Reliable training and estimation of variance networks

NeurIPS 2019 paper reproduction

Ramon Emiliani, Emilio Botero, Isaac Ahouma, Lawrence Abdulnour, Louis Preville-Ratelle,

Mila (Quebec Artificial Intelligence Institute)

Objective

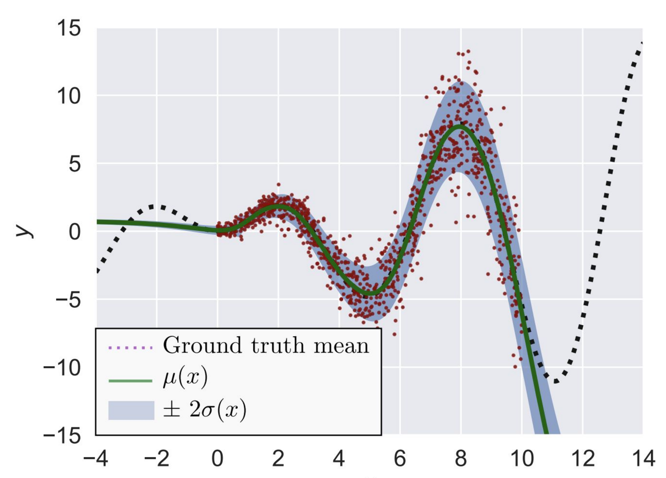
To reproduce a NeurIPS paper on the predictive *uncertainty* of neural networks. The paper proposes a new set of complementary methodologies for estimating the predictive variance in regression tasks.

Introduction

When doing regression, we typically focus on the mean prediction. Consider a toy regression dataset of the form

$$y = x\,sin(x) + \epsilon_1 + x\,\epsilon_2 \text{ where } \epsilon_1, \epsilon_2 \sim \mathcal{N}(0,1)$$

And say we want MLE of $\mathcal{N}(\mu(x), \sigma^2(x))$ where $\mu(x), \sigma^2(x)$ are neural networks.



Notice that the variance is underestimated, and doesn't increase outside data support. Authors claim these problems are general and propose methods to solve them.

Methodology

Preliminaries

Assume that datasets contain i.i.d observations

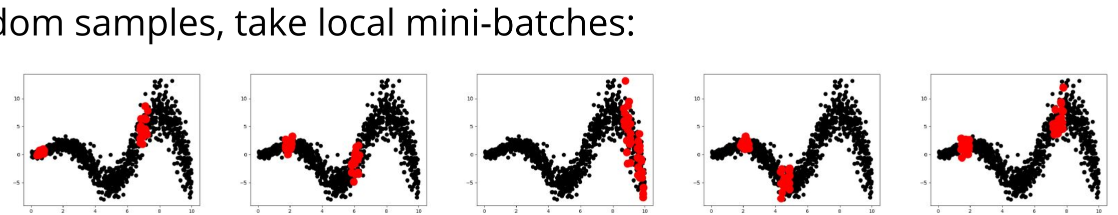
$$\mathcal{D} = \{\boldsymbol{x}_i, y_i\}_{i=1}^N \text{ where } \boldsymbol{x}_i \in \mathbb{R}^d, y_i \in \mathbb{R}$$

$$p_\theta(y|\boldsymbol{x}) \sim \mathcal{N}(\mu(\boldsymbol{x}), \sigma^2(\boldsymbol{x}))$$

where $\mu(\cdot), \sigma^2(\cdot)$ continuous functions parametrized by $\theta = \{\theta_\mu, \theta_{\sigma^2}\}$

The locality sampler

Instead of training our neural networks using standard mini-batches of random samples, take local mini-batches:



$$\sum_{i=1}^{N}\Big[-\frac{1}{2}\log\sigma^2(\boldsymbol{x}_i) - \frac{(y_i-\mu(\boldsymbol{x}_i))^2}{2\sigma^2(\boldsymbol{x}_i)}\Big] \approx \sum_{\boldsymbol{x}_j\in\mathcal{O}}\frac{1}{\pi_j}\Big[-\frac{1}{2}\log\sigma^2(\boldsymbol{x}_j) - \frac{(y_j-\mu(\boldsymbol{x}_j))^2}{2\sigma^2(\boldsymbol{x}_j)}\Big]$$

Mean variance split training

First train only for $\mu(x)$. Then alternate the training of $\mu(x)$ and $\sigma^2(x)$ to avoid problems in low data region.

Estimating distributions of variance

In low data regions, it is better to be Bayesian. Instead of calculating $\sigma^2(x)$ directly, they train two neural networks $\alpha(x)$ and $\beta(x)$ where α and β are the two parameters of the Inverse-Gamma distribution, which is the conjugate prior of σ^2 when the data is Gaussian.

$$\log p_\theta(y_i) = \log\int \mathcal{N}(y_i|\mu_i, \sigma_i^2)\frac{\beta^\alpha}{\Gamma(\alpha)}(\sigma_i^2)^{-\alpha+1}\exp\Big(-\frac{\beta_i}{\sigma_i^2}\Big)\,d\sigma_i^2$$
$$= \log t_{\mu_i,\alpha_i,\beta_i}(y_i)$$

Extrapolation

To bound the variance, they impose the variance to tend to a chosen value when out of distribution. Similar to sparse GP, take points $\{c_i\}_{i=1}^L$ that that represent training data and let

$$\hat{\sigma}^2(x_0) = (1-\nu(\delta(x_0)))\hat{\sigma}_\theta^2 + \eta\nu(\delta(x_0))$$

where $\nu(x) = \mathrm{sigmoid}((x+a)/(\gamma))$

and $\delta(x_0) = \min_i ||c_i - x_0||$

and the c_i are initialized using k-means.

Baselines

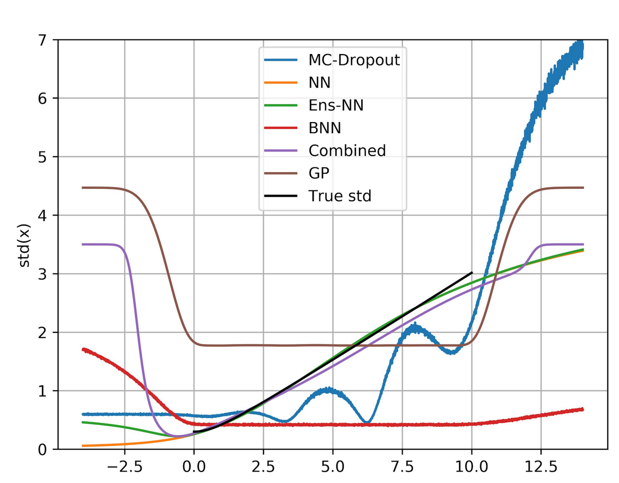
- Gaussian Process (GP)
- Sparse GP
- Bayesian Neural Networks (BNN)
- Monte Carlo Dropout (MCD)
- Neural Networks (NN)
- Ensembles of NN (ENN)

Datasets

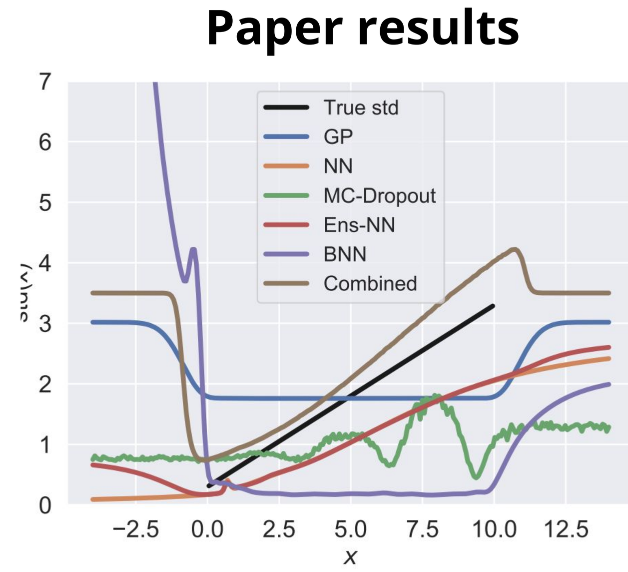
- Toy sine dataset
- Weather data
- Boston House Prices

Results

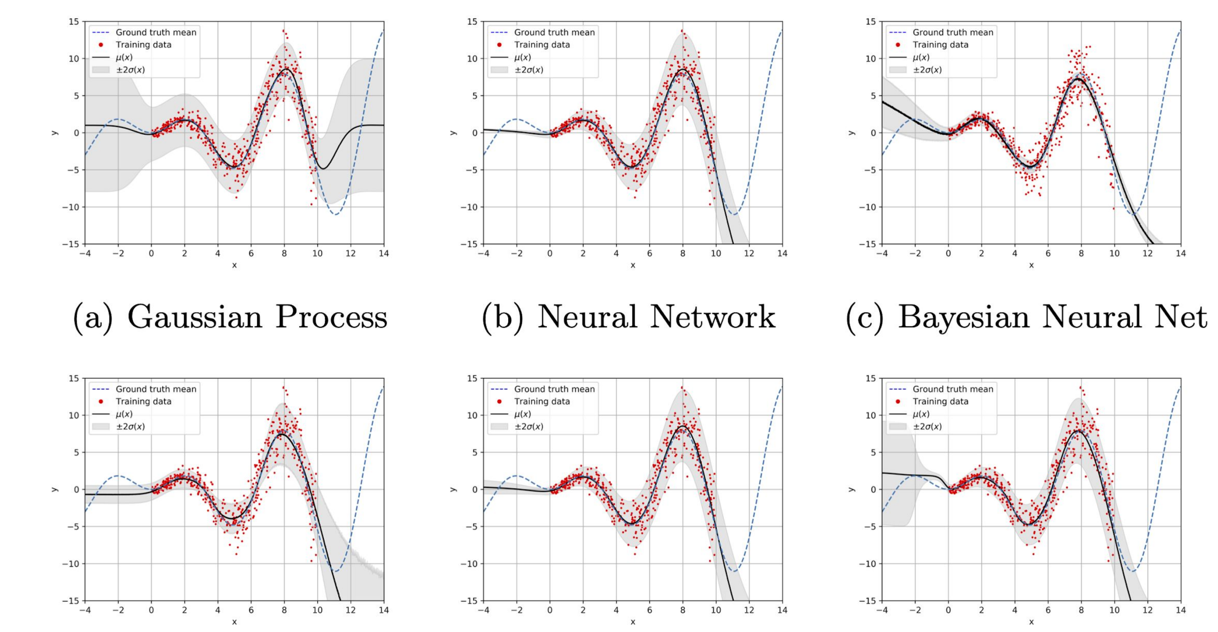
Our results



Paper results



Our results



(a) Gaussian Process (b) Neural Network (c) Bayesian Neural Net

(d) Monte Carlo Dropout (e) Ensemble of NNs (f) Combined

Figure 1: Regression on toy dataset

Paper results

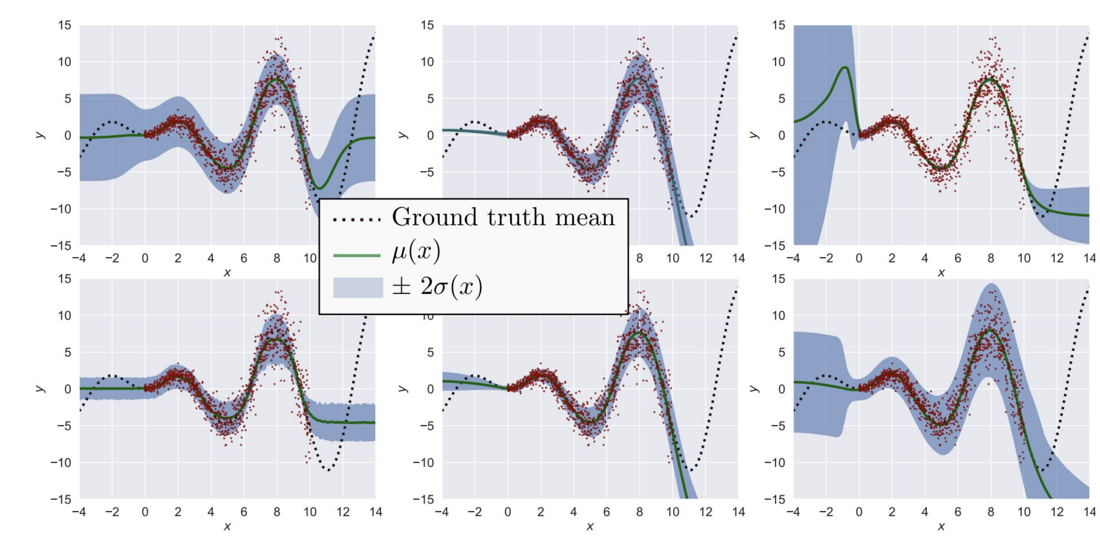
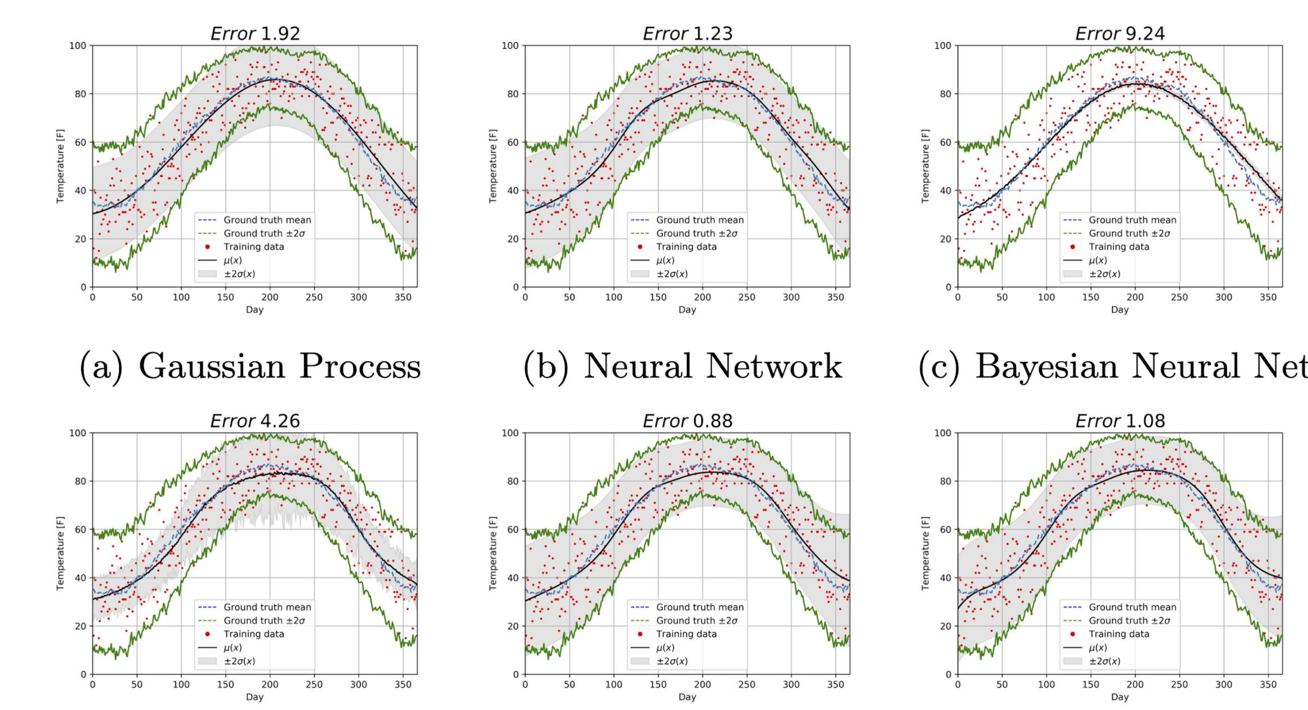


Figure 2: From top left to bottom right: GP, NN, BNN, MC-Dropout, Ens-NN, Combined.

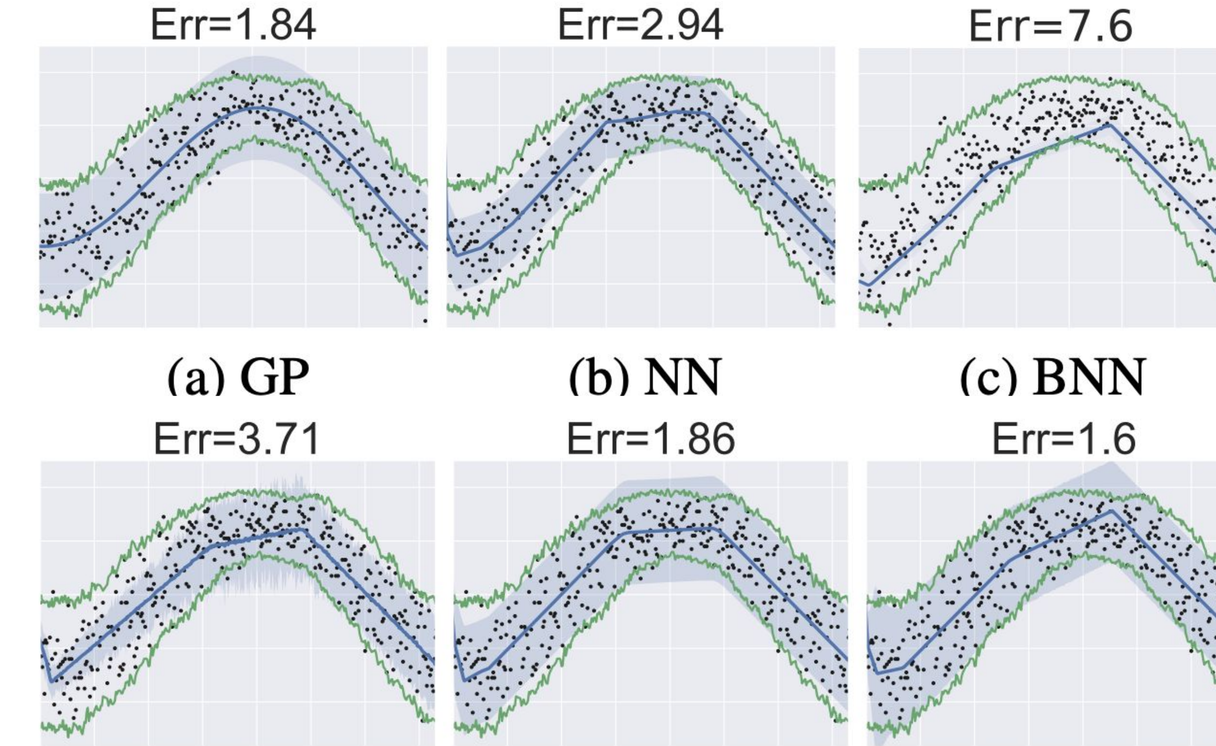
Weather dataset



(a) Gaussian Process (b) Neural Network (c) Bayesian Neural Net

(d) Monte Carlo Dropout (e) Ensemble of NNs (f) Combined

Figure 1: Regression on toy dataset



(a) GP (b) NN (c) BNN

(d) MC-Dropout (e) Ens-NN (f) Combined

Mean Test log-likelihoods on the Boston UCI Regression Dataset		
	Ours	Paper
sGP	**-1.93**	-1.85
GP	-2.01	**-1.76**
NN	-4.34	-3.64
BNN	-	-2.59
MC-Dropout	-4.23	-2.51
Ens-NN	-4.22	-2.45
Combined	-3.53	-2.09

Table 1: Mean Test Log-likelihoods on the Boston Dataset

Conclusions

- We successfully reproduced the paper's method and results on their main regression tasks.
- The paper is based in four complementary methodologies that result in better uncertainty estimates.
- When out of distribution, the paper aims to replicate a GP's behaviour.
- Need to test on higher dimension data.