10,000, batch_size=100, #networks=100. |
E Additional Results

E.1 Log Joints versus Log Marginals

We visualize the scatter plot of the joint log-likelihoods and the marginal log-likelihoods in Figure 9. We observe that the joint log-likelihood is positively correlated with the marginal log-likelihood.

![Figure 9: Scatters of log joint likelihoods versus log marginal likelihoods where each point represents one method. The log joints are computed over random batches with 5 points.](image)

E.2 Average Rank in TAL (Table 4)

Table 4 shows the results of mixing and matching a wider variety of training and selection models. In general, we observe that regardless of which model is used for training, the best results are obtained when queries are selected using the most accurate models, rather than the same models used for training. We believe this experiment directly indicates that high-quality posterior predictive distributions are useful for data selection, above and beyond the benefits from making better predictions from a fixed training set.

![Table 4: Average rank of each method’s LLD and RMSE on TAL at the last iteration with different prediction models. We use red to highlight the best ones, and blue for the worst ones.](image)

E.3 XLL on UCI datasets (Table 5)

Table 5: The average values of XLL.

| Dataset/Method | (SV)GP-RBF | BBB | NNG | HMC | FBNN | Dropout | Ensemble |
|----------------|------------|-----|-----|-----|------|---------|----------|
| Boston         | -3.217 (0.134) | -3.316 (0.156) | -3.202 (0.133) | -3.177 (0.133) | -3.237 (0.138) | -3.456 (0.160) | -3.202 (0.139) |
| Concrete       | -3.342 (0.015) | -3.394 (0.018) | -3.351 (0.016) | -3.336 (0.015) | -3.344 (0.015) | -3.615 (0.029) | -3.340 (0.015) |
| Energy         | -1.382 (0.065) | -1.430 (0.068) | -1.437 (0.068) | -1.378 (0.064) | -1.384 (0.064) | -1.434 (0.067) | -1.386 (0.065) |
| Wine           | -2.062 (0.115) | -2.112 (0.108) | -2.074 (0.102) | -2.126 (0.118) | -2.011 (0.103) | -2.014 (0.096) | -1.998 (0.102) |
| Yacht          | 0.902 (0.031)  | 0.892 (0.031)  | 0.890 (0.032)  | 0.902 (0.031)  | 0.901 (0.031)  | 0.796 (0.032)  | 0.897 (0.031)  |
| Kin8nm         | 6.853 (0.172)  | 6.795 (0.176)  | 6.811 (0.175)  | 6.882 (0.166)  | 6.870 (0.171)  | 6.971 (0.163)  | 6.920 (0.173)  |
| Naval          | 2.793 (0.015)  | 2.812 (0.018)  | 2.821 (0.019)  | 2.801 (0.017)  | 2.806 (0.017)  | 2.828 (0.015)  | 2.796 (0.016)  |

We use red to highlight the best ones, and blue for the worst ones.
E.4 More Comparisons between Data Acquisition Functions

We present here the results using different data acquisition functions on synthetic datasets and on UCI datasets, where we can observe that TAL acquisition functions consistently outperform other criterions.

![Graph showing RMSE and LLD performance of TAL on UCI datasets](image)

**Figure 10:** UCI dataset: RMSE and LLD performance of TAL with the pre-trained NKN kernel (Oracle).

![Graph showing RMSE and LLD performance of TAL on synthetic datasets](image)

**Figure 11:** Synthetic dataset: RMSE and LLD performance of TAL with the Oracle model.
E.5 TAL Results of Different Models on Synthetic Datasets

To evaluate how each model performs on TAL, we compare them with BatchMIG and TIG on the synthetic datasets. The results are presented in Figure 12 and Figure 13 respectively.

Figure 12: BatchMIG on toy datasets, with fixed observation variance.
Figure 13: TIG on toy datasets, with fixed observation variance.
F  A Theoretical Connection between Log Likelihoods and Predictive Correlations

To understand why XLL directly reflects the accuracy of the correlations, consider the following distributions:

\[ p(y|X) = \mathcal{N}(y|\mu_{gen}, \text{diag}(\sigma_{gen}^2)C_{gen}), \]
\[ q(y|X) = \mathcal{N}(y|\mu_{ref}, \text{diag}(\sigma_{ref}^2)C), \]
\[ p_m(y|X) = \mathcal{N}(y|\mu_{gen}, \sigma_{gen}^2), \]
\[ q_m(y|X) = \mathcal{N}(y|\mu_{ref}, \sigma_{ref}^2), \]
\[ p_c(y|X) = \mathcal{N}(y|0, C_{gen}), \]
\[ q_c(y|X) = \mathcal{N}(y|0, C) \]

(8)

where \( p(y|X) \) is the data generating distribution, and \( \mu_{gen}, \sigma_{gen}^2 \) and \( C_{gen} \) are the ground-truth mean, variance and correlations respectively. Observe that \( -KL(p||q) \) is the quantity that XLL is approximating using samples (up to a constant), while \( KL(p_c||q_c) \) is a measure of dissimilarity between the correlation matrices or the LogDet divergence between two positive semidefinite matrices \( C_{gen} \) and \( C \). We now show that, if the reference marginals (i.e., \( \mu_{ref}, \sigma_{ref} \)) are close to the ground truth marginals, then \( KL(p||q) \) approximately equals \( KL(p_c||q_c) \). Hence, XLL can be seen as a measure of the accuracy of the predictive correlations.

**Theorem 1.** Let the predictive distributions be defined above, and let \( b \) be the number of points for evaluation, \( \lambda \) denote the smallest eigenvalue of \( C \) and \( \xi = KL(p_m||q_m) \). If \( \xi \ll 1 \), then we have:

\[ |KL(p||q) - KL(p_c||q_c)| = O\left(\frac{b^{3/2}}{\lambda \sqrt{\xi}}\right). \]

(9)

**Remark 1.** Because the expected joint log-likelihood \( E_{p(y|X)} \log q(y|X) = E_{p(y|X)} \log p(y|X) - KL(p||q) \), this theorem illustrates that, for nearly-optimal reference marginals, the expected joint log-likelihood reflects the quality of the predictive correlations. This validates the reliability of XLL.

**Remark 2.** In practice, the predictive covariance is \( \Sigma + \sigma_n^2 I \), where \( \sigma_n^2 \) is the variance of the modeled observation noise and \( \Sigma \) is the predictive covariance for the underlying function. In general, \( \sigma_n^2 \) and the predictive variances in \( \Sigma \) are in the same order of magnitude. Therefore, the smallest eigenvalue \( \lambda \) of the correlation matrix \( C \) is not much smaller than 1. Furthermore, \( b \) is small because we evaluate XLL and XLLR over mini-batches (\( b = 5 \) in our experiments).

As suggested by the theorem, the ideal reference model would be the oracle, i.e. the true data generating distribution. However, in practice we only have access to models which imperfectly match the distribution. Fortunately, we found that the relative order of XLL values do not appear to be overly sensitive to the choice of reference model. Therefore, to further avoid favoring any particular model as the reference, we propose to iterate through every candidate model to act as the reference model once. Then, for each candidate model, we average its XLL or XLLR across all reference models. Empirically, we found that XLL and XLLR align well with the corresponding performance in TAL benchmarks as well as the oracle-based meta-correlations. In below, we provide the proof of Theorem 1:
Proof. We first define:
\[
d := \frac{\mu_{\text{gen}} - \mu_{\text{ref}}}{\sigma_{\text{ref}}} , \quad r := \frac{\sigma_{\text{gen}}}{\sigma_{\text{ref}}} ,
\]
and let \(1 \in \mathbb{R}^{b \times b}\) be the all-ones matrix and \(I\) be the identity matrix, then we have:
\[
2\text{KL}(p||q) = \log \frac{\text{diag}(\sigma_{\text{gen}})C\text{diag}(\sigma_{\text{gen}})}{\text{diag}(\sigma_{\text{ref}})C_{\text{gen}}\text{diag}(\sigma_{\text{ref}})} - b + \text{tr} (C^{-1}\text{diag}(r)C_{\text{gen}}\text{diag}(r)) + d^T C^{-1} d
\]
\[
= \sum_{i=1}^{b} \log \frac{\sigma_{\text{gen}}^2 i}{\sigma_{\text{ref}}^2 i} + \log \frac{|C|}{|C_{\text{gen}}|} - b + r^T (C^{-1} \circ C_{\text{gen}}) r + d^T C^{-1} d
\]
\[
= \log \frac{|C|}{|C_{\text{gen}}|} - b + \text{tr}((C^{-1} \circ C_{\text{gen}}) 1) - \sum_{i=1}^{b} \log r_i^2 + d^T d + r^T r - b
\]
\[
+ \text{tr}((C^{-1} \circ C_{\text{gen}}) (rr^T - 1)) + \text{tr}((C^{-1} - I) dd^T) + (b - r^T r). \tag{11}
\]
Therefore, we have
\[
2 |\text{KL}(p_c||q_c) - \text{KL}(p||q)| \leq 2\text{KL}(p_m||q_m) + \left| \frac{1}{2} \right| + \left| \frac{2}{2} \right| + \left| \frac{3}{2} \right|. \tag{12}
\]
Given that the marginal KL divergence is upper bounded by,
\[
2\text{KL}(p_m||q_m) = -\sum_{i=1}^{b} \log r_i^2 + d^T d + r^T r - b \leq 2\xi , \tag{13}
\]
and since \(\forall x , x - 1 - \log x \geq 0\), we have
\[
0 \leq -\sum_{i=1}^{b} \log r_i^2 + r^T r - b \leq 2\xi . \tag{14}
\]
Then \(\forall i , r_i^2 - \log r_i^2 - 1 \leq 2\xi\), which means \(r_i = 1 + \mathcal{O}(\sqrt{\xi})\). As a result, we have the following bounds,
\[
\| dd^T \|_F = d^T d \leq 2\xi , \tag{15}
\]
\[
\| b - r^T r \| = \mathcal{O}(b\sqrt{\xi}) , \tag{16}
\]
\[
\| rr^T - 1 \|_F = \mathcal{O}(b\sqrt{\xi}) . \tag{17}
\]
We further let \(\lambda := \lambda_{\text{min}}(C)\) be the smallest eigenvalue of \(C\). Then, we have \(\| C^{-1} \|_2 = \frac{1}{\lambda} \). Because \(C_{\text{gen}}\) is a correlation matrix, \(\| C_{\text{gen}} \|_\infty = 1\). Because
\[
(\text{tr}(A^T B))^2 \leq \text{tr}(A^T A) \text{ tr}(B^T B) = \| A \|_F^2 \| B \|_F^2 , \tag{18}
\]
which gives us the upper bound of \(|1|\):
\[
\frac{1}{2} = \left| \text{tr}((C^{-1} \circ C_{\text{gen}}) (rr^T - 1)) \right|
\]
\[
\leq \| C^{-1} \circ C_{\text{gen}} \|_F \| rr^T - 1 \|_F
\]
\[
\leq \| C \|_F \| rr^T - 1 \|_F
\]
\[
\leq \frac{\sqrt{b}}{\lambda} \| rr^T - 1 \|_F ,
\]
\[
= \frac{\sqrt{b}}{\lambda} \mathcal{O}(b\sqrt{\xi}) . \tag{19}
\]
Similarly, we can further bound \(|2\)| by:

\[
|2| = |\text{tr}((C^{-1} - I)dd^\top)| \\
\leq \|C^{-1} - I\|_F \|dd^\top\|_F \\
\leq \sqrt{2 + 2\|C^{-1}\|^2_F \|dd^\top\|_F} \\
\leq \sqrt{2 + \frac{2b}{\lambda^2} \|dd^\top\|_F} \\
\leq \sqrt{2 + \frac{2b}{\lambda^2} 2\xi}. \quad (20)
\]

Lastly, we can bound:

\[
|3| = |b - r^\top r| = O(b\sqrt{\xi}). \quad (21)
\]

Overall, since \(\xi \ll 1\), we have

\[
2|\text{KL}(p_c||q_c) - \text{KL}(p||q)| \leq 2\text{KL}(p_m||q_m) + \frac{\sqrt{b}}{\lambda} O(b\sqrt{\xi}) + \sqrt{2 + \frac{2b}{\lambda^2} 2\xi} + O(b\sqrt{\xi}) \\
\leq 2\xi + \frac{\sqrt{b}}{\lambda} O(b\sqrt{\xi}) + \sqrt{2 + \frac{2b}{\lambda^2} 2\xi} + O(b\sqrt{\xi}) \\
= O\left(\frac{b^{3/2}}{\lambda} \sqrt{\xi}\right). \quad (22)
\]