Energy functionals, kernel discrepancy, minimizing measures and kernel herding

Anatoly Zhigljavsky & Luc Pronzato

Marseille, May 3, 2018

Some notation

$\mathcal{X}$ is a (compact) subset of $\mathbb{R}^d$.
Kernel K is a function $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$.
$\mathcal{M}$ is the set of finite signed measures μ on $\mathcal{X}$.
$\mathcal{M}_1 = \{\mu \in \mathcal{M} : \ \mu(\mathcal{X}) = 1\}$.

Energy: $\Phi(\mu) = \int_{\mathcal{X}} \int_{\mathcal{X}} K(x, x')\mu(dx)\mu(dx'), \ \mu \in \mathcal{M}$.

Minimal energy: $\Phi^* = \inf_{\mu \in \mathcal{M}_1} \Phi(\mu)$.
Capacity of $\mathcal{X}$ (with respect to K): $C^* = 1/\Phi^*$.
Minimizing (equilibrium) measure (may not exist):

$$\mu^* = \arg \min_{\mu \in \mathcal{M}_1} \Phi(\mu) \, .$$

$-\Phi(\mu)$ is also known as Quadratic Rao's Entropy of μ.

Kernel K in $\Phi(\mu) = \int\int K(x, x')\mu(dx)\mu(dx')$

PD: positive definite: $K_N = \|K(x_i, x_j)\|_{i,j=1}^N \geq 0$, $\forall x_1, \ldots, x_N \in \mathcal{X}$
CPD: conditionally positive definite: $K_N \geq 0$, $\sum_{i=1}^N x_i = 0$
SPD: strictly positive definite: $K_N > 0$ $(x_i \neq x_j)$
CSPD: conditionally strictly positive definite
SPD $\Rightarrow$ PD; (S)PD $\Rightarrow$ C(S)PD

Other classes of kernels (possibly unbounded)
IPD: integrally positive definite: $\Phi(\mu) \geq 0$, $\forall \mu \in \mathcal{M}$
ISPD: integrally strictly positive definite: $\Phi(\mu) > 0$, $\forall \mu \neq 0$
CI(S)PD: $\Phi(\mu) \geq 0$, $\forall \mu \in \mathcal{M}$, $\mu(\mathcal{X}) = 0$
CI(S)PD of order k: $\Phi(\mu) \geq 0$, $\forall \mu : \int P_k(x)\mu(dx) = 0$
ISPD $\Rightarrow$ IPD; I(S)PD $\Rightarrow$ CI(S)PD

K is CI(S)PD $\Rightarrow$ Φ is (strictly) convex on $\{\mu \in \mathcal{M}_1 : \Phi(\mu) < \infty\}$

Proof of convexity (C.R.Rao, early 1980th)

$$
\begin{aligned}
&\Phi((1-\alpha)\mu+\alpha\nu) = \\
= &\quad \int\int K(x,x')[(1-\alpha)\mu+\alpha\nu](dx)\cdot[(1-\alpha)\mu+\alpha\nu](dx') \\
= &\quad (1-\alpha)^2\Phi(\mu)+\alpha^2\Phi(\nu)+2\alpha(1-\alpha)\phi(\mu,\nu) \\
= &\quad (1-\alpha)\Phi(\mu)+\alpha\Phi(\nu)-\alpha(1-\alpha)\Phi(\mu-\nu) \\
\leq (<) &\quad (1-\alpha)\Phi(\mu)+\alpha\Phi(\nu)\,.
\end{aligned}
$$

Here

$$
\begin{aligned}
\phi(\mu,\nu) &= \int\int K(x,x')\mu(dx)\nu(dx')\,; \\
\Phi(\mu-\nu) &= \int\int K(x,x')[\mu-\nu](dx)\cdot[\mu-\nu](dx') \\
&= \Phi(\mu)+\Phi(\nu)-2\phi(\mu,\nu)
\end{aligned}
$$

Kernel (maximum mean) discrepancy

If K is CISPD (= 'characteristic') then

$$\gamma_K(\mu, \nu) = \sqrt{\Phi(\mu - \nu)}$$

defines a proper distance (metric) on the space $\mathcal{P}$ of probability measures.
It appears that $\Phi(\mu - \nu)$ is Bregman divergence associated with Φ.
If K is bounded and hence defines RKHS $\mathcal{H}$ then C-S inequality:

$$\left| \int f(x)\mu(dx) - \int f(x)\nu(dx) \right| \leq \|f\|_{\mathcal{H}} \cdot \gamma_K(\mu, \nu)$$

for all $f \in \mathcal{H}$ and $\mu, \nu \in \mathcal{P}$ (Koksma-Hlawka type inequality).

Ref: Sriperumbudur et al. (2010); Sejdinovic et al. (2013)

Example of the kernel discrepancy: distance covariance

$$K(x, x') = \|x\|^\delta + \|x'\|^\delta - \|x - x'\|^\delta, \quad 0 < \delta < 2.$$

This kernel is CISPD. The corresponding energy and discrepancy was studied in many papers by Székely and Székely & Rizzo. All proofs are direct and very technical.

The main case is $\delta = 1$, where the squared kernel discrepancy $\gamma_K^2(\mu, \nu)$ is called distance covariance and Brownian distance covariance.

Székely & Rizzo considered many statistical applications of this discrepancy.

The fact that the theory of the distance covariance by Székely and Székely & Rizzo can be generalized to general kernels was understood by R. Lyons, AoP (2013) (partial generalization) and Sejdinovic et al., AoS (2013).

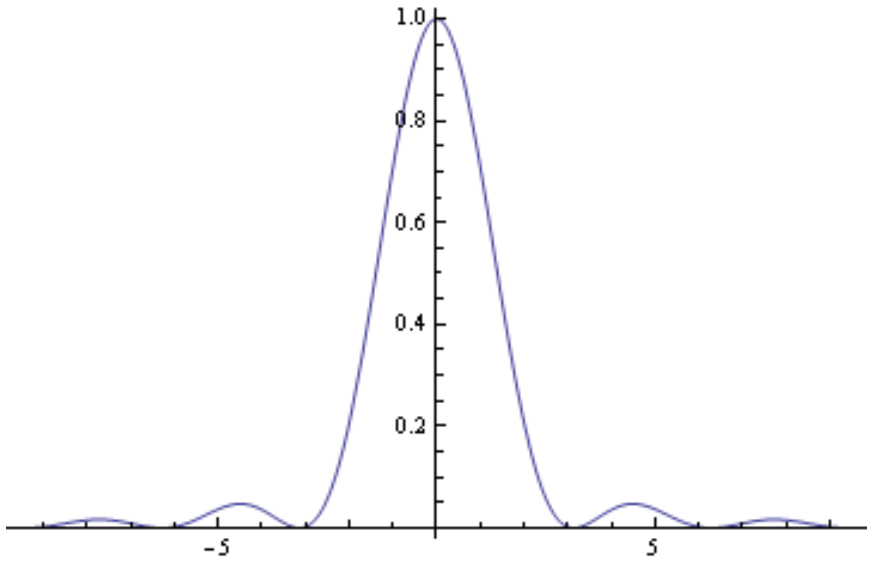
Correlation kernels

In probability, K is usually assumed to be classically PD (bounded).

Random processes and fields with singular correlation kernels appear, for example, in models of quantum gravity.

(C)SPD does not imply I(C)SPD. An example is the sinc squared kernel

$$K(x, x') = \frac{\sin^2(\beta(x - x'))}{(x - x')^2}; \quad \beta > 0.$$



Kernels $K(x, x') = k(x - x')$

Let $K(x, x') = k(x - x')$ with k bounded, continuous PD function and $\Lambda(\omega)$ be the spectral measure of k:

$$k(x) = \int_{\mathbb{R}^d} e^{-ix^T\omega} \, d\Lambda(\omega) \, .$$

Then K is CISPD if and only if the support of Λ coincides with $\mathbb{R}^d$, see Sejdinovic et al., AoS (2013), Th. 9. In this case,

$$\gamma_K^2(\mu, \nu) = \Phi(\mu - \nu) = \int_{\mathbb{R}^d} |\varphi_\mu(\omega) - \varphi_\nu(\omega)|^2 d\Lambda(\omega)$$

where $\varphi_\mu(\omega)$ and $\varphi_\nu(\omega)$ are characteristic functions of probability measures μ and ν.
For the squared sinc kernel $K(x, x') = \sin^2(\beta(x - x'))/(x - x')^2$, the support of $\Lambda(\omega) = \max\{0, 1 - |\omega|/\beta\}$ is $[-\beta, \beta]$.

Directional derivative and potential

Directional derivative of Φ at μ in the direction ν:

$$\begin{aligned} F(\mu;\nu) &= \lim_{\alpha\to 0^+} \frac{\Phi[(1-\alpha)\mu+\alpha\nu]-\Phi_K(\mu)}{\alpha} \\ &= 2\left[\int\int K(x,x')\,d\mu(x')d\nu(x)-\Phi(\mu)\right] \\ &= 2\left[\int P_\mu(x)d\nu(x)-\Phi(\mu)\right] \end{aligned}$$

where

$$P_\mu(x)=\int K(x,x')\,d\mu(x')$$

is the potential of μ at x.

Optimality theorems

Assume K is CISPD and $P_\mu(x) = \int K(x, x')\, d\mu(x')$.

(i) μ_* is the minimum-energy probability measure if and only if

$$P_{\mu^*}(x) \geq \Phi(\mu^*)\,, \quad \forall\, x \in \mathcal{X}\,;$$

we also have $P_{\mu^*}(x) = \Phi(\mu^*)$ on the support of μ^*.

(ii) $\mu_* \in \mathcal{M}_1$ is the minimum-energy signed measure with total mass 1 if and only if

$$P_{\mu^*}(x) = \Phi(\mu^*)\,, \quad \forall\, x \in \mathcal{X}\,.$$

Hajek (1956): if $d = 1$, $K(x, x') = k(x - x')$, k convex then the minimizing signed measure is necessarily a probability measure. LP has noticed that this fact is true in a much more general case when k is subharmonic (any d). Does not require proof but is based on deep results of general potential theory.

Regression with correlated errors, BLUE

Consider a linear regression model:

$$y(x) = \theta_1 f_1(x) + \ldots + \theta_m f_m(x) + \varepsilon(x) = \boldsymbol{\theta}^T f(x) + \varepsilon(x) \ ,$$

where $x \in \mathcal{X} \subset \mathbb{R}^d$, $f(x) = (f_1(x), \ldots, f_m(x))^T$,
$\boldsymbol{\theta} = (\theta_1, \ldots, \theta_m)^T$, $E[\varepsilon(x)] = 0$, $K(x, x') = \mathbb{E}[\varepsilon(x)\varepsilon(x')]$.
For observations at $\{x_1, \ldots, x_N\}$, the BLUE is

$$\widehat{\boldsymbol{\theta}}_{BLUE} = (\mathbf{X}^T \boldsymbol{\Sigma}^{-1} \mathbf{X})^{-1} \mathbf{X}^T \boldsymbol{\Sigma}^{-1} \mathbf{Y},$$

where $\mathbf{X} = (f_i(x_j))_{j=1,\ldots,N}^{i=1,\ldots,m}$ and $\boldsymbol{\Sigma} = (K(x_i, x_j))_{i,j=1,\ldots,N}$.

The BLUE, continuous version

General linear estimator which uses $y(x)$:

$$\hat{\boldsymbol{\theta}}_\zeta = \int y(x)\zeta(dx)\,,$$

where $\zeta(dx)$ is a signed vector-measure.

Th. *If ζ is a vector-measure such that $\int f(x)\zeta^T(dx) = I_m$ (the unbiasedness condition) and there exists matrix D such that*

$$P_\zeta(x) = \int K(x,x')\zeta(dx') = Df(x), \quad \forall x \in X,$$

then ζ defines the BLUE; its covariance matrix is D.

Corollary. Set $m = 1$, $f(x) = 1$ (location-scale model), then the unbiasedness condition becomes $\int \zeta(dx) = 1$ and the optimality condition for a signed measure ζ is $P_\zeta(x) = \mathrm{const}$, $\forall x \in \mathcal{X}$.

Differentiable kernels (joint work with Holger and Andrey)

If K is differentiable then $y(x)$ is differentiable and the general linear estimator uses derivatives, e.g.

$$\hat{\boldsymbol{\theta}}_\zeta = \int y(x)\zeta_0(dx) + \int y'(x)\zeta_1(dx)\,.$$

In this case ($m = 1$, $f(x) = 1$) the BLUE optimality condition becomes

$$\int K(x,x')\zeta_0(dx') + \int \frac{\partial K(x,x')}{\partial x}\zeta_1(dx') = \text{const}, \quad \forall x \in \mathcal{X},$$

Corollary: If $\zeta_1 \neq 0$ (for all optimal measures) then the BLUE, which uses $y(x)$ only, does not exist.

Which functions (e.g. 1) belong to the RKHS?

If K is bounded then potentials

$$P_\mu(x) = \int K(x', x)\, \mu(dx')$$

belong to the RKHS, a space containing all functions $\sum_i a_i K(x, x_i)$ and their limits. What about functions

$$\int K(x', x)\mu_0(dx') + \int K^{(1)}(x', x)\mu_1(dx') + \ldots ?$$

Here

$$K^{(i)}(x, x') = \frac{\partial^i K(x, x')}{\partial x^i}\,.$$

Standard explanation: *'functions in RKHS are as smooth as the kernel'* could be misleading. For example, 1 does not belong to the RKHS of the Gaussian kernel $K(x, x') = \exp\{-\|x - x'\|^2\}$.

$1 \notin$ RKHS for $K(x, x') = \exp\{-|x - x'|^2\}$; Implications for computer experiments. (Based on discussions with Holger)

I.Steinwart, A. Christmann. SVM (2008), Corollary 4.44 (thanks to Bertrand for finding this).
This also follows from the results of W.Hu & M.Stein (2017), who studied the behaviour of $X_N^T W_N^{-1} X_N$ for $W_N = (K(i/N, j/N))_{i,j=1}^N$ and $X_N = (f(i/N))_{i=1}^N$, $f(x) = x^p$, $x \in [0, 1]$.

Computer experiments: Under the assumption that f is a realization of a random field with covariance kernel $\sigma^2 K(x, x')$, the MLE of σ^2 is

$$\hat{\sigma^2} = \frac{1}{N} X_N^T W_N^{-1} X_N = \frac{1}{N \cdot \text{ variance of discrete BLUE}}$$

This can tend to 0, a constant or ∞ depending on the rate of convergence of $X_N^T W_N^{-1} X_N$ to 0 as $N \to \infty$.

$K(x, x') = \exp\{-|x - x'|^2\}$, $\mathcal{X} = [0, 1]$, $N = 100$.

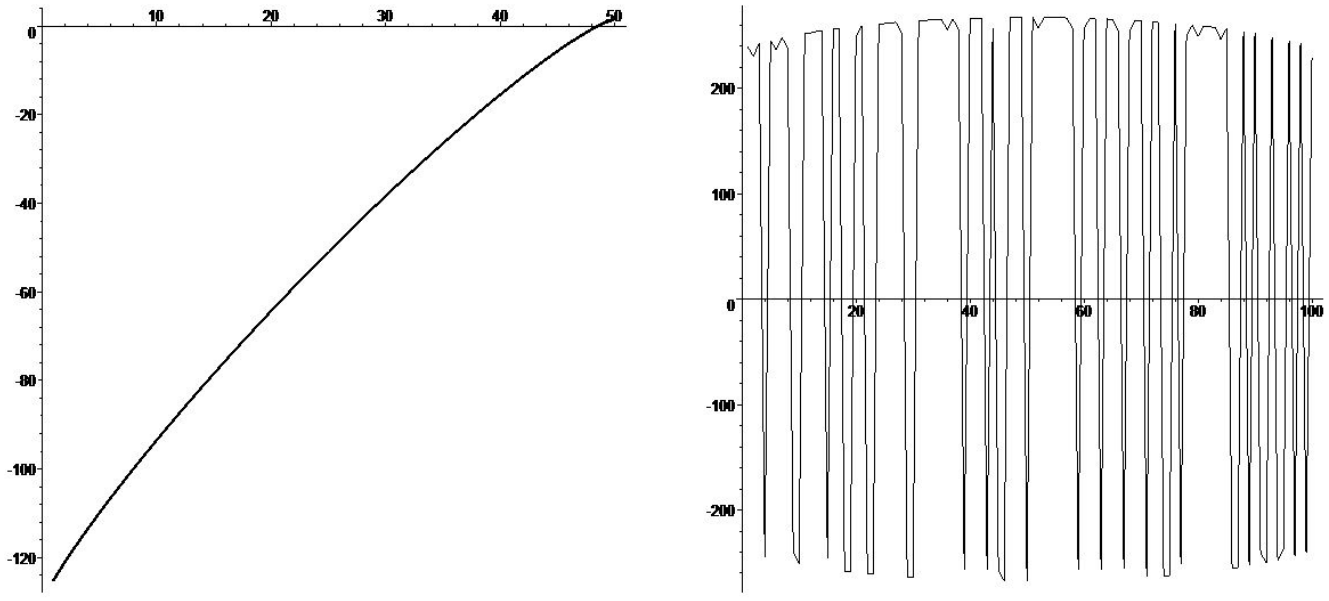


Figure: *Left:* $\log_{10} \lambda_i$ *for the eigenvalues the matrix* $(K(i/N, j/N))_{i,j=1}^N$ *for* $K(x, x') = \exp\{-|x - x'|^2\}$, $\mathcal{X} = [0, 1]$, $N = 100$. *Right: sign times decimal logarithm for the optimal BLUE weights for approximation of 1*

$K(x, x') = \exp\{-|x - x'|^\rho\}$, $\mathcal{X} = [0, 1]$, $N = 100$, $\rho < 2$.

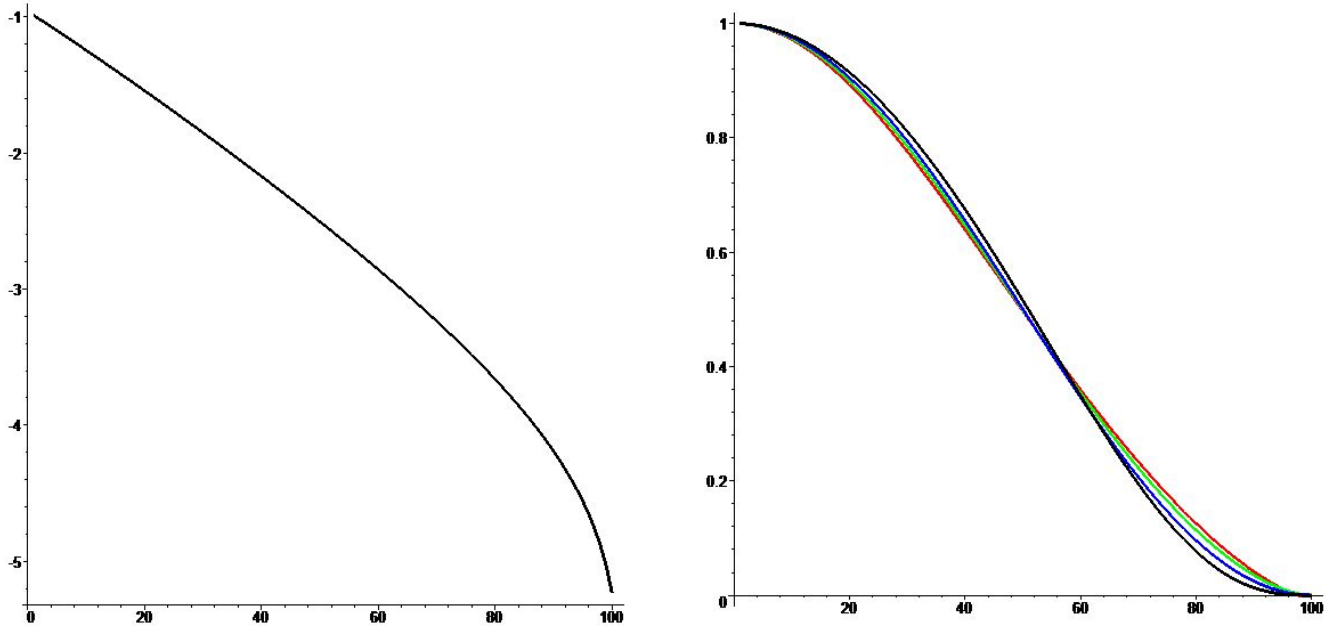


Figure: *Eigenvalues of the matrix* $(K(i/N, j/N))_{i,j=1}^N$ *for different values of* ρ. *Left: decimal logarithm of* $\lambda_{\min}$ ($\rho \in [0.5, 1.999]$). *Right: normalised eigenvalues* $(\lambda_i/\lambda_{\min})^{-1/\rho}$: ρ *=0.5 (black),* ρ *=1 (blue),* ρ *=1.5 (green),* ρ *=1.9999 (red)*

Note $\log_{10}(\lambda_{\min}) \simeq -125$ for $\rho = 2$ and $N = 100$.

Riesz kernel, potential theory

$$K(x,x') = \begin{cases} 1/\|x-x'\|^s\,, & 0<s<d \quad \text{(ISPD)} \\ -\log\|x-x'\|\,, & s=0 \quad \text{(CISPD)} \end{cases}$$

Standard advice (see e.g. Sriperumbudur et al, 2010) is to approximate it with the 'inverse multiquadratic' kernel ($s>0$)

$$K(x,x') = \frac{1}{(\|x-x'\|^2+\varepsilon)^{s/2}} \quad \text{(ISPD)}$$

Riesz kernel: two different approximations

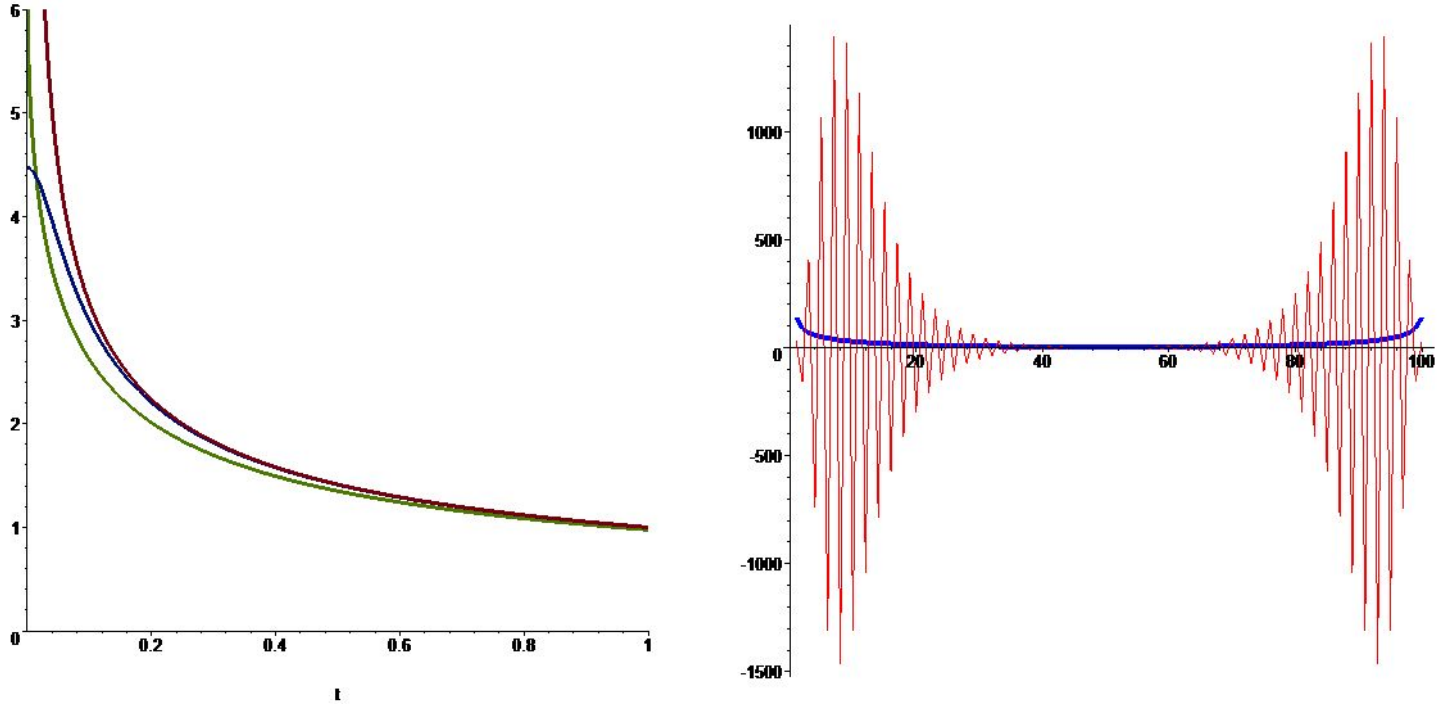


Figure: Left: Brown: $1/\sqrt{x}$; blue: multiquadratic $1/\sqrt{x^2+0.005}$; $x > 0$. Right: BLUE weights for multiquadratic approximation (red) and (scaled) true weights

Completely monotone & Bernstein functions

$f : (0, \infty) \to \mathbb{R}$ is CM if $(-1)^k f^{(k)}(t) > 0$, $\forall t > 0$, $k = 0, 1, \ldots$
$g : (0, \infty) \to \mathbb{R}$ is BF if $g(0) = 0$ and g' is CM
f is CM $\Rightarrow$ Kernel $K(x, x') = f(\|x - x'\|)$ is PD
g is BF $\Rightarrow$ Kernel $K(x, x') = g(\|x - x'\|)$ is CPD
If we take k derivatives, we get CPD of order k.
CM functions may have singularity at 0: $f(t) = t^{-\alpha}, 0 < \alpha < 1$.
A long list of BFs is contained in R.Schilling et all, Bernstein Functions (2010).

Assume g is BF and $f = g'$ is singular at 0. Then $\forall \varepsilon > 0$,

$$f_\varepsilon(t) = \frac{g(t + \varepsilon) - g(t)}{\varepsilon} \qquad (f_\varepsilon(0) = g(\varepsilon)/\varepsilon < \infty)$$

is CM, bounded and can be used as an approximation of f.

Relation between PD and CPD: Schoenberg theorem

Assume the kernel is $K(x, x') = k(x - x')$.
For bounded functions (kernels) we have:
Th. (Schoenberg) A function $k : \mathbb{R}^d \to \mathbb{R}$ is CPD if and only if for all $\beta \in (0, \beta_0)$ the functions $e^{\beta k(\cdot)}$ are PD.

Tomos Phillips has extended this theorem to the case of (possibly unbounded) IPD functions, see his poster.

Example: if $k(x) = -\log(\|x\|)$ then $e^{\beta k(x)} = \|x\|^{-\beta}$.

Eigenvalues of differentiable kernels; Mercer theorem

H. Weyl (1912), J.Reade (1979, 1983-86), C.-W. Ha (1986), J.Cochran (1976-88), etc:
Roughly: if $K \in C^k$ then $\lambda_n \asymp n^{-k-1}$ as $n \to \infty$.

Minimizing measure and the Mercer theorem. Let $\phi_1, \phi_2, \ldots$ be the eigenfunctions of the integral operator $h \to \int K(t, \cdot) h(x) dx$ and f is in the RKHS. Then $f(x) = \sum_{i=1}^{\infty} q_i \phi_i(x)$ for some $\{q_i\}$, $\|f\|^2 = \sum_{i=1}^{\infty} \lambda_i^{-1} q_i^2$, and the BLUE measure is

$$\mu^*(dx) = \|f\|^{-2} \sum_{i=1}^{\infty} \lambda_i^{-1} q_i \phi_i(x) dx \,,$$

if the sum converges for all $x \in \mathcal{X}$.

Kernel herding

This is the problem of approximating a given measure μ by a sequence of N-point (nested) measures $\{\mu_N\}$ so that

$$\gamma_K(\mu, \mu_N) = \sqrt{\Phi(\mu - \mu_N)}$$

is small for all N. Some standard algorithms for construction of optimal designs (like the vertex exchange and some other discussed by Radoslav) can be applied.
Reduced kernel:

$$\tilde{K}(x, x') = K(x, x') - P_\mu(x) - P_\mu(x') + \Phi(\mu)$$

For this kernel, μ is the minimizing measure and $\tilde{\Phi}(\mu) = 0$; the kernel herding can be thought of as a problem of constructing a sequence of N-point measures $\{\mu_N\}$ with small values of $\tilde{\Phi}(\mu_N)$.

Thank you for listening