Kernel Interpolation for Scalable Online Gaussian Processes

NEW YORK UNIVERSITY

Sam Stanton, Wesley Maddox, Ian Delbridge, Andrew Gordon Wilson

Paper: https://arxiv.org/abs/2103.01454

Motivation

- We often need to update our models and make decisions as we are acquiring data in an online (streaming) fashion.
- Predictive *distributions* are especially useful for online decision making, and are a hallmark of Gaussian process models (used in Bayesian optimization, RL, active learning).
- But updating the predictive distributions are computationally quite expensive.

Contribution:

- We propose **WISKI** (Woodbury interpolation with SKI) 🥃 that uses SKI kernel matrices to enable exact GP online updates in time constant in the number of data points.

Background

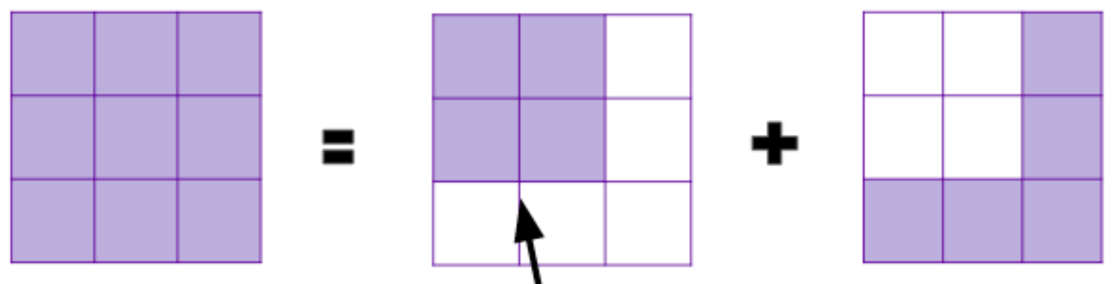
- GP predictive equations require at least $\mathcal{O}(N^2)$ space and computation (with iterative methods)

$$p(y|\mathbf{x}^*, \mathcal{D}, \theta) = \mathcal{N}(y; \mu(\mathbf{x}^*), \Sigma(\mathbf{x}^*))$$

$$\mu(\mathbf{x}^*) = K_{\mathbf{x}^* X}(K_{XX} + \sigma^2 I)^{-1}\mathbf{y}$$

$$\Sigma(\mathbf{x}^*) = K_{\mathbf{x}^*\mathbf{x}^*} - K_{\mathbf{x}^* X}(K_{XX} + \sigma^2 I)^{-1} K_{X\mathbf{x}^*}$$

Adding a new data point expands the kernel matrix, costing at least $\mathcal{O}(N)$ time



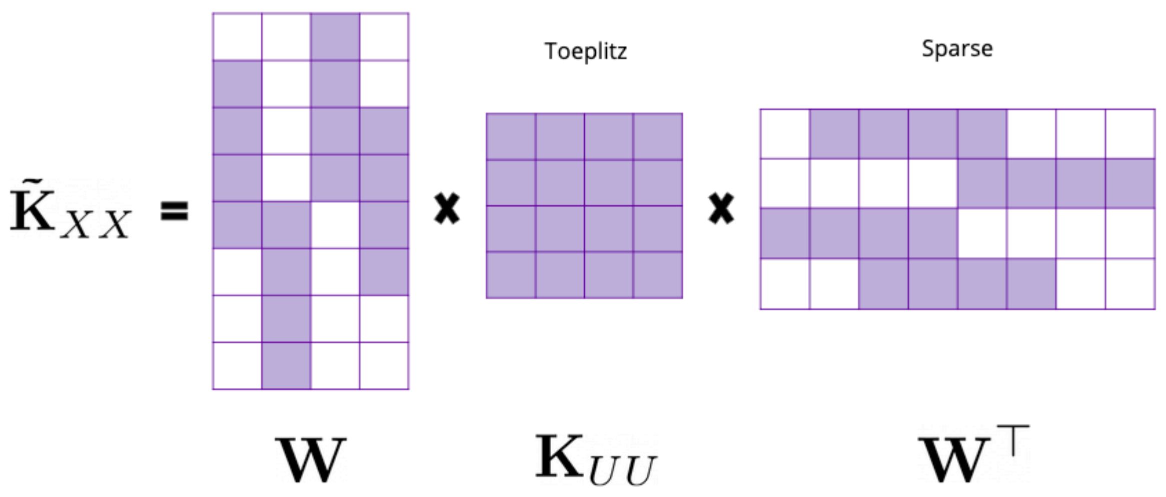
$(\mathbf{K}_{X'X'} + \sigma^2 I)$ $\quad$ $(\mathbf{K}_{XX} + \sigma^2 I)$ $\quad$ $\begin{bmatrix} \mathbf{0} & k(X, \mathbf{x}') \\ k(\mathbf{x}', X) & k(\mathbf{x}', \mathbf{x}') + \sigma^2 \end{bmatrix}$

Time complexities: M is the # of inducing points

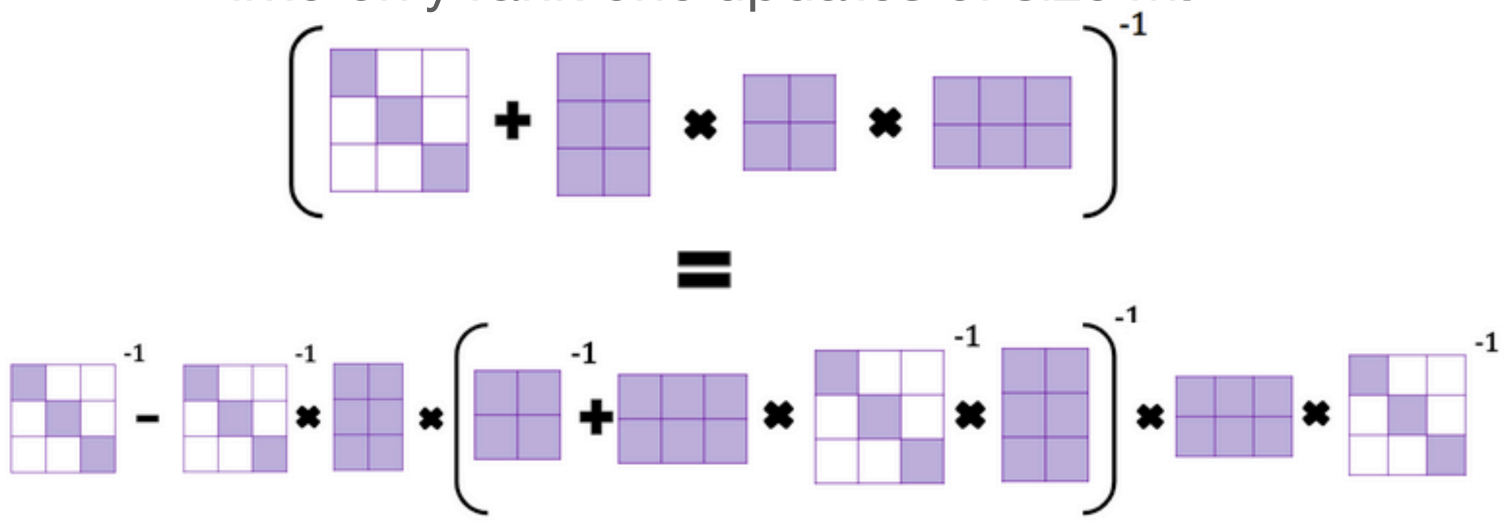
	Exact GP	WISKI GP	O-SVGP (Bui et al, '17)
Parameter Inference	$\mathcal{O}(N^3)$	$\mathcal{O}(M^2)$	$\mathcal{O}(M^3)$
Conditioning on New Data	$\mathcal{O}(N^2)$	$\mathcal{O}(M^2)$	–
Querying Test Points	$\mathcal{O}(N^2)$	$\mathcal{O}(M^2)$	$\mathcal{O}(M^2)$
Data Storage	$\mathcal{O}(N)$	$\mathcal{O}(1)$	$\mathcal{O}(1)$

Methodology

Through a careful combination of caching and structured kernel interpolation (SKI), we enable online updates in *constant time* with respect to the number of data points n, while *retaining exact inference*.



We use Woodbury's matrix identity to invert the SKI kernel matrix and convert it into only rank one updates of size m.



$$(A + UCV)^{-1} = A^{-1} - A^{-1}U(C^{-1} + VA^{-1}U)^{-1}VA^{-1}$$

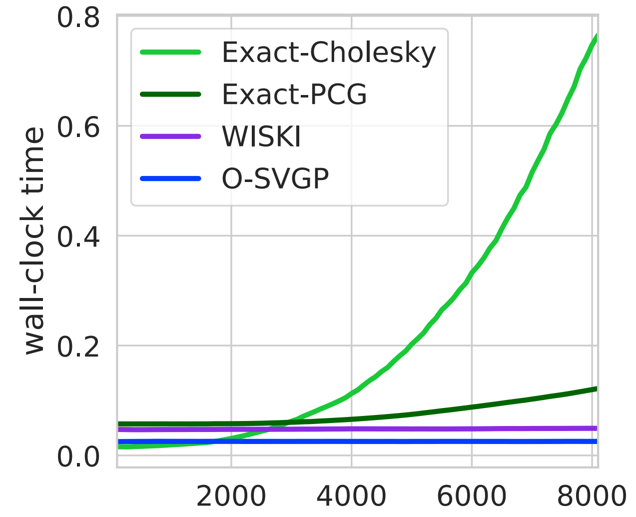
References

Pleiss et al, Constant Time Predictive Distributions for Gaussian Processes, ICML, 2018.
Gardner et al, GPyTorch: Blackbox Matrix Matrix Gaussian Process Inference, NeurIPS, 2018.
Wilson & Nickisch, Kernel Interpolation for Scalable Structured Gaussian Processes, ICML, 2015.
Bui et al, Streaming Sparse Variational Approximations, NeurIPS, 17.

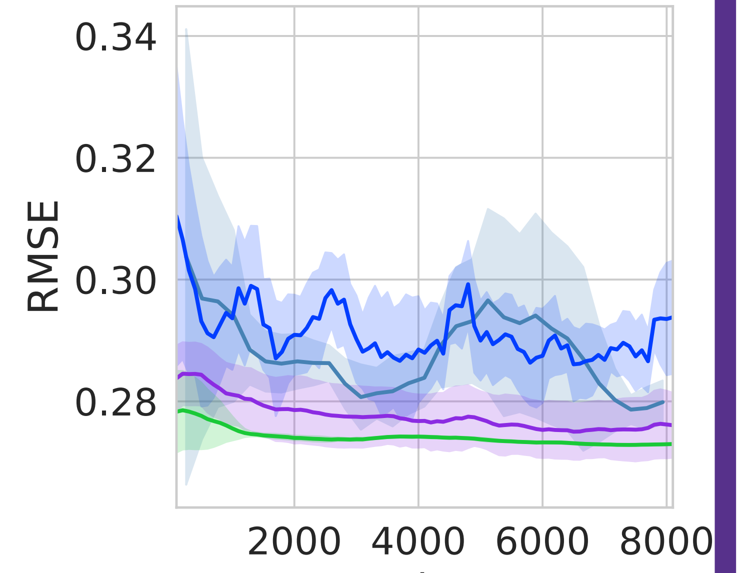
Code at https://github.com/wjmaddox/online_gp

Results

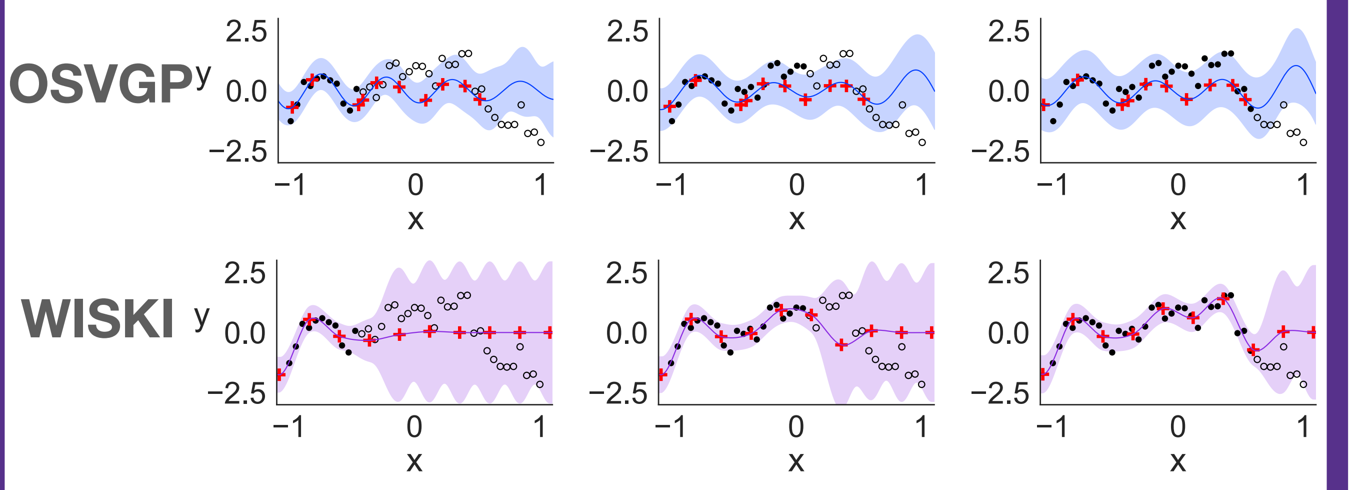
WISKI has constant time inference like variational methods, but unlike traditional exact GPs.



WISKI performs particularly well for incremental regression tasks.



WISKI maintains the advantages of exact inference in the non i.i.d setting



Finally, WISKI makes it uniquely possible to find an optimal set of new locations for experimental design in constant time.