Linear Algebra Perspectives on Inducing Point Selection

David Bindel

14 May 2024

Collaborators

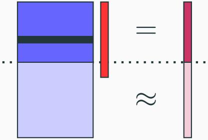
Xinran Zhu
(Cadence)

Jake Gardner
(U Penn)

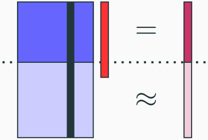
(+ many other past collaborators)

Kernel-Based Regression: Four Stories

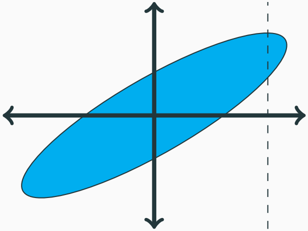
Feature map



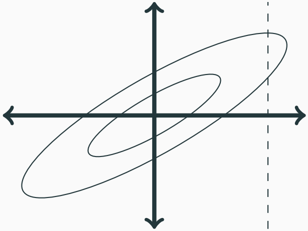
Data-dependent basis



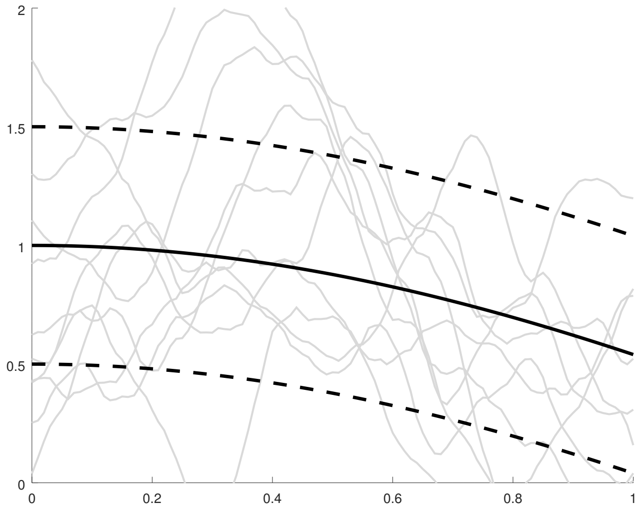
Energy minimization



Gaussian process



Basic ingredient: Gaussian Processes (GPs)



Basic ingredient: Gaussian Processes (GPs)

Our favorite continuous distributions over

$\mathbb{R}$:	$\text{Normal}(\mu, \sigma^2)$,	$\mu, \sigma^2 \in \mathbb{R}$
$\mathbb{R}^n$:	$\text{Normal}(\mu, C)$,	$\mu \in \mathbb{R}^n, C \in \mathbb{R}^{n \times n}$
$\mathbb{R}^d \to \mathbb{R}$:	$\text{GP}(\mu, k)$,	$\mu : \mathbb{R}^d \to \mathbb{R}$, $k : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$

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

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

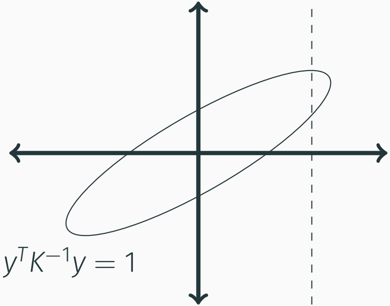
$$\text{have } f_X \sim N(\mu_X, K_{XX}), \text{ where}$$

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

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

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

Being Bayesian



Let $Y = (Y_1, Y_2) \sim N(0, K)$. Given $Y_1 = y_1$, what is Y_2?

Posterior distribution: $(Y_2 | Y_1 = u_1) \sim N(w, S)$ where

$$w = K_{21} K_{11}^{-1} y_1$$

$$S = K_{22} - K_{21} K_{11}^{-1} K_{12}$$

Being Bayesian

Consider a (zero-mean) GP prior with kernel k:

$$f \sim \mathrm{GP}(0, k)$$

Measure at X with noise, apply Bayes to get posterior:

$$(f \,|\, y = f_X + \epsilon) \sim \mathrm{GP}(\mu, \tilde{k})$$

where

$$\begin{aligned} \mu(x) &= k_{xX} c, \quad \hat{K}_{XX} c = y \\ \tilde{k}(x, y) &= k(x, x) - k_{xX} \hat{K}_{XX}^{-1} k_{Xy} \\ \hat{K}_{XX} &= K_{XX} + \eta I \end{aligned}$$

Specifically,

$$(f(x)|y = f_X + \epsilon) \sim N\left(k_{xX} c, k(x, x) - k_{xX} \hat{K}_{XX}^{-1} k_{Xx}\right)$$

Scandalous Scaling

Can we go faster than the naive costs?

- Fitting and hyperparameter selection: $O(N^3)$
- Evaluating: $O(N)$
- Evaluating uncertainty: $O(N^2)$

Idea: Approximate via $m \ll N$ *inducing points*.

Where Do We Go Now?

- (Corrected) Nyström matrix and operator approximation
- Matrix and quasimatrix forward selection
- Getting the right predictive uncertainty

(Corrected) Nyström

Approximate via inducing points $U \subset X$:

$$K_{XX} + \eta I \approx K_{XU} K_{UU}^{-1} K_{UX} + D,$$

where $D = \eta I$ (SoR), or plus some additional correction (FITC).

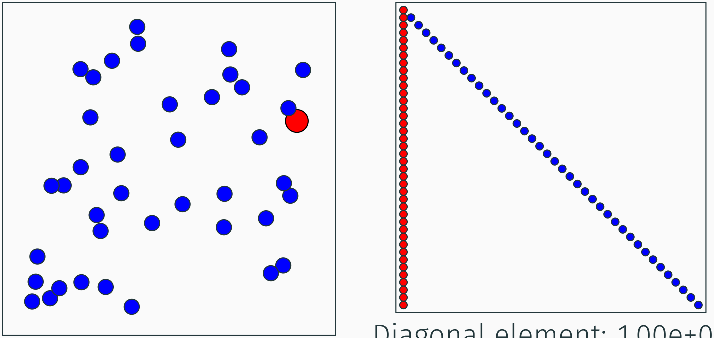
A good exercise: solve $(K_{XU} K_{UU}^{-1} K_{UX} + D)c = y$ by

- Minimize $\left\| \begin{bmatrix} D^{-1/2} K_{XU} \\ K_{UU} \end{bmatrix} \lambda - \begin{bmatrix} D^{-1/2} y \\ 0 \end{bmatrix} \right\|$
- Recover $c = D^{-1}(y - K_{XU}\lambda)$ if desired
- Prediction $K_{xU} K_{UU}^{-1} K_{UX} c = K_{xU} \lambda$.

Can be a good preconditioner even when not great alone.
Things like log determinants are also simple to compute.

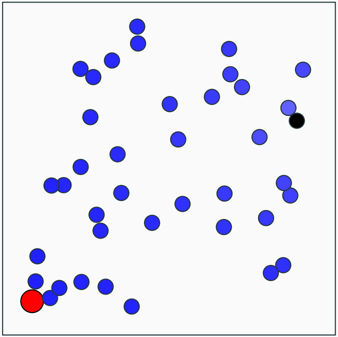
Greedy Selection and Choosy Cholesky

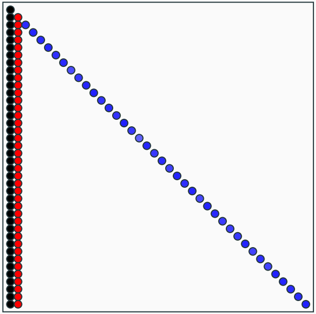
Greedy choice of inducing points U for smooth case:
Left-looking partial pivoted Cholesky



Diagonal element: 1.00e+00

Greedy choice of inducing points U for smooth case:
Left-looking partial pivoted Cholesky

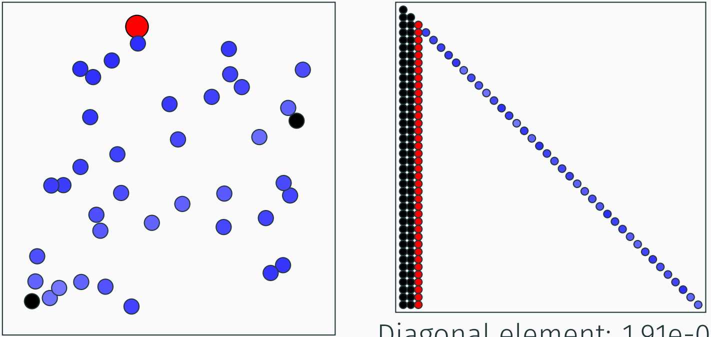




Diagonal element: 6.77e-02

Greedy Selection and Choosy Cholesky

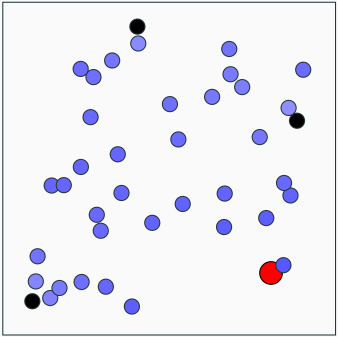
Greedy choice of inducing points U for smooth case:
Left-looking partial pivoted Cholesky

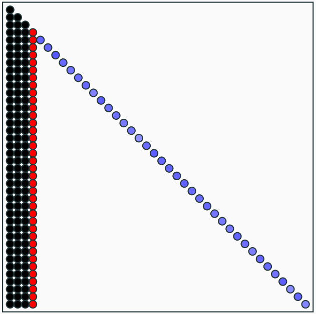


Diagonal element: 1.91e-02

Greedy Selection and Choosy Cholesky

Greedy choice of inducing points U for smooth case:
Left-looking partial pivoted Cholesky

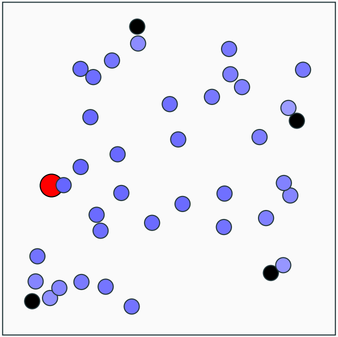


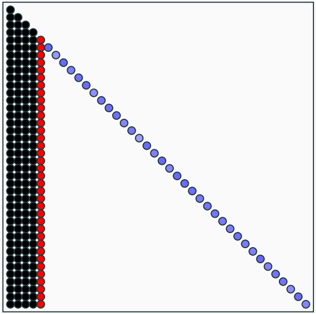


Diagonal element: 5.11e-04

Greedy Selection and Choosy Cholesky

Greedy choice of inducing points U for smooth case:
Left-looking partial pivoted Cholesky

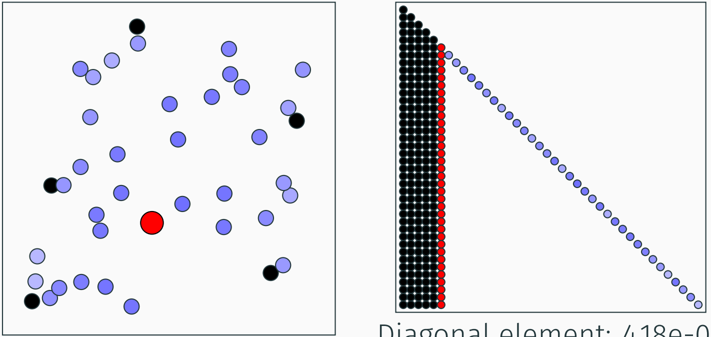




Diagonal element: 1.19e-04

Greedy Selection and Choosy Cholesky

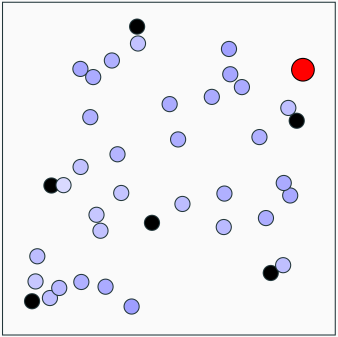
Greedy choice of inducing points U for smooth case:
Left-looking partial pivoted Cholesky

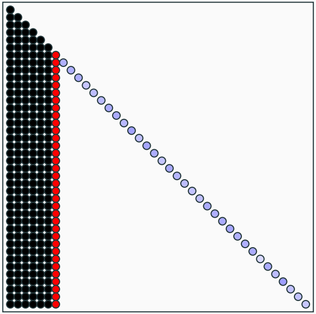


Diagonal element: 4.18e-05

Greedy Selection and Choosy Cholesky

Greedy choice of inducing points U for smooth case:
Left-looking partial pivoted Cholesky

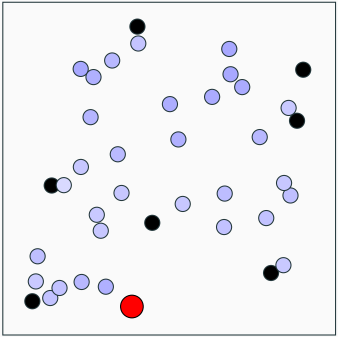


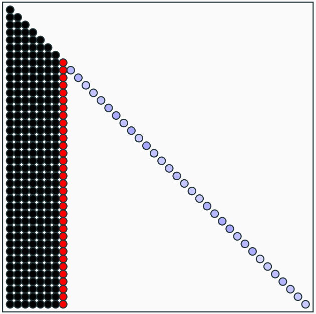


Diagonal element: 8.54e-07

Greedy Selection and Choosy Cholesky

Greedy choice of inducing points U for smooth case:
Left-looking partial pivoted Cholesky

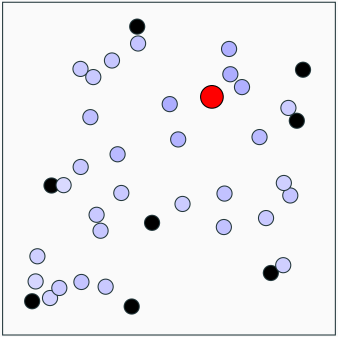


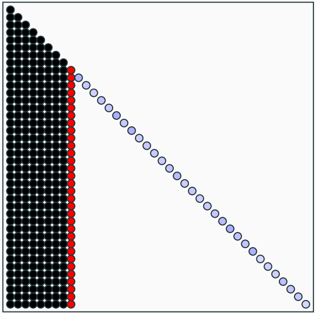


Diagonal element: 3.58e-07

Greedy Selection and Choosy Cholesky

Greedy choice of inducing points U for smooth case:
Left-looking partial pivoted Cholesky





Diagonal element: 1.92e-07

Greedy Selection and Choosy Cholesky

What if we can choose new sample points (or fake data)?

- Continuous pivoted Cholesky: next point maximizes the posterior variance:

$$v(x) = k_{xx} - k_{xX}\hat{K}_{XX}^{-1}k_{Xx}$$

- Same optimization, just over continuous vs discrete set!
- Limiting case of several Bayesian optimization methods
- May want to re-optimize kernel hypers between samples

Function Values?

- So far, focused on approximating kernel matrix/operator.
- … but we did not use the observations f_X!
- What if we focus on approximating f_X?

Forward Selection

Goal:

$$\text{minimize } \|K_{XU}c - f_X\|^2 \text{ over } U \subset X \text{ of size } m, c \in \mathbb{R}^m$$

Stepwise regression with forward selection:

- Initialize $r = f_X$
- Select next point u to maximize $|k_{Xu}^T r| / \|k_{Xu}\|^2$
- Update residual and repeat

Similar to pivoted QR on $\begin{bmatrix} f_X & K_{XX} \end{bmatrix}$.

Continuous Forward Selection

- Why not choose $U \not\subset X$?
  - Gradient-based maximization of $|k_{Xu}^T r| / \|k_{Xu}\|$.
  - Use a discrete set $\hat{U}$ of starting guesses
- Given initial guess (e.g. from greedy approach) can refine with variable projection approach:

$$\min_U \|(I - K_{XU} K_{XU}^\dagger) f_X\|^2$$

See Zhu, Gardner, B, NeurIPS 2022 Workshop on GPs.
(Also: Cornell CS 4220 project 3, Spring 2022)

What About the Distribution?

- Started focused on approximating kernel matrix/operator.
- Then we paid direct attention to $s_X \approx f_X$.
- What about trying to match the uncertainty ($v(x)$)?

Probabilistic Perspective

Usual GP inference:

- Prior $p(f_X, f_*)$ on training values and test values
- Condition on observations y
- Marginalize out f_X

Inducing points:

- Prior $p(f_X, f_*, u)$ on training, test, *inducing* values
- Assume conditional independence of f_X, f_* given u
- Marginalize out f_X and u

Perspective unifies many inducing point schemes (Quiñonera-Candela and Rasmussen, 2006).

Sparse GP framework

	Training $(f_X \vert f_U)$	Test $(f_* \vert f_U)$
DTC	$\mathcal{N}(K_{XU}K_{UU}^{-1}f_U, 0)$	$\mathcal{N}(K_{*U}K_{UU}^{-1}f_U, \tilde{K}_{**})$
FITC	$\mathcal{N}(K_{XU}K_{UU}^{-1}f_U, \text{diag}(\tilde{K}_{XX}))$	$\mathcal{N}(K_{*U}K_{UU}^{-1}f_U, \tilde{K}_{**})$
SVGP	$\mathcal{N}(K_{XU}K_{UU}^{-1}f_U, \tilde{K}_{XX})$	$\mathcal{N}(K_{*U}K_{UU}^{-1}f_U, \tilde{K}_{**})$

Variational Inference

Desiderata: choos inducing point locations (and other params) to maximize log-likelihood $\log p(y)$ – but hard!

Basic idea:

$$p(y) = \int p(y|f_X)p(f_X)$$

$$p(y|f_U) = \int p(y|f_X)p(f_X|f_U)$$

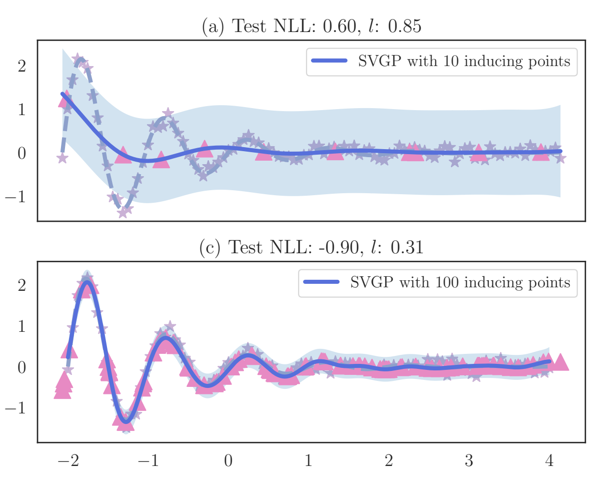
Jensen's inequality

$$\log p(y|f_U) \geq \int \log p(y|f_X)p(f_X|f_U)$$

Yields *evidence lower bound* – maximize that. Like minimizing KL divergence between true posterior and parametric approximation.

(See Blei, Kucukelbir, McAuliffe, 2018)

Variational GPs



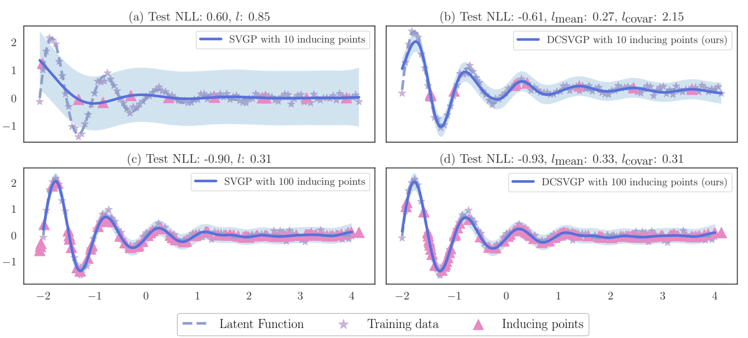
Maybe works poorly with too few inducing points?

Sparse GP Framework

	Training $(f_X \vert f_U)$	Test $(f_* \vert f_U)$
DTC	$\mathcal{N}(K_{XU}K_{UU}^{-1}f_U, 0)$	$\mathcal{N}(K_{*U}K_{UU}^{-1}f_U, \tilde{K}_{**})$
FITC	$\mathcal{N}(K_{XU}K_{UU}^{-1}f_U, \mathbf{diag}(\tilde{K}_{XX}))$	$\mathcal{N}(K_{*U}K_{UU}^{-1}f_U, \tilde{K}_{**})$
SVGP	$\mathcal{N}(K_{XU}K_{UU}^{-1}f_U, \tilde{K}_{XX})$	$\mathcal{N}(K_{*U}K_{UU}^{-1}f_U, \tilde{K}_{**})$
DCSVGP	$\mathcal{N}(Q_{XU}Q_{UU}^{-1}f_U, \tilde{K}_{XX})$	$\mathcal{N}(Q_{*U}Q_{UU}^{-1}f_U, \tilde{K}_{**})$

Not obliged to capture conditional mean and covariance with same kernels!

What do we get?



Zhu, Wu, Maus, Gardner, B, NeurIPS 2023

Decoupling mean and covariance approximation (via separate length scales for predictive mean and covariance).

Concluding notes

- Common idea: approximate kernel approximations via a few inducing points
- Reduces cost of fitting the approximation and computation of predictive mean and variance
- Different “glasses” give different approaches to inducing points
  - NLA: Pivoted factorizations!
  - **Function approximation**: Forward selection
  - **Distribution approximation**: Variational inference

Refs: Zhu, Gardner, B, NeurIPS 2022 Workshop on GPs;
Zhu, Wu, Maus, Gardner, B, NeurIPS 2023