

Numerical Methods for Data Science: Scalable Kernel Methods, Part II

David Bindel

19 June 2019

Department of Computer Science
Cornell University

Kernel Approximation

Goal: Approximate $f : \Omega \subset \mathbb{R}^d \rightarrow \mathbb{R}$.

Input: (Possibly noisy) samples $y_i = f(x_i) + \epsilon_i$.

Approximation scheme: $f(x) \approx s(x)$ where

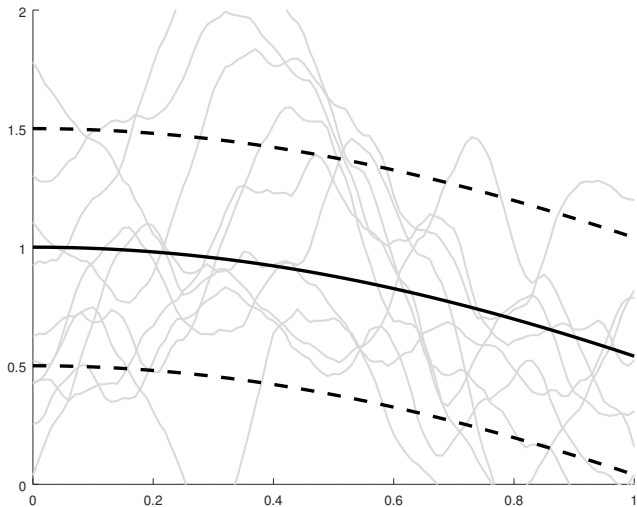
$$s(x) = \sum_{j=1}^n k(x, x_j) c_j + \sum_{j=1}^m b_j(x) \lambda_j$$

with a discrete orthogonality condition for each b_j :

$$\sum_{j=1}^m b_i(x_j) c_j = 0.$$

Last hour: several explanations for why this makes sense.

Picture Today: Gaussian Processes (GPs)



Gaussian Processes (GPs)

Our favorite continuous distributions over

$$\mathbb{R}: \quad \text{Normal}(\mu, \sigma^2), \quad \mu, \sigma^2 \in \mathbb{R}$$

$$\mathbb{R}^n: \quad \text{Normal}(\mu, C), \quad \mu \in \mathbb{R}^n, C \in \mathbb{R}^{n \times n}$$

$$\mathbb{R}^d \rightarrow \mathbb{R}: \quad \text{GP}(\mu, k), \quad \mu: \mathbb{R}^d \rightarrow \mathbb{R}, k: \mathbb{R}^d \times \mathbb{R}^d \rightarrow \mathbb{R}$$

More technically, define GPs by looking at finite sets of points:

$$\forall X = (x_1, \dots, x_n), x_i \in \mathbb{R}^d,$$

have $f_X \sim N(\mu_X, K_{XX})$, where

$$f_X \in \mathbb{R}^n, \quad (f_X)_i \equiv f(x_i)$$

$$\mu_X \in \mathbb{R}^n, \quad (\mu_X)_i \equiv \mu(x_i)$$

$$K_{XX} \in \mathbb{R}^{n \times n}, \quad (K_{XX})_{ij} \equiv k(x_i, x_j)$$

When X is unambiguous, we will sometimes just write K .

Kernel Functions

Call the *kernel* (or *covariance*) function k . Required¹ property:

- **Pos def:** K_{XX} is always positive definite

Often desirable:

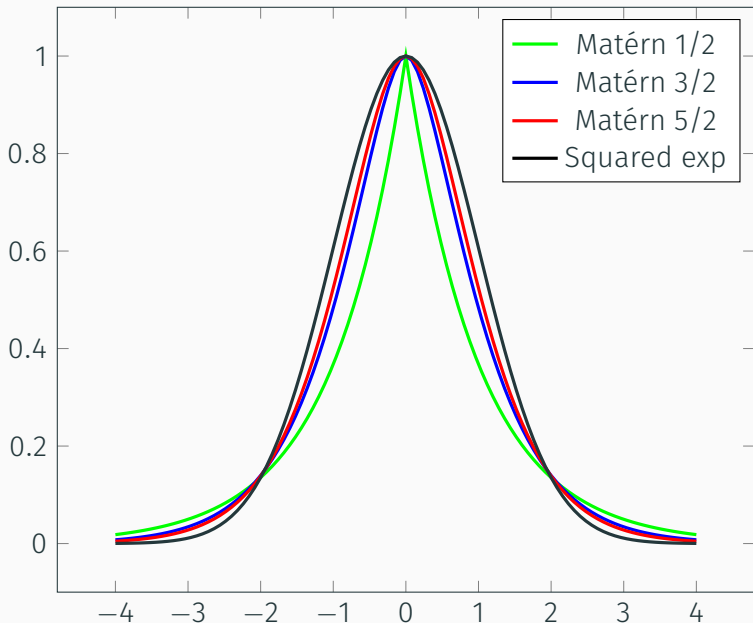
- **Stationary:** $k(x, y)$ depends only on $x - y$
- **Isotropic:** $k(x, y)$ depends only on x and $\|x - y\|$

Often want both (sloppy notation: $k = k(r)$).

Common examples (e.g. Matérn, SE) also depend on *hyper-parameters* θ — suppressed in notation unless needed.

¹Conditionally pos def is interesting, but we'll skip it today.

Matérn and SE kernels



Choosing Kernels

Kernel is *chosen by modeler*

- Choose Matérn / SE for simplicity
- Rarely have the intuition to pick the “right” kernel
- Common choices are *universal* — can recover anything
 - Requires less data for “good” choice (inductive bias)
 - Roughly: smooth kernels good for smooth functions
- Common kernel families have *hyperparameters*
 - Length scale (almost all)
 - Smoothness parameter (Matérn)
- Also have hyperparameter for noise level (“nugget”)

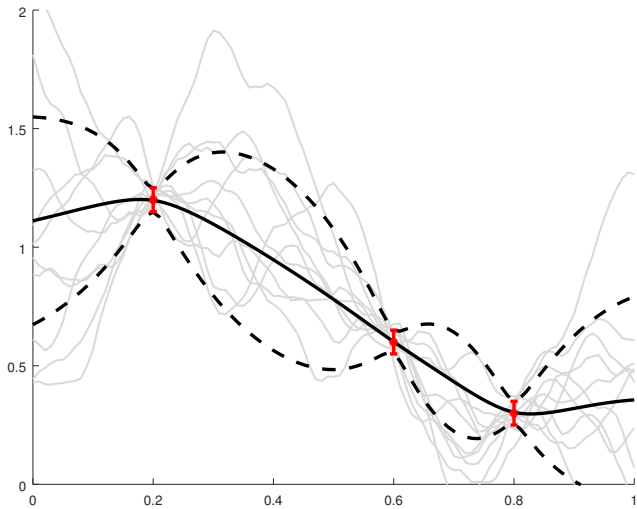
Would like to choose hyperparameters using data.

Basic tasks in function approximation via a kernel scheme

- **Select a kernel and hyperparameters (length scale, noise)**
- **Use data to fit model parameters**
- Evaluate the model at new points
- Evaluate predictive variance at new points

For optimization: want gradients of everything

Being Bayesian



Being Bayesian

Now consider prior of $f \sim \text{GP}(\mu, k)$, noisy measurements

$$f_X \sim y + \epsilon, \quad \epsilon \sim N(0, W), \quad \text{typically } W = \sigma^2 I$$

Posterior is $f \sim \text{GP}(\mu', k')$ with

$$\begin{aligned} \mu'(x) &= \mu(x) + K_{xx}c & \tilde{K} &= K_{xx} + W \\ k'(x, x') &= K_{xx'} - K_{xx}\tilde{K}^{-1}K_{xx'} & c &= \tilde{K}^{-1}(y - \mu_X) \end{aligned}$$

The expensive bit: solves with \tilde{K} .

Dealing with Derivatives

Derivative information \implies multi-output GP:

$$\mu^\nabla(x) = \begin{bmatrix} \mu(x) \\ \partial_x \mu(x) \end{bmatrix}, \quad k^\nabla(x, x') = \begin{bmatrix} k(x, x') & \nabla'_x k(x, x') \\ \partial_x k(x, x') & \partial_{x, x'}^2 k(x, x') \end{bmatrix}$$

Requires a little care, but ideas in this talk generalize.

NB: Other (linear) measurements also allowed – e.g. integrals.

Kernel Hyper-parameters: MLE

How to estimate *hyper-parameters* θ ?

- Bayesian approach? Expensive...
- Usually just do maximum likelihood estimation (MLE)

Likelihood function is same as for a multivariate normal:

$$\ell(\theta|y) = \frac{1}{\sqrt{\det(2\pi\tilde{K})}} \exp\left(-\frac{1}{2}(y - \mu_X)^T \tilde{K}^{-1}(y - \mu_X)\right).$$

Of course, we usually work with *log-likelihood* and derivatives.

Kernel Hyper-parameters: MLE

How to estimate *hyper-parameters* θ ?

- Bayesian approach? Expensive...
- Usually just do maximum likelihood estimation (MLE)

Log-likelihood function for kernel hypers θ

$$\mathcal{L}(\theta|y) = \mathcal{L}_y + \mathcal{L}_{|K|} - \frac{n}{2} \log(2\pi)$$

where (again with $c = \tilde{K}^{-1}(y - \mu_X)$)

$$\mathcal{L}_y = -\frac{1}{2}(y - \mu_X)^T c, \quad \frac{\partial \mathcal{L}_y}{\partial \theta_i} = \frac{1}{2} c^T \left(\frac{\partial \tilde{K}}{\partial \theta_i} \right) c$$

$$\mathcal{L}_{|K|} = -\frac{1}{2} \log \det \tilde{K}, \quad \frac{\partial \mathcal{L}_{|K|}}{\partial \theta_i} = -\frac{1}{2} \text{tr} \left(\tilde{K}^{-1} \frac{\partial \tilde{K}}{\partial \theta_i} \right)$$

Alternative to MLE: Leave One Out Cross-Validation (LOOCV)

$$LOOCV = \sum_{i=1}^n \left(r^{(-i)} \right)^2$$

where

$$r^{(-i)} = y_i - s^{(-i)}(x)$$

$s^{(-i)}$ = fit to all data points but x_i

Kernel Hyper-parameters: LOOCV Trick

Fitting problem for all the data:

$$\tilde{K}c = y - \mu_X$$

Leave out point i : $c_i^{(-i)} = 0$ and $r^{(-i)} = y_i - \mu(x_i) - K_{x_i X} c^{(-i)}$.

Write as

$$\begin{bmatrix} \tilde{K} & e_i \\ e_i^T & 0 \end{bmatrix} \begin{bmatrix} c^{(-i)} \\ r^{(-i)} \end{bmatrix} = \begin{bmatrix} y \\ 0 \end{bmatrix}$$

where e_i denotes column i of an identity matrix.

Gaussian elimination gives

$$r^{(-i)} = \frac{c_i}{\left[\tilde{K}^{-1}\right]_{ii}}$$

NB: $\left[\tilde{K}^{-1}\right]_{ii}$ is the leave one out predictive variance.

Generalized Cross-Validation

Problem with LOOCV: no weights on residuals!

Alternative is Generalized Cross-Validation (Golub, Heath, Wahba 79):

$$v(\lambda) = \frac{\frac{1}{n} \|(I - A(\lambda))(y - \mu_X)\|^2}{\left[\frac{1}{n} \text{tr}(I - A(\lambda))\right]^2}$$

where $A(\lambda)$ is the influence matrix.

In case of $\tilde{K} = K + \sigma^2 I$, have

$$I - A(\lambda) = I - K\tilde{K}^{-1} = \sigma^2 \tilde{K}^{-1}$$

Therefore

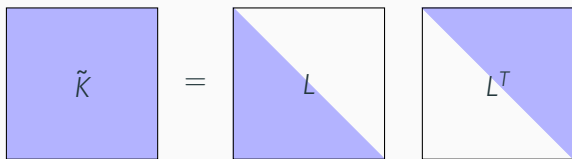
$$V = n \frac{\|c\|^2}{\text{tr}(\tilde{K}^{-1})^2}$$

Derivative of $V = n\|c\|^2/\text{tr}(\tilde{K}^{-1})^2$ is tedious, not difficult:

$$\delta V = n \frac{\|c\|^2 \text{tr}(\tilde{K}^{-1}(\delta\tilde{K})\tilde{K}^{-1}) - (c^T \tilde{K}^{-1}(\delta\tilde{K})c) \text{tr}(\tilde{K}^{-1})}{\text{tr}(\tilde{K}^{-1})^3}$$

Computational issues for GCV are similar to those for MLE.
Focus on latter.

Learning Parameters and Hypers: Small n


$$\tilde{K} = L L^T$$

The diagram illustrates the Cholesky decomposition of a symmetric positive definite matrix \tilde{K} . It is represented as the product of a lower triangular matrix L and its transpose L^T . The matrix \tilde{K} is shown as a solid blue square. The matrix L is shown as a square with a blue lower triangular region and a white upper triangular region. The matrix L^T is shown as a square with a white lower triangular region and a blue upper triangular region.

In an optimization loop:

```
1  L = chol(Ktilde(theta));  
2  c = L' \ (L \ (y - mu));  
3  liky   = -0.5 * ((y - mu)' * c);  
4  likK   = -sum(log(diag(L)));  
5  for i = 1:len(theta)  
6      dliky = 0.5 * (c' * dK(theta, i) * c);  
7      dlikK = -0.5 * trace(L' \ (L \ dk(theta, i)));  
8  end
```

Scalability Bottlenecks

Consider n data points

- Straightforward regression: factor \tilde{K} at $O(n^3)$ cost
- Kernel hyper MLE requires multiple $O(n^3)$ ops
 - To compute $\log \det \tilde{K}$ is $O(n^3)$ per step
 - To compute $\text{tr} \left(\tilde{K}^{-1} \frac{\partial \tilde{K}}{\partial \theta_i} \right)$ is $O(n^3)$ per hyper per step
- Cost of kernel hyper GCV is similar (lots of costly traces)

Two possible work-arounds

- Data-sparse factorization methods
- Methods that avoid factorization (e.g. iterative solvers)
 - Q: how to handle determinants and traces?

Several possible approaches:

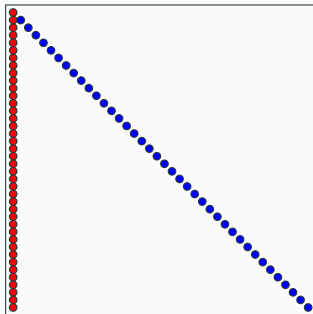
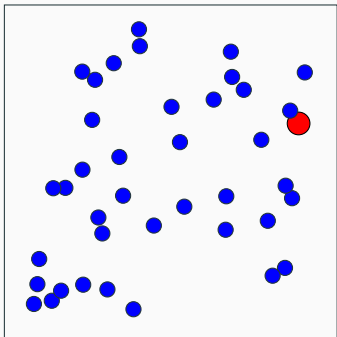
- Very smooth kernels \implies low rank K
- Compactly supported kernels \implies sparse K
- Can piece together local kernel approximations
- Rank-structured factorization methods
- Black-box multiplication (via FMM or FFT)

We will focus on first and last.

Scaling GPs: Factorization approach

Simplest Data-Sparse Approach

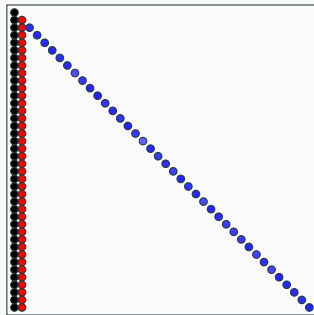
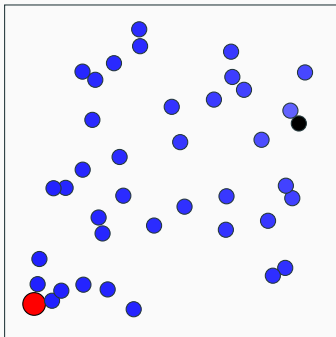
For K (nearly) low rank: *partial pivoted Cholesky*



Diagonal element: 1.00e+00

Simplest Data-Sparse Approach

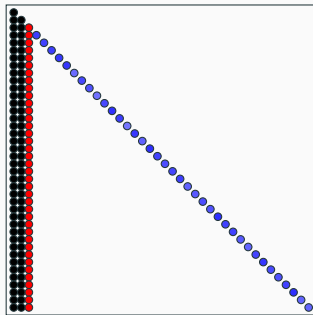
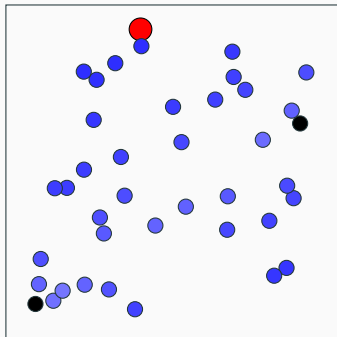
For K (nearly) low rank: *partial pivoted Cholesky*



Diagonal element: 6.77e-02

Simplest Data-Sparse Approach

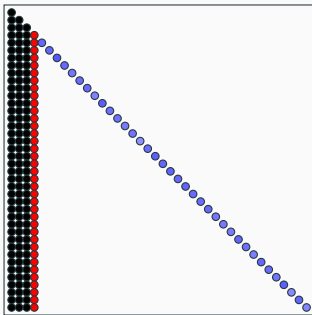
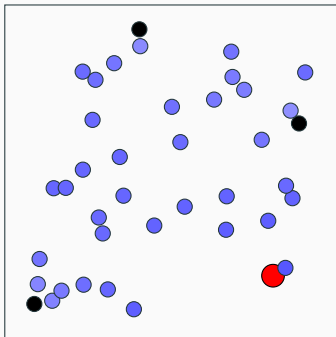
For K (nearly) low rank: *partial pivoted Cholesky*



Diagonal element: 1.91e-02

Simplest Data-Sparse Approach

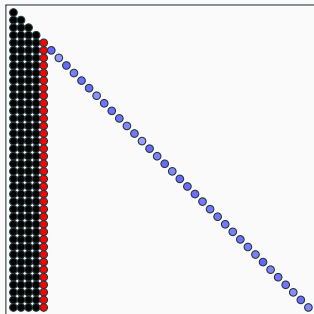
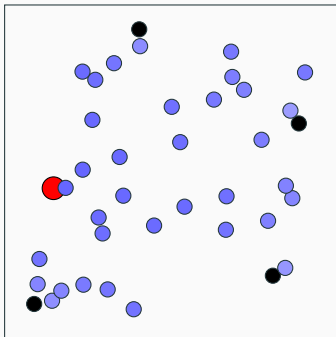
For K (nearly) low rank: *partial pivoted Cholesky*



Diagonal element: 5.11e-04

Simplest Data-Sparse Approach

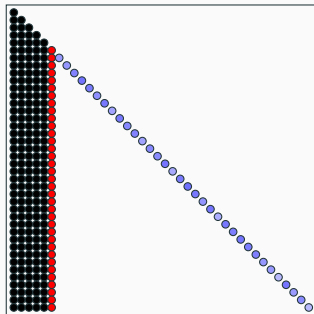
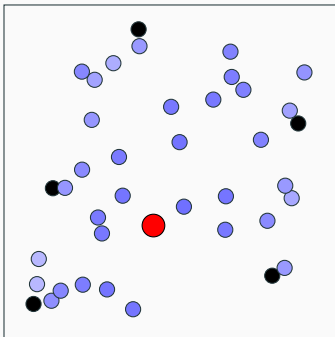
For K (nearly) low rank: *partial pivoted Cholesky*



Diagonal element: 1.19e-04

Simplest Data-Sparse Approach

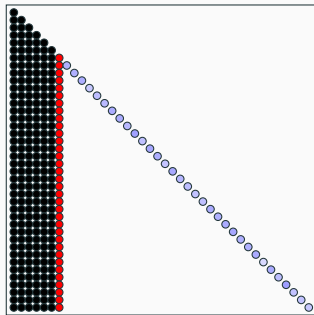
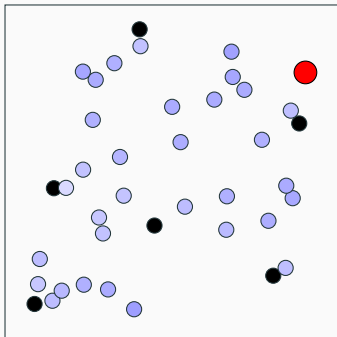
For K (nearly) low rank: *partial pivoted Cholesky*



Diagonal element: 4.18e-05

Simplest Data-Sparse Approach

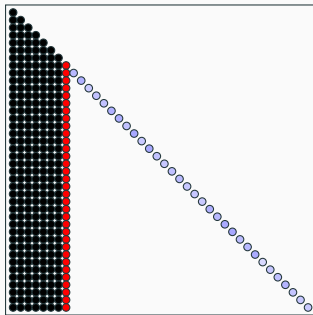
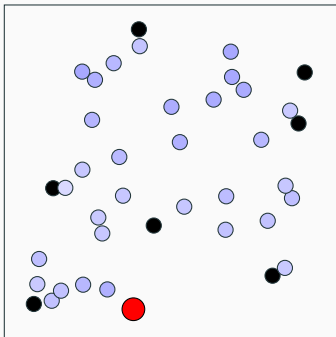
For K (nearly) low rank: *partial pivoted Cholesky*



Diagonal element: $8.54\text{e-}07$

Simplest Data-Sparse Approach

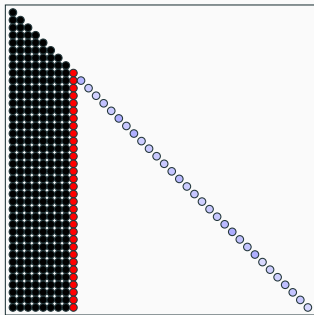
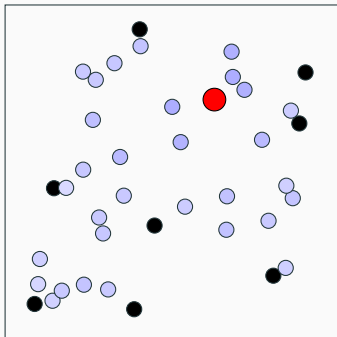
For K (nearly) low rank: *partial pivoted Cholesky*



Diagonal element: 3.58e-07

Simplest Data-Sparse Approach

For K (nearly) low rank: *partial pivoted Cholesky*



Diagonal element: 1.92e-07

Simplest Data-Sparse Approach

Smooth kernel and long length scales: K nearly rank $m \ll n$

- $P(LL^T)P^T$ = partial pivoted Cholesky (select m inducing points); does *not* require forming K_{XX}
- Solve $P(LL^T + \sigma^2 I)P^T c = f_X$ stably by reformulating c as a scaled regularized least squares residual:

$$\text{minimize } \|f_X - Lw\|^2 + \sigma^2 \|w\|^2, \quad c = \sigma^{-2} (f_X - Lw)$$

- Compute $\log \det \tilde{K} = \log \det(L^T L + \sigma^2 I) + 2(n - m) \log \sigma$; similar cheap rearrangement for derivatives.
- Prediction and predictive variance are also cheap.

- Don't have to be purely greedy (swaps as in CUR/ID)
- Can make it depend on data (LARS or Lasso approaches)
- Piv Chol good for *preconditioning* black-box approaches

Beyond Low Rank

If K is not low rank, can still use *rank-structured* factorization.

Example (1D):

$$\tilde{K} \approx \begin{bmatrix} \tilde{K}_{11} & U_1 U_2^T \\ U_2 U_1^T & \tilde{K}_{22} \end{bmatrix}$$

Can solve by recognizing “disguised sparsity”:

$$\tilde{K}c = y \quad \Rightarrow \quad \begin{bmatrix} \tilde{K}_{11} & 0 & U_1 & 0 \\ 0 & \tilde{K}_{22} & 0 & U_2 \\ U_1 & 0 & 0 & -I \\ 0 & U_2 & -I & 0 \end{bmatrix} \begin{bmatrix} c_1 \\ c_2 \\ \gamma_1 \\ \gamma_2 \end{bmatrix} = \begin{bmatrix} y_1 \\ y_2 \\ 0 \\ 0 \end{bmatrix}$$

Apply recursively for an $O(n \log n)$ solver. More sophisticated options available, extensions to multi-dimensional, etc.

Scaling GPs: Black box approach

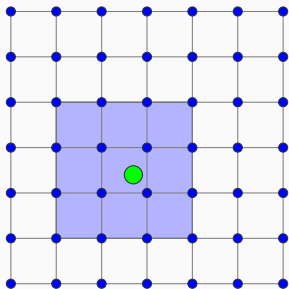
Basic ingredients for “black box” approach

- Fast MVMs with kernel matrices
- Krylov methods for linear solves and matrix functions
- Stochastic estimators: trace, diagonal, and other

Kernel approximations

- Low-rank approximation (via *inducing variables*)
 - Non-smooth kernels, small length scales \implies large rank
 - Only semi-definite
- Sparse approximation
 - OK with SE kernels and short length scales
 - Less good with heavy tails or long length scales
 - May again lose definiteness
- More sophisticated: fast multipole, Fourier transforms
 - Same picture as in integral eq world (FMM, PFFT)
 - Main restriction: low dimensional spaces (2-3D)
- Kernel a model choice — how does approx affect results?

Example: Structured Kernel Interpolation (SKI)



Write $K_{XX} \approx W^T K_{UU} W$ where

- U is a uniform mesh of m points
- K_{UU} has Toeplitz or block Toeplitz structure
- Sparse W interpolates values from X to U

Apply K_{UU} via FFTs in $O(m \log m)$ time.

The power of fast MVMs

With MVMs alone, natural to explore nested *Krylov subspaces*:

$$\mathcal{K}_{d+1}(\tilde{K}, b) = \text{span}\{b, \tilde{K}b, \tilde{K}^2b, \dots, \tilde{K}^db\} = \{p(\tilde{K})b : p \in \mathcal{P}_k\}$$

Lanczos process: *expansion + Gram-Schmidt*

$$\beta_j q_{j+1} = \tilde{K}q_j - \alpha_j q_j - \beta_{j-1} q_{j-1}$$

Lanczos factorization: $\tilde{K}Q_k = Q_k \bar{T}_k$ where

$$Q_k = \begin{bmatrix} q_1 & q_2 & \dots & q_k \end{bmatrix},$$
$$\bar{T}_k = \begin{bmatrix} \alpha_1 & \beta_1 & & & \\ \beta_1 & \alpha_2 & \beta_2 & & \\ & \beta_2 & \alpha_3 & \beta_3 & \\ & & \ddots & \ddots & \ddots \\ & & & \beta_{k-1} & \alpha_k \\ \hline & & & & \beta_k \end{bmatrix} = \begin{bmatrix} T_k \\ \beta_k e_k^T \end{bmatrix}$$

The power of fast MVMs

Fast MVM with symmetric $\tilde{K} \implies$ try Lanczos!

- Incrementally computes $\tilde{K}Q = QT$ where
 - Q has orthonormal columns
 - Leading k columns span k -dim Krylov space
 - T is tridiagonal
- Building block for
 - Solving linear systems (CG)
 - Approximating eigenvalues
 - Approximating matrix functions: $f(\tilde{K})b$
 - Quadrature vs spectral measure for \tilde{K}
- Fast (three-term recurrence) and elegant...
- ... but not forward stable in finite precision

Function application via Lanczos

A computational kernel: $f(\tilde{K})b$

- Run Lanczos from starting vector $b/\|b\|$
- At n steps in exact arithmetic,

$$f(\tilde{K})b = Qf(T)Q^Tb = \|b\|Qf(T)e_1$$

- Truncate at $k \ll n$ steps, use

$$f(\tilde{K})b \approx \|b\|Q_kf(T_k)e_1$$

- Error analysis hinges on quality of poly approx

$$\min_{f \in P_k} \max_{\lambda \in \Lambda(\tilde{K})} |f(\lambda) - \hat{f}(\lambda)|$$

- Compare: Chebyshev methods just use $[\lambda_{\min}, \lambda_{\max}]$

CG is a special case corresponding to $f(z) = z^{-1}$.

CG solves systems with \tilde{K} ; problem terms are

$$\mathcal{L}_{|K|} = -\frac{1}{2}\text{tr} \left(\log \tilde{K} \right) \qquad \frac{\partial \mathcal{L}_{|K|}}{\partial \theta_i} = -\frac{1}{2}\text{tr} \left(\tilde{K}^{-1} \frac{\partial \tilde{K}}{\partial \theta_i} \right)$$

Q: How do we parley fast MVMs into trace computations?

Tractable traces

Stochastic trace estimation trick:

- $z \in \mathbb{R}^n$ has independent random entries
- $\mathbb{E}[z_i] = 0$ and $\mathbb{E}[z_i^2] = 1$

Then

$$\mathbb{E}[z^T A z] = \sum_{i,j} a_{ij} \mathbb{E}[z_i z_j] = \text{tr}(A).$$

NB: $\mathbb{E}[z \odot A z] = \text{diag}(A)$.

Standard choices for the probe vector z :

- Hutchinson: $z_i = \pm 1$ with probability 0.5
- Gaussian: $z_i \sim N(0, 1)$

See Avron and Toledo review, JACM 2011.

Putting it together

For each probe vector z until error bars small enough:

- Run Lanczos from $z/\|z\|$
- Use Lanczos to estimate $\tilde{K}^{-1}z$ and $\log(\tilde{K})z$
- Dot products yield estimators:

$$\mathcal{L}_{|K|} = -\frac{1}{2}\mathbb{E}\left[z^T \log(\tilde{K})z\right]$$
$$\frac{\partial \mathcal{L}_{|K|}}{\partial \theta_i} = -\frac{1}{2}\mathbb{E}\left[(\tilde{K}^{-1}z)^T \left(\frac{\partial \tilde{K}}{\partial \theta_i}z\right)\right]$$

Cost per probe:

- One Lanczos process
- One matvec per parameter with derivative

Quite effective in practice! And amenable to preconditioning.

“There is No New Thing Under the Sun”

“Generalized Cross-Validation for Large-Scale Problems”

Golub and Von Matt (1997)

- Treats least squares and GCV (vs kernel methods and MLE)
- But the same Lanczos + stochastic trace estimator combo

Pivoted Cholesky preconditioning

Let $M = P(LL^T + \sigma^2 I)P^T \approx \tilde{K}$ with $L \in \mathbb{R}^{n \times m}$, $m \ll n$:

- *Preconditioned* CG: works (implicitly) with $M^{-1}\tilde{K}$
- Note $\log \det \tilde{K} = \log \det M + \log \det M^{-1}\tilde{K}$
- Know how to do fast direct solves and $\log \det$ with M
- All boils down to generalized Lanczos with (\tilde{K}, M)
- Smooth kernels (e.g. SE) and long length scales \implies convergence in few steps

Blocking for performance

Generalized Lanczos per probe vector involves

- One matvec with \tilde{K} per step
- One solve with M per step
- Barrier between steps
- Low *arithmetic intensity* (flops / memory access)
- Limited opportunities for parallelism

Idea: Lanczos for several probes in parallel

- Multiply \tilde{K} or M^{-1} by *panel* of vectors / step
- Improves cache use and parallelism

The whole package

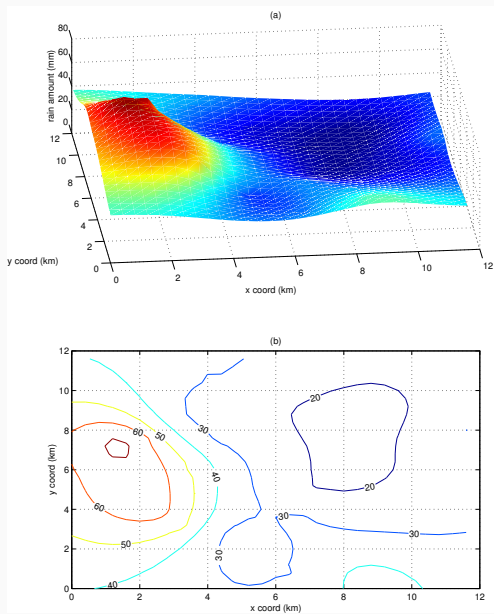
So we have

- Stochastic estimators + Krylov iterations
- Preconditioning to reduce steps to convergence
- Blocking to reduce time per step
- GPU acceleration speeds things up further

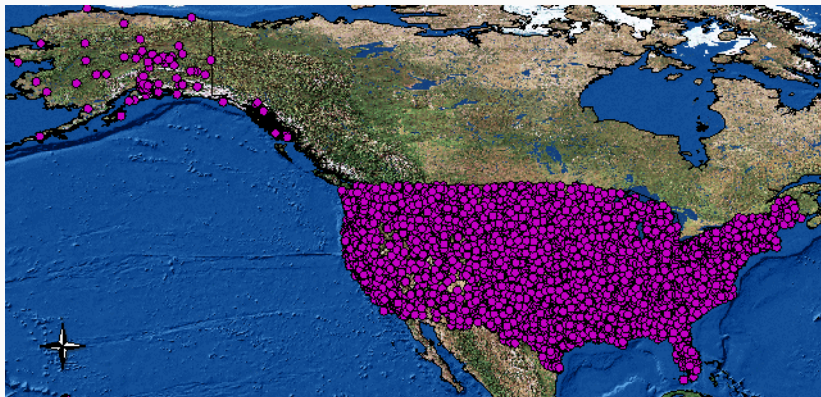
For all the tricks together: <https://gpytorch.ai>

Examples

Example: Rainfall



Example: Rainfall



Map generated by NOAA's National Climatic Data Center, 2007

0 698mi

Example: Rainfall

Method	n	m	MSE	Time [min]
Lanczos	528k	3M	0.613	14.3
Scaled eigenvalues	528k	3M	0.621	15.9
Exact	12k	-	0.903	11.8

- Data: Hourly precipitation data at 5500 weather stations
- Aggregate into daily precipitation
- Total data: 628K entries
- Train on 100K data points, test on remainder
- Use SKI with 100 points per spatial dim, 300 in time
- Comparison: scaled eigenvalues approx, exact solve

NB: This is with an older MATLAB code (GPML), not GPyTorch

Example: Hickory data

Can build other stochastic processes via GPs

- Example: Log-Gaussian Cox process model
 - Models count data (e.g. events in spatial bins)
 - Poisson conditional on intensity function
 - Log intensity drawn from a GP
- Laplace approximation for posterior
- Data set is point pattern of 703 hickory trees in Michigan

Example: Hickory data

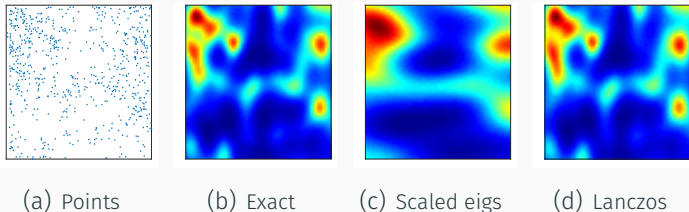
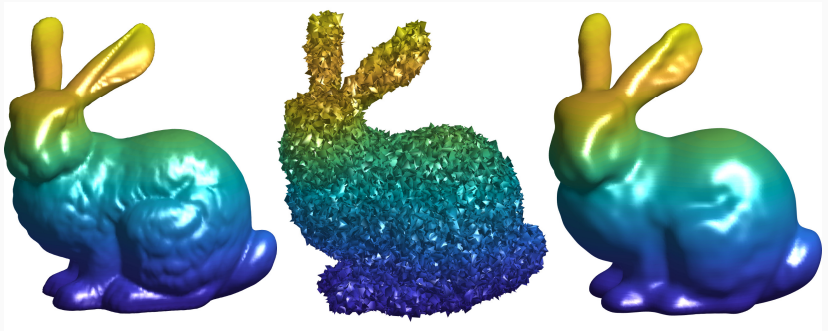


Figure 1: Prediction by different methods on the Hickory dataset.

Method	s_f	ℓ_1	ℓ_2	$-\log p(y \theta)$	Time [s]
Exact	0.696	0.063	0.085	1827.56	465.9
Lanczos	0.693	0.066	0.096	1828.07	21.4
Scaled eigs	0.543	0.237	0.112	1851.69	2.5

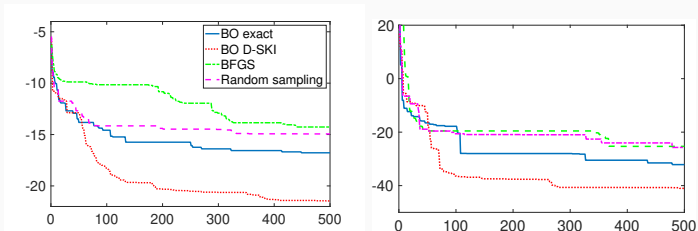
Table 1: Hyper-parameters recovered by different methods

Example: Surface reconstruction



Recovering the Stanford bunny model from 25K noisy normals.

Example: Bayesian optimization



- Learn active subspace of high-dimensional space
- Fit GP model with derivative in subspace
- Optimize expected improvement
 - Need predictive variance (use stochastic estimator)
 - Use low-rank approximation for control variate
- Works surprisingly well — still lots to do!

For more...

Eriksson, Dong, Lee, B., Wilson. Scaling Gaussian Process Regression with Derivatives. NeurIPS 2018.

Gardner, Pleiss, Weinberger, B., Wilson. GPyTorch: Blackbox Matrix-Matrix Gaussian Process Inference with GPU Acceleration. NeurIPS 2018.

K. Dong, D. Eriksson, H. Nickisch, D. Bindel, and A. G. Wilson, Scalable Log Determinants for Gaussian Process Kernel Learning. NIPS 2017.

GPyTorch: <https://gpytorch.ai>