# Low Rank Approximation and Gaussian Conditioning

## Gaussian Conditioning

The strategy we employ here is transforming the problem into a form where independence simplifies the conditioning. So let

$$\begin{bmatrix} Y \\ X \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} \mu_Y \\ \mu_X \end{bmatrix}, \begin{bmatrix} \Sigma_{YY} & \Sigma_{YX} \\ \Sigma_{XY} & \Sigma_{XX} \end{bmatrix}\right)$$

with $\Sigma_{XX}$ invertible. Define

$$W = Y - \Sigma_{YX}\Sigma_{XX}^{-1}X.$$

Since

$$\mathrm{Cov}(W, X) = \Sigma_{YX} - \Sigma_{YX}\Sigma_{XX}^{-1}\Sigma_{XX} = 0,$$

$W$ and $X$ are independent. Hence, we can write

$$Y = \Sigma_{YX}\Sigma_{XX}^{-1}X + W.$$

Conditioning on $X = x$ yields

$$Y \mid X = x \sim \mathcal{N}\Big(\mu_Y + \Sigma_{YX}\Sigma_{XX}^{-1}(x - \mu_X),\ \Sigma_{YY} - \Sigma_{YX}\Sigma_{XX}^{-1}\Sigma_{YX}^T\Big).$$

## Gaussian Process Regression

Assume a joint model

$$(X, Y) \sim N\left(0, \begin{bmatrix} K & K \\ K & K + \Sigma \end{bmatrix}\right)$$

where $K$ is a covariance kernel. Use Gaussian conditioning to take

$$X \mid Y = y \sim N\left(K(K+\Sigma)^{-1}y, K - K(K+\Sigma)^{-1}K\right)$$

We can augment the covariance matrix to also predict on new points with uncertainty.
Two commmon kernels.

1. Gaussian kernel:

$$K(x, y) := \exp\left(-\|x - y\|^2 / l^2\right)$$

where $l$ is a hyperparameter, usually called the length scale.

2. Polynomial kernel. For $p \in \mathbb{N}$,

$$K(x, y) = (1 + x^\top y)^p$$

which is equivalent to performing degree $p$ polynomial interpolation.

## Nyström Approximation

- Given a PSD matrix $A \in \mathbb{R}^{N\times N}$ and an arbitrary $N \times k$ "test" matrix $\Omega$ (with $k < N$) the Nyström approximation is defined as

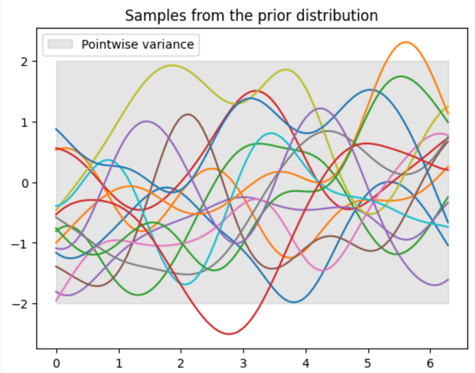
$$A\langle\Omega\rangle = A\Omega\,(\Omega^T A \Omega)^{-1}\,\Omega^T A$$

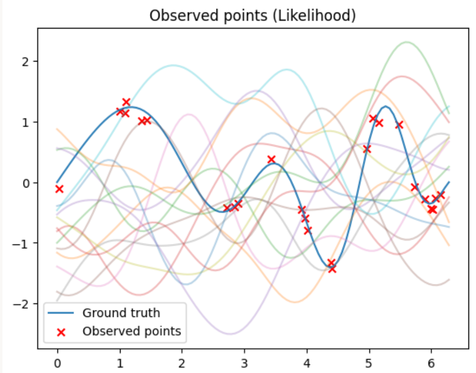
## *Theorem*

*The Nyström is the best positive semi-definite under-approximation of $A$ ($A - A\langle\Omega\rangle \succeq 0$): It minimizes the trace norm $\mathrm{trace}(A - \hat{A})$ over all PSD matrices $\hat{A}$ with positive residual $A - \hat{A}$ and spanned by the columns of $A\Omega$.*

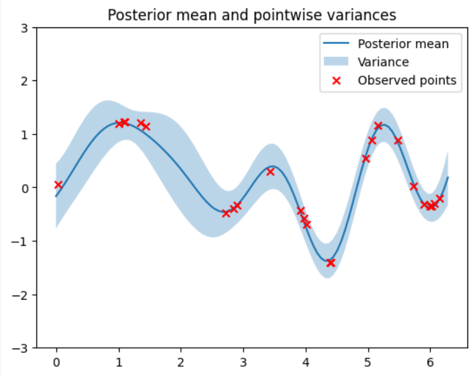
- **Important case:** If $\Omega = \begin{bmatrix} -I_k- \\ -0- \end{bmatrix}$ then $A\Omega$ is simply the first $k$ columns of $A$, known as column subset selection matrix.

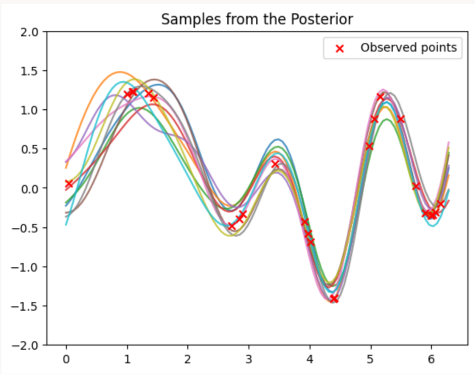
## Gaussian Process Regression, Visualized




- We may obtain prior functions by sampling the values of the function $f \sim \mathcal{N}(\mu, \Sigma)$ at a finite number of the $x$ coordinates.
- With noisy ground truth observations, we condition on these measurements to obtain a **posterior distribution** representing our updated beliefs.




## Cholesky Factorization and Randomly Pivoted Cholesky

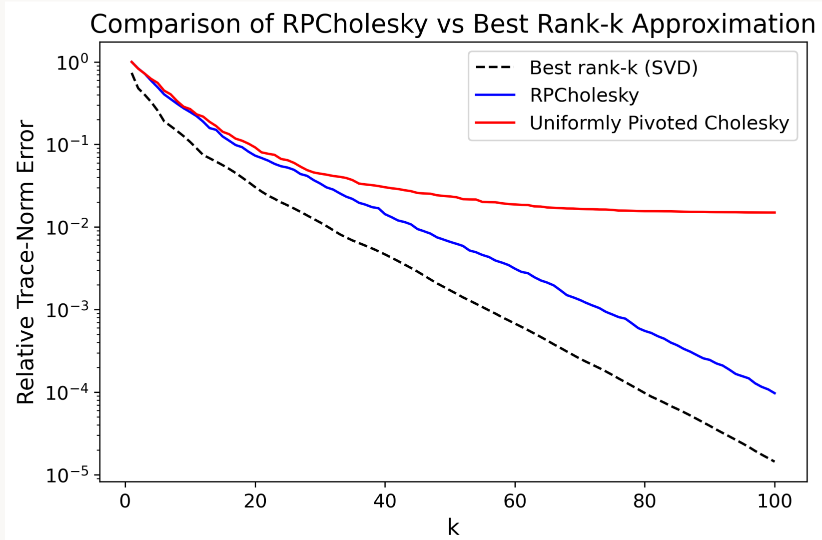
The Cholesky Factorization is a decomposition of a symmetric positive definite matrix $A \in \mathbb{R}^{N\times N}$, into $A = LL^*$ where $L \in \mathbb{R}^{N\times N}$ is a lower triangular matrix. We construct $L$ column by column, accounting for each column of $A$ from left to right.
Rather than working with the columns of $A$ in order, we can choose specific pivot columns and use the Cholesky algorithm to construct a low-rank approximation to $A$. **This algorithm updates a running Nyström approximation one column at a time.** RPCholesky is a smart way of choosing which column to add next to reduce the error of the approximation.

**Algorithm** RPCholesky

**Input:** Psd matrix $A \in \mathbb{C}^{N\times N}$; approximation rank $k$
**Output:** Pivot set $S = \{s_1, \ldots, s_k\}$; matrix $F \in C^{N\times k}$ defining Nyström approximation $\hat{A} = FF^*$

Initialize $F \leftarrow 0_{N\times k}$ and $\mathbf{d} \leftarrow \mathrm{diag}\, A$
**for** $i = 1$ to $k$ **do**
  Sample pivot $s_i \sim \mathbf{d}/\sum_{j=1}^N \mathbf{d}(j)$ ▷ Pick pivot
  $\mathbf{g} \leftarrow A(:, s_i)$ ▷ Evaluate column $s$ of input matrix
  $\mathbf{g} \leftarrow \mathbf{g} - F(:, 1:i-1)F(s_i, 1:i-1)^*$
  $F(:, i) \leftarrow \mathbf{g}/\sqrt{\mathbf{g}(s_i)}$ ▷ Update approximation
  $\mathbf{d} \leftarrow \mathbf{d} - |F(:, i)|^2$ ▷ Update diagonal of residual matrix
  $\mathbf{d} \leftarrow \max\{\mathbf{d}, \mathbf{0}\}$ ▷ Ensure diagonal remains nonnegative
**end for**

## Convergence of RPCholesky





**Alex Undergrad, Raaga, Chenab**  **Mentored by: Alex Hsu, Bamdad Hosseini**

University of Washington