Variable Selection with BNN
Wenying Deng
3/13/2020

In the last meeting, we validated the ability of Random Fourier Feature in approximating Gaussian Process, specifically using RBF kernel (with length-scale and variance to be 1). Last time, I computed the posterior covariance using the whole sample at once, since there is closed form relationship between the kernel and random Fourier Feature by Equation (2) from the RFF original paper:

$$k(\mathbf{x}-\mathbf{y}) = \int_R^d p(\omega) e^{j\omega'(\mathbf{x}-\mathbf{y})} d\omega = E_\omega[\zeta_\omega(\mathbf{x})\zeta_\omega(\mathbf{x})^*].$$

This week I tried to use Woodbury identity formula to update the covariance, to reduce time complexity. A GP (say f) under RFF approximation is written as $f = \Phi\beta$, where Φ are the Fourier Features and β is the regression coefficients, in this way we can get GP posterior by performing posterior inference on β, which in our case is the same as performing posterior inference for Bayesian linear regression.

Specifically, the posterior variance of β is $\Sigma_\beta = (\Phi^\top\Phi + \sigma^2\mathbf{I})^{-1}$. Given a mini-batch of Φ, we can update the covariance matrix using Woodbury identity formula:

$$\Sigma_{\beta,t+1} = \Sigma_{\beta,t} - \Sigma_{\beta,t}\Phi^\top(\mathbf{I} + \Phi\Sigma_{\beta,t}\Phi^\top)^{-1}\Phi\Sigma_{\beta,t}.$$

After we get the posterior variance Σ_β, the GP posterior for prediction mean $f_{new} = \Phi_{new}\beta$ is $\Phi_{new}\Sigma_\beta\Phi_{new}^\top$.

In this way we can reduce time complexity from $O(N^3)$ to $O(N)$, where N is sample size.

Figure 1 shows the random sample plots of posterior predictive distribuion: Black points are training samples, with blue curve the posterior mean prediction and the shaded area the 95% CI. For each row, I use GP and RFF to fit the same dataset. And the training sample sizes are 40, 200 and 1200 respectively.

Figure 2 shows four metrics comparing GP and RFF: sample size of the first two subfigures ranges from 40 to 200, and the last two from 200 to 1200. They are all on the same scale. Note:

- GP and RFF perform similarly on the test log-likelihood and RMSE.
- In small training sample size scenario, the PICP of RFF is lower than expected. But as training sample size increases, the PICP of RFF performs better.
- Although the difference is tiny, the MPIW of RFF is generally smaller than GP's.

I will try to explore the reason why MPIW of RFF is generally smaller than GP's.

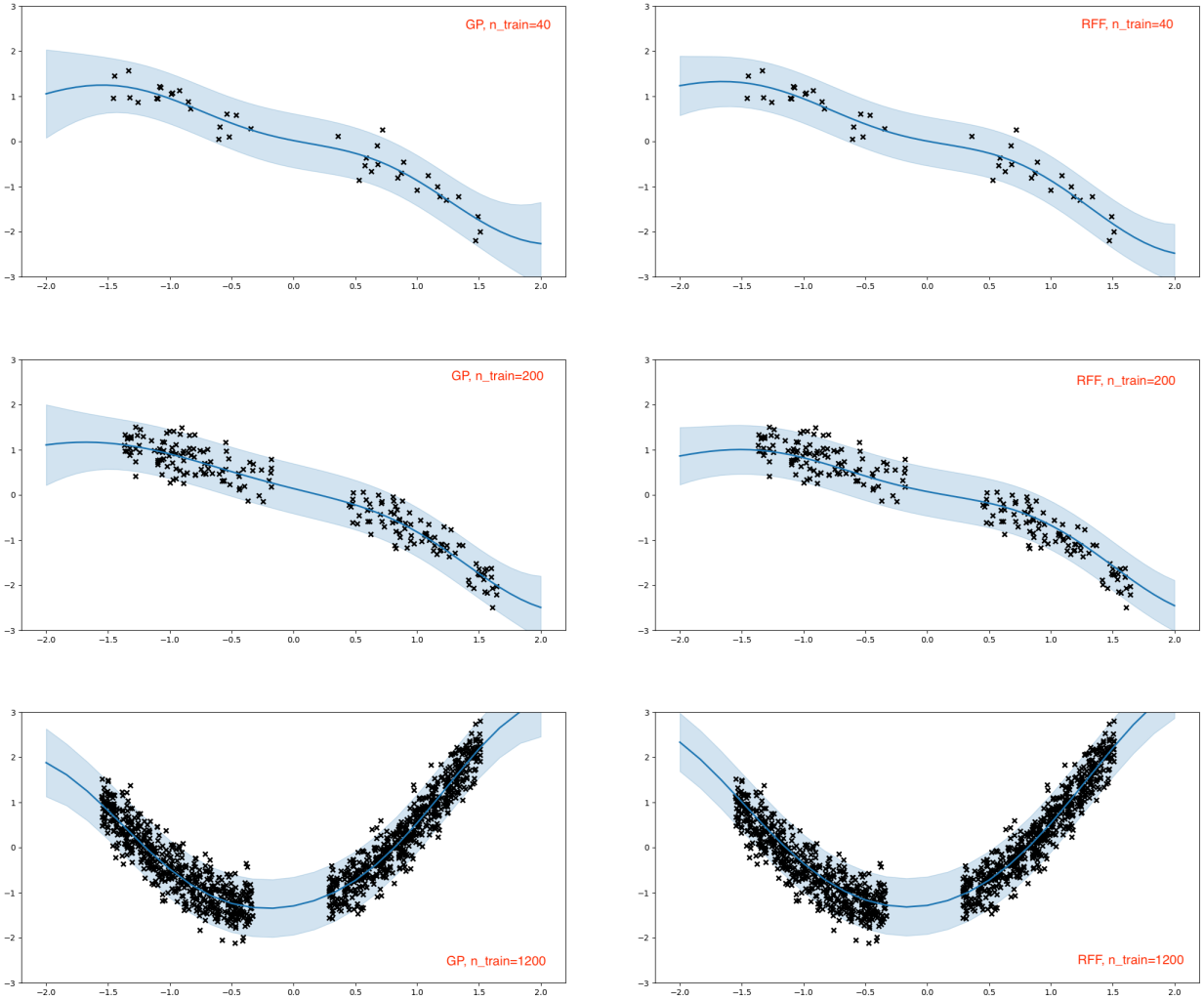
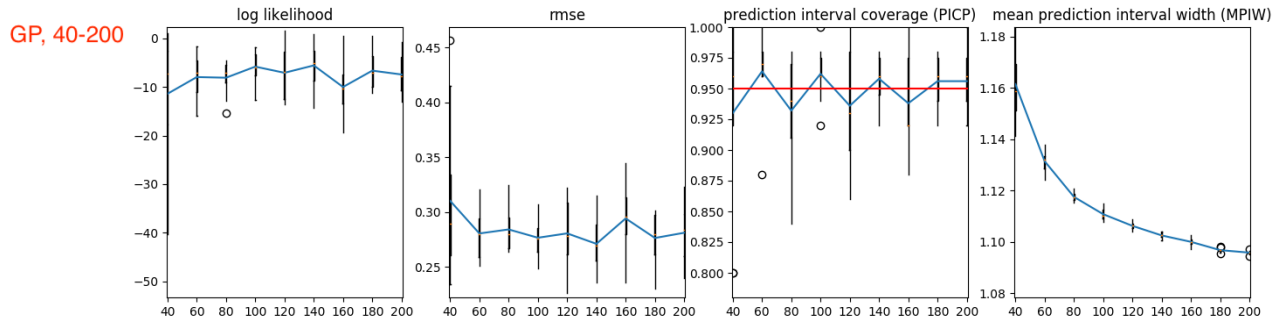
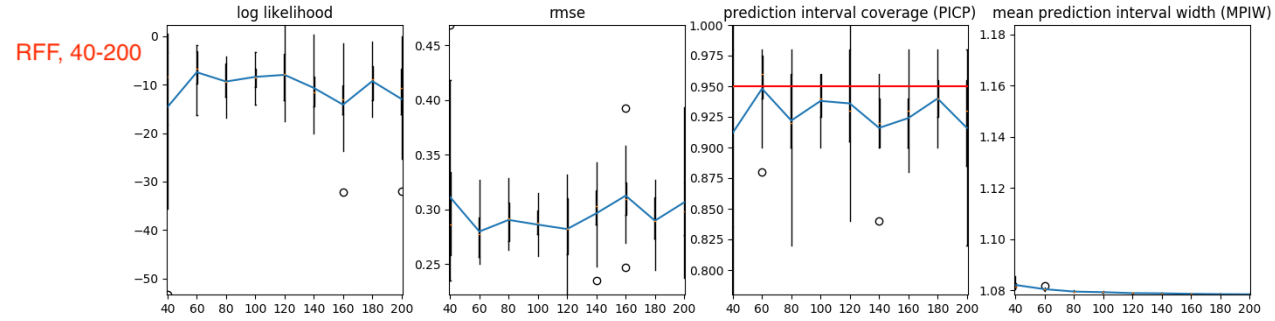


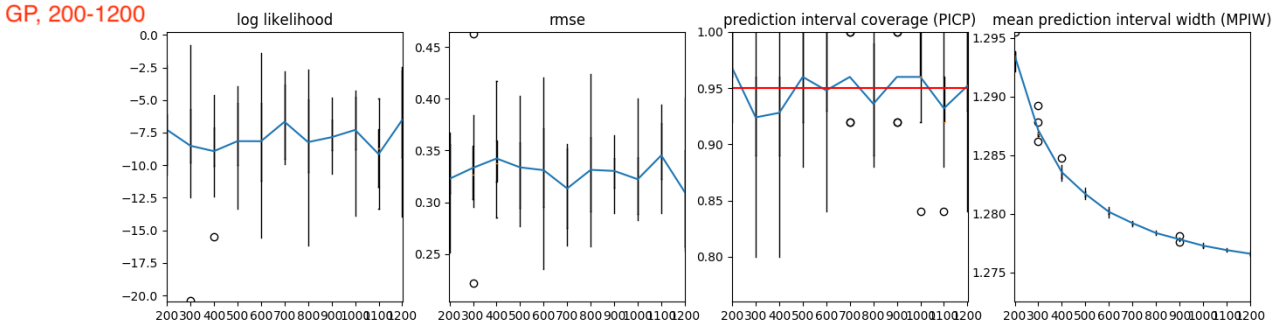
Figure 1: Sample plots of posterior predictive dist: Black points are training samples, with blue curve the posterior mean prediction and the shaded area the 95% CI.



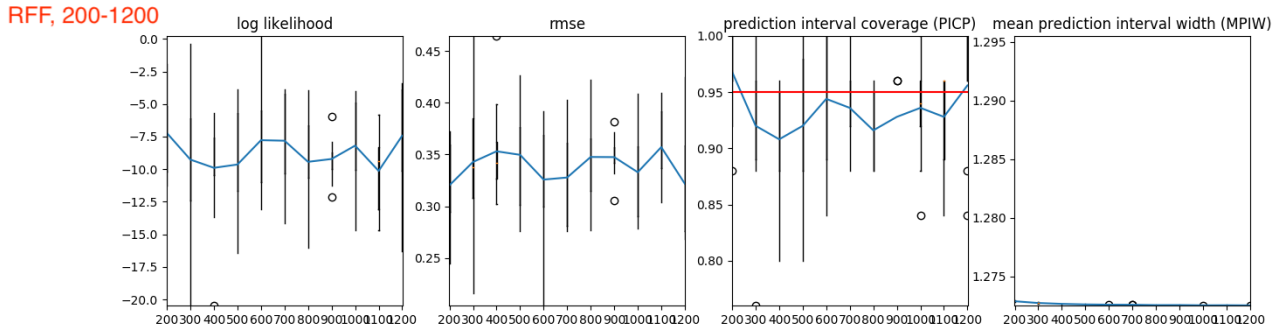
(a)



(b)



(c)



(d)

Figure 2: Four metrics comparing GP and RFF: Sample size of the first two subfigures ranges from 40 to 200, and the last two from 200 to 1200.