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

David Bindel

19 June 2019

Department of Computer Science
Cornell University

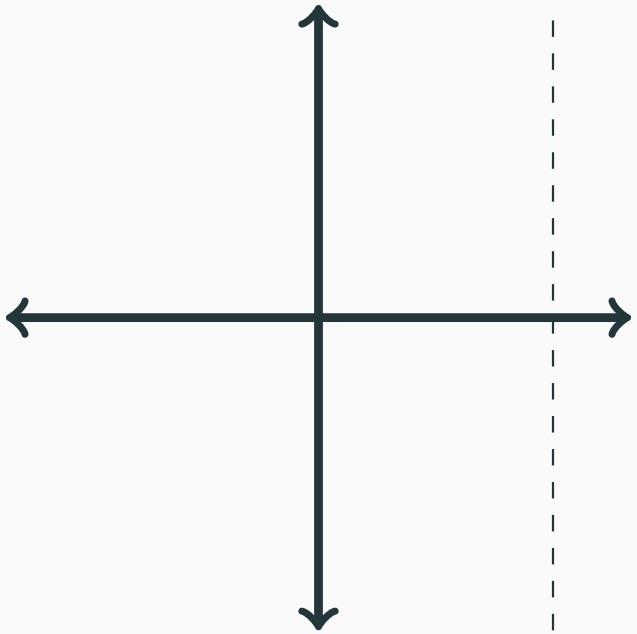
Lecture plan

Three threads from “lay of the land” to current research:

- Monday: Latent Factor Models
- Wednesday: Scalable Kernel Methods
  - 1:30-2:30: Structure and interpretation of kernels
  - 3:00-4:00: Making kernel methods scale
- Friday: Spectral Network Analysis

Slides posted on web page (linked from my Cornell page).

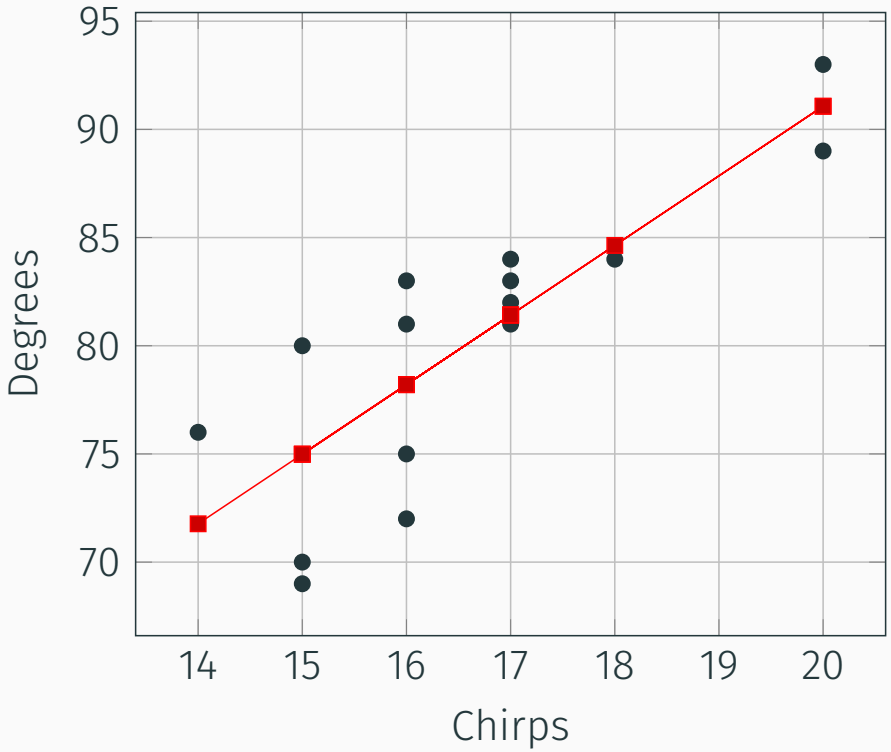
Simple and Impossible



Let $u = (u_1, u_2)$. Given an estimate of u_1, what is u_2?

We need an assumption!

Basic Regression Task



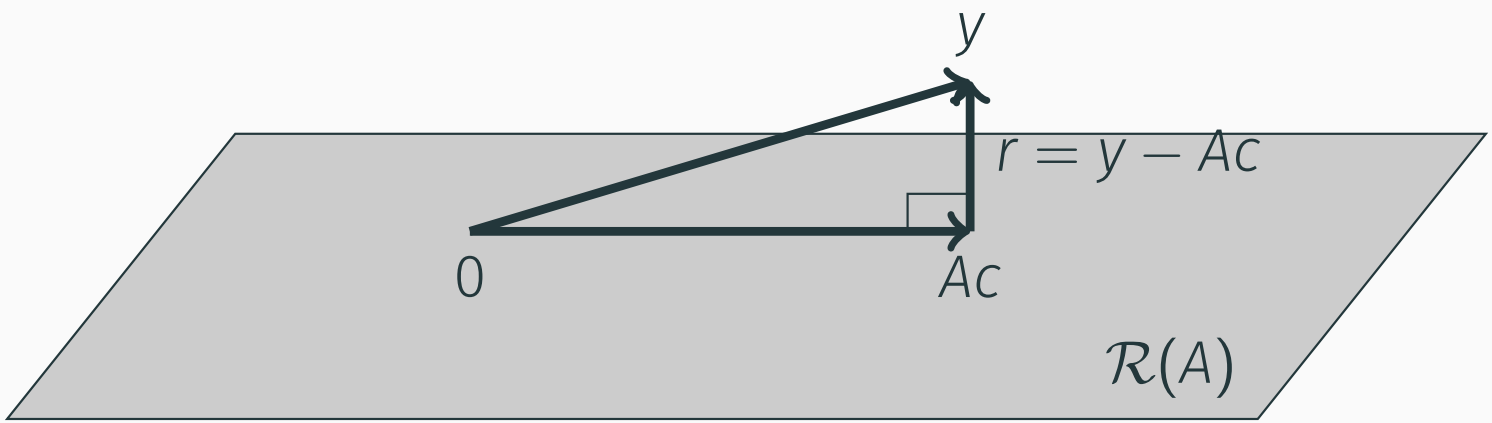
Standard example for today:

Unknown: $f : \Omega \subset \mathbb{R}^d \to \mathbb{R}$

Data: $y_j = f(x_j) + \epsilon_j$ for $x_j \in \Omega$

Goal: Estimate $f(x)$ for general x

The Linear Least Squares Picture



Least squares approach gives good predictions assuming

- *Consistency*: Space contains good approximations
- *Stability*: Space is not sensitive to basis perturbations

In statistical terms, care about *bias* and *variance*.

The Linear Least Squares Picture

Simplest model:

$$f(x) \approx c^T x$$

If overdetermined ($n > d$), solve

$$\text{minimize } \frac{1}{2}\|Ac - y\|_2^2 \text{ where } A^T = \begin{bmatrix} x_1 & x_2 & \dots & x_n \end{bmatrix}$$

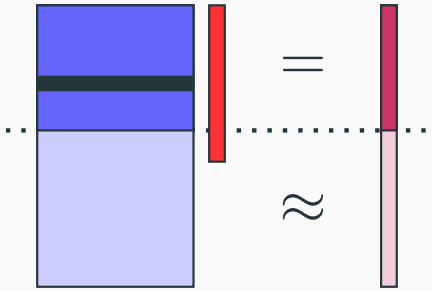
Potential problems:

- Want a different loss? LS still a good building block.
- A poorly conditioned? May need to regularize.
- Model space is not rich enough.

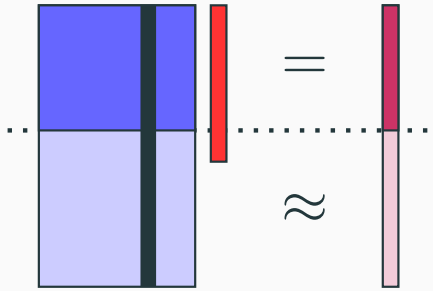
Kernel methods give us richer model spaces.
Usually so rich that we are no longer overdetermined.

Kernel-Based Regression: Four Stories

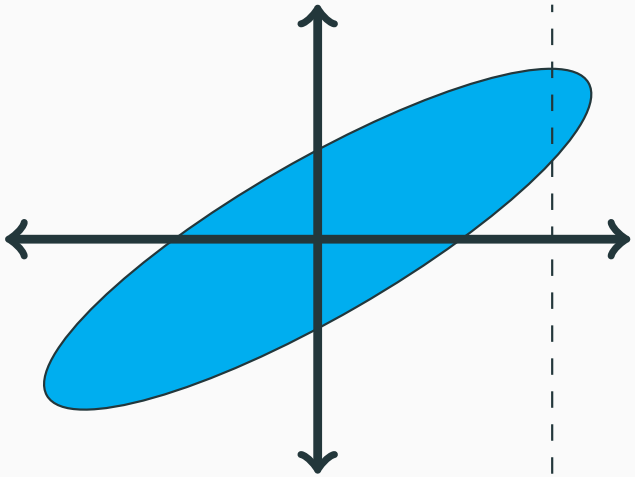
Feature map



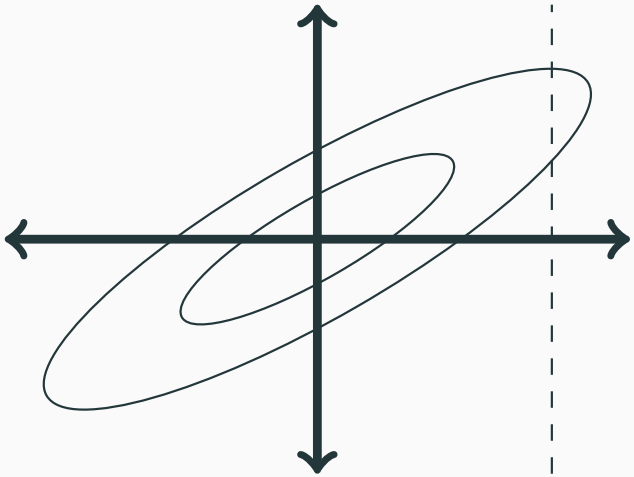
Data-dependent basis



Energy minimization



Gaussian process



Feature Maps and Data-Dependent Bases

Feature Maps

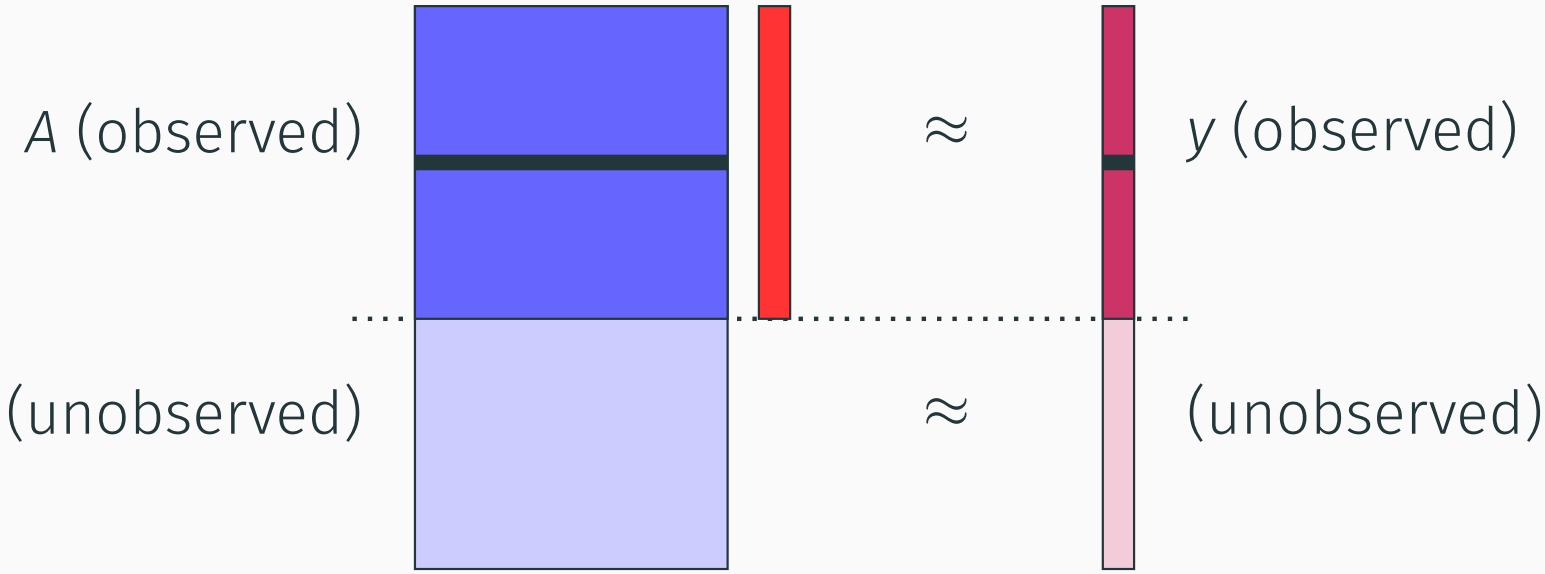
$$\begin{bmatrix} x \\ y \end{bmatrix} \mapsto \begin{bmatrix} 1 \\ x \\ y \\ x^2 \\ xy \\ y^2 \end{bmatrix}$$

Augment simple linear model ($c^T x$) with feature map:

$$f(x) \approx \langle d, \psi(x) \rangle$$

where $\psi : \Omega \to \mathcal{F}$ and $d \in \mathcal{F}$, some Hilbert space $\mathcal{F}$.

Feature Maps and Dimensionality

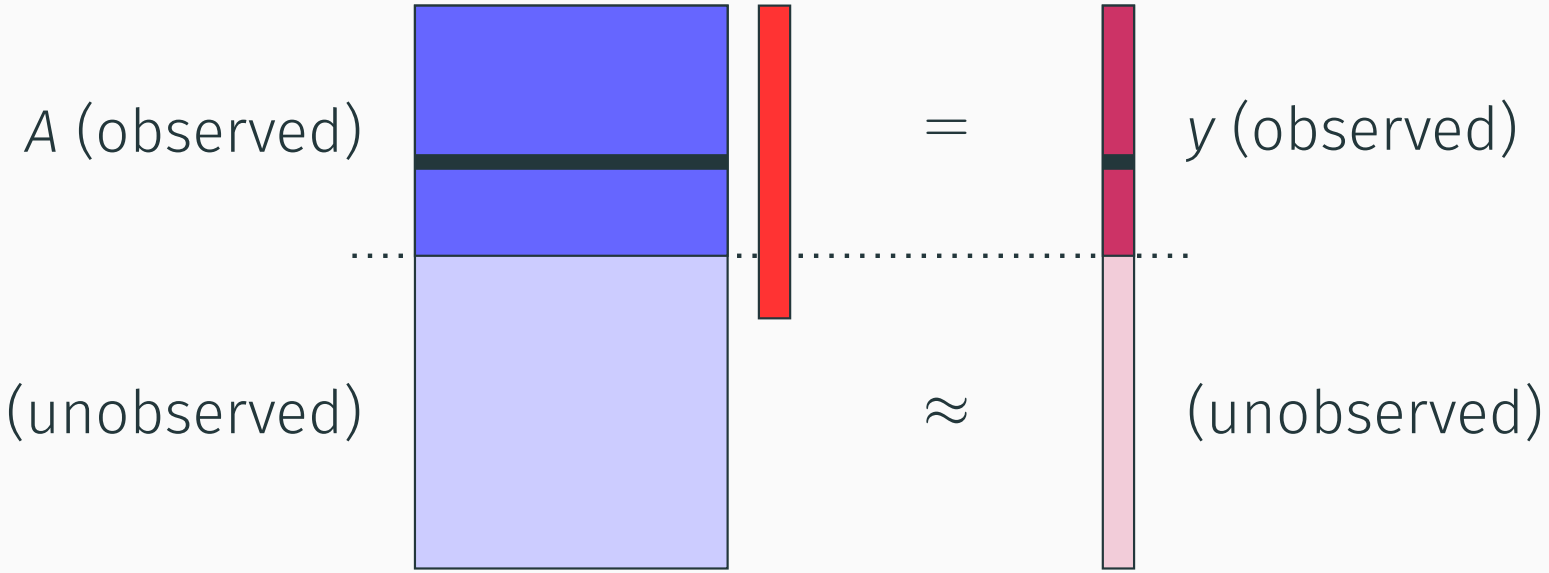


Usual idea with $A^T = \begin{bmatrix} \psi(x_1) & \dots & \psi(x_n) \end{bmatrix}^T$:

- $\dim \mathcal{F} < n$: same least squares as before
- $\dim \mathcal{F} = n$: interpolation problem

May be ill-posed for general scattered (multidimensional) case (Mairhuber-Curtis theorem)

Feature Maps and Dimensionality



Underdetermined ($\dim \mathcal{F} > n$): seek *minimal norm* solution.
For standard inner product (ℓ^2):

$$d = A^\dagger y = A^T(AA^T)^{-1}y$$

$$f(x) \approx \psi(x)^T d = \psi(x)^T A^T (AA^T)^{-1} y$$

Implicit preference for some models over others.

The Kernel Trick (Noise-Free)

Formula:

$$A^T = \begin{bmatrix} \psi(x_1) & \dots & \psi(x_n) \end{bmatrix}$$
$$f(x) \approx s(x) \equiv \psi(x)^T A^T (AA^T)^{-1} y$$

In terms of *kernel* $k(x, y) = \langle \psi(x), \psi(y) \rangle$:

$$(AA^T)_{ij} = k(x_i, x_j) = (K_{XX})_{ij}$$
$$K_{XX} c = y = f_X$$
$$s(x) = K_{xX} c = \sum_{j=1}^{n} k(x, x_j) c_j$$

Subscripts to denote vectors/matrices of function evaluations.

The Kernel Trick

May still want regularization in underdetermined case!

$$\text{minimize } \frac{1}{2}\|r\|^2 + \frac{\eta}{2}\|d\|^2 \quad \text{s.t.} \quad r = y - Ad$$

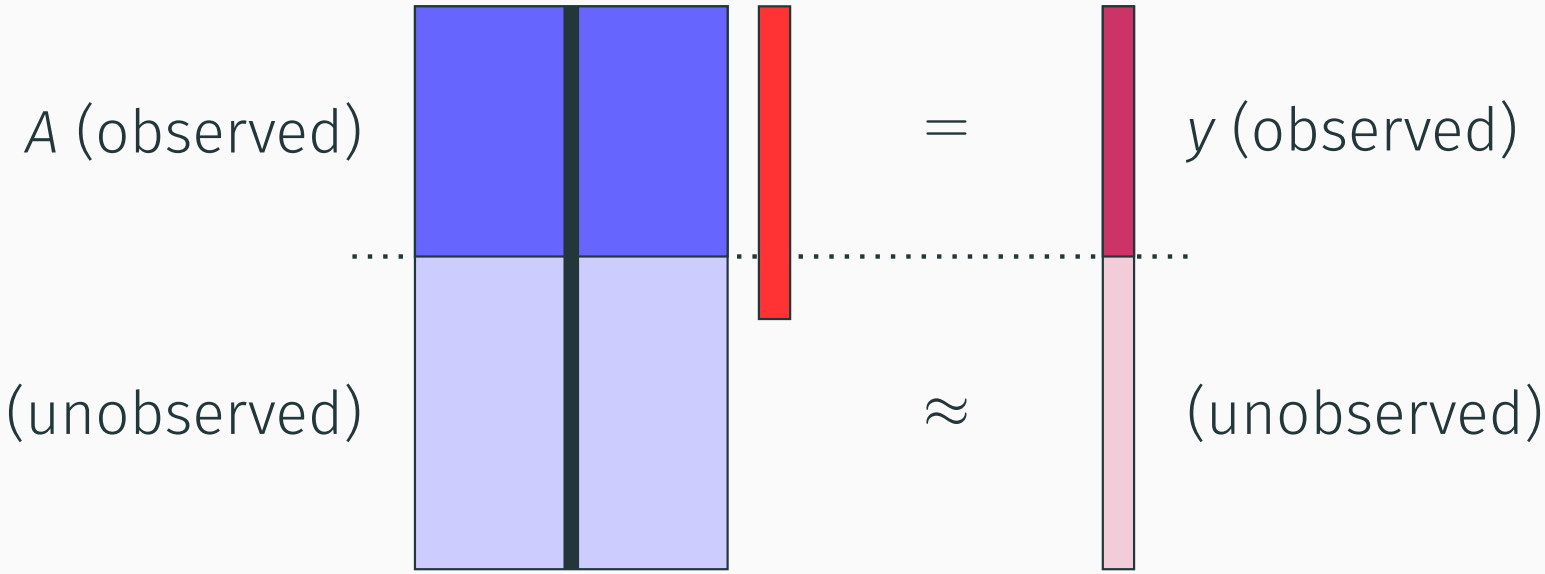
KKT conditions with Lagrange multiplier μ give

$$\begin{aligned} r - \mu &= 0 \\ \eta d - A^T \mu &= 0 \qquad & A^T r &= \eta d \\ y - Ad - r &= 0 & r &= y - Ad \end{aligned}$$

Result can again be expressed entirely via kernel:

$$\begin{aligned} (AA^T + \eta I) r &= \eta y & (K_{XX} + \eta I) r &= \eta y \\ (AA^T + \eta I) c &= y & (K_{XX} + \eta I) c &= y \\ s(x) = \psi(x)^T d &= \psi(x)^T A^T c & s(x) &= K_{xX} c \end{aligned}$$

Rows or Columns?



Change perspective: feature vectors $\psi(x_i)$ to basis vectors $\psi_j(x)$

- Native space: combinations of basis (with $d \in \mathcal{F}$)
- Sampling-dependent subspace of min-norm interpolants:

$$\mathcal{S}_X = \left\{ x \mapsto \sum_{j=1}^{n} k(x, x_j) c_j \right\}$$

Feature Maps to RKHS

Basis vectors $\psi_j(x)$ form orthonormal basis for a Hilbert space:

$$s(x) = \sum_i d_i \psi_i(x), \quad \|s\|_{\mathcal{H}}^2 = \|d\|^2$$

For $x \in \Omega$, define $k_x \in \mathcal{H}$ as

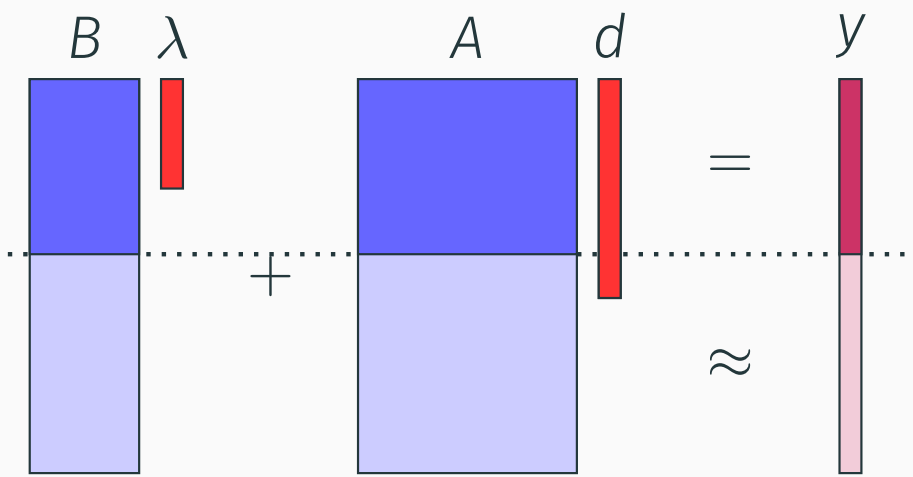
$$k_x(y) = \sum_i \psi_i(x)\psi_i(y)$$

Then k_x defines a point evaluation functional in $\mathcal{H}$

$$\begin{aligned}\langle k_x, s\rangle_{\mathcal{H}} &= \sum_i \psi_i(x) \left\langle \psi_i, \sum_j d_j \psi_j \right\rangle_{\mathcal{H}} \\ &= \sum_i \psi_i(x) d_i = s(x)\end{aligned}$$

This is a *reproducing kernel Hilbert space* (RKHS).

Role of Residual



Can also make d as small as possible for fitting a residual:

$$\text{minimize } \frac{1}{2}\|d\|^2 \text{ s.t. } B\lambda + Ad = y$$

KKT conditions (with c a Lagrange multiplier):

$$\begin{bmatrix} K_{XX} & B \\ B^T & 0 \end{bmatrix} \begin{bmatrix} c \\ \lambda \end{bmatrix} = \begin{bmatrix} y \\ 0 \end{bmatrix}$$

Note: Need B nonsingular for well-posedness.

Beyond the Basis

Beyond the Basis

- Story so far involves explicit feature maps.
- Computations only require kernel (inner products).

Putting the Kernel before the Feature Map

Start with symmetric kernel function $k : \Omega \times \Omega \to \mathbb{R}$.
k positive definite if K_{XX} spd for all samples X.

Often assume positive definite and:

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

Both: $k(x, y) = \phi(\|x - y\|)$, ϕ a *radial basis function*.

Have Mercer!

Associate integral operator with continuous spd kernel k:

$$(\mathcal{K}f)(x) = \int k(x,y)f(y)\,dy$$

$\mathcal{K}$ compact (actually Hilbert-Schmidt), so have

$$\mathcal{K} = \sum_{j=1}^{\infty} \lambda_j \psi_j \psi_j^*$$

and features are $\sqrt{\lambda_j}\psi_j(x)$.

But features are not really needed! Focus on the kernel.

From Kernels to Inner Products

To build RKHS without an explicit feature map:

- Observe that $\langle k_x, k_y \rangle_{\mathcal{H}} = k(x, y)$
- For $u(x) = \sum_{i=1}^{N} c_i k(x_i, x)$ and $v(x) = \sum_{i=1}^{N} d_i k(x_i, x)$, have

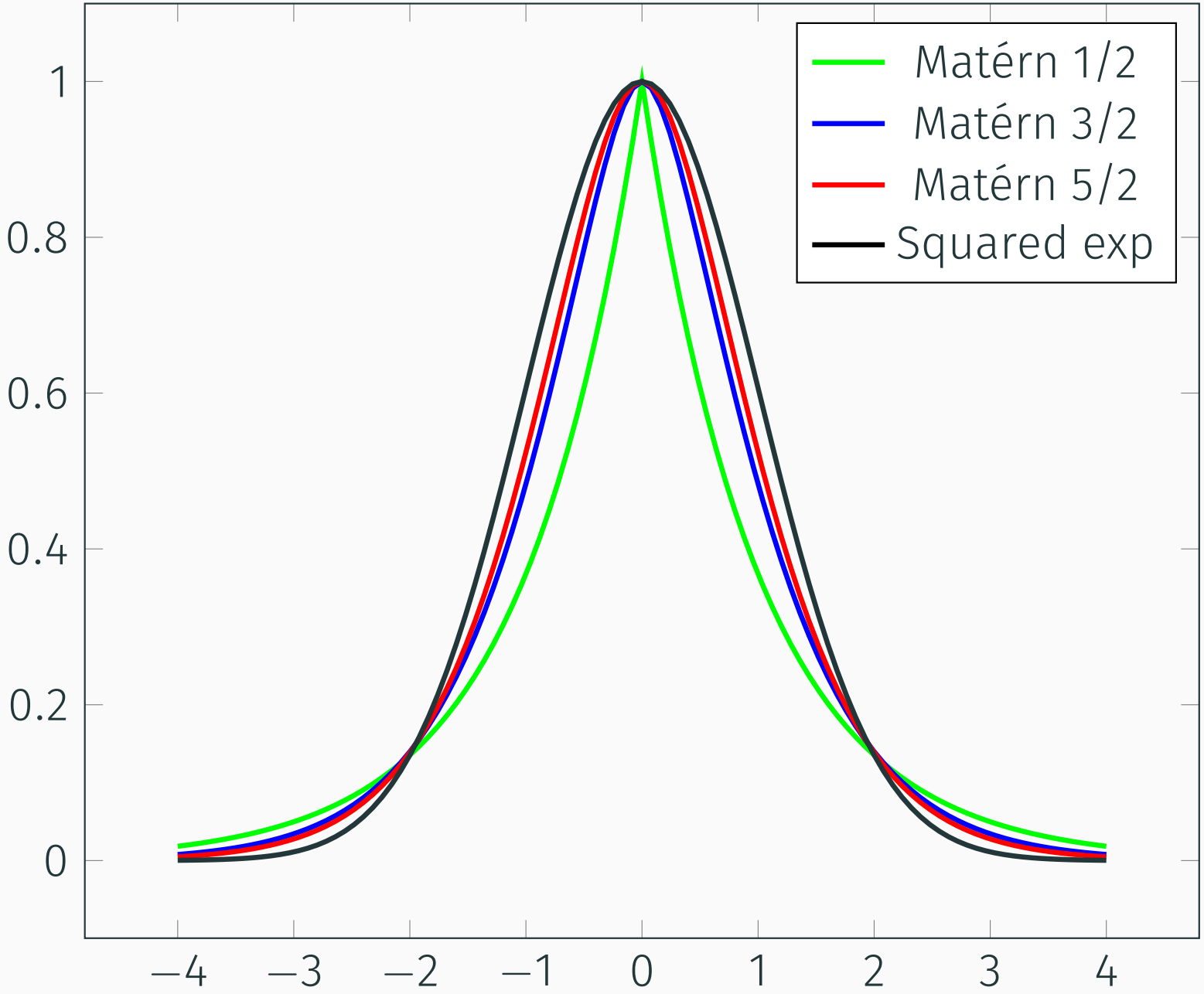
$$\langle u, v \rangle_{\mathcal{H}} = \left\langle \sum_i c_i k_{x_i}, \sum_j d_j k_{x_j} \right\rangle_{\mathcal{H}} = \sum_{i,j} c_i k(x_i, x_j) d_j = d^T K_{XX} c.$$

  Note:

$$\langle u, v \rangle_{\mathcal{H}} = v_X^T K_{XX}^{-1} u_X$$

- Gives pre-Hilbert structure, close to get Hilbert space.

Common Kernels



Common Kernels

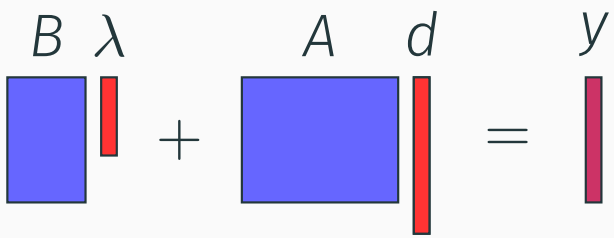
Kernel is *chosen by modeler*

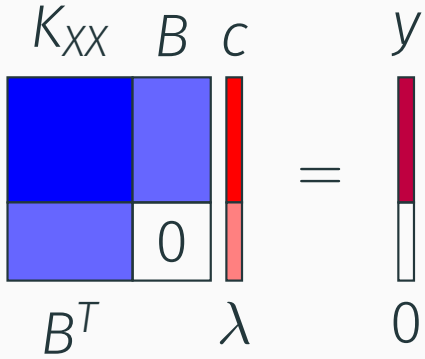
- Choose Matérn / RBF for regularity and simplicity
- Rarely have the intuition to pick the "right" kernel
- Different kernels generate different RKHS
- Common choices are *universal* (RKHS dense in $C(\Omega)$)
  - ... though with less data for a "good" choice

Properties of kernel matrices:

- Positive definite by design, but not well conditioned!
- Weyl: $k(r) \in C^\nu \implies |\lambda_n| = o(n^{-\nu-1/2})$
- RBF case: eigenvalues decay exponentially
- Adding regularization "wipes out" small eigenvalues

Conditionally Positive Definite Case





Consider kernelized "minimize $\mathcal{H}$-norm of residual" picture:

- Mental picture: $K_{XX} = AA^T$ (implicitly)
- But system with $K_{XX} - BMB^T$ gives *same answer* (for any symmetric M)
- And predictions do not depend on changes in B directions:

$$\begin{aligned} s(x) &= K_{xX}c + b(x)^T\lambda \\ &= (K_{xX} + \mu(x)^T B^T)c + b(x)^T\lambda \end{aligned}$$

Conditionally Positive Definite Case

If we have a polynomial fit + minimize $\mathcal{H}$-norm of residual, OK to "cheat" on the kernel definiteness:

- Symmetric $k : \Omega \times \Omega \to \mathbb{R}$
- $\{p_j\}$ a basis for $\mathcal{P}_{m-1}(\Omega)$ (poly of degree $< m$)
- *k conditionally positive definite of order m* if

$$c \neq 0, \Pi_X^T c = 0 \quad \Longrightarrow \quad c^T K_{XX} c > 0$$

where $[\Pi_X]_{ij} = p_j(x_i)$.

Well-posed problem if Π_X nonsingular.
Jargon: need X well-poised (for polynomial interpolation).

More Common Kernels

	$\phi(r)$	Order
Cubic	r^3	2
Thin-plate	$r^2 \log r$	2
Multiquadric	$-\sqrt{\gamma^2 + r^2}$	1
Inverse multiquadric	$(\gamma^2 + r^2)^{-1/2}$	0
Gaussian	$\exp(-r^2/\gamma^2)$	0

An Aside

General picture: kernel-based approximation + fixed-basis approximation

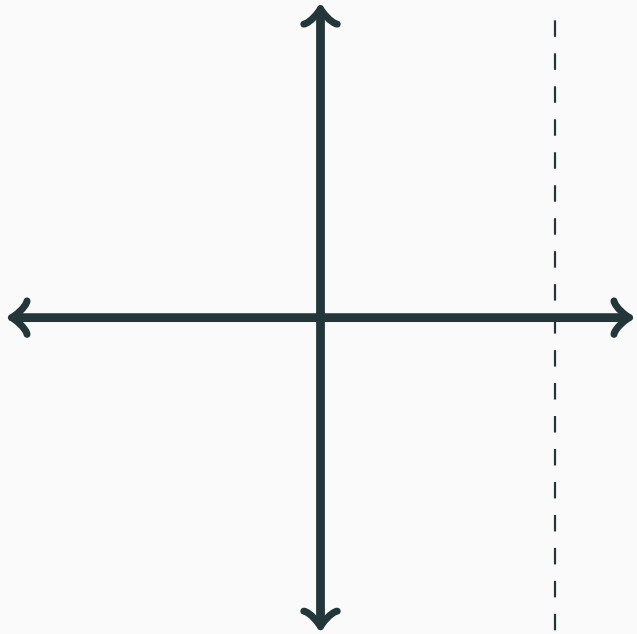
$$s(x) = k_{xX}c + b(x)^T\lambda$$

In CPD case, $b(x)$ is a basis for a polynomial space (the "tail")

- For CPD kernels we *need* a polynomial tail
- Can incorporate polynomial (or other) tails in other cases!
- A "tail" is a good way of putting in domain knowledge

Optimal Recovery vs Gaussian Processes

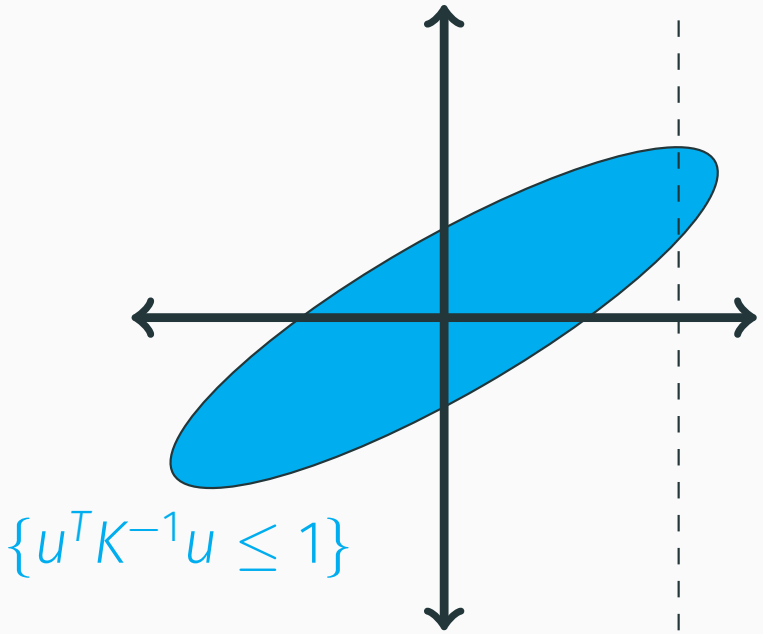
Simple and Impossible



Let $u = (u_1, u_2)$. Given u_1, what is u_2?

We need an assumption! Two different standard takes.

Being Bounded



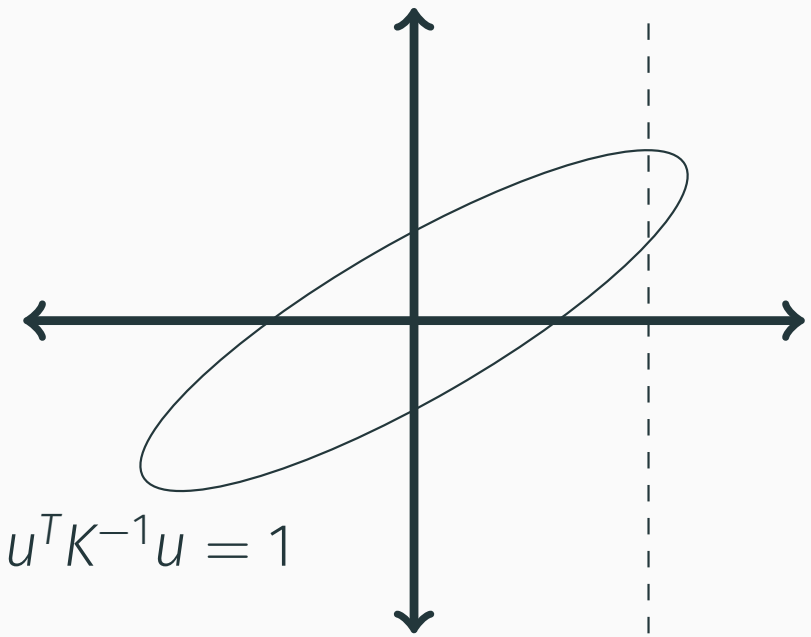
Let $u = (u_1, u_2)$ s.t. $\|u\|^2_{K^{-1}} \leq 1$. Given u_1, what is u_2?

Optimal recovery: $\|u_2 - w\|^2_{S^{-1}} \leq 1 - \|u_1\|^2_{(K_{11})^{-1}}$

$$w = K_{21} K_{11}^{-1} u_1$$

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

Being Bayesian



Let $U = (U_1, U_2) \sim N(0, K)$. Given $U_1 = u_1$, what is U_2?

Posterior distribution: $(U_2 | U_1 = u_1) \sim N(w, S)$ where

$$w = K_{21} K_{11}^{-1} u_1$$

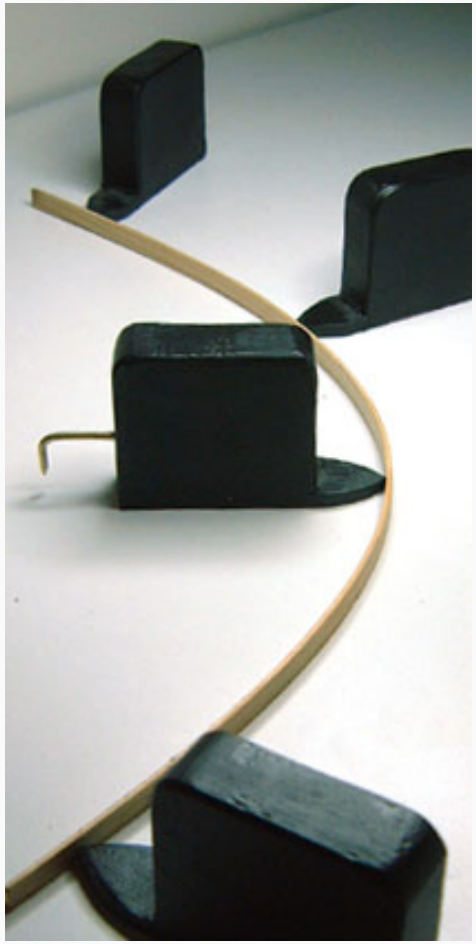
$$S = K_{22} - K_{21} K_{11}^{-1} K_{12}$$

Bounded vs Bayesian

Rest of lecture exploring these two different takes

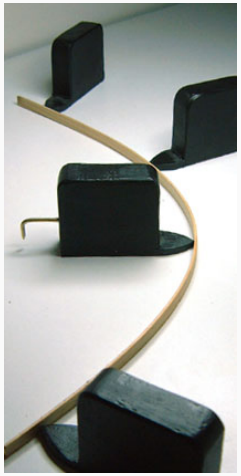
- Optimal recovery
  - Prior: bound in terms of a pos def quadratic
  - Prediction minimizes *worst-case* error given bound + data
  - Error analysis framework: worst-case bounds
- Bayesian inference
  - Prior: joint normal distribution (pos def covariance)
  - Prediction minimizes *average-case* error given prior + data
  - Error analysis framework: posterior distribution

From Energy to Error



http://www.duckworksmagazine.com/03/r/articles/splineducks/splineDucks.htm

Cubic Splines



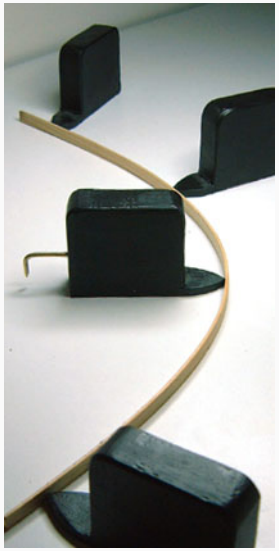
http://www.duckworksmagazine.com/03/r/articles/splineducks/splineDucks.htm

- $\phi(r) = r^3$ is conditionally positive definite of order 2
- Squared (semi-)norm is bending energy:

$$\|s\|_{\mathcal{H}}^2 \propto \frac{1}{2} \int_{\Omega} s''(x)^2 \, dx$$

- Linear polynomial tail = rigid body modes

Force, Displacement, Stiffness



Target function $f \in \mathcal{H}^2$, known bending energy

$$E[f] = \frac{1}{2} \int_\Omega f''(x)^2 \, dx$$

Cubic spline minimizes $E[s]$ s.t. $s(x_i) = f(x_i)$, so

$$E[s] \leq E[f]$$

- $f(x_i)$ as displacement, c_i as corresponding force
- Kernel matrix K_{XX} is compliance (force $\mapsto$ displacement)
- Residual compliance (inverse stiffness) at x is $P_X(x)^{-2}$
- Energy bound for error at X

$$P_X(x)^{-2} \left(s(x) - f(x)\right)^2 \leq E[f] - E[s]$$

General Picture

Interpolant is

$$s(x) = K_{xX}c + b(x)^T\lambda$$

Can compute *power function* $P_X(x)$ from factorization; SPD case:

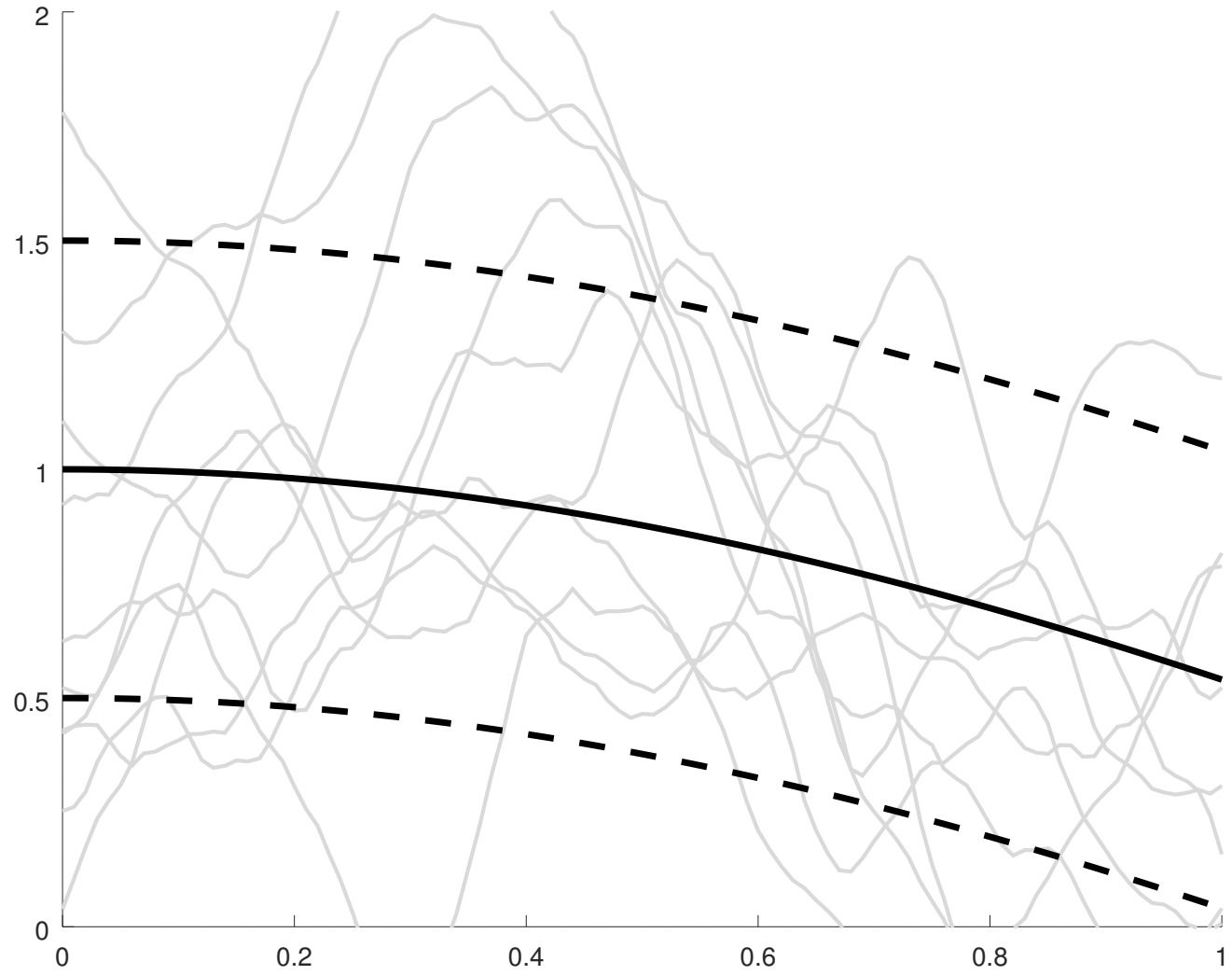
$$P_X(x)^2 = \phi(0) - K_{xX}K_{XX}^{-1}K_{Xx}$$

Bound is

$$|s(x) - f(x)| \leq P_X(x)\sqrt{\|f\|_{\mathcal{H}}^2 - \|s\|_{\mathcal{H}}^2}$$

Only thing that is hard to compute generally: $\|f\|_{\mathcal{H}}^2$.

Basic ingredient: Gaussian Processes (GPs)



Basic ingredient: Gaussian Processes (GPs)

Our favorite continuous distributions over

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

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

$$\begin{aligned} \forall X = (x_1, \ldots, x_n),\ x_i \in \mathbb{R}^d, & \\ \text{have } f_X \sim N(\mu_X, K_{XX}),\ & \text{where} \\ f_X \in \mathbb{R}^n, \quad & (f_X)_i \equiv f(x_i) \\ \mu_X \in \mathbb{R}^n, \quad & (\mu_X)_i \equiv \mu(x_i) \\ K_{XX} \in \mathbb{R}^{n \times n}, \quad & (K_{XX})_{ij} \equiv k(x_i, x_j) \end{aligned}$$

Being Bayesian

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

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

Measure at X, apply Bayes to get posterior:

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

where

$$\mu(x) = k_{xX} c$$
$$\tilde{k}(x, y) = k(x, x) - k_{xX} K_{XX}^{-1} k_{Xy}$$

Specifically, posterior for $f(x)$ at given x is

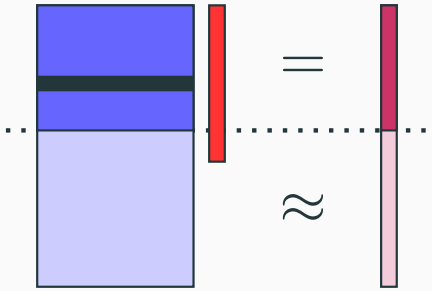
$$N(k_{xX} c, k(x, x) - k_{xX} K_{XX}^{-1} k_{Xx})$$

Predictive variance = squared power function!

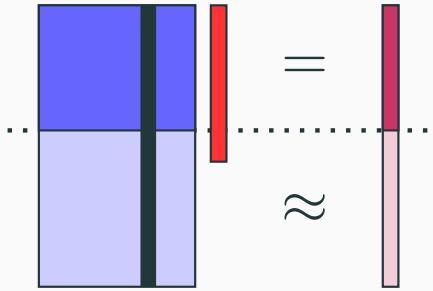
Summary and Wrap-Up

Kernel-Based Regression: Four Stories

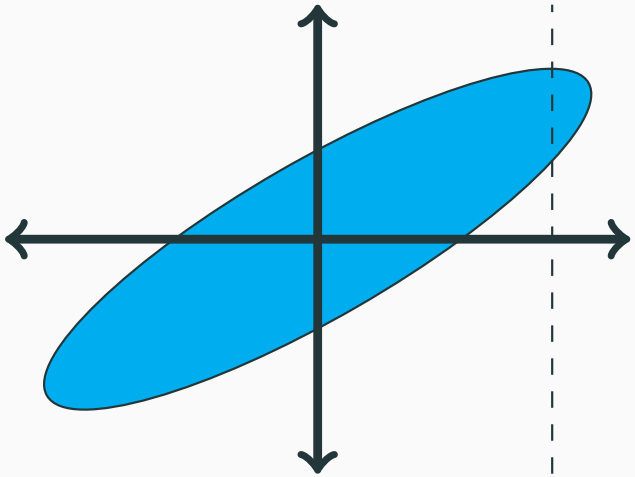
Feature map



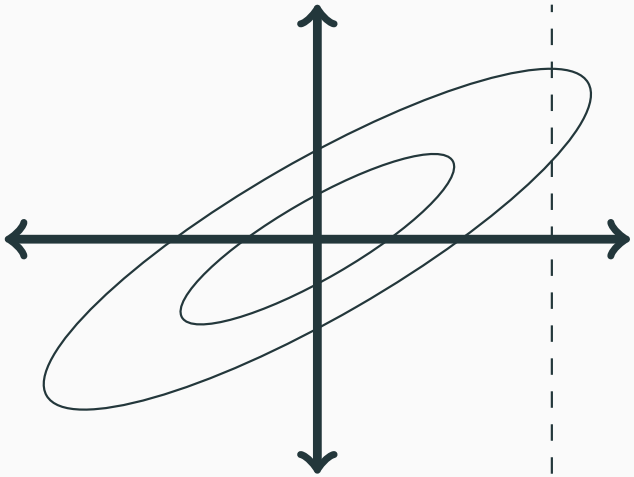
Data-dependent basis



Energy minimization



Gaussian process



The Power of Different Lenses

- "Kernel trick" used to go basis-free
  - But there is power in thinking with a basis, too!
  - Comes up as a computational tool (next time)
- Kernels can correspond to physics!
  - Ex: Cubic spline and thin-plate spline
  - Kernel as a Green's function for an elliptic PDE
  - Physical interpretation helps understand error analysis
- Optimal recovery and GP interpretation mostly coincide
  - But *only* when data is linear functionals of f
  - Ex: Different predictions for non-negativity constraints!
- CPD kernels popular in RBF literature (optimal recovery)
  - But also works for Bayesian interp — improper GP priors
  - Does appear in Wahba's work, but often overlooked
  - Tails are useful even in pos def case

Plan for Next Time

- This hour: the analysis half of numerical analysis
- Up next: the numerical side!
- Start with standard $O(n^3)$ approaches for
  - Fitting with a fixed kernel
  - Kernel selection (max likelihood and cross-validation)
  - Adaptive sampling and Bayesian optimization
- Then the fun stuff: how to scale better than $O(n^3)$!