RFF approx GP
Wenying Deng
3/20/2020

Last time we validated the ability of Random Fourier Feature in approximating Gaussian Process, specifically using RBF kernel (with length-scale and variance to be 1). 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 [1]:

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

These two weeks 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.

On the other hand, [2] proposed Orthogonal Random Features (ORF), which can significantly reduced the kernel approximation error. The idea of ORF is to impose orthogonality on the approximate kernel using RFF. Jeremiah also implemented ORF in the original RFF and we see that it indeed has better performance.

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

Figure 2 shows the random sample plots of posterior predictive distribution: Black points are training samples, with blue curve the posterior mean prediction and the shaded area the 95% CI. For each column, I use GP, ORF-Woodbury and ORF-Population to fit the same dataset. And the training sample sizes are 40 and 200 respectively.

Figure 3 shows four metrics comparing GP, RFF-Woodbury and RFF-Population. They are all on the same scale. Note:

- GP and RFF perform similarly in the test log-likelihood, RMSE and MPIW.
- PICP of RFF is lower than expected.

Figure 4 shows four metrics comparing GP, ORF-Woodbury and ORF-Population. They are all on the same scale. Note:

- GP and ORF perform similarly in the test log-likelihood, RMSE and MPIW.
- PICP of ORF is lower than expected.
- Comparing the scale with RFF's, we see there are improvement in the test log-likelihood, RMSE and MPIW.

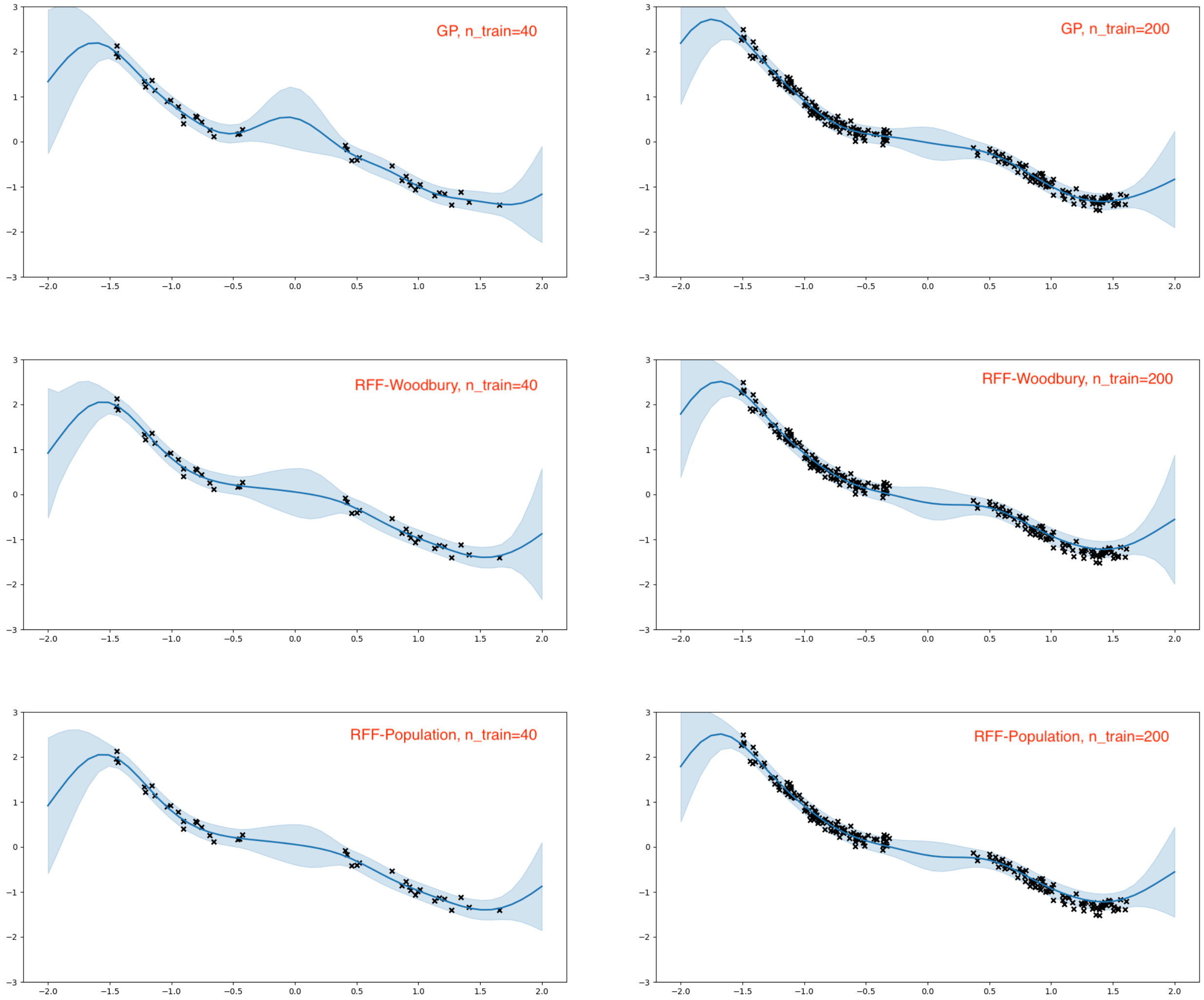
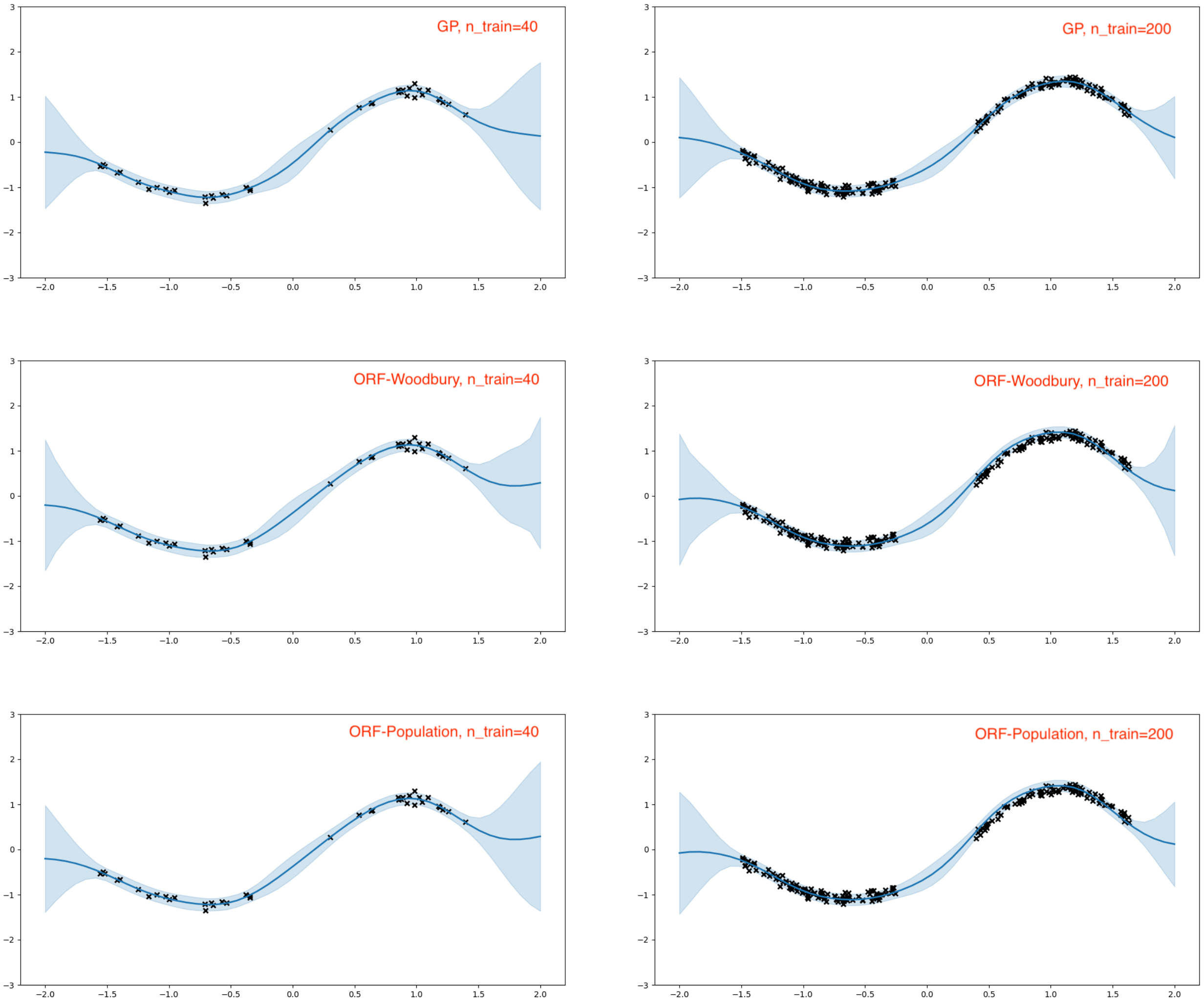


Figure 1: Sample plots of posterior predictive dist using RFF: Black points are training samples, with blue curve the posterior mean prediction and the shaded area the 95% CI.


GP, n_train=40

GP, n_train=200

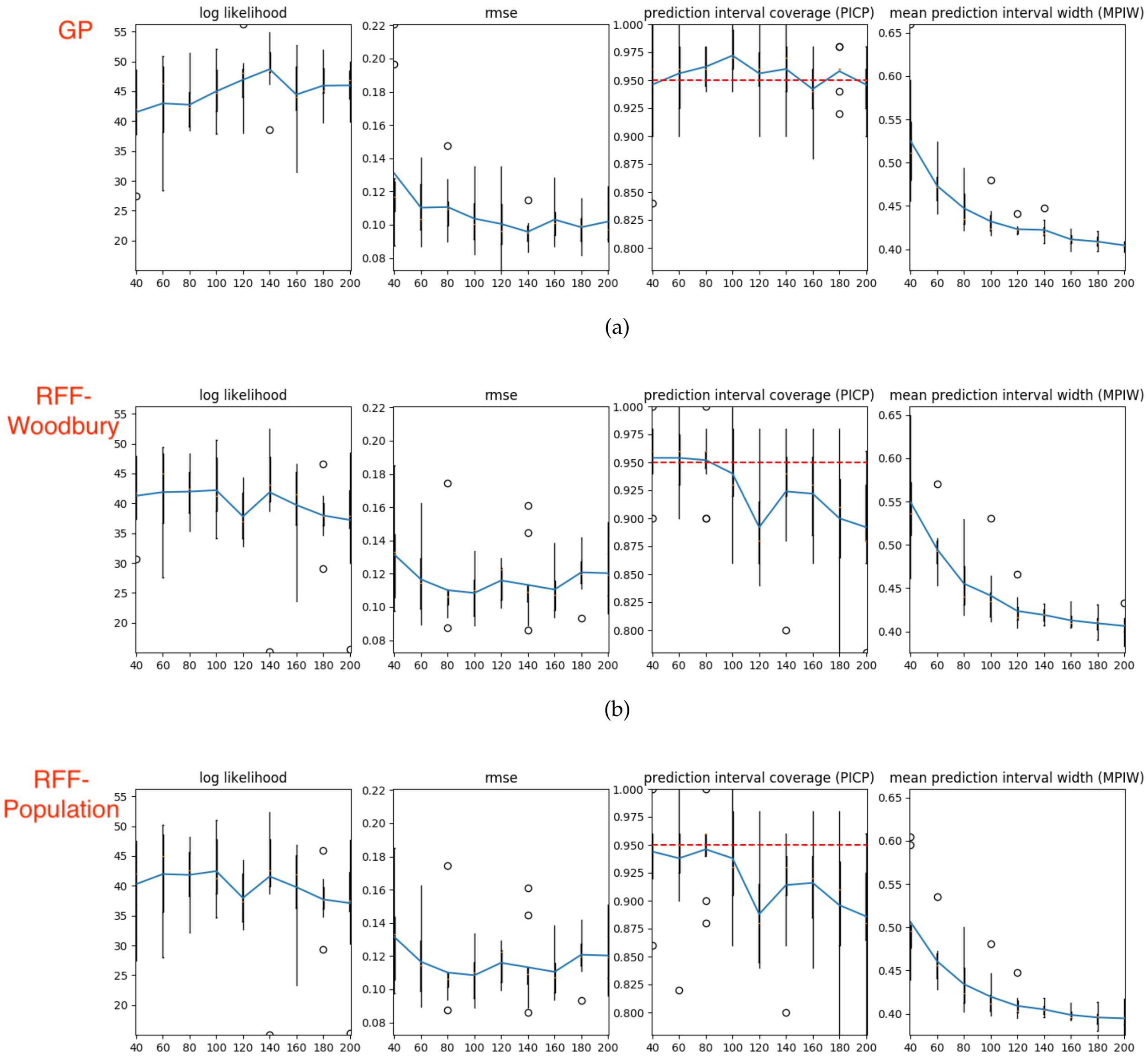
ORF-Woodbury, n_train=40

ORF-Woodbury, n_train=200

ORF-Population, n_train=40

ORF-Population, n_train=200

Figure 2: Sample plots of posterior predictive dist using ORF: Black points are training samples, with blue curve the posterior mean prediction and the shaded area the 95% CI.

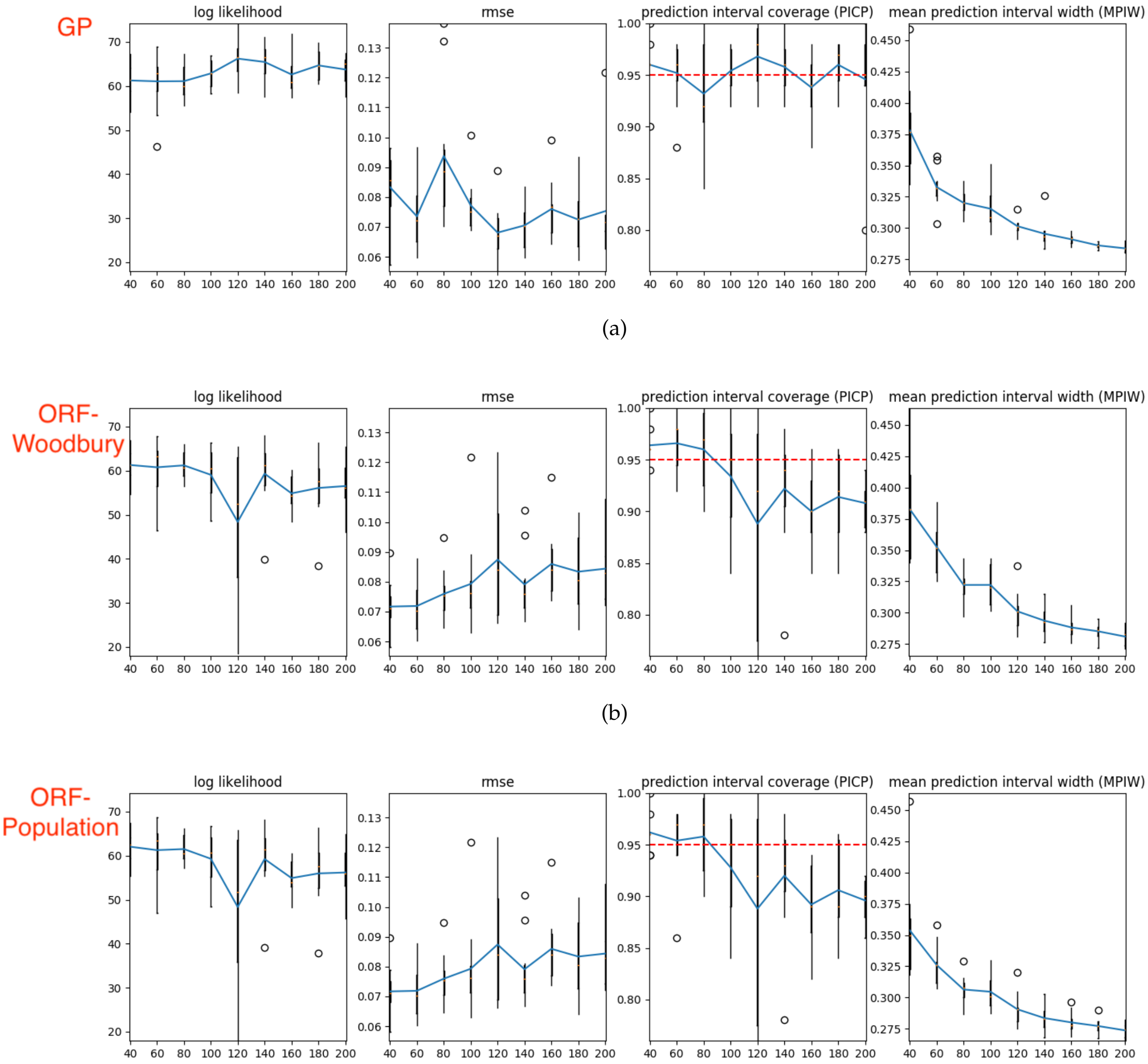


(a)

(b)

(c)

Figure 3: Four metrics comparing GP, RFF-Woodbury and RFF-Population.



(a)

(b)

(c)

Figure 4: Four metrics comparing GP, ORF-Woodbury and ORF-Population.

References

[1] Ali Rahimi and Ben Recht. Random Features for Large-Scale Kernel Machines. page 8.

[2] Felix Xinnan X Yu, Ananda Theertha Suresh, Krzysztof M Choromanski, Daniel N Holtmann-Rice, and Sanjiv Kumar. Orthogonal Random Features. In D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems 29*, pages 1975–1983. Curran Associates, Inc., 2016.