Chebyshev Featurization

David Bindel
Department of Computer Science
Cornell University

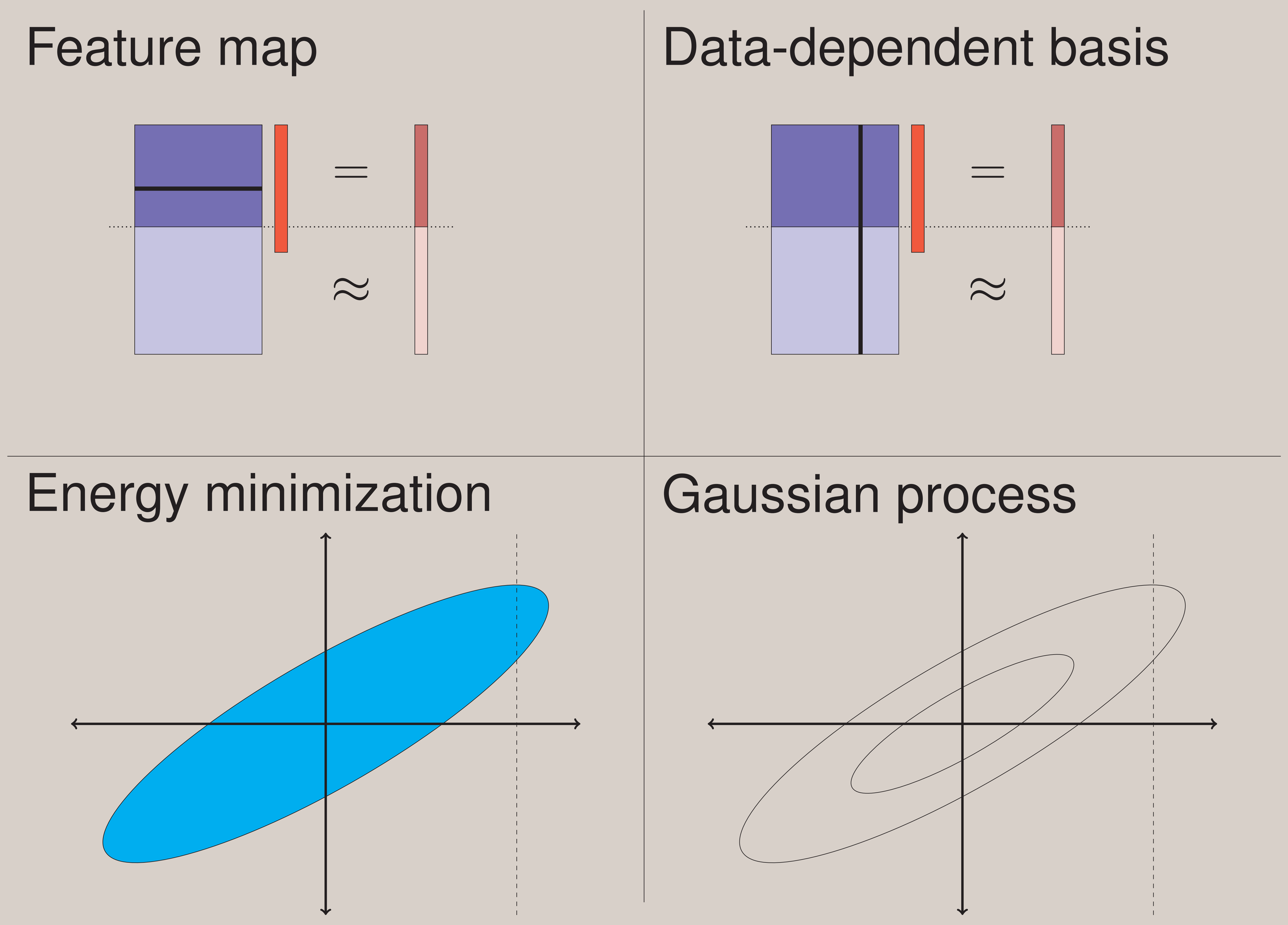
Function Approximation from Scattered Data

Goal: Approximate $f : \Omega \to \mathbb{R}$ from $f_X = \left[f(x_1) \dots f(x_n)\right]^T$.
Approach: Choose $s(x) = \sum_{i=1}^n k(x, x_i) c_i$ with *kernel* $k : \Omega \times \Omega \to \mathbb{R}$.
(often $k(x,y) = \phi(\|x-y\|)$ for some *radial basis function* ϕ)

To fit: solve $(K_{XX} + \lambda I)c = f_X$ where $(K_{XX})_{ij} = k(x_i, x_j)$.

- Computational issue: K_{XX} is dense and ill-conditioned.
- Theoretical issue: How to choose kernel?

Kernel Regression Stories



Minimize

$$\lambda \|s\|^2_{\mathcal{H}} + \|s_X - f_X\|^2$$

where $s(x) = \langle d, \psi(x) \rangle_{\mathcal{H}}$ for some *feature map* $\psi : \Omega \to \mathcal{H}$.
Gives $d = \sum_{j=1}^n c_j \psi(x_j)$, kernel is $k(x,y) = \langle \psi(x), \psi(y) \rangle_{\mathcal{H}}$.

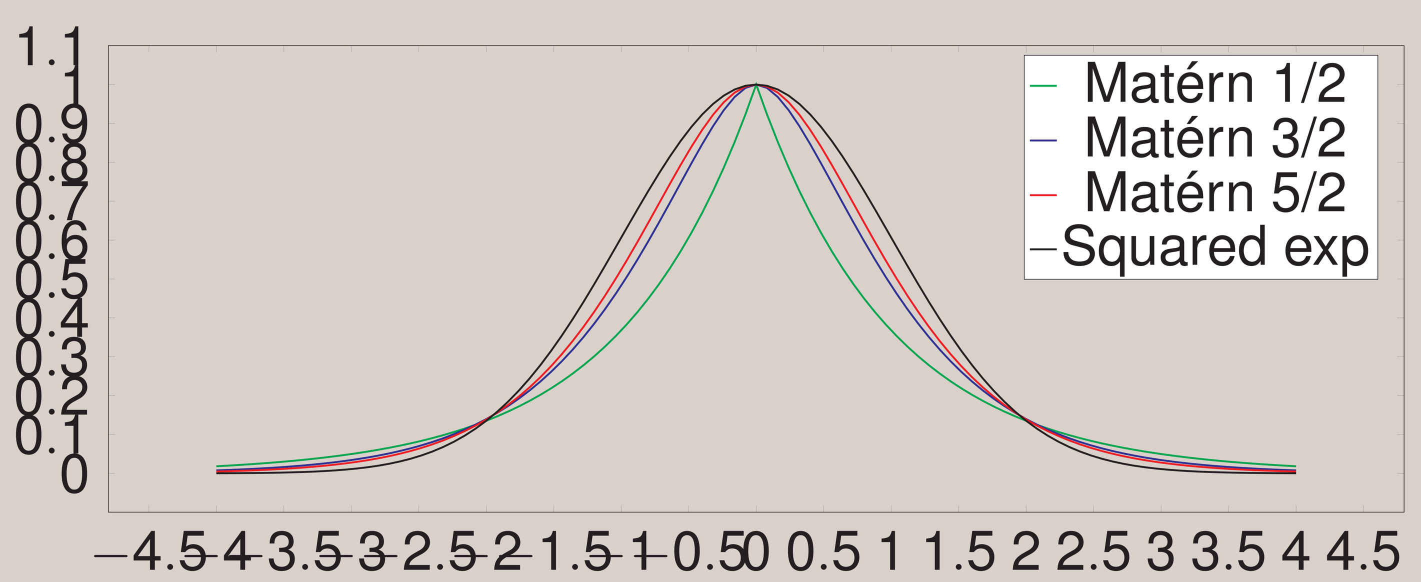
Can reconstruct features if needed from eigenpairs of

$$\mathcal{K} u = \int_\Omega k(x,y) u(y) \, d\Omega(y).$$

Or treat as regularized regression with a *data-dependent basis* determined by sample locations (overcomes Mairhuber-Curtis).

Or Gaussian process: Gaussian random variables indexed by Ω, kernel gives covariance, regression gives posterior mean.

Matérn and SE kernels



Low-Rank Approximation of Kernels

Smooth kernels $\implies$ eigenvalues of K_{XX} decay fast.
Approximate $K_{XX} = UU^T$, regression $\equiv$ regularized LS with U:

$$(U^T U + \lambda I)d = U^T f_X, \quad c = \lambda^{-1}(f_X - Ud).$$

Useful idea: approximate kernel *function*, not kernel *matrix*.
(Or devise an approximate feature map, like rows of U.)

Examples:

- Use inducing points: $k(x,y) = k_{xZ} K_{ZZ}^{-1} k_{Zy}$
- Leading eigenpairs of associated integral operator $\mathcal{K}$ (Mercer)
- Random Fourier features: $k(x,y) = \mathbb{E}_\omega[\exp(\imath \omega^T x) \exp(\imath \omega^T y)^*]$, $\omega \sim$ Fourier transform of (scaled) kernel. Then MC quadrature.

For each: reduced approximation space $\mathcal{U} \subset \mathcal{H}$ and inner product on $\mathcal{U}$ depend on kernel.

Approximation by Chebyshev Features

Alternate idea: Use a kernel-independent $\mathcal{U} \subset \mathcal{H}$ – but kernel determines the inner product.

Concrete 1D case: $k(x,y) = \phi(x-y) = T(x)^T M T(y)$, where

- $T(x) = \left[T_0(x) \; T_1(x) \dots\right]^T$ (Chebyshev features)
- M determined from k

Truncated expansion gives polynomial $s(x) = T(x)d$ with

$$(T_X^T T_X + \lambda M^{-1})d = T_X^T f_X.$$

Constructing the Inner Product

Goal: $\phi(x-y) = T(x)^T M T(y)$.

Approach: Compute $D_k : \ell^2 \to \ell^2$ s.t. $T_k((x-y)/2) = T(x)^T D_k T(y)$. Then

$$\phi(x-y) = \sum_{k=0}^\infty \alpha_k T_k((x-y)/2) = T(x) \left(\sum_{k=0}^\infty \alpha_k D_k \right) T(y).$$

Rewrite recurrence on $T_k(x)$ as operator on $T(x)$ vector:

$$xT_k(x) = \frac{1}{2} \begin{cases} T_{k+1}(x) + T_{k-1}(x), & k > 0 \\ 2T_1(x), & k = 0 \end{cases}$$

$$xT(x) = \frac{1}{2} S T(x), S \equiv \mathrm{tridiag} \begin{pmatrix} 2 \; 1 \; 1 \dots \\ 0 \; 0 \; 0 \; 0 \dots \\ 1 \; 1 \; 1 \dots \end{pmatrix}$$

Then $T_{k+1}(z) = 2zT_k(z) - T_{k-1}(z)$ for $z = (x-y)/2$ yields

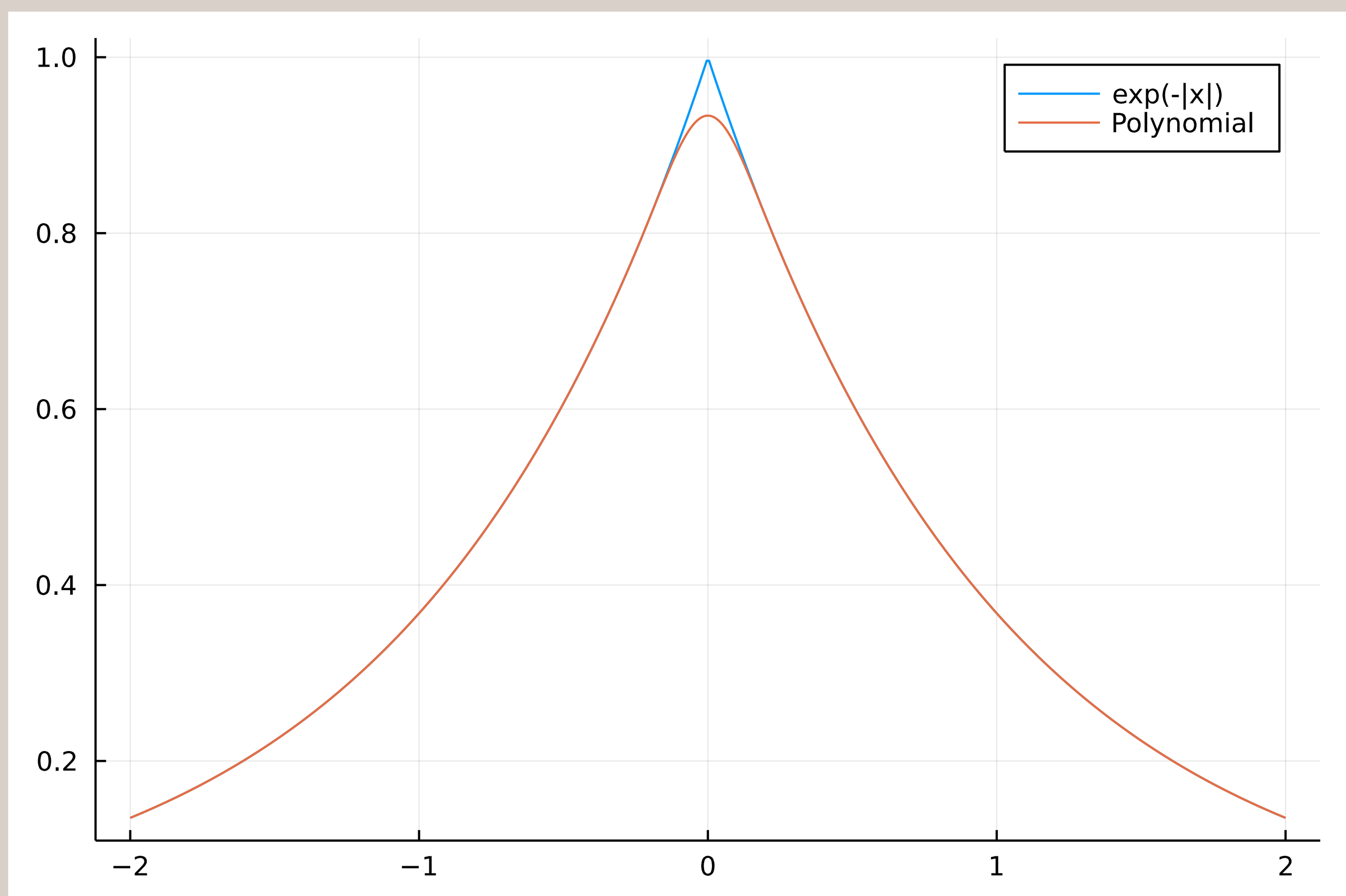
$$T_{k+1}((x-y)/2) = T(x) \left(\frac{1}{2} S^T D_k - \frac{1}{2} D_k S - D_{k-1} \right) T(y)$$

$$\implies D_{k+1} = \frac{1}{2} S^T D_k - \frac{1}{2} D_k S - D_{k-1}$$

with starting values

$$D_0(0:0, 0:0) = 1, \quad D_1(0:1, 0:1) = \frac{1}{2} \begin{bmatrix} 0 & 1 \\ -1 & 0 \end{bmatrix}$$

Splitting the Kernel



Common case: not low rank, lacks regularity near zero. Write

$$\phi(r) \approx \phi_{\mathrm{smooth}}(r) + \phi_{\mathrm{cpt}}(r)$$

where

- $\phi_{\mathrm{smooth}}(r)$ is an even polynomial (treat as above)
- $\phi_{\mathrm{cpt}}(r)$ is supported only near origin

Resulting kernel matrix looks like

$$K_{XX} \approx T_X M T_X + B,$$

where first term is low rank (as above), second term is sparse.

http://www.cs.cornell.edu/~bindel/
bindel@cornell.edu