Scaled eigenvalue method

- Approximately eigenvalues \( \lambda_i \) of \( K_{XX} \) using the \( n \) largest eigenvalues \( \mu_i \) of \( K_{YY} \) on a full grid with \( m \) points such that \( X \subset Y \):
  \[
  \log |K_{XX} + \alpha I| = \sum \log |\lambda_i + \alpha^2| = \sum \log \left( \frac{n}{m} \mu_i + \alpha^2 \right)
  \]
- Can handle non-Gaussian likelihoods via the Fielder bound

Stochastic trace estimation

- Our goal is to estimate, for a symmetric positive definite matrix \( \mathcal{K} \):
  \[
  \mathcal{K}_{[K]} = \frac{1}{2} \text{Tr}(\text{log}(\hat{\mathcal{K}})) \quad \text{and} \quad \frac{\partial \mathcal{K}_{[K]}}{\partial \theta} = -\frac{1}{2} (\partial K/\partial \theta)^T \hat{\mathcal{K}}^{-1} (\partial K/\partial \theta) \]
- Stochastic expression for \( \mathcal{K}_{[K]} \) and first derivatives:
  \[
  \mathcal{K}_{[K]} = \frac{1}{2} \mathbb{E} [z^T (\text{log}(\hat{\mathcal{K}})) z] \quad \frac{\partial \mathcal{K}_{[K]}}{\partial \theta} = -\frac{1}{2} (\partial K/\partial \theta)^T \hat{\mathcal{K}}^{-1} (\partial K/\partial \theta) \]
- Common choices of probe vectors:
  - Hutchinson: \( z_i \sim \mathcal{N}(0, 1) \)
  - Gaussian: \( z_i \sim \mathcal{N}(0, 1) \)
- Estimate via sample means with several random probe vectors
- Need to multiply log(\( \hat{\mathcal{K}} \)) by probe vectors efficiently

Lanczos

- Function application with fast MVMs \( \implies \text{try Lanczos} \)
- Lanczos on \( \mathcal{K} \) computes partial tri diagonalization:
  \[
  \mathcal{K}_Q = Q_0 \mathcal{K} Q_0^T \quad Q_0 = (q_0 \ldots q_n) \quad \mathcal{K} = (\beta_0 \ldots \beta_n) \quad \mathcal{K}_Q = (\mathcal{K}_{Q_0} \ldots \mathcal{K}_{Q_n})
  \]
  - \( \mathcal{K}_Q \) is the Cholesky factor of \( \mathcal{K} \) via bi-orthogonalization

Chebyshev

- Based on a polynomial approximation of the log
- Minimizes the worst-case error over an interval
- Lanczos is sensitive to the locations of the eigenvalues and tends to yield better accuracy

Fast MVMs

- Our experiments use structured kernel interpolation (SKI)
  \[
  K_{XX} \approx W K_{UW} W^T
  \]
- \( W \) is a sparse \( n \times m \) matrix of interpolation weights
- The points \( U \) are referred to as inducing points
- \( > 1D \): rectilinear grid, product covariance \( \implies \text{Kronecker structure} \)
- \( 1D \): regular grid, stationary covariance \( \implies \text{Toeplitz structure} \)
- \( > 1D \): regular grid, stationary covariance \( \implies \text{BTB structure} \)

Diagonal correction

- SKI may provide a poor estimate of the diagonal entries
- Modify the SKI approximation to add a diagonal matrix \( D \):
  \[
  K_{XX} \approx W K_{UW} W^T + D \quad \text{diag}(K_{XX}) = \text{diag}(W K_{UW} W^T + D)
  \]
- Not supported by scaled eigenvalues, works with our MVM based approach