Provably Reliable Large-Scale Sampling from Gaussian Processes

Anthony Stephenson[1], Robert Allison[1,2], Edward Pyzer-Knapp[3]

1 University of Bristol
2 Part funded by National Cyber Security Centre (UK)
3 IBM Research

IBM **Research** Europe

NEURAL INFORMATION PROCESSING SYSTEMS

Motivation

Usually, approximate GP methods are evaluated on a toy synthetic dataset at small scale (e.g., $n \approx \mathcal{O}(10^3)$) and a limited set of real benchmark-datasets. We believe this leaves a gap in the analysis; careful assessment of performance at scale on data adhering exactly to a GP model.

Unfortunately, (naïvely) generating a sample of size n from a GP model is a task of complexity $\mathcal{O}(n^3)$. Since we want to do this at scale, this is clearly infeasible; as a result, we wish to find a way to generate *approximate* samples that are *extremely close* in some sense to "real" samples.

"Indistinguishable" distributions

Definition (ϵ-indistinguishable). P_0 and P_1 are ϵ-indistinguishable if the optimal Bayesian decision process (see [5]) has $\Pr(\text{error}) \geq \frac{1}{2} - \epsilon$.
Lemma (ϵ-indistinguishable). P_0 and P_1 are ϵ-indistinguishable if $\mathcal{TV}(P_0, P_1) \leq 2\epsilon$. (TV = Total variation distance).

Experiments

To empirically test our results, we ran 1000 repeat experiments generating samples using RFF and CIQ of different sizes, with varying hyperparameters and measured how often a Cràmer von Mises test rejected the null hypothesis (that the data came from an $\mathcal{N}(0, I_n)$ after applying an exact whitening transformation) as a function of the fidelity parameters D, J.

Random Fourier Features (RFF)

Random Fourier Features were introduced as a method of approximating kernels at large scales in Support Vector Machines and Kernel Ridge Regression problems in [2]. One of the appealing features of the RFF approximation for sampling from a GP is the fact that we don't need to form the full Gram matrix in order to generate samples.

To construct the full approximate matrix, we form the product ZZ^T of $Z \in \mathbb{R}^{n \times D}$ matrices. To generate samples, we need only construct a single Z matrix and simply transform an $w \sim \mathcal{N}(0, I_D)$ variable to get $\hat{f} = Zw$. This shows that we have a method of complexity $\mathcal{O}(nD)$ to produce an approximate sample of size n.

Lemma (RFF). To generate a sample of size n whose marginal distribution differs from the true marginal distribution from a given GP by a total variation distance ($\mathcal{TV}$) of at most ϵ, with probability $1 - \delta$ it is sufficient to use D RFFs, where $D \geq 8\log\left(\frac{n}{\sqrt{\delta}}\right)\frac{n^2}{8\epsilon^2\sigma_\xi^4}$ for some $\delta > 0$.

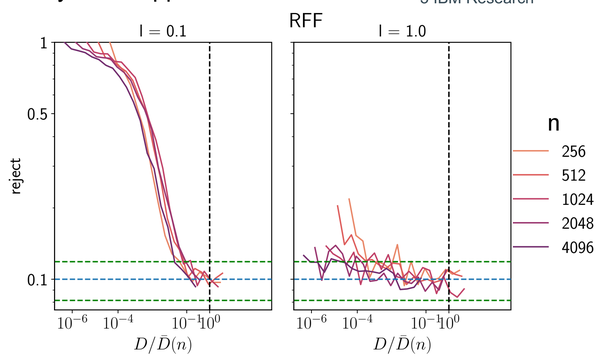

Figure 1. Rejection rate convergence with # RFFs D. Significance level is shown by a blue dashed line and the 95% CI (for converged results) is in green. The range of results obtained from running a Cholesky benchmark is shown by the grey bar. D is rescaled on the x-axis by the upper bound derived above. Vertical black dashed line is at 1.0 indicating where we reach that bound. $\bar{D}(n) = n^2 \log n$.

Contour Integral Quadrature (CIQ)

CIQ is a quadrature algorithm designed to exploit the Cauchy integral formula to approximate functions of square matrices. The most pertinent example in the literature can be found in [1], which derives efficient implementations for matrix-vector products of the form $A^{\pm 1/2}u$.

Lemma (CIQ). To generate a sample of size n satisfying the requirements outlined in the RFF lemma, it is sufficient to use Q quadrature points and J Lanczos iterations, where $Q \geq \mathcal{O}\left(\log\left({}^{n}/_{\eta\sigma_\xi^2}\right)\left(-\log \delta_Q\right)\right)$ and $J \geq \tilde{\mathcal{O}}\left(\frac{\sqrt{n}}{\sqrt{\eta}\sigma_\xi}\log\frac{n}{\sigma_\xi(\epsilon\sigma_\xi\sqrt{1-\eta}-\delta_Q)}\right)$ with $0 < \delta_Q < \epsilon\sigma_\xi\sqrt{1-\eta}$.

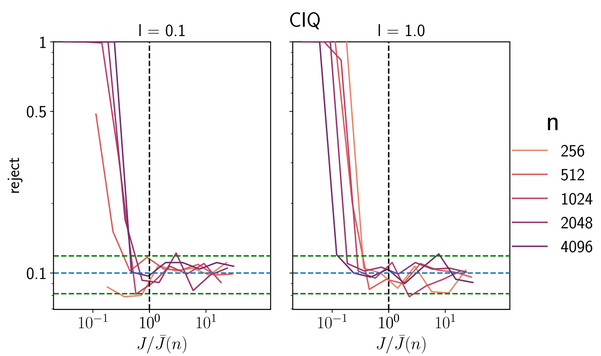

Figure 2. Rejection rate convergence with # Lanczos iterations J. Significance level is in blue and the 95% CI (for converged results) is in green. J is rescaled on the x-axis by the upper bound derived above. The black dashed line indicates this bound. $\bar{J}(n) = \sqrt{n}\log n$.

Since J relies on the condition number of K we expect a preconditioning to improve the efficiency of the algorithm. We call this version PCIQ.

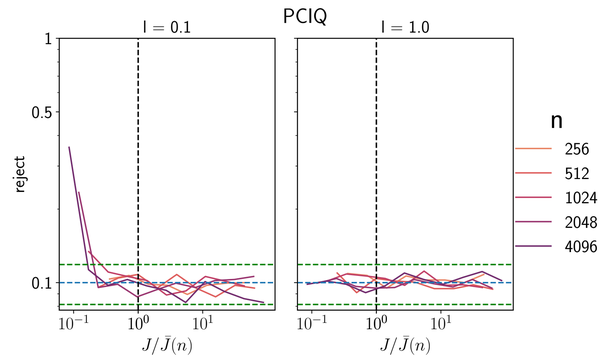

Figure 3. Rejection rate convergence with # Lanczos iterations J. Significance level is in blue and the 95% CI (for converged results) is in green. J is rescaled on the x-axis by the upper bound derived above. The black dashed line indicates this bound. $\bar{J}(n) = n^{3/8}\log n$.

Lemma (PCIQ). To generate a size n sample satisfying the same requirements as before, making use of a rank-k Nyström preconditioner, $J \geq 1 + \frac{\sqrt{\lambda_{k+1}}n^{\frac{3}{8}}}{\sqrt{\eta}\sigma_\xi}\left(\frac{5}{4}\log n - \log\left(\epsilon\sigma_\xi\sqrt{1-\eta} - \delta_Q\right) + C'\right)$ Lanczos iterations will be sufficient, for some constant $C' > 0$. See [5] for further refinements.

Conclusion

We show how to generate approximate samples from any Gaussian Process that, with high probability, cannot be distinguished from a draw from the assumed GP. Bounds on time and space complexity are for the methods considered are given in the table below.

Method	Time	Space
Cholesky	$\mathcal{O}(n^3)$	$\mathcal{O}(n^2)$
RFF	$\mathcal{O}(n^3 \log n)^*$	$\mathcal{O}(n)$
CIQ	$\mathcal{O}\left(n^{5/2}\log n\right)^*$	$\mathcal{O}(n \log n)$
PCIQ	$\mathcal{O}(n^{2.375}\log n)^*$	$\mathcal{O}(n \log n)$

Table 1: Time and space complexity of competing methods of generating draws from a GP. P=with preconditioning. Methods with superscript * represent possibly loose upper bounds we expect can be tightened, particularly in the case of RFF.

References

[1] G. Pleiss, M. Jankowiak, D. Eriksson, A. Damle, and J. Gardner. Fast matrix square roots with applications to gaussian processes and Bayesian optimization. Advances in Neural Information Processing Systems, 33:22268–22281, 2020.
[2] A. Rahimi and B. Recht. Random features for large-scale kernel machines. Advances in neural information processing systems, 20, 2007.
[3] Danica J. Sutherland and Jeff Schneider. On the error of random fourier features. In Proceedings of the Thirty-First Conference on Uncertainty in Artificial Intelligence, UAI'15, page 862–871 Arlington, Virginia, USA, 2015. AUAI Press.
[4] Mikhail Belkin. Approximation beats concentration? An approximation view on inference with smooth radial kernels. https://arxiv.org/pdf/1801.03437, 2018.
[5] A. Stephenson, R. Allison and E. Pyzer-Knapp. Provably reliable large-scale sampling from Gaussian processes. NeurIPS Proceedings 2022.