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

Chaoqi Wang*

University of Chicago

Shengyang Sun*

University of Toronto

Vector Institute

Roger Grosse

University of Toronto

Vector Institute

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 guid-

ing exploration and making robust decisions. Most works on uncertainty estimation have focused on *marginal uncertainty*, i.e. the posterior predictive variance $\text{Var}(f(x)|\mathcal{D}_{tr})$ of a function f at a single location x conditioned on the training data \mathcal{D}_{tr} . Marginal uncertainty is used in many active learning and Bayesian optimization algorithms (Hernández-Lobato & Adams, 2015; Guo et al., 2017; Cobb et al., 2018), as high uncertainty at a point indicates it is favorable to query.

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 $\rho(f(x), f(x')|\mathcal{D}_{tr})$, where ρ denotes the Pearson correlation. Algorithms that exploit PPCs include transductive experiment design (Yu et al., 2006), 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 (Hennig & Schuler, 2012; Frazier et al., 2009). 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 (Guo et al., 2017) 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.

*Equal contributions. Proceedings of the 24th International Conference on Artificial Intelligence and Statistics (AISTATS) 2021, San Diego, California, USA. PMLR: Volume 130. Copyright 2021 by the author(s).

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 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 (MacKay, 1992). Hence, better PPCs ought to lead to better TAL performance. We validate the TAL task itself as a benchmark for PPCs, 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.

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 re-

gression problem with Gaussian observation noise (Gal & Ghahramani, 2016; Hernández-Lobato & Adams, 2015; Lakshminarayanan et al., 2017), i.e., the target y is a noisy observation of the function value, $y = f(\mathbf{x}) + \epsilon, \epsilon \sim \mathcal{N}(0, \sigma_n^2(\mathbf{x}))$ ¹. We use \mathcal{D}_{tr} and \mathcal{D}_{te} to denote the training set and the test set, respectively. For active learning problems, we also use \mathcal{D}_{pl} to denote the pool set. Given the training set, for n locations $\mathbf{X} \in \mathbb{R}^{n \times d}$, the predictive distribution for $f(\mathbf{X})|\mathcal{D}_{\text{pl}}$ is represented as a multivariate Gaussian $\mathcal{N}(\boldsymbol{\mu}(\mathbf{X}), \boldsymbol{\Sigma}(\mathbf{X}, \mathbf{X}))$. Here $\boldsymbol{\mu}(\mathbf{X}) = [\mu_{\mathbf{x}}]_{\mathbf{x} \in \mathbf{X}}$ is the predictive mean and $\boldsymbol{\Sigma}(\mathbf{X}, \mathbf{X}) = [\Sigma(\mathbf{x}, \mathbf{x}')]_{\mathbf{x}, \mathbf{x}' \in \mathbf{X}}$ is the predictive covariance. Then the predictive variance $\sigma_{\mathbf{x}}^2 = \Sigma(\mathbf{x}, \mathbf{x})$ and the predictive correlation satisfy

$$\rho(\mathbf{x}, \mathbf{x}') = \Sigma(\mathbf{x}, \mathbf{x}') / (\sigma_{\mathbf{x}} \sigma_{\mathbf{x}}'). \quad (1)$$

2.1 Models

We consider the following models in our experiments:

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 $\{\mathbf{x}_1, \dots, \mathbf{x}_n\} \subset \mathcal{X}$, the function values $\mathbf{f} = (f(\mathbf{x}_1), f(\mathbf{x}_2), \dots, f(\mathbf{x}_n))$ have a multivariate Gaussian distribution $\mathbf{f} \sim \mathcal{N}(\boldsymbol{\mu}, \mathbf{K}_{XX})$, where $\boldsymbol{\mu} := [\mu(\mathbf{x}_1), \dots, \mu(\mathbf{x}_n)]^\top$ is computed from the mean function $\mu(\cdot)$, and \mathbf{K}_{XX} denotes the matrix $[k(\mathbf{x}_i, \mathbf{x}_j)]_{i,j}$ using the kernel function $k(\cdot, \cdot)$. For Gaussian likelihoods with variance σ_n^2 , we can make predictions $p(y_*|\mathbf{x}_*, \mathcal{D}_{\text{tr}})$ in closed form:

$$p(y_*|\mathbf{x}_*, \mathcal{D}_{\text{tr}}) = \mathcal{N}(\mathbf{K}_{*X}(\mathbf{K}_{XX} + \sigma_n^2 \mathbf{I})^{-1} \mathbf{y}, \boldsymbol{\Sigma}),$$

$$\boldsymbol{\Sigma} := \mathbf{K}_{**} - \mathbf{K}_{*X}(\mathbf{K}_{XX} + \sigma_n^2 \mathbf{I})^{-1} \mathbf{K}_{X*} + \sigma_n^2.$$

Exact posterior inference of GPs suffers from a computational cost that scales cubically with the number of training data. Thus, for large scale problems, Stochastic Variational Gaussian Processes (SVGPs) (Titsias, 2009; Hensman et al., 2013) are usually adopted.

Bayesian Neural Networks (BNNs). Given an L -layer neural network, the weights $\mathbf{w} = \{\mathbf{w}_l\}_{l=1}^L$ are 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 $\mathbf{a}_l = \mathbf{w}_l^\top \mathbf{z}_{l-1} / \sqrt{V_{l-1} + 1}$, 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 experiments, we use $p(\mathbf{w}) = \mathcal{N}(\mathbf{0}, \eta \mathbf{I})$, where the scalar η is the prior variance.

¹Following the original papers of each model, the observation variance $\sigma_n^2(\mathbf{x})$ is constant for all models other than Deep Ensemble (Lakshminarayanan et al., 2017). If Deep Ensemble uses a constant $\sigma_n^2(\mathbf{x})$, severe overfitting occurs in our experiments.

HMC BNNs. The posterior over weights $p(\mathbf{w}|\mathcal{D})$ is intractable in BNNs. Hamiltonian Monte Carlo (HMC) (Neal et al., 2011) incorporates the score function and momentum in the proposals of Markov Chain Monte Carlo (MCMC), which makes it more efficient in high dimensional parameter spaces. However, Neal et al. (2011) points out that jointly sampling weights and the prior variance η makes it difficult to mix; he suggests performing Gibbs sampling for prior variances combined with HMC for weights. Instead, we adopt Monte-Carlo Expectation-Maximization (Delyon et al., 1999) for optimizing hyperparameters: the prior variance η and the observation variance σ_n^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. They fit an approximate posterior $q(\mathbf{w})$ to maximize the evidence lower bound:

$$\mathcal{L}_q = \mathbb{E}_q[\log p(\mathcal{D}_{\text{tr}}|\mathbf{w})] - \text{KL}(q(\mathbf{w})||p(\mathbf{w})).$$

Specifically, Bayes-by-backprop (BBB) (Blundell et al., 2015) uses a fully factorized Gaussian $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\sigma}^2)$ to approximate the posterior. Using the reparameterization trick (Kingma & Welling, 2013), the gradients towards $(\boldsymbol{\mu}, \boldsymbol{\sigma}^2)$ can be computed by backpropagation, and then be used for updates. We fix the prior variance $\eta = 1$ because we did not find a benefit to optimizing it.

Noisy Natural Gradients (NNG). Among the variational Bayesian families, the noisy natural gradient (NNG) (Zhang et al., 2018) is an efficient method to fit multivariate Gaussian posteriors by adding adaptive weight noise to ordinary natural gradient updates. Assuming $q(\mathbf{w}) = \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ and $\mathcal{D}\mathbf{w} = \nabla_{\mathbf{w}} \log p(y|\mathbf{x}, \mathbf{w})$, then the update rules are:

$$\begin{aligned} \mathbf{F} &\leftarrow (1 - \beta)\mathbf{F} + \beta (\mathcal{D}\mathbf{w}\mathcal{D}\mathbf{w}^\top), \\ \boldsymbol{\mu} &\leftarrow \boldsymbol{\mu} + \alpha \left(\mathbf{F} + \frac{1}{N\eta} \mathbf{I} \right)^{-1} \left(\mathcal{D}\mathbf{w} - \frac{1}{N\eta} \mathbf{w} \right). \end{aligned}$$

In particular, when using a Kronecker-factored approximation (Martens & Grosse, 2015) for the Fisher matrix \mathbf{F} , NNG is equivalent to imposing a matrix-variate Gaussian distribution for the variational posterior $q(\mathbf{w})$. Similarly to BBB, we fix the prior variance $\eta = 1$ because we did not find a benefit to optimizing it.

Functional BNNs (FBNNs) The weights usually have a complicated relationship with the corresponding function, making it difficult to specify a meaningful prior over weights. Functional BNNs (Sun et al., 2019) propose to directly apply priors over the function and perform functional variational inference. For estimating

the functional KL divergence $\text{KL}(q(f)||p(f))$, they use mini-batch marginal KL divergences,

$$\mathcal{L}_q = \mathbb{E}_q[\log p(\mathcal{D}_{\text{tr}}|f)] - \lambda \text{KL} \left(q \left(\begin{bmatrix} \mathbf{f}^{\mathcal{D}_s} \\ \mathbf{f}^M \end{bmatrix} \right) \| p \left(\begin{bmatrix} \mathbf{f}^{\mathcal{D}_s} \\ \mathbf{f}^M \end{bmatrix} \right) \right),$$

where $\mathbf{f}^{\mathcal{D}_s}$ and \mathbf{f}^M are the function values on locations within a random mini-batch \mathcal{D}_s and on M random locations from a heuristic distribution, respectively. Across all experiments, we set $\lambda = |\mathcal{D}_{\text{tr}}| / |\mathcal{D}_s|$ and we use a GP-RBF prior. Following Shi et al. (2019), we use a RBF random feature network (Rahimi & Recht, 2008) 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 \mathbf{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 (Srivastava et al., 2014) is a technique employed in training neural networks wherein each unit of each layer is possibly discarded independently with probability p . Typically, at test time, the trained network is made deterministic by scaling the outputs of each layer by $1-p$ rather than randomly zeroing them out. For obtaining uncertainty, Gal & Ghahramani (2016) propose to keep dropout stochastic at test time and use many forward passes to compute the means and variances. They show that dropout can be seen as implicitly optimizing a variational objective.

Deep Ensemble. The Deep Ensemble (Lakshminarayanan et al., 2017) consists of training multiple networks independently under the same objective. Then the predictions of all networks can be aggregated to compute the predictive means and variances. Notably, the networks in the deep ensemble output not only the predicted mean $\mu_i(\mathbf{x})$, but also the predicted variance $\sigma_i^2(\mathbf{x})$. Each $\sigma_i^2(\mathbf{x})$ corresponds to the aleatoric variance; each $\mu_i(\mathbf{x})$ can be seen as a random function sample, so the epistemic uncertainty can be computed through the variance of $\{\mu_i(\mathbf{x})\}_{i=1}^m$.

2.2 Covariance Computation

The posterior predictive correlations (PPCs) can be obtained from the predictive covariance $\boldsymbol{\Sigma}(\mathbf{X}, \mathbf{X})$ by $\rho(\mathbf{x}, \mathbf{x}') = \boldsymbol{\Sigma}(\mathbf{x}, \mathbf{x}') / (\sigma_{\mathbf{x}} \sigma_{\mathbf{x}'})$. 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,

$$K(\mathbf{X}, \mathbf{X}) - K(\mathbf{X}, \mathbf{X}_{\text{tr}}) (K(\mathbf{X}_{\text{tr}}, \mathbf{X}_{\text{tr}}) + \sigma_n^2 \mathbb{I})^{-1} K(\mathbf{X}_{\text{tr}}, \mathbf{X}),$$

For other models such as BNNs, we need to draw samples to estimate the covariance matrix. For two points \mathbf{x}, \mathbf{x}' , we use a Monte Carlo estimate of the covariance:

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

where $f_i \sim p(f|D_{\text{tr}})$ are random function samples from the posterior, and $\widehat{\mu}_{\mathbf{x}} := \frac{1}{m} \sum_{i=1}^m f_i(\mathbf{x})$.

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 given 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

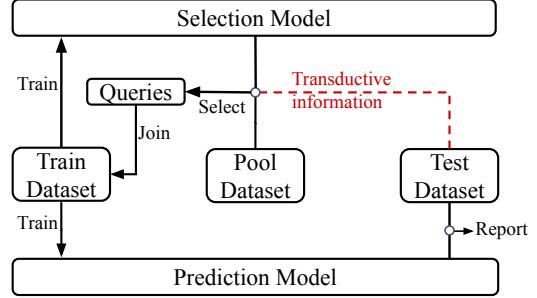


Figure 1: A diagram for the pipeline of (transductive) active learning.

is the only part of the active learning algorithm that uses the uncertainty estimates.

The **Total Information Gain** (TIG) (MacKay, 1992) is one acquisition function that has been used to benchmark predictive marginal uncertainty (Cohn et al., 1996; Houlsby et al., 2011; Hernández-Lobato et al., 2014; Zhang et al., 2018; Gal et al., 2017; Beluch et al., 2018). TIG computes the informativeness of each point by measuring the mutual information between it and the model parameters \mathbf{w} : $\text{TIG}(\mathbf{x}) := \mathbb{I}(y_{\mathbf{x}}; \mathbf{w} | D_{\text{tr}})$, where \mathbb{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 (1992), TIG is problematic since it favors points 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 (1992) presents the **Marginal Information Gain** (MIG) acquisition function, which measures the information gained about the region of interest by querying a new point. Supposing \mathbf{x}_u is the point of interest², $\text{MIG}(\mathbf{x}; \mathbf{x}_u) := \mathbb{I}(y_{\mathbf{x}}; f(\mathbf{x}_u) | D_{\text{tr}})$. We prefer MIG to TIG both because it's been found to avoid TIG's outlier problem (MacKay, 1992) and because it makes direct use of PPCs.

In practice, selecting one point at a time, and re-training 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 and redundant points. To encourage diversity in a batch, we introduce the **Batch Marginal Information Gain** (BatchMIG) acquisition function, inspired by the BatchBALD algorithm (Kirsch et al.,

²For a test set $\{\mathbf{x}_u^i\}_i$ of interest, we adopt the mean marginal information gain (MacKay, 1992).

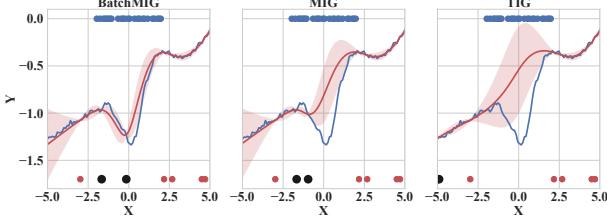


Figure 2: A visual comparison between BatchMIG, MIG and TIG. The red curve is the prediction after query, blue curve is the ground-truth, the • on top are the testing points, the • are the training points, and the • are the queried points of the corresponding acquisition function.

2019) in active learning. For a batch $\mathbf{x}_{1:q}$,

$$\text{BatchMIG}(\mathbf{x}_{1:q}; \mathbf{x}_u) := \mathbb{I}(\mathbf{y}_{\mathbf{x}_{1:q}}; f(\mathbf{x}_u) | \mathcal{D}_{\text{tr}}). \quad (2)$$

BatchMIG quantifies the amount of information carried by the selected batch. Though selecting the optimal batch $\mathbf{x}_{1:q}^*$ for BatchMIG is intractable, we adopt a greedy algorithm (Kirsch et al., 2019) to approximate it. We note that BatchMIG exploits PPCs more fully than MIG: in addition to using transductive information, it also uses PPCs between candidate query points to encourage diversity. Hence, it appears especially well-suited for benchmarking PPC estimators.

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 PPCs 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 $\sum_{i=1}^n \log p(y_i | \mathbf{x}_i)$ 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, $\log p(y_1, \dots, y_b | \mathbf{x}_1, \dots, \mathbf{x}_b)$. However, this is unsatisfying for two reasons:

Impact of predictive marginals. We've found the joint log-likelihood scores to be determined almost en-

tirely 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 $\{(\mathbf{x}_i, y_i)\}_{i=1}^b$ 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 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 $\{\mathbf{X}, \mathbf{y}\}$, we define

$$\begin{aligned} \text{XLL}(\mathbf{y} | \mathbf{X}, \mathcal{M}, \mathcal{M}_{\text{ref}}) \\ = \log \mathcal{N}(\mathbf{y} | \boldsymbol{\mu}_{\text{ref}}, \text{diag}(\boldsymbol{\sigma}_{\text{ref}}) \mathbf{C}_{\mathcal{M}} \text{diag}(\boldsymbol{\sigma}_{\text{ref}})). \end{aligned} \quad (3)$$

Here, \mathcal{M} and \mathcal{M}_{ref} denote the candidate and reference model, $\boldsymbol{\mu}_{\text{ref}}$ and $\boldsymbol{\sigma}_{\text{ref}}^2$ are the predictive mean and variance given by \mathcal{M}_{ref} , and $\mathbf{C}_{\mathcal{M}}$ is the predictive correlation matrix computed by \mathcal{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 $\mathbf{C}_{\mathcal{M}}$ match the ground truth correlations \mathbf{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 $\text{LogDet}(\mathbf{C}_{\text{gt}}, \mathbf{C}_{\mathcal{M}})$ denote the LogDet divergence (Kulis et al., 2006); then,

$$\text{LogDet}(\mathbf{C}_{\text{gt}}, \mathbf{C}_{\mathcal{M}}) = -\mathbb{E}_{\mathbf{y} | \mathbf{X}} \text{XLL} + \mathcal{O}\left(\frac{b^{3/2}}{\lambda} \sqrt{\xi}\right) + c,$$

where c is a constant, b is the batch size (set to 5 in our experiments). Here, ξ denotes the KL divergence between the reference marginal $(\mu_{\text{ref}}, \sigma_{\text{ref}})$ and the true marginal $(\mu_{\text{gt}}, \sigma_{\text{gt}})$. λ denotes the smallest eigenvalue of $\mathbf{C}_{\mathcal{M}}$, which is usually not much smaller than 1 due 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 (Hernández-Lobato et al., 2014) are used extensively for evaluating Bayesian neural networks. Calibration metrics (Guo et al., 2017; Kuleshov et al., 2018) are used for testing whether the predictions are over- or under-confident. Snoek et al. (2019) 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. (2018) developed a contextual bandit benchmark which uses marginal uncertainty to balance exploitation and exploration. Diabetic retinopathy diagnosis (Filos et al., 2019) was also used for comparing uncertainty estimates. However, all of these benchmarks are on evaluating marginal uncertainty estimations.

Algorithms exploiting PPCs. Although less attention has been paid to PPCs, there are still several algorithms that exploit PPCs in their designs. In transductive active learning (MacKay, 1992; Yu et al., 2006), 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 (Hennig & Schuler, 2012; Hernández-Lobato et al., 2014; Wang & Jegelka, 2017) make the query to acquire the most information about the optimum of the underlying function. Knowledge gradient (Frazier et al., 2009) 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 taht can indirectly acquire information about more expensive ones (Hennig & Schuler, 2012; Swersky et al., 2013). 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 (Neal, 1995). 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 \mathbf{x} is then $y = f(\mathbf{x}) + \epsilon, \epsilon \sim \mathcal{N}(0, 0.01)$. For all models, we used 1,000 epochs for training and the true observation variance³. 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 (Hernández-Lobato et al., 2014; Gal & Ghahramani, 2016; Zhang et al., 2018; Lakshminarayanan et al., 2017; Sun et al., 2019). This includes five

³For the synthetic setting, each network in the Deep Ensemble only predicts the mean.

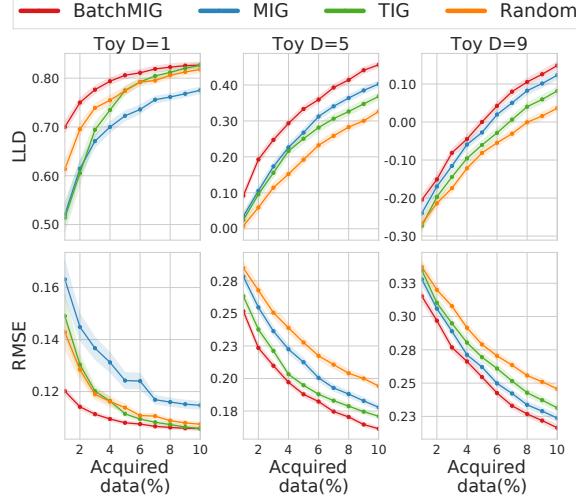


Figure 3: Results on synthetic datasets using BatchMIG, MIG, TIG and random selection with the oracle model.

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 \mathcal{D}_{tr} , 20% as the test set \mathcal{D}_{te} , and the remaining 60% as the pool set \mathcal{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.

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 (Sun et al., 2018), 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 (Hernández-Lobato et al., 2014).

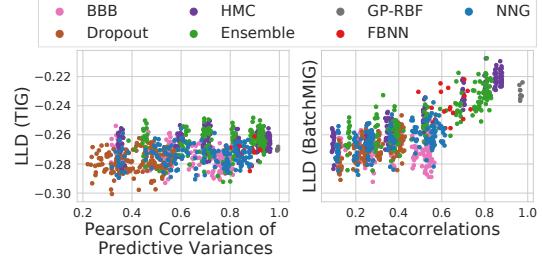


Figure 4: Left: Active learning. TIG performance vs. the quality of predictive variances. Right: Transductive active learning. BatchMIG performance vs. the quality of predictive correlations (metacorrelations). The BatchMIG performance and the metacorrelations are strongly correlated with a coefficient **0.762**.

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 first 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 12 (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 study 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 PPCs. 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 PPCs.

Posterior Predictive Correlations

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).

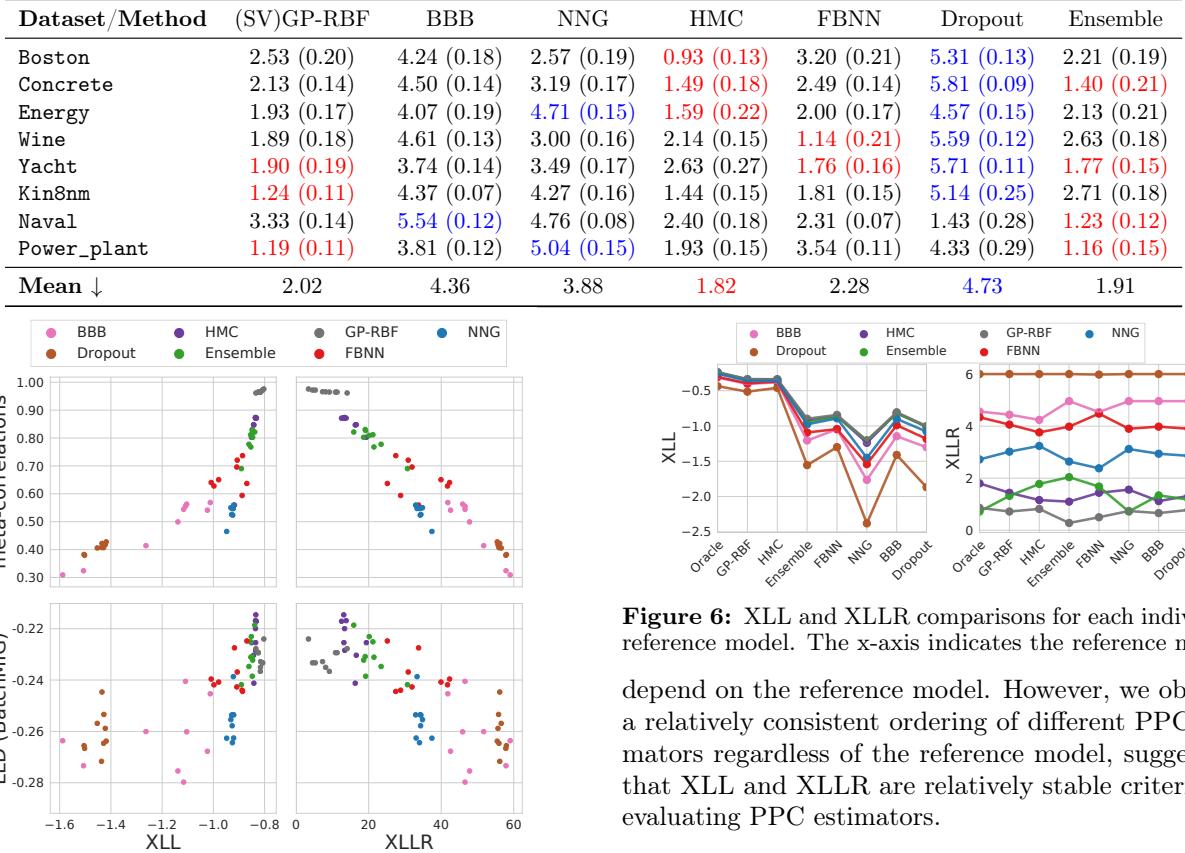


Figure 5: Metacorrelations and BatchMIG performance versus XLL and XLLR.

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.

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 absolute XLL values

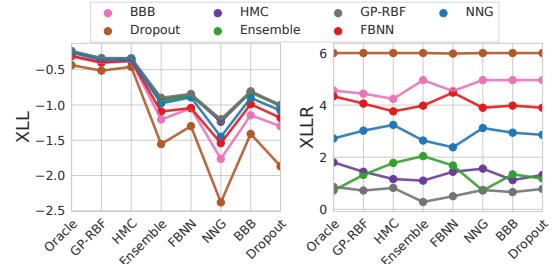


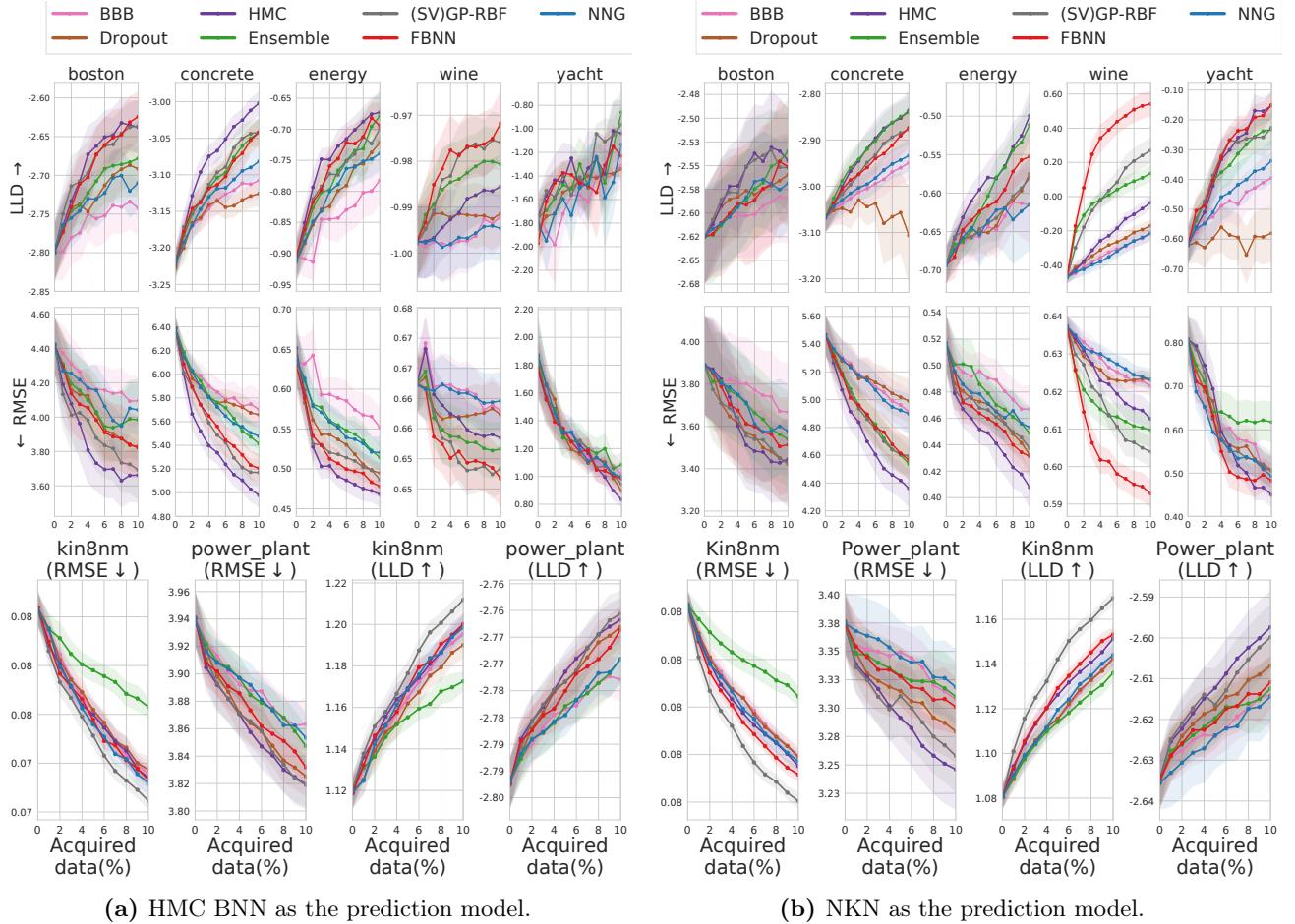
Figure 6: XLL and XLLR comparisons for each individual reference model. The x-axis indicates the reference model.

depend on the reference model. However, we observe a relatively consistent ordering of different PPC estimators regardless of the reference model, suggesting that XLL and XLLR are relatively stable criteria for evaluating PPC estimators.

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 XXL/XLLR and TAL on UCI datasets. We first talk about the empirical results on both benchmarks, and then discuss how do 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 (Riquelme et al., 2018). 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 terms of log marginal likelihoods.

**Figure 7:** BatchMIG-TAL using different selection methods, with (a) HMC BNN and (b) NKN as the prediction model.

Transductive Active Learning. Figures 7a and 7b show the TAL performance where various algorithms are used for query selection but predictions are made using the HMC and NKN models, respectively. (We omitted `naval`, since active learning appears to provide no benefit; see Figure 12 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 13).

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

- Beluch, W. H., Genewein, T., Nürnberger, A., and Köhler, J. M. The power of ensembles for active learning in image classification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 9368–9377, 2018.
- Blundell, C., Cornebise, J., Kavukcuoglu, K., and Wierstra, D. Weight uncertainty in neural network. In *International Conference on Machine Learning*, pp. 1613–1622, 2015.
- Cobb, A. D., Roberts, S. J., and Gal, Y. Loss-calibrated approximate inference in bayesian neural networks. *arXiv preprint arXiv:1805.03901*, 2018.
- Cohn, D. A., Ghahramani, Z., and Jordan, M. I. Active learning with statistical models. *Journal of artificial intelligence research*, 4:129–145, 1996.
- Delyon, B., Lavielle, M., and Moulines, E. Convergence of a stochastic approximation version of the em algorithm. *Annals of statistics*, pp. 94–128, 1999.
- Filos, A., Farquhar, S., Gomez, A. N., Rudner, T. G. J., Kenton, Z., Smith, L., Alizadeh, M., de Kroon, A., and Gal, Y. Benchmarking bayesian deep learning with diabetic retinopathy diagnosis. <https://github.com/OATML/bdl-benchmarks>, 2019.
- Frazier, P., Powell, W., and Dayanik, S. The knowledge-gradient policy for correlated normal beliefs. *INFORMS journal on Computing*, 21(4):599–613, 2009.
- Gal, Y. and Ghahramani, Z. Dropout as a Bayesian approximation: Representing model uncertainty in deep learning. In *International Conference on Machine Learning*, pp. 1050–1059, 2016.
- Gal, Y., Islam, R., and Ghahramani, Z. Deep bayesian active learning with image data. In *International Conference on Machine Learning*, pp. 1183–1192, 2017.
- Guo, C., Pleiss, G., Sun, Y., and Weinberger, K. Q. On calibration of modern neural networks. In *International Conference on Machine Learning*, pp. 1321–1330, 2017.
- Hennig, P. and Schuler, C. J. Entropy search for information-efficient global optimization. *Journal of Machine Learning Research*, 13(Jun):1809–1837, 2012.
- Hensman, J., Fusi, N., and Lawrence, N. D. Gaussian processes for big data. In *Uncertainty in Artificial Intelligence*, pp. 282, 2013.
- Hernández-Lobato, J. M. and Adams, R. Probabilistic backpropagation for scalable learning of Bayesian neural networks. In *International Conference on Machine Learning*, pp. 1861–1869, 2015.
- Hernández-Lobato, J. M., Hoffman, M. W., and Ghahramani, Z. Predictive entropy search for efficient global optimization of black-box functions. In *Advances in Neural Information Processing Systems*, pp. 918–926, 2014.
- Houlsby, N., Huszár, F., Ghahramani, Z., and Lengyel, M. Bayesian active learning for classification and preference learning. *arXiv preprint arXiv:1112.5745*, 2011.
- Kingma, D. P. and Welling, M. Auto-encoding variational Bayes. *arXiv preprint arXiv:1312.6114*, 2013.
- Kirsch, A., van Amersfoort, J., and Gal, Y. Batchbald: Efficient and diverse batch acquisition for deep bayesian active learning, 2019.
- Kuleshov, V., Fenner, N., and Ermon, S. Accurate uncertainties for deep learning using calibrated regression. In *International Conference on Machine Learning*, pp. 2796–2804, 2018.
- Kulis, B., Sustik, M., and Dhillon, I. Learning low-rank kernel matrices. In *Proceedings of the 23rd international conference on Machine learning*, pp. 505–512, 2006.
- Lakshminarayanan, B., Pritzel, A., and Blundell, C. Simple and scalable predictive uncertainty estimation using deep ensembles. In *Advances in Neural Information Processing Systems*, pp. 6402–6413, 2017.
- MacKay, D. J. Information-based objective functions for active data selection. *Neural computation*, 4(4):590–604, 1992.
- Martens, J. and Grosse, R. Optimizing neural networks with kronecker-factored approximate curvature. In *International conference on machine learning*, pp. 2408–2417, 2015.
- Neal, R. M. *Bayesian Learning for Neural Networks*. PhD thesis, University of Toronto, 1995.
- Neal, R. M. et al. MCMC using Hamiltonian dynamics. *Handbook of Markov Chain Monte Carlo*, 2(11), 2011.
- Rahimi, A. and Recht, B. Random features for large-scale kernel machines. In *Advances in Neural Information Processing Systems*, pp. 1177–1184, 2008.
- Riquelme, C., Tucker, G., and Snoek, J. Deep Bayesian bandits showdown: An empirical comparison of

- Bayesian deep networks for thompson sampling. In *International Conference on Learning Representations*, 2018.
- Shi, J., Khan, M. E., and Zhu, J. Scalable training of inference networks for gaussian-process models. *arXiv preprint arXiv:1905.10969*, 2019.
- Snoek, J., Ovadia, Y., Fertig, E., Lakshminarayanan, B., Nowozin, S., Sculley, D., Dillon, J., Ren, J., and Nado, Z. Can you trust your model's uncertainty? evaluating predictive uncertainty under dataset shift. In *Advances in Neural Information Processing Systems*, pp. 13969–13980, 2019.
- Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., and Salakhutdinov, R. Dropout: a simple way to prevent neural networks from overfitting. *The journal of machine learning research*, 15(1):1929–1958, 2014.
- Sun, S., Zhang, G., Wang, C., Zeng, W., Li, J., and Grosse, R. Differentiable compositional kernel learning for Gaussian processes. In *International Conference on Machine Learning*, pp. 4828–4837, 2018.
- Sun, S., Zhang, G., Shi, J., and Grosse, R. Functional variational bayesian neural networks. *arXiv preprint arXiv:1903.05779*, 2019.
- Swersky, K., Snoek, J., and Adams, R. P. Multi-task bayesian optimization. In *Advances in neural information processing systems*, pp. 2004–2012, 2013.
- Titicaias, M. Variational learning of inducing variables in sparse Gaussian processes. In *Artificial Intelligence and Statistics*, pp. 567–574, 2009.
- Wang, Z. and Jegelka, S. Max-value entropy search for efficient Bayesian optimization. In *International Conference on Machine Learning*, pp. 3627–3635, 2017.
- Yu, K., Bi, J., and Tresp, V. Active learning via transductive experimental design. In *Proceedings of the 23rd international conference on Machine learning*, pp. 1081–1088. ACM, 2006.
- Zhang, G., Sun, S., Duvenaud, D., and Grosse, R. Noisy natural gradient as variational inference. In *International Conference on Machine Learning*, pp. 5852–5861, 2018.