Maximum Mean Discrepancy, Bayesian integration and kernel herding for space-filling design

Luc Pronzato

(joint work with Anatoly Zhigljavsky, Cardiff Univ.)

Université Côte d'Azur, CNRS, France

Dec. 2, 2021

1 Space-filling design

Objective: approximate $f(\cdot)$ over $\mathscr{X}$ (a compact subset of $\mathbb{R}^d$)
using pairs $(\mathbf{x}_i, f(\mathbf{x}_i))$, $i = 1, 2, \ldots, n$ $\rightarrow$ observe "everywhere"

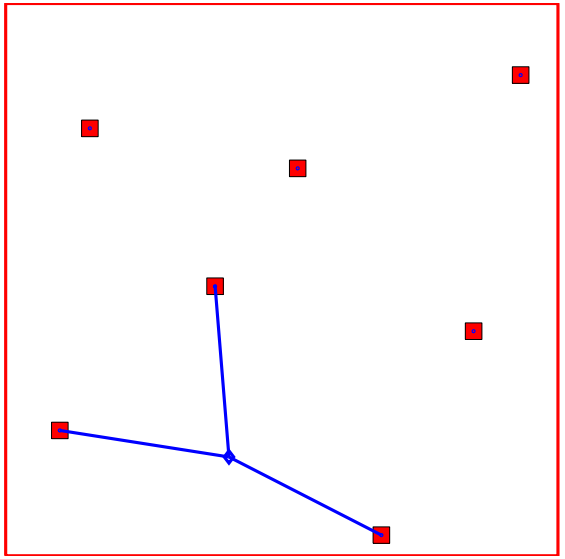
Design $\mathbf{X}_n = \{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$

1 Space-filling design

Objective: approximate $f(\cdot)$ over $\mathscr{X}$ (a compact subset of $\mathbb{R}^d$)
using pairs $(\mathbf{x}_i, f(\mathbf{x}_i))$, $i = 1, 2, \ldots, n$ ⟶ observe "everywhere"
Design $\mathbf{X}_n = \{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$

Covering radius $\mathrm{CR}(\mathbf{X}_n) \triangleq \max_{\mathbf{x} \in \mathscr{X}} \min_{\mathbf{x}_i} \|\mathbf{x} - \mathbf{x}_i\|$

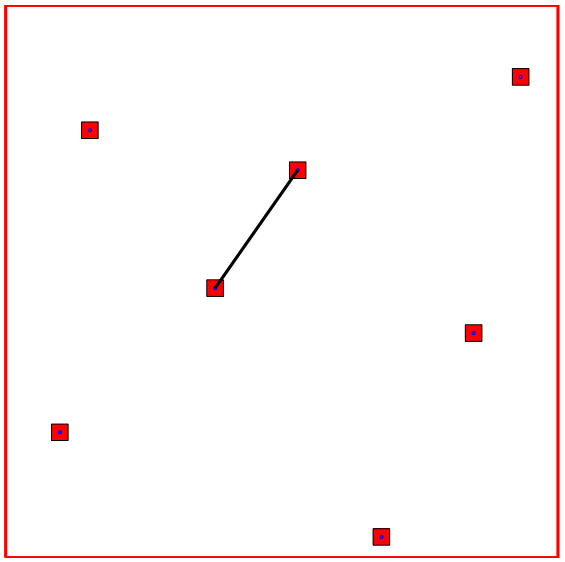
$\mathrm{CR}(\mathbf{X}_n)$ = fill distance = dispersion = miniMax distance criterion



⟶ we are never far from a design point

Packing radius $\mathrm{PR}(\mathbf{X}_n) \triangleq \frac{1}{2} \min_{i \neq j} \|\mathbf{x}_i - \mathbf{x}_j\|$

$$\mathrm{PR}(\mathbf{X}_n) = \text{separation radius} = \frac{1}{2} \text{ Maximin distance criterion}$$

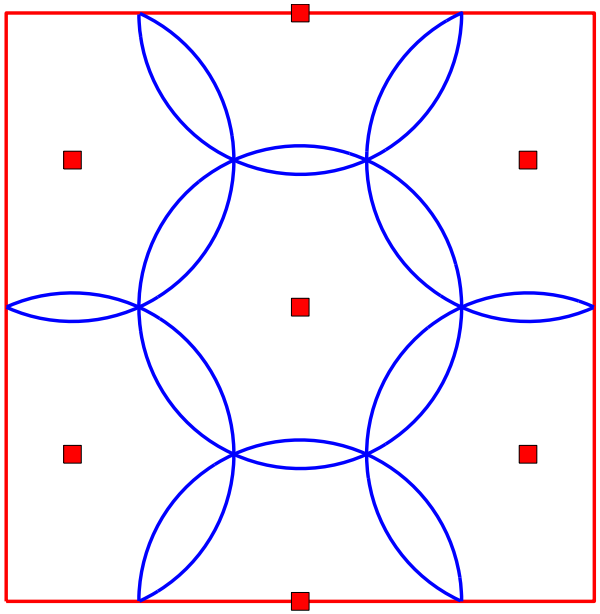


→ easier to compute, but pushes points to the boundary of $\mathscr{X}$

Examples:

① Covering, miniMax

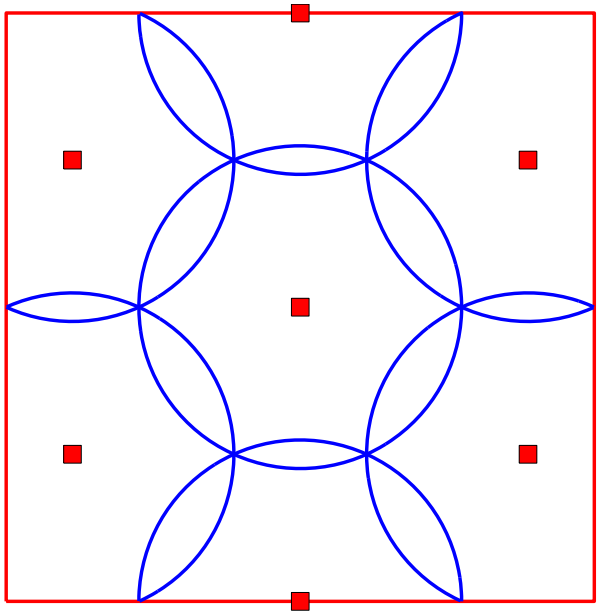
$d = 2, n = 7$ (radius=CR($\mathbf{X}_n$))



Examples:

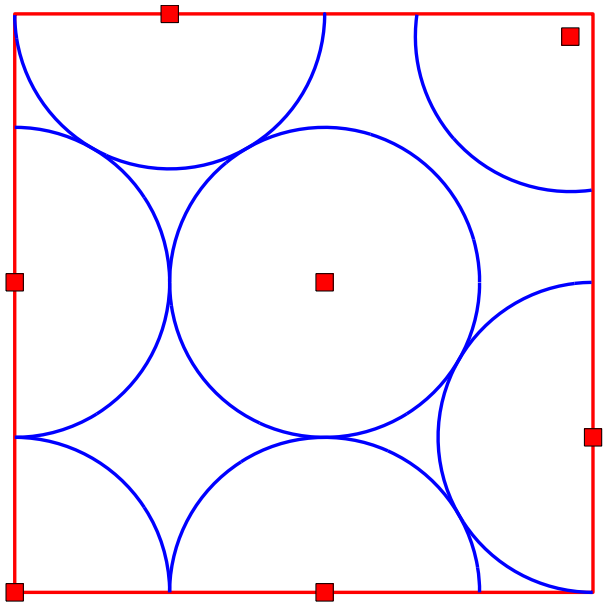
① Covering, miniMax

$d = 2, n = 7$ (radius=CR($\mathbf{X}_n$))



② Packing, Maximin

$d = 2, n = 7$ (radius=P($\mathbf{X}_n$))



→ Minimising CR is preferable to maximising PR; both are difficult

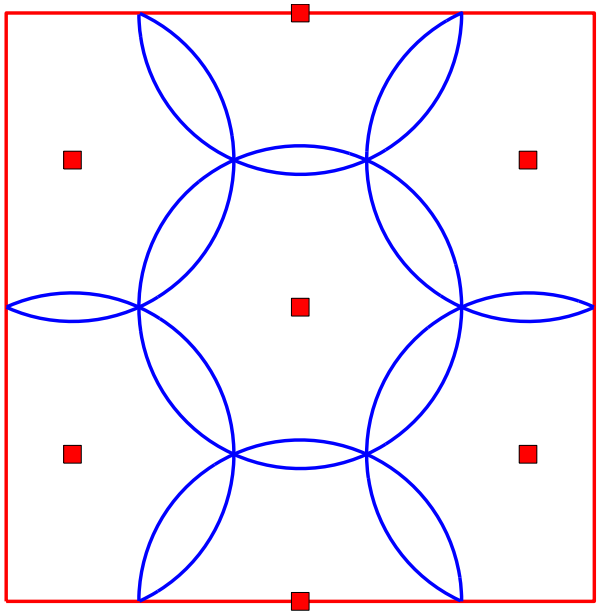
Bounds on approximation error are related to CR:

$$\|\text{error}\|_{L_\infty(\mathscr{X})} \leq \text{CR}(\mathbf{X}_n)^{\alpha - d/2} \text{ when } f \in \text{Sobolev space } W_2^\alpha(\mathscr{X})$$

Examples:

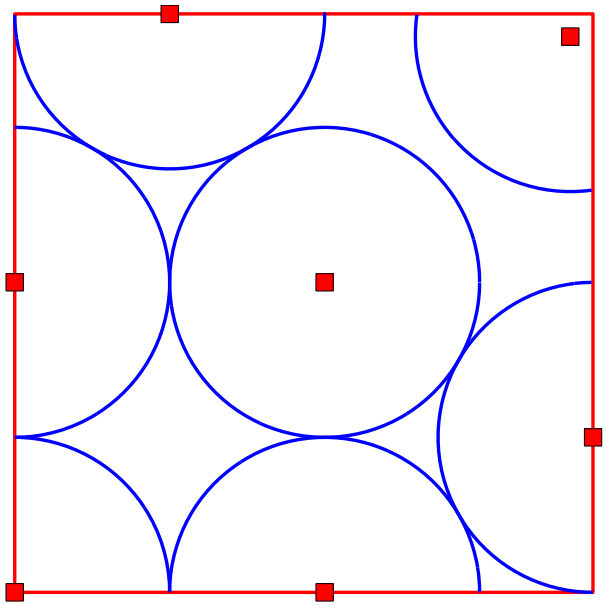
① Covering, miniMax

$d = 2, n = 7$ (radius=CR($\mathbf{X}_n$))



② Packing, Maximin

$d = 2, n = 7$ (radius=P($\mathbf{X}_n$))



→ Minimising CR is preferable to maximising PR; both are difficult

Bounds on approximation error are related to CR:

$$\|\text{error}\|_{L_\infty(\mathscr{X})} \leq \text{CR}(\mathbf{X}_n)^{\alpha - d/2} \text{ when } f \in \text{Sobolev space } W_2^\alpha(\mathscr{X})$$

Any-time design: good-space-filling properties for $\mathbf{X}_k \subset \mathbf{X}_{k+1} \subset \mathbf{X}_{k+3} \subset \ldots$?

Can we apply a greedy gradient-type descent algorithm to a suitable criterion?

2 Bayesian integration

Notation:

- $\mathscr{M} = \mathscr{M}[\mathscr{X}]$ set of finite signed Borel measures on $\mathscr{X}$
- $\mathscr{M}(\alpha) = \{\mu \in \mathscr{M} : \mu(\mathscr{X}) = \alpha\}$ (signed measures with total mass $\alpha \in \mathbb{R}$)
- $\mathscr{M}^+$ positive measures in $\mathscr{M}$
- $\mathscr{M}^+(1)$ probability measures on $\mathscr{X}$

Objective: integrate f with respect to $\mu \in \mathscr{M}^+(1)$

$\rightarrow$ estimate $\boxed{I_\mu(f) = \mathsf{E}_\mu\{f(\mathbf{X})\} \triangleq \int_{\mathscr{X}} f(\mathbf{x})\,\mathrm{d}\mu(\mathbf{x})}$

2 Bayesian integration

Notation:

- $\mathscr{M} = \mathscr{M}[\mathscr{X}]$ set of finite signed Borel measures on $\mathscr{X}$
- $\mathscr{M}(\alpha) = \{\mu \in \mathscr{M} : \mu(\mathscr{X}) = \alpha\}$ (signed measures with total mass $\alpha \in \mathbb{R}$)
- $\mathscr{M}^+$ positive measures in $\mathscr{M}$
- $\mathscr{M}^+(1)$ probability measures on $\mathscr{X}$

Objective: integrate f with respect to $\mu \in \mathscr{M}^+(1)$

$\rightarrow$ estimate $\boxed{I_\mu(f) = \mathrm{E}_\mu\{f(\mathbf{X})\} \triangleq \int_{\mathscr{X}} f(\mathbf{x})\,\mathrm{d}\mu(\mathbf{x})}$

Suppose $\boxed{f(\mathbf{x}) = \beta + Z_x}$, with Z_x a Gaussian RF,
$\mathbb{E}\{Z_x\} = 0$ and $\mathbb{E}\{Z_x Z_{x'}\} = \sigma^2\, K(\mathbf{x}, \mathbf{x}')$
and a vague prior on β ($\sim \mathcal{N}(0, \sigma^2\, A)$ with $A \to +\infty$)

(K symmetric positive definite $\rightarrow$ defines a RKHS $\mathcal{H}_K$)

$\rightarrow$ observe f at $\mathbf{X}_n = \{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$

Bayesian integration of f:

Denote $\mathbf{1}_n = (1, \ldots, 1)^T$, $\mathbf{y}_n = (f(\mathbf{x}_1), \ldots, f(\mathbf{x}_n))^T$, $\{\mathbf{K}_n\}_{i,j} = K(\mathbf{x}_i, \mathbf{x}_j)$,

$$P_\mu(\mathbf{x}) \triangleq \int_{\mathscr{X}} K(\mathbf{x}, \mathbf{x}')\, \mathrm{d}\mu(\mathbf{x}') \text{ (= kernel imbedding of } \mu \text{ into } \mathcal{H}_K)$$

$$\mathscr{E}_K(\mu) \triangleq \int_{\mathscr{X}^2} K(\mathbf{x}, \mathbf{x}')\, \mathrm{d}\mu(\mathbf{x}) \mathrm{d}\mu(\mathbf{x}') \text{ (= energy of } \mu)$$

Bayesian integration of f:

Denote $\mathbf{1}_n = (1, \ldots, 1)^T$, $\mathbf{y}_n = (f(\mathbf{x}_1), \ldots, f(\mathbf{x}_n))^T$, $\{\mathbf{K}_n\}_{i,j} = K(\mathbf{x}_i, \mathbf{x}_j)$,

$$P_\mu(\mathbf{x}) \triangleq \int_{\mathscr{X}} K(\mathbf{x}, \mathbf{x}')\, \mathrm{d}\mu(\mathbf{x}') \text{ (= kernel imbedding of } \mu \text{ into } \mathcal{H}_K)$$

$$\mathscr{E}_K(\mu) \triangleq \int_{\mathscr{X}^2} K(\mathbf{x}, \mathbf{x}')\, \mathrm{d}\mu(\mathbf{x}) \mathrm{d}\mu(\mathbf{x}') \text{ (= energy of } \mu)$$

$I_\mu(f)$ has the normal posterior $\mathcal{N}(\widehat{I}_n, \sigma^2 s_n^2)$ (O'Hagan, 1991)

$$\begin{aligned} \widehat{I}_n &= \hat{\beta}^n + \mathbf{p}_n(\mu)^T \mathbf{K}_n^{-1}(\mathbf{y}_n - \hat{\beta}^n \mathbf{1}_n) \\ &= \hat{\mathbf{w}}_n^T \mathbf{y}_n \quad \text{(the estimation is linear)} \\ s_n^2 &= \mathscr{E}_K(\mu) - \mathbf{p}_n^T(\mu) \mathbf{K}_n^{-1} \mathbf{p}_n(\mu) + \frac{(1 - \mathbf{p}_n^T(\mu) \mathbf{K}_n^{-1} \mathbf{1}_n)^2}{\mathbf{1}_n^T \mathbf{K}_n^{-1} \mathbf{1}_n} \end{aligned}$$

where $\hat{\beta}^n = \dfrac{\mathbf{1}_n^T \mathbf{K}_n^{-1} \mathbf{y}_n}{\mathbf{1}_n^T \mathbf{K}_n^{-1} \mathbf{1}_n}$ (= BLUE) and $\mathbf{p}_n(\mu) = (P_\mu(\mathbf{x}_1), \ldots, P_\mu(\mathbf{x}_n))^T$

The approximation of $f(\mathbf{x}_0)$ is a particular case!

Take $\mu = \delta_{\mathbf{x}_0}$ → $\boxed{P_\mu(\mathbf{x}) = K(\mathbf{x}, \mathbf{x}_0),\ \mathscr{E}_K(\mu) = K(\mathbf{x}_0, \mathbf{x}_0)}$

and $f(\mathbf{x}_0)$ has the normal posterior $\mathcal{N}(\hat{\eta}_n(\mathbf{x}_0), \sigma^2\, \rho_n^2(\mathbf{x}_0))$ with

$$\begin{aligned}
\hat{\eta}_n(\mathbf{x}_0) &= \hat{\beta}^n + \mathbf{k}_n^T(\mathbf{x}_0)\mathbf{K}_n^{-1}(\mathbf{y}_n - \hat{\beta}^n \mathbf{1}_n) \\
\rho_n^2(\mathbf{x}_0) &= K(\mathbf{x}_0, \mathbf{x}_0) - \mathbf{k}_n^T(\mathbf{x}_0)\mathbf{K}_n^{-1}\mathbf{k}_n(\mathbf{x}_0) + \frac{(1 - \mathbf{k}_n^T(\mathbf{x}_0)\mathbf{K}_n^{-1}\mathbf{1}_n)^2}{\mathbf{1}_n^T\mathbf{K}_n^{-1}\mathbf{1}_n}
\end{aligned}$$

where $\{\mathbf{k}_n(\mathbf{x})\}_i = K(\mathbf{x}, \mathbf{x}_i)$

$\equiv$ ordinary kriging

→ minimise $\boxed{\mathrm{IMSPE}(\mathbf{X}_n) = \sigma^2 \int_{\mathscr{X}} \rho_n^2(\mathbf{x})\, \mathrm{d}\mu(\mathbf{x})}$

The approximation of $f(\mathbf{x}_0)$ is a particular case!

Take $\mu = \delta_{\mathbf{x}_0}$ ➞ $\boxed{P_\mu(\mathbf{x}) = K(\mathbf{x}, \mathbf{x}_0),\ \mathscr{E}_K(\mu) = K(\mathbf{x}_0, \mathbf{x}_0)}$

and $f(\mathbf{x}_0)$ has the normal posterior $\mathcal{N}(\hat{\eta}_n(\mathbf{x}_0), \sigma^2\, \rho_n^2(\mathbf{x}_0))$ with

$$\begin{aligned}
\hat{\eta}_n(\mathbf{x}_0) &= \hat{\beta}^n + \mathbf{k}_n^T(\mathbf{x}_0)\mathbf{K}_n^{-1}(\mathbf{y}_n - \hat{\beta}^n \mathbf{1}_n) \\
\rho_n^2(\mathbf{x}_0) &= K(\mathbf{x}_0, \mathbf{x}_0) - \mathbf{k}_n^T(\mathbf{x}_0)\mathbf{K}_n^{-1}\mathbf{k}_n(\mathbf{x}_0) + \frac{(1 - \mathbf{k}_n^T(\mathbf{x}_0)\mathbf{K}_n^{-1}\mathbf{1}_n)^2}{\mathbf{1}_n^T\mathbf{K}_n^{-1}\mathbf{1}_n}
\end{aligned}$$

where $\{\mathbf{k}_n(\mathbf{x})\}_i = K(\mathbf{x}, \mathbf{x}_i)$

$\equiv$ ordinary kriging

➞ minimise $\boxed{\mathrm{IMSPE}(\mathbf{X}_n) = \sigma^2 \int_{\mathscr{X}} \rho_n^2(\mathbf{x})\, \mathrm{d}\mu(\mathbf{x})}$

➠ For $\underline{\mu \text{ uniform on } \mathscr{X}}$, a design minimising s_n^2 should be well spread

➠ Minimise $s_n^2 = \mathscr{E}_K(\mu) - \mathbf{p}_n^T(\mu)\mathbf{K}_n^{-1}\mathbf{p}_n(\mu) + \frac{(1-\mathbf{p}_n^T(\mu)\mathbf{K}_n^{-1}\mathbf{1}_n)^2}{\mathbf{1}_n^T\mathbf{K}_n^{-1}\mathbf{1}_n}$

instead of $\int_{\mathscr{X}} \rho_n^2(\mathbf{x})\, \mathrm{d}\mu(\mathbf{x})$ for space-filling design?

3 Kernel discrepancy, energy & potentials

Very much based on:

Damelin, S., Hickernell, F., Ragozin, D., Zeng, X., 2010. On energy, discrepancy and group invariant measures on measurable subsets of Euclidean space. *J. Fourier Anal. Appl.* 16, 813–839.

Sriperumbudur, B., Gretton, A., Fukumizu, K., Schölkopf, B., Lanckriet, G., 2010. Hilbert space embeddings and metrics on probability measures. *Journal of Machine Learning Research* 11 (Apr), 1517–1561.

Sejdinovic, S., Sriperumbudur, B., Gretton, A., Fukumizu, K., 2013. Equivalence of distance-based and RKHS-based statistics in hypothesis testing. *The Annals of Statistics* 41 (5), 2263–2291.

Pronzato, L., Zhigljavsky, A., 2020. Bayesian quadrature, energy minimization and space-filling design. SIAM/ASA J. Uncertainty Quantification 8 (3), 959–1011.

3 Kernel discrepancy, energy & potentials

Very much based on:

Damelin, S., Hickernell, F., Ragozin, D., Zeng, X., 2010. On energy, discrepancy and group invariant measures on measurable subsets of Euclidean space. *J. Fourier Anal. Appl.* 16, 813–839.

Sriperumbudur, B., Gretton, A., Fukumizu, K., Schölkopf, B., Lanckriet, G., 2010. Hilbert space embeddings and metrics on probability measures. *Journal of Machine Learning Research* 11 (Apr), 1517–1561.

Sejdinovic, S., Sriperumbudur, B., Gretton, A., Fukumizu, K., 2013. Equivalence of distance-based and RKHS-based statistics in hypothesis testing. *The Annals of Statistics* 41 (5), 2263–2291.

Pronzato, L., Zhigljavsky, A., 2020. Bayesian quadrature, energy minimization and space-filling design. SIAM/ASA J. Uncertainty Quantification 8 (3), 959–1011.

K = a symmetric positive definite kernel, defines a RKHS $\mathcal{H}_K$

For ν a signed measure,

$$\mathscr{E}_K(\nu) \triangleq \int_{\mathscr{X}^2} K(\mathbf{x},\mathbf{x}')\,\mathrm{d}\nu(\mathbf{x})\mathrm{d}\nu(\mathbf{x}') = \text{energy of } \nu$$

$$P_\nu(\mathbf{x}) \triangleq \int_{\mathscr{X}} K(\mathbf{x},\mathbf{x}')\,\mathrm{d}\nu(\mathbf{x}') = \text{potential of } \nu \text{ at } \mathbf{x}$$

$$[P_\nu(\cdot) = \text{kernel imbedding of } \nu \text{ into } \mathcal{H}_K]$$

Suppose $\underline{f \in \mathcal{H}_K}$, μ, ν prob. measures with finite energy

$K_{\mathbf{x}}(\cdot) = K(\mathbf{x}, \cdot)$, reproducing property $f(\mathbf{x}) = \langle f, K_{\mathbf{x}} \rangle_{\mathcal{H}_K} \quad \Rightarrow$

$$|I_\mu(f) - I_\nu(f)| = \left| \int_{\mathscr{X}} \langle f, K_{\mathbf{x}} \rangle_{\mathcal{H}_K} \, \mathrm{d}(\mu - \nu)(\mathbf{x}) \right| = |\langle f, P_\mu - P_\nu \rangle_{\mathcal{H}_K}|$$

CS inequality ➞ a Koksma-Hlawka type inequality:

$$\left| \int_{\mathscr{X}} f(\mathbf{x}) \, \mathrm{d}\nu(\mathbf{x}) - \int_{\mathscr{X}} f(\mathbf{x}) \, \mathrm{d}\mu(\mathbf{x}) \right| \leq \|f\|_{\mathcal{H}_K} \gamma_K(\mu, \nu)$$

where $\boxed{\gamma_K^2(\mu, \nu) \triangleq \|P_\mu - P_\nu\|_{\mathcal{H}_K}^2 = \mathscr{E}_K(\nu - \mu)}$

$\gamma_K(\mu, \nu) = \sup_{\|f\|_{\mathcal{H}_K} = 1} |\int_{\mathscr{X}} f(\mathbf{x}) \, \mathrm{d}\nu(\mathbf{x}) - \int_{\mathscr{X}} f(\mathbf{x}) \, \mathrm{d}\mu(\mathbf{x})|$

$=$ **Maximum Mean Discrepancy** between (MMD) μ and ν

(Sriperumbudur et al., 2010; Sejdinovic et al., 2013)

Suppose $\underline{f \in \mathcal{H}_K}$, μ, ν prob. measures with finite energy

$K_{\mathbf{x}}(\cdot) = K(\mathbf{x}, \cdot)$, reproducing property $f(\mathbf{x}) = \langle f, K_{\mathbf{x}} \rangle_{\mathcal{H}_K} \Rightarrow$

$$|I_\mu(f) - I_\nu(f)| = \left| \int_{\mathscr{X}} \langle f, K_{\mathbf{x}} \rangle_{\mathcal{H}_K} \, \mathrm{d}(\mu - \nu)(\mathbf{x}) \right| = |\langle f, P_\mu - P_\nu \rangle_{\mathcal{H}_K}|$$

CS inequality ➞ a Koksma-Hlawka type inequality:

$$\left| \int_{\mathscr{X}} f(\mathbf{x}) \, \mathrm{d}\nu(\mathbf{x}) - \int_{\mathscr{X}} f(\mathbf{x}) \, \mathrm{d}\mu(\mathbf{x}) \right| \leq \|f\|_{\mathcal{H}_K} \gamma_K(\mu, \nu)$$

where $\boxed{\gamma_K^2(\mu, \nu) \triangleq \|P_\mu - P_\nu\|_{\mathcal{H}_K}^2 = \mathscr{E}_K(\nu - \mu)}$

$\gamma_K(\mu, \nu) = \sup_{\|f\|_{\mathcal{H}_K}=1} |\int_{\mathscr{X}} f(\mathbf{x}) \, \mathrm{d}\nu(\mathbf{x}) - \int_{\mathscr{X}} f(\mathbf{x}) \, \mathrm{d}\mu(\mathbf{x})|$

$=$ **Maximum Mean Discrepancy** between (MMD) μ and ν

(Sriperumbudur et al., 2010; Sejdinovic et al., 2013)

Space-filling design: take μ uniform on $\mathscr{X}$

➞ find ξ_n (with n support points) minimising $\mathscr{E}_K(\xi_n - \mu) = \mathrm{MMD}^2(\xi_n, \mu)$

$\gamma_K(\cdot,\cdot)$ defines a pseudo-metric on $\mathscr{M}^+(1)$
Does it define a metric? ($\Leftrightarrow$ K is **characteristic**)

Definition

K is Integrally Strictly Positive Definite (ISPD) on $\mathscr{M}$ when $\mathscr{E}_K(\nu) > 0$ for any nonzero $\nu \in \mathscr{M}$

Definition

K is Conditionally Integrally Strictly Positive Definite (CISPD) on $\mathscr{M}$ when it is ISPD on $\mathscr{M}(0)$; that is, when $\mathscr{E}_K(\nu) > 0$ for all nonzero $\nu \in \mathscr{M}$ with $\nu(\mathscr{X}) = 0$

$\gamma_K(\cdot,\cdot)$ defines a pseudo-metric on $\mathscr{M}^+(1)$
Does it define a metric? ($\Leftrightarrow$ K is **characteristic**)

Definition

K is Integrally Strictly Positive Definite (ISPD) on $\mathscr{M}$ when $\mathscr{E}_K(\nu) > 0$ for any nonzero $\nu \in \mathscr{M}$

Definition

K is Conditionally Integrally Strictly Positive Definite (CISPD) on $\mathscr{M}$ when it is ISPD on $\mathscr{M}(0)$; that is, when $\mathscr{E}_K(\nu) > 0$ for all nonzero $\nu \in \mathscr{M}$ with $\nu(\mathscr{X}) = 0$

Sriperumbudur et al. (2010):

- K bounded & ISPD $\Rightarrow$ K is strictly positive definite
  ($\rightarrow$ defines a RKHS $\mathcal{H}_K$)
- if K uniformly bounded: characteristic $\Leftrightarrow$ CISPD

$\gamma_K(\cdot,\cdot)$ defines a pseudo-metric on $\mathscr{M}^+(1)$
Does it define a metric? ($\Leftrightarrow$ K is **characteristic**)

Definition

K is Integrally Strictly Positive Definite (ISPD) on $\mathscr{M}$ when $\mathscr{E}_K(\nu) > 0$ for any nonzero $\nu \in \mathscr{M}$

Definition

K is Conditionally Integrally Strictly Positive Definite (CISPD) on $\mathscr{M}$ when it is ISPD on $\mathscr{M}(0)$; that is, when $\mathscr{E}_K(\nu) > 0$ for all nonzero $\nu \in \mathscr{M}$ with $\nu(\mathscr{X}) = 0$

Sriperumbudur et al. (2010):

- K bounded & ISPD $\Rightarrow$ K is strictly positive definite
  ($\rightarrow$ defines a RKHS $\mathcal{H}_K$)
- if $\underbrace{K \text{ uniformly bounded}}_{\text{assumed in the following}}$: characteristic $\Leftrightarrow$ CISPD

- Spectral interpretation (Sriperumbudur et al., 2010):
  if K translation invariant, $K(\mathbf{x}, \mathbf{x}') = \psi(\mathbf{x} - \mathbf{x}')$, with ψ bounded, continuous, pos. def., $\psi(\mathbf{x}) = \int_{\mathbb{R}^d} e^{-i\mathbf{x}^T\boldsymbol{\omega}} \, \mathrm{d}\Lambda(\boldsymbol{\omega})$ (Bochner)

  $\gamma_K(\mu, \nu) = L_2$ distance between characteristic functions of μ and ν:

$$\gamma_K(\mu, \nu) = \left[\int_{\mathbb{R}^d} |\Phi_\mu(\boldsymbol{\omega}) - \Phi_\nu(\boldsymbol{\omega})|^2 \, \mathrm{d}\Lambda(\boldsymbol{\omega}) \right]^{1/2}$$

  K characteristic $\Leftrightarrow$ Λ supported on $\mathbb{R}^d$
  (for example, Matérn kernels)

For many kernels K:

- $\gamma_K(\cdot,\cdot)$ defines a metric for probability measures
- $\mathscr{E}_K(\cdot)$ is strictly convex

For many kernels K:

- $\gamma_K(\cdot,\cdot)$ defines a metric for probability measures
- $\mathscr{E}_K(\cdot)$ is strictly convex

Examples:

- Matérn kernels, e.g., $K_{1/2,\theta}(\mathbf{x},\mathbf{x}') = \exp(-\theta\,\|\mathbf{x}-\mathbf{x}'\|)$,

$$\begin{aligned} K_{3/2,\theta}(\mathbf{x},\mathbf{x}') &= (1+\sqrt{3}\theta\,\|\mathbf{x}-\mathbf{x}'\|)\,\exp(-\sqrt{3}\theta\,\|\mathbf{x}-\mathbf{x}'\|) \\ K_{5/2,\theta}(\mathbf{x},\mathbf{x}') &= [1+\sqrt{5}\theta\,\|\mathbf{x}-\mathbf{x}'\|+5\theta^2\,\|\mathbf{x}-\mathbf{x}'\|^2/3]\,\exp(-\sqrt{5}\theta\,\|\mathbf{x}-\mathbf{x}'\|) \\ &\ldots \end{aligned}$$

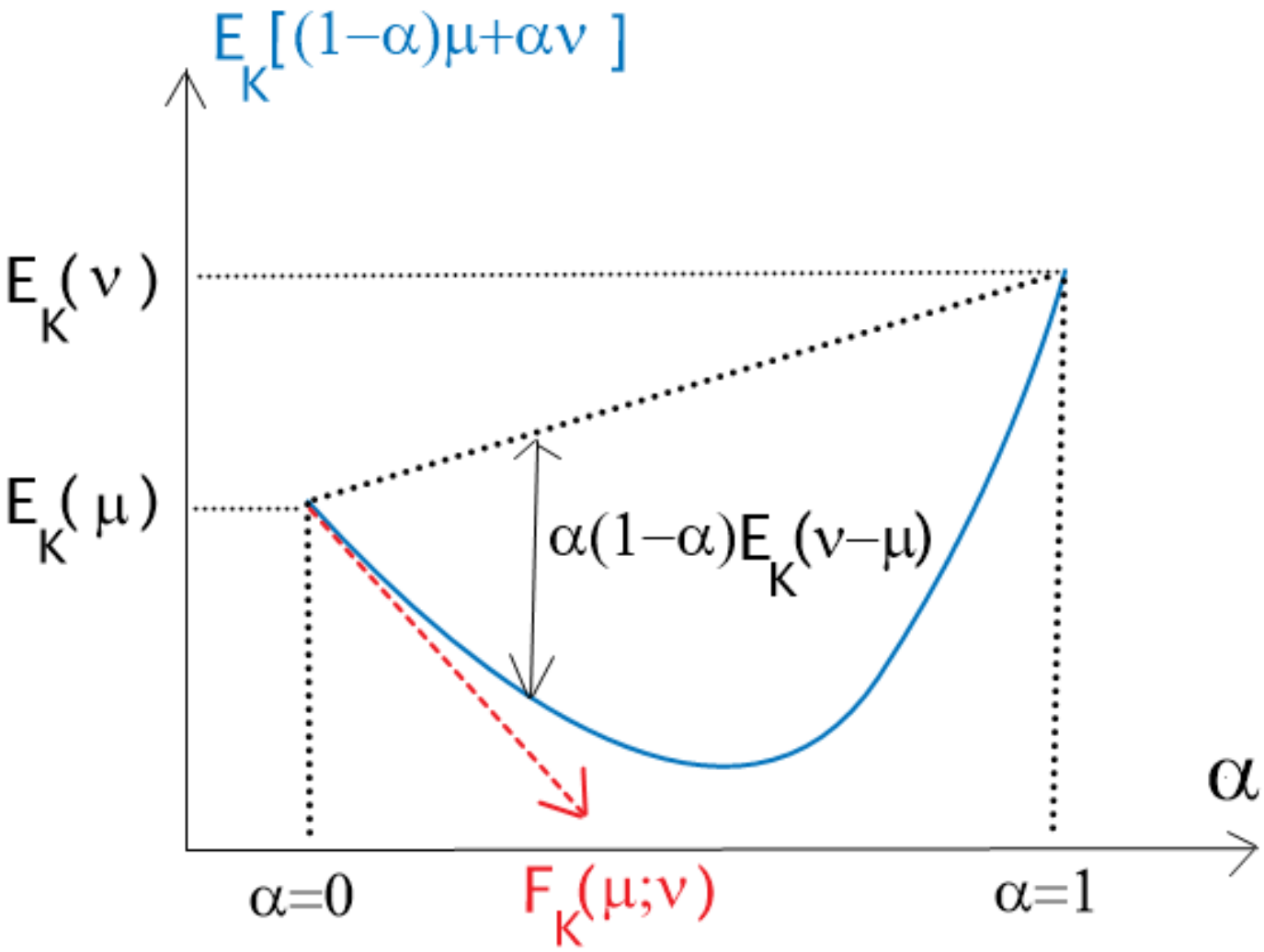
- distance-induced kernels (→ energy distance of Székely and Rizzo (2013)):

$$K'_{(s)}(\mathbf{x},\mathbf{x}') = \|\mathbf{x}\|^s + \|\mathbf{x}'\|^s - \|\mathbf{x}-\mathbf{x}'\|^s\,,\ 0<s<2$$

- Riesz kernels (unbounded):

$$K_{(s)}(\mathbf{x},\mathbf{x}') = \|\mathbf{x}-\mathbf{x}'\|^{-s}\,,\ s\in(0,d),\ K_{(0)}(\mathbf{x},\mathbf{x}') = -\log\|\mathbf{x}-\mathbf{x}'\|$$

- ...


$E_K[(1-\alpha)\mu+\alpha\nu]$
$E_K(\nu)$
$E_K(\mu)$
$\alpha(1-\alpha)E_K(\nu-\mu)$
α
$\alpha=0$
$F_K(\mu;\nu)$
$\alpha=1$

For $\alpha \in (0,1)$, $\mu, \nu \in \mathscr{M}$:

$$\boxed{(1-\alpha)\,\mathscr{E}_K(\mu) + \alpha\,\mathscr{E}_K(\nu) - \mathscr{E}_K[(1-\alpha)\mu + \alpha\nu] = \alpha(1-\alpha)\,\mathscr{E}_K(\nu - \mu)}$$

- K ISPD $\Leftrightarrow$ $\mathscr{E}_K(\cdot)$ strictly convex on $\mathscr{M}$
- K CISPD $\Leftrightarrow$ $\mathscr{E}_K(\cdot)$ strictly convex on $\mathscr{M}(1)$ (or $\mathscr{M}^+(1)$)

For $\alpha \in (0,1)$, $\mu, \nu \in \mathscr{M}$:

$$\boxed{(1-\alpha)\,\mathscr{E}_K(\mu) + \alpha\,\mathscr{E}_K(\nu) - \mathscr{E}_K[(1-\alpha)\mu + \alpha\nu] = \alpha(1-\alpha)\,\mathscr{E}_K(\nu - \mu)}$$

- K ISPD $\Leftrightarrow$ $\mathscr{E}_K(\cdot)$ strictly convex on $\mathscr{M}$
- K CISPD $\Leftrightarrow$ $\mathscr{E}_K(\cdot)$ strictly convex on $\mathscr{M}(1)$ (or $\mathscr{M}^+(1)$)

<u>Directional derivative</u> of $\mathscr{E}_K(\cdot)$ at μ in the direction ν:

$$\begin{aligned} F_K(\mu;\nu) &= \lim_{\alpha \to 0^+} \frac{\mathscr{E}_K[(1-\alpha)\mu + \alpha\nu] - \mathscr{E}_K(\mu)}{\alpha} \\ &= 2\left[\int_{\mathscr{X}^2} K(\mathbf{x},\mathbf{x}')\,\mathrm{d}\nu(\mathbf{x})\mathrm{d}\mu(\mathbf{x}') - \mathscr{E}_K(\mu)\right] \end{aligned}$$

For $\alpha \in (0,1)$, $\mu, \nu \in \mathscr{M}$:

$$\boxed{(1-\alpha)\,\mathscr{E}_K(\mu) + \alpha\,\mathscr{E}_K(\nu) - \mathscr{E}_K[(1-\alpha)\mu + \alpha\nu] = \alpha(1-\alpha)\,\mathscr{E}_K(\nu - \mu)}$$

- K ISPD $\Leftrightarrow$ $\mathscr{E}_K(\cdot)$ strictly convex on $\mathscr{M}$
- K CISPD $\Leftrightarrow$ $\mathscr{E}_K(\cdot)$ strictly convex on $\mathscr{M}(1)$ (or $\mathscr{M}^+(1)$)

<u>Directional derivative</u> of $\mathscr{E}_K(\cdot)$ at μ in the direction ν:

$$\begin{aligned} F_K(\mu;\nu) &= \lim_{\alpha \to 0^+} \frac{\mathscr{E}_K[(1-\alpha)\mu + \alpha\nu] - \mathscr{E}_K(\mu)}{\alpha} \\ &= 2\left[\int_{\mathscr{X}^2} K(\mathbf{x},\mathbf{x}')\,\mathrm{d}\nu(\mathbf{x})\mathrm{d}\mu(\mathbf{x}') - \mathscr{E}_K(\mu)\right] \end{aligned}$$

$\Rightarrow$ $F_K(\mu;\delta_{\mathbf{x}})$ is related to the potential $P_\mu(\mathbf{x})$:

$$\boxed{P_\mu(\mathbf{x}) = \tfrac{1}{2}\,F_K(\mu;\delta_{\mathbf{x}}) + \mathscr{E}_K(\mu)}$$

Suppose $\mathscr{X}$ compact

Theorem (Equivalence theorem in DOE)

If $\mathscr{E}_K(\cdot)$ strictly convex on $\mathscr{M}^+(1)$, $\mu_K^* \in \mathscr{M}^+(1)$ is the minimum-energy probability measure on $\mathscr{X}$ iff $\forall\, \mathbf{x} \in \mathscr{X}$, $P_{\mu_K^*}(\mathbf{x}) \geq \mathscr{E}_K(\mu_K^*)$; moreover, $P_{\mu_K^*}(\mathbf{x}) = \mathscr{E}_K(\mu_K^*)$ on the support of μ_K^*

Suppose $\mathscr{X}$ compact

Theorem (Equivalence theorem in DOE)

If $\mathscr{E}_K(\cdot)$ strictly convex on $\mathscr{M}^+(1)$, $\mu_K^* \in \mathscr{M}^+(1)$ is the minimum-energy probability measure on $\mathscr{X}$ iff $\forall\, \mathbf{x} \in \mathscr{X}$, $P_{\mu_K^*}(\mathbf{x}) \geq \mathscr{E}_K(\mu_K^*)$; moreover, $P_{\mu_K^*}(\mathbf{x}) = \mathscr{E}_K(\mu_K^*)$ on the support of μ_K^*

Potential theory (K singular, ISPD):

$$\mu_K^* \in \mathscr{M}^+(1) = \textit{equilibrium measure} \text{ of } \mathscr{X}$$
$$C_K = [\inf_{\mu \in \mathscr{M}^+(1)} \mathscr{E}_K(\mu)]^{-1} \geq 0 = \textit{capacity} \text{ of } \mathscr{X}$$

Consider now (signed) measures in $\mathscr{M}(1)$

Theorem

If $\mathscr{E}_K(\cdot)$ strictly convex on $\mathscr{M}(1)$, $\widetilde{\mu}_K^* \in \mathscr{M}(1)$ is the minimum-energy signed measure on $\mathscr{X}$ with total charge 1 iff $\forall\, \mathbf{x} \in \mathscr{X}$, $P_{\widetilde{\mu}_K^*}(\mathbf{x}) = \mathscr{E}_K(\widetilde{\mu}_K^*)$

Consider now (signed) measures in $\mathscr{M}(1)$

Theorem

If $\mathscr{E}_K(\cdot)$ strictly convex on $\mathscr{M}(1)$, $\widetilde{\mu}_K^* \in \mathscr{M}(1)$ is the minimum-energy signed measure on $\mathscr{X}$ with total charge 1 iff $\forall\, \mathbf{x} \in \mathscr{X}$, $P_{\widetilde{\mu}_K^*}(\mathbf{x}) = \mathscr{E}_K(\widetilde{\mu}_K^*)$

. . . but $\widetilde{\mu}_K^*$ does not always exist!

When minimum-energy signed measures are probability measures?

Theorem

Assume that K is ISPD and translation invariant, with $K(\mathbf{x}, \mathbf{x}') = \Psi(\mathbf{x} - \mathbf{x}')$ and Ψ continuous, twice differentiable except at the origin, with Laplacian $\Delta_\Psi(\mathbf{x}) = \sum_{i=1}^d \partial^2 \Psi(\mathbf{x}) / \partial x_i^2 \geq 0$, $\mathbf{x} \neq \mathbf{0}$. Then there exists a unique $\widetilde{\mu}_K^*$ in $\mathscr{M}(1)$, and $\widetilde{\mu}_K^* \in \mathscr{M}^+(1)$

(Generalises (Hájek, 1956) to $d > 1$, but $\Psi(\cdot)$ must have a singularity at zero.)

minimise $\mathscr{E}_K(\xi - \mu)$=MMD$^2(\xi, \mu)$

Relation with (continuous) BLUE of β in $\boxed{f(\mathbf{x}) = \beta + Z_x}$

with Z_x a Gaussian RF ($\mathbb{E}\{Z_x\} = 0$, $\mathbb{E}\{Z_x Z_{x'}\} = \sigma^2 K(\mathbf{x}, \mathbf{x}')$)
Näther (1985, Sect. 4.2):

- Any linear estimator of β: $\hat{\beta} = \hat{\beta}(\xi) = \int_{\mathscr{X}} f(\mathbf{x})\, \mathrm{d}\xi(\mathbf{x})$, $\xi \in \mathscr{M}$
- $\hat{\beta}(\xi)$ unbiased: $\xi \in \mathscr{M}(1)$
- Variance: $V_\xi = \mathbb{E}\{(\hat{\beta}(\xi) - \beta)^2\} = \sigma^2\, \mathscr{E}_K(\xi)$

Relation with (continuous) BLUE of β in $\boxed{f(\mathbf{x}) = \beta + Z_x}$

with Z_x a Gaussian RF ($\mathbb{E}\{Z_x\} = 0$, $\mathbb{E}\{Z_x Z_{x'}\} = \sigma^2 K(\mathbf{x}, \mathbf{x}')$)
Näther (1985, Sect. 4.2):

- Any linear estimator of β: $\hat{\beta} = \hat{\beta}(\xi) = \int_{\mathscr{X}} f(\mathbf{x})\, \mathrm{d}\xi(\mathbf{x})$, $\xi \in \mathscr{M}$
- $\hat{\beta}(\xi)$ unbiased: $\xi \in \mathscr{M}(1)$
- Variance: $V_\xi = \mathbb{E}\{(\hat{\beta}(\xi) - \beta)^2\} = \sigma^2\, \mathscr{E}_K(\xi)$

$\rightarrow$ Existence of a minimum-energy signed measure $\widetilde{\mu}_K^* \in \mathscr{M}(1)$
$\Leftrightarrow$ existence of the BLUE $\hat{\beta}^* = \hat{\beta}(\widetilde{\mu}_K^*)$, with variance $\sigma^2\, \mathscr{E}_K(\widetilde{\mu}_K^*)$

(The property $[\forall\, \mathbf{x} \in \mathscr{X}\ ,\ P_{\widetilde{\mu}_K^*}(\mathbf{x}) = \mathscr{E}_K(\widetilde{\mu}_K^*)]$ corresponds to Wiener-Hopf equation in Grenander's theorem (1950)).

We do not want to minimise $\mathscr{E}_K(\nu)$ but $\mathscr{E}_K(\xi-\mu)$=$\mathrm{MMD}^2(\xi,\mu)$
for a given μ

Kernel reduction

Damelin et al. (2010): for μ with total mass 1, consider the kernel

$$\boxed{\widetilde{K}_\mu(\mathbf{x},\mathbf{x}') \triangleq K(\mathbf{x},\mathbf{x}') - P_\mu(\mathbf{x}) - P_\mu(\mathbf{x}') + \mathscr{E}_K(\mu)}$$

- $\forall\,\xi$ with total mass 1 , $\boxed{\underbrace{\mathscr{E}_K(\xi-\mu)}_{\gamma_K^2(\xi,\mu)} = \mathscr{E}_{\widetilde{K}_\mu}(\xi-\mu) = \mathscr{E}_{\widetilde{K}_\mu}(\xi)}$

For μ,ξ having total mass 1,

minimising MMD $\gamma_K(\xi,\mu)$ w.r.t. ξ $\Leftrightarrow$ minimising MMD $\gamma_{\widetilde{K}_\mu}(\xi,\mu)$
$\Leftrightarrow$ minimising $\mathscr{E}_{\widetilde{K}_\mu}(\xi)$

We do not want to minimise $\mathscr{E}_K(\nu)$ but $\mathscr{E}_K(\xi-\mu)$=MMD$^2(\xi,\mu)$
for a given μ

Kernel reduction

Damelin et al. (2010): for μ with total mass 1, consider the kernel

$$\boxed{\widetilde{K}_\mu(\mathbf{x},\mathbf{x}') \triangleq K(\mathbf{x},\mathbf{x}') - P_\mu(\mathbf{x}) - P_\mu(\mathbf{x}') + \mathscr{E}_K(\mu)}$$

- $\forall\,\xi$ with total mass 1 , $\boxed{\underbrace{\mathscr{E}_K(\xi-\mu)}_{\gamma_K^2(\xi,\mu)} = \mathscr{E}_{\widetilde{K}_\mu}(\xi-\mu) = \mathscr{E}_{\widetilde{K}_\mu}(\xi)}$

For μ,ξ having total mass 1,

minimising MMD $\gamma_K(\xi,\mu)$ w.r.t. ξ $\Leftrightarrow$ minimising MMD $\gamma_{\widetilde{K}_\mu}(\xi,\mu)$
$\Leftrightarrow$ minimising $\mathscr{E}_{\widetilde{K}_\mu}(\xi)$

[For $\mu=\delta_{\mathbf{x}_0}$ (function approximation):
$\widetilde{K}_\mu(\mathbf{x},\mathbf{x}') \triangleq K(\mathbf{x},\mathbf{x}') - K(\mathbf{x},\mathbf{x}_0) - K(\mathbf{x}',\mathbf{x}_0) + K(\mathbf{x}_0,\mathbf{x}_0)$]

We do not want to minimise $\mathscr{E}_K(\nu)$ but $\mathscr{E}_K(\xi-\mu)$=$\mathrm{MMD}^2(\xi,\mu)$ for a given μ

Kernel reduction

Damelin et al. (2010): for μ with total mass 1, consider the kernel

$$\boxed{\widetilde{K}_\mu(\mathbf{x},\mathbf{x}') \triangleq K(\mathbf{x},\mathbf{x}') - P_\mu(\mathbf{x}) - P_\mu(\mathbf{x}') + \mathscr{E}_K(\mu)}$$

- $\forall\,\xi$ with total mass 1 , $\boxed{\underbrace{\mathscr{E}_K(\xi-\mu)}_{\gamma_K^2(\xi,\mu)} = \mathscr{E}_{\widetilde{K}_\mu}(\xi-\mu) = \mathscr{E}_{\widetilde{K}_\mu}(\xi)}$

For μ,ξ having total mass 1,

minimising MMD $\gamma_K(\xi,\mu)$ w.r.t. ξ $\Leftrightarrow$ minimising MMD $\gamma_{\widetilde{K}_\mu}(\xi,\mu)$

$\Leftrightarrow$ minimising $\mathscr{E}_{\widetilde{K}_\mu}(\xi)$

[For $\mu=\delta_{\mathbf{x}_0}$ (function approximation):
$\widetilde{K}_\mu(\mathbf{x},\mathbf{x}') \triangleq K(\mathbf{x},\mathbf{x}') - K(\mathbf{x},\mathbf{x}_0) - K(\mathbf{x}',\mathbf{x}_0) + K(\mathbf{x}_0,\mathbf{x}_0)$]

compute $\widetilde{K}_\mu(\mathbf{x},\mathbf{x}')$

Back to the (continuous) BLUE of β

$\boxed{f(\mathbf{x}) = \beta + Z_x}$, Z_x a Gaussian RF, $\mathbb{E}\{Z_x\} = 0$, $\mathbb{E}\{Z_x Z_{x'}\} = \sigma^2 K(\mathbf{x}, \mathbf{x}')$

Denote $\mathfrak{p} \triangleq$ orthogonal projection of $L^2(\mathscr{X}, \mu)$ onto the linear space spanned by constant 1

$$a)\ f(\mathbf{x}) = \beta + \underbrace{Z_x}_{\mathfrak{p} Z_x + (\mathrm{Id}_{L^2} - \mathfrak{p}) Z_x} \quad \to \sigma^2 K$$

a) $\mathrm{var}\{BLUE(\beta)\} = \sigma^2 \mathscr{E}_K(\widetilde{\mu}_K^*)$ (if BLUE(β) exists...)

Back to the (continuous) BLUE of β

$\boxed{f(\mathbf{x}) = \beta + Z_x}$, Z_x a Gaussian RF, $\mathbb{E}\{Z_x\} = 0$, $\mathbb{E}\{Z_x Z_{x'}\} = \sigma^2 K(\mathbf{x}, \mathbf{x}')$

Denote $\mathfrak{p} \triangleq$ orthogonal projection of $L^2(\mathscr{X}, \mu)$ onto the linear space spanned by constant 1

$$a)\ f(\mathbf{x}) = \beta + \underbrace{Z_x}_{\mathfrak{p}Z_x + (\mathrm{Id}_{L^2} - \mathfrak{p})Z_x} \quad \to \sigma^2 K$$

$$b)\ f(\mathbf{x}) = \underbrace{\beta'}_{\beta + \mathfrak{p}Z_x} + \underbrace{\widetilde{Z}_x}_{(\mathrm{Id}_{L^2} - \mathfrak{p})Z_x} \quad \to \boxed{\sigma^2 \widetilde{K}_\mu}$$

a) $\mathrm{var}\{BLUE(\beta)\} = \sigma^2 \mathscr{E}_K(\widetilde{\mu}_K^*)$ (if BLUE(β) exists...)

b) $\mathrm{var}\{BLUE(\beta')\} = \sigma^2 \mathscr{E}_{\widetilde{K}_\mu}(\mu) = 0$ (with BLUE(β') $= I_\mu(f)$)

How to compute $\widetilde{K}_\mu(\mathbf{x},\mathbf{x}') \triangleq K(\mathbf{x},\mathbf{x}') - P_\mu(\mathbf{x}) - P_\mu(\mathbf{x}') + \mathscr{E}_K(\mu)$?

How to compute $\widetilde{K}_\mu(\mathbf{x},\mathbf{x}') \triangleq K(\mathbf{x},\mathbf{x}') - P_\mu(\mathbf{x}) - P_\mu(\mathbf{x}') + \mathscr{E}_K(\mu)$?

Tensor product kernels:

Take K separable on $\mathscr{X} = \times_{i=1}^d \mathscr{X}_i$:

$$K(\mathbf{x},\mathbf{x}') = K^{(t)}(\mathbf{x},\mathbf{x}') = \prod_{i=1}^d K_i(x_i, x_i')$$

Szabó and Sriperumbudur (2018): each K_i continuous, bounded, translation invariant and CISPD on $\mathscr{M}^{(i)} = \mathscr{M}[\mathscr{X}_i] \Rightarrow K^{(t)}$ is CISPD on $\mathscr{M}[\mathscr{X}]$

How to compute $\widetilde{K}_\mu(\mathbf{x},\mathbf{x}') \triangleq K(\mathbf{x},\mathbf{x}') - P_\mu(\mathbf{x}) - P_\mu(\mathbf{x}') + \mathscr{E}_K(\mu)$?

Tensor product kernels:

Take K separable on $\mathscr{X} = \times_{i=1}^d \mathscr{X}_i$:

$$K(\mathbf{x},\mathbf{x}') = K^{(t)}(\mathbf{x},\mathbf{x}') = \prod_{i=1}^d K_i(x_i, x_i')$$

Szabó and Sriperumbudur (2018): each K_i continuous, bounded, translation invariant and CISPD on $\mathscr{M}^{(i)} = \mathscr{M}[\mathscr{X}_i] \Rightarrow K^{(t)}$ is CISPD on $\mathscr{M}[\mathscr{X}]$

When $\mu = \otimes_{i=1}^d \mu^{(i)}$ is a product measure on $\mathscr{X} = \times_{i=1}^d \mathscr{X}_i$, then

$$\mathscr{E}_{K^{(t)}}(\mu) = \prod_{i=1}^d \mathscr{E}_{K_i}(\mu^{(i)}), \quad P_\mu(\mathbf{x}) = \prod_{i=1}^d P_{\mu^{(i)}}(x_i)$$

and $\mathscr{E}_{K_i}(\mu^{(i)})$ and $P_{\mu^{(i)}}(x)$ easily computed (analytical expressions available for some $\mu^{(i)}$ and K_i on $\mathscr{X}_i = [a,b]$ — Matérn, Riesz...)

4 Design of experiments

$\boxed{\mu \text{ is uniform on } \mathscr{X}}$

Minimise $\mathscr{E}_K(\xi - \mu) = \mathrm{MMD}^2(\xi, \mu)$ w.r.t. $\xi \rightarrow \xi = \mu$, not very useful...

4 Design of experiments

μ is uniform on $\mathscr{X}$

Minimise $\mathscr{E}_K(\xi - \mu)$=MMD$^2(\xi, \mu)$ w.r.t. $\xi \rightarrow \xi = \mu$, not very useful...

We want a discrete measure !

- $\mathbf{X}_n = \{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$ a n-point design ($\mathbf{x}_i \in \mathscr{X}$ for all i)
- $\xi_n = \sum_{i=1}^n w_i \, \delta_{\mathbf{x}_i}$ a finite signed measure supported on $\mathbf{X}_n$, $\mathbf{w}_n = (w_1, \ldots, w_n)^\top$

$\rightarrow$ $\boxed{\mathscr{E}_K(\xi_n - \mu) = \mathscr{E}_K(\mu) - 2\mathbf{w}_n^\top \mathbf{p}_n(\mu) + \mathbf{w}_n^\top \mathbf{K}_n \mathbf{w}_n}$

where $\mathbf{p}_n(\mu) = (P_\mu(\mathbf{x}_1), \ldots, P_\mu(\mathbf{x}_n))^\top$ and $\{\mathbf{K}_n\}_{i,j} = K(\mathbf{x}_i, \mathbf{x}_j)$

4 Design of experiments

μ is uniform on $\mathscr{X}$

Minimise $\mathscr{E}_K(\xi - \mu)$=MMD$^2(\xi, \mu)$ w.r.t. ξ ➞ $\xi = \mu$, not very useful...

We want a discrete measure !

- $\mathbf{X}_n = \{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$ a n-point design ($\mathbf{x}_i \in \mathscr{X}$ for all i)
- $\xi_n = \sum_{i=1}^n w_i \, \delta_{\mathbf{x}_i}$ a finite signed measure supported on $\mathbf{X}_n$, $\mathbf{w}_n = (w_1, \ldots, w_n)^\top$

➞ $\boxed{\mathscr{E}_K(\xi_n - \mu) = \mathscr{E}_K(\mu) - 2\mathbf{w}_n^\top \mathbf{p}_n(\mu) + \mathbf{w}_n^\top \mathbf{K}_n \mathbf{w}_n}$

where $\mathbf{p}_n(\mu) = (P_\mu(\mathbf{x}_1), \ldots, P_\mu(\mathbf{x}_n))^\top$ and $\{\mathbf{K}_n\}_{i,j} = K(\mathbf{x}_i, \mathbf{x}_j)$

➞ $\boxed{\mathscr{E}_K(\xi_n - \mu) = \mathscr{E}_{\widetilde{K}_\mu}(\xi_n) = \mathbf{w}_n^\top \widetilde{\mathbf{K}}_n \mathbf{w}_n}$

$\mathbf{X}_n$ fixed, optimal weights $\hat{\mathbf{w}}_n$ summing to 1: Bayesian integration

$$\hat{\mathbf{w}}_n = \frac{\widetilde{\mathbf{K}}_n^{-1}\mathbf{1}_n}{\mathbf{1}_n^\top \widetilde{\mathbf{K}}_n^{-1}\mathbf{1}_n}$$

$\mathbf{X}_n$ fixed, optimal weights $\hat{\mathbf{w}}_n$ summing to 1: Bayesian integration

$$\hat{\mathbf{w}}_n = \frac{\widetilde{\mathbf{K}}_n^{-1}\mathbf{1}_n}{\mathbf{1}_n^\top \widetilde{\mathbf{K}}_n^{-1}\mathbf{1}_n}$$

$$\begin{cases} \hat{\mathbf{w}}_n & = \text{weights} \\ \sigma^2 \mathscr{E}_K(\hat{\xi}_n - \mu) & = \text{variance} \end{cases}$$ for **Bayesian integration w.r.t.** μ

in the model

$\boxed{f(\mathbf{x}) = \beta + Z_x}$ with $\mathrm{E}\{Z_x\} = 0$ and $\mathrm{E}\{Z_x Z_{x'}\} = \sigma^2 K(\mathbf{x}, \mathbf{x}')$

→ Posterior mean: $\widehat{I}_n = \hat{\mathbf{w}}_n{}^\top \mathbf{y}_n$, with $\mathbf{y}_n = (f(\mathbf{x}_1), \ldots, f(\mathbf{x}_n))^\top$

→ Posterior variance $\sigma^2 s_n^2 = \sigma^2 \mathscr{E}_K(\hat{\xi}_n - \mu) = \sigma^2 \mathscr{E}_{\widetilde{K}_\mu}(\hat{\xi}_n) = \dfrac{\sigma^2}{\mathbf{1}_n^\top \widetilde{\mathbf{K}}_n^{-1}\mathbf{1}_n}$

▸▸ all weights $= 1/n$

. . . ?!? initially we had

$$\begin{aligned}
\hat{\mathbf{w}}_n &= \left(\mathbf{K}_n^{-1} - \frac{\mathbf{K}_n^{-1}\mathbf{1}_n\mathbf{1}_n^T\mathbf{K}_n^{-1}}{\mathbf{1}_n^T\mathbf{K}_n^{-1}\mathbf{1}_n}\right)\mathbf{p}_n(\mu) + \frac{\mathbf{K}_n^{-1}\mathbf{1}_n}{\mathbf{1}_n^T\mathbf{K}_n^{-1}\mathbf{1}_n} \\
s_n^2 &= \mathscr{E}_K(\mu) - \mathbf{p}_n^T(\mu)\mathbf{K}_n^{-1}\mathbf{p}_n(\mu) + \frac{(1-\mathbf{p}_n^T(\mu)\mathbf{K}_n^{-1}\mathbf{1}_n)^2}{\mathbf{1}_n^T\mathbf{K}_n^{-1}\mathbf{1}_n}
\end{aligned}$$

. . . ?!? initially we had

$$\hat{\mathbf{w}}_n = \left(\mathbf{K}_n^{-1} - \frac{\mathbf{K}_n^{-1}\mathbf{1}_n\mathbf{1}_n^T\mathbf{K}_n^{-1}}{\mathbf{1}_n^T\mathbf{K}_n^{-1}\mathbf{1}_n}\right)\mathbf{p}_n(\mu) + \frac{\mathbf{K}_n^{-1}\mathbf{1}_n}{\mathbf{1}_n^T\mathbf{K}_n^{-1}\mathbf{1}_n}$$

$$s_n^2 = \mathscr{E}_K(\mu) - \mathbf{p}_n^T(\mu)\mathbf{K}_n^{-1}\mathbf{p}_n(\mu) + \frac{(1-\mathbf{p}_n^T(\mu)\mathbf{K}_n^{-1}\mathbf{1}_n)^2}{\mathbf{1}_n^T\mathbf{K}_n^{-1}\mathbf{1}_n}$$

$\mathfrak{p} \triangleq$ orthogonal projection of $L^2(\mathscr{X},\mu)$ onto the linear space spanned by 1

$$f(\mathbf{x}) = \beta + \underbrace{Z_x}_{\mathfrak{p}Z_x+(\mathrm{Id}_{L^2}-\mathfrak{p})Z_x \;\to\sigma^2 K} = \underbrace{\beta'}_{\beta+\mathfrak{p}Z_x} + \underbrace{\widetilde{Z}_x}_{(\mathrm{Id}_{L^2}-\mathfrak{p})Z_x \;\to\sigma^2\widetilde{K}_\mu}$$

New trend β', but the predictions are not modified when $K \to \widetilde{K}_\mu$

. . . ?!? initially we had

$$\hat{\mathbf{w}}_n = \left(\mathbf{K}_n^{-1} - \frac{\mathbf{K}_n^{-1}\mathbf{1}_n\mathbf{1}_n^T\mathbf{K}_n^{-1}}{\mathbf{1}_n^T\mathbf{K}_n^{-1}\mathbf{1}_n}\right)\mathbf{p}_n(\mu) + \frac{\mathbf{K}_n^{-1}\mathbf{1}_n}{\mathbf{1}_n^T\mathbf{K}_n^{-1}\mathbf{1}_n}$$

$$s_n^2 = \mathscr{E}_K(\mu) - \mathbf{p}_n^T(\mu)\mathbf{K}_n^{-1}\mathbf{p}_n(\mu) + \frac{(1-\mathbf{p}_n^T(\mu)\mathbf{K}_n^{-1}\mathbf{1}_n)^2}{\mathbf{1}_n^T\mathbf{K}_n^{-1}\mathbf{1}_n}$$

$\mathfrak{p} \triangleq$ orthogonal projection of $L^2(\mathscr{X},\mu)$ onto the linear space spanned by 1

$$f(\mathbf{x}) = \beta + \underbrace{Z_x}_{\mathfrak{p}Z_x+(\mathrm{Id}_{L^2}-\mathfrak{p})Z_x \;\to\sigma^2 K} = \underbrace{\beta'}_{\beta+\mathfrak{p}Z_x} + \underbrace{\widetilde{Z}_x}_{(\mathrm{Id}_{L^2}-\mathfrak{p})Z_x \;\to\sigma^2\widetilde{K}_\mu}$$

New trend β', but the predictions are not modified when $K \to \widetilde{K}_\mu$ . . . and $\mathscr{E}_{\widetilde{K}_\mu}(\mu) = 0$, $\widetilde{\mathbf{p}}_n(\mu) = \mathbf{0}$

$$\hat{\mathbf{w}}_n = \left(\widetilde{\mathbf{K}}_n^{-1} - \frac{\widetilde{\mathbf{K}}_n^{-1}\mathbf{1}_n\mathbf{1}_n^T\widetilde{\mathbf{K}}_n^{-1}}{\mathbf{1}_n^T\widetilde{\mathbf{K}}_n^{-1}\mathbf{1}_n}\right)\widetilde{\mathbf{p}}_n(\mu) + \frac{\widetilde{\mathbf{K}}_n^{-1}\mathbf{1}_n}{\mathbf{1}_n^T\widetilde{\mathbf{K}}_n^{-1}\mathbf{1}_n}$$

$$s_n^2 = \mathscr{E}_{\widetilde{K}_\mu}(\mu) - \widetilde{\mathbf{p}}_n^T(\mu)\widetilde{\mathbf{K}}_n^{-1}\widetilde{\mathbf{p}}_n(\mu) + \frac{(1-\widetilde{\mathbf{p}}_n^T(\mu)\widetilde{\mathbf{K}}_n^{-1}\mathbf{1}_n)^2}{\mathbf{1}_n^T\widetilde{\mathbf{K}}_n^{-1}\mathbf{1}_n}$$

. . . ?!? initially we had

$$\hat{\mathbf{w}}_n = \left(\mathbf{K}_n^{-1} - \frac{\mathbf{K}_n^{-1}\mathbf{1}_n\mathbf{1}_n^T\mathbf{K}_n^{-1}}{\mathbf{1}_n^T\mathbf{K}_n^{-1}\mathbf{1}_n}\right)\mathbf{p}_n(\mu) + \frac{\mathbf{K}_n^{-1}\mathbf{1}_n}{\mathbf{1}_n^T\mathbf{K}_n^{-1}\mathbf{1}_n}$$

$$s_n^2 = \mathscr{E}_K(\mu) - \mathbf{p}_n^T(\mu)\mathbf{K}_n^{-1}\mathbf{p}_n(\mu) + \frac{(1-\mathbf{p}_n^T(\mu)\mathbf{K}_n^{-1}\mathbf{1}_n)^2}{\mathbf{1}_n^T\mathbf{K}_n^{-1}\mathbf{1}_n}$$

$\mathfrak{p} \triangleq$ orthogonal projection of $L^2(\mathscr{X},\mu)$ onto the linear space spanned by 1

$$f(\mathbf{x}) = \beta + \underbrace{Z_x}_{\mathfrak{p}Z_x+(\mathrm{Id}_{L^2}-\mathfrak{p})Z_x \;\to\sigma^2 K} = \underbrace{\beta'}_{\beta+\mathfrak{p}Z_x} + \underbrace{\widetilde{Z}_x}_{(\mathrm{Id}_{L^2}-\mathfrak{p})Z_x \;\to\sigma^2\widetilde{K}_\mu}$$

New trend β', but the predictions are not modified when $K \to \widetilde{K}_\mu$ . . . and $\mathscr{E}_{\widetilde{K}_\mu}(\mu) = 0$, $\widetilde{\mathbf{p}}_n(\mu) = \mathbf{0}$

$$\hat{\mathbf{w}}_n = \frac{\widetilde{\mathbf{K}}_n^{-1}\mathbf{1}_n}{\mathbf{1}_n^T\widetilde{\mathbf{K}}_n^{-1}\mathbf{1}_n}$$

$$s_n^2 = \frac{1}{\mathbf{1}_n^T\widetilde{\mathbf{K}}_n^{-1}\mathbf{1}_n}$$

Equivalence with estimation in a location model

$$\begin{cases} \hat{\mathbf{w}}_n & = \text{ weights} \\ \sigma^2 \mathscr{E}_K(\hat{\xi}_n - \mu) & = \text{ variance} \end{cases}$$ for **(discrete) BLUE** of β' in the model

$$\boxed{f(\mathbf{x}) = \beta' + \widetilde{Z}_x}$$ with $\mathrm{E}\{\widetilde{Z}_x\} = 0$, $\mathrm{E}\{\widetilde{Z}_x \widetilde{Z}_{x'}\} = \sigma^2 \widetilde{K}_\mu(\mathbf{x}, \mathbf{x}')$

→ $\widehat{I}_n = \hat{\beta'}^n$ = (discrete) BLUE of β'
→ $\sigma^2 s_n^2 = \mathrm{var}[\hat{\beta'}^n]$

(design for) Bayesian integration ⇔ (design for) parameter estimation in a model with errors having a modified covariance

Equivalence with estimation in a location model

$$\begin{cases} \hat{\mathbf{w}}_n & = \text{ weights} \\ \sigma^2 \mathscr{E}_K(\hat{\xi}_n - \mu) & = \text{ variance} \end{cases}$$ for **(discrete) BLUE** of β' in the model

$$\boxed{f(\mathbf{x}) = \beta' + \widetilde{Z}_x} \text{ with } \mathrm{E}\{\widetilde{Z}_x\} = 0,\ \mathrm{E}\{\widetilde{Z}_x \widetilde{Z}_{x'}\} = \sigma^2\, \widetilde{K}_\mu(\mathbf{x}, \mathbf{x}')$$

→ $\widehat{I}_n = \hat{\beta'}^n$ = (discrete) BLUE of β'
→ $\sigma^2 s_n^2 = \mathrm{var}[\hat{\beta'}^n]$

(design for) Bayesian integration ⇔ (design for) parameter estimation in a model with errors having a modified covariance

[$\mu = \delta_{\mathbf{x}_0}$ yields a similar equivalence between function approximation (at $\mathbf{x}_0$) and parameter estimation]

1/ Fix the weights to $1/n$, optimise w.r.t. $\mathbf{X}_n$

→ choose $\mathbf{X}_n$ that minimises $\mathscr{E}_K(\xi_n - \mu) = \mathbf{1}_n^\top \widetilde{\mathbf{K}}_n \mathbf{1}_n / n^2$

→ *Projection kernel* of Mak and Joseph (2017);
Support points of Mak and Joseph (2018) based on energy distance;
L_2 *discrepancy* of QMC (symmetric, centred, wrap-around...) based on tensorised kernels related to variants of Brownian-motion covariance (Hickernell, 1998)

2/ **Nested designs (extensible point sequences)**

Choose n points sequentially among a finite set $\mathscr{X}_Q = \{\mathbf{s}_1, \ldots, \mathbf{s}_Q\} \subset \mathscr{X}$
A measure ξ on $\mathscr{X}_Q \Leftrightarrow$ a vector $\boldsymbol{\omega} \in \mathbb{R}^Q$

2/ **Nested designs (extensible point sequences)**

Choose n points sequentially among a finite set $\mathscr{X}_Q = \{\mathbf{s}_1, \ldots, \mathbf{s}_Q\} \subset \mathscr{X}$
A measure ξ on $\mathscr{X}_Q \Leftrightarrow$ a vector $\boldsymbol{\omega} \in \mathbb{R}^Q$

<u>Vertex Direction</u> (="kernel herding" in machine learning):
$\rightarrow$ minimise $\mathscr{E}_K(\xi - \mu)$, quadratic in $\boldsymbol{\omega}$,
using Frank-Wolfe conditional gradient (Wynn's algorithm):

$$\xi^{(n+1)} = \frac{n}{n+1}\,\xi^{(n)} + \frac{1}{n+1}\,\delta_{\mathbf{x}_{n+1}}$$

with $\mathbf{x}_{n+1} \in \operatorname{Arg\,min}_{\mathbf{x}\in\mathscr{X}_Q} \left[\sum_{i=1}^n w_i^{(n)} K(\mathbf{x}, \mathbf{x}_i) - P_\mu(\mathbf{x})\right]$
(well defined also for singular kernels, e.g., Riesz kernels)

2/ **Nested designs (extensible point sequences)**

Choose n points sequentially among a finite set $\mathscr{X}_Q = \{\mathbf{s}_1, \ldots, \mathbf{s}_Q\} \subset \mathscr{X}$
A measure ξ on $\mathscr{X}_Q \Leftrightarrow$ a vector $\boldsymbol{\omega} \in \mathbb{R}^Q$

<u>Vertex Direction</u> (="kernel herding" in machine learning):
$\rightarrow$ minimise $\mathscr{E}_K(\xi - \mu)$, quadratic in $\boldsymbol{\omega}$,
using Frank-Wolfe conditional gradient (Wynn's algorithm):

$$\xi^{(n+1)} = \frac{n}{n+1}\,\xi^{(n)} + \frac{1}{n+1}\,\delta_{\mathbf{x}_{n+1}}$$

with $\mathbf{x}_{n+1} \in \operatorname{Arg\,min}_{\mathbf{x}\in\mathscr{X}_Q} \left[\sum_{i=1}^n w_i^{(n)} K(\mathbf{x}, \mathbf{x}_i) - P_\mu(\mathbf{x})\right]$
(well defined also for singular kernels, e.g., Riesz kernels)

- $\xi^{(1)} = \delta_{\mathbf{x}_1} \rightarrow$ all weights $w_i^{(n)}$ in $\xi^{(n)}$ equal $1/n$
- Rate of decrease of $\mathscr{E}_K(\xi^{(n)} - \mu) = \mathcal{O}(1/n)$
  Cost $\mathcal{O}(Q)$ at each iteration ($\mathcal{O}(nQ)$ for n iterations) $\boxed{\rightarrow \mathbf{X}_n^{VD}}$

Minimum-Norm-point algorithm of (Bach et al., 2012):

replace $\xi^{(n)}$ (uniform on its support) by $\hat{\xi}^{(n)}$ having the same support but optimal weights, positive with sum $= 1$

➠ We use $\hat{\mathbf{w}}_n$ from Bayesian quadrature

(extra comput. cost ➞ $\mathcal{O}(n^2 Q)$ for n iterations (P., 2021)) $\boxed{\rightarrow \mathbf{X}_n^{MN}}$

Minimum-Norm-point algorithm of (Bach et al., 2012):

replace $\xi^{(n)}$ (uniform on its support) by $\hat{\xi}^{(n)}$ having the same support but optimal weights, positive with sum $= 1$

➠ We use $\hat{\mathbf{w}}_n$ from Bayesian quadrature

(extra comput. cost → $\mathcal{O}(n^2 Q)$ for n iterations (P., 2021)) $\boxed{\rightarrow \mathbf{X}_n^{MN}}$

Comparison with:

- (scrambled) Sobol' sequence $\boxed{\rightarrow \mathbf{X}_n^{sS}}$
- Extensible Lattice sequence $\mathbf{X}_n^{EL}$: $\mathbf{x}_k = \{k\mathbf{g}\}$ ($\mathbf{g}$ has irrational components independent over the rationals, $\{\cdot\}$ = fractional part) We use $\mathbf{g} = (1/\varphi_d, 1/\varphi_d^2, \ldots, 1/\varphi_d^d)^\top$, with φ_d the unique positive root of $x^{d+1} = x + 1$, see http://extremelearning.com.au/ $\boxed{\rightarrow \mathbf{X}_n^{EL}}$

  (but not competitive for large d)

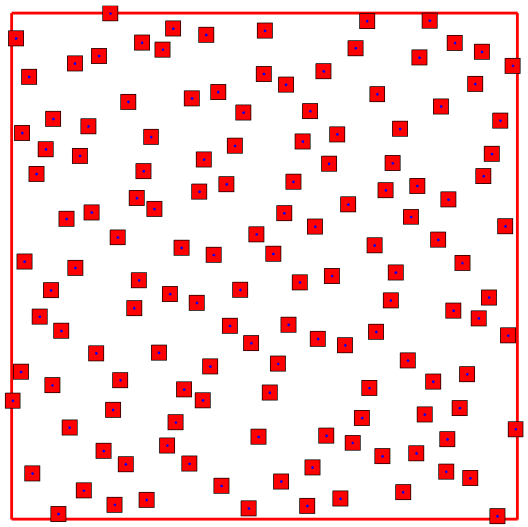
Example: $d = 2$, $\mathscr{X} = [0,1]^2$, $\mathscr{X}_Q =$ regular grid 64×64

$n_{\max} = 140$, $\mathbf{X}_1 = (0.5, 0.5)$, $\xi_{(1)} = \delta_{\mathbf{x}_1}$

$K =$ tensor product of Matérn 3/2:

$K_{3/2,\theta}(x, x') = (1 + \sqrt{3}\theta|x - x'|)\exp(-\sqrt{3}\theta|x - x'|)$, $\theta = 10$ $(n_{\max}^{1/d} \simeq 11.8)$

scrambled Sobol' LDS $\mathbf{X}_{140}^{sS}$



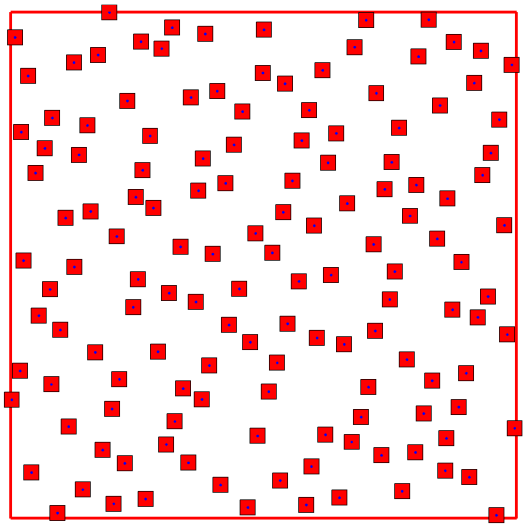
Example: $d = 2$, $\mathscr{X} = [0,1]^2$, $\mathscr{X}_Q =$ regular grid 64×64

$n_{\max} = 140$, $\mathbf{X}_1 = (0.5, 0.5)$, $\xi_{(1)} = \delta_{\mathbf{x}_1}$

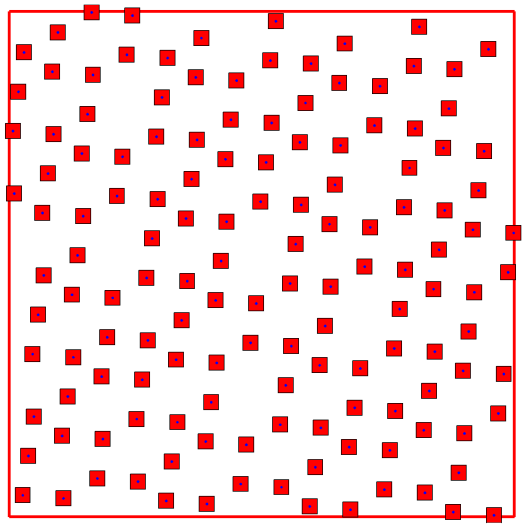
$K =$ tensor product of Matérn 3/2:

$K_{3/2,\theta}(x, x') = (1 + \sqrt{3}\theta|x - x'|)\exp(-\sqrt{3}\theta|x - x'|)$, $\theta = 10$ ($n_{\max}^{1/d} \simeq 11.8$)

scrambled Sobol' LDS $\mathbf{X}_{140}^{sS}$



Extensible Lattice sequence $\mathbf{X}_{140}^{EL}$



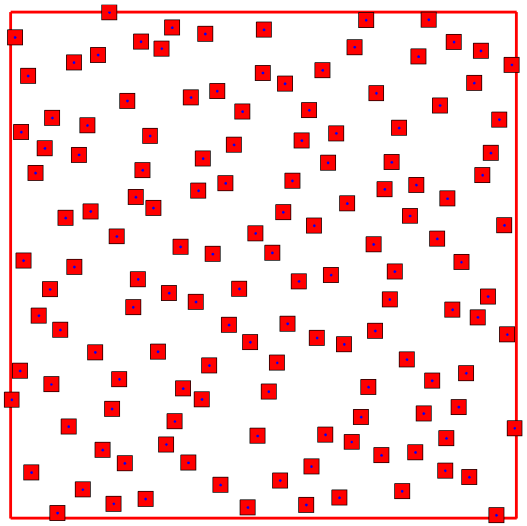
Example: $d = 2$, $\mathscr{X} = [0,1]^2$, $\mathscr{X}_Q =$ regular grid 64×64

$n_{\max} = 140$, $\mathbf{X}_1 = (0.5, 0.5)$, $\xi_{(1)} = \delta_{\mathbf{x}_1}$

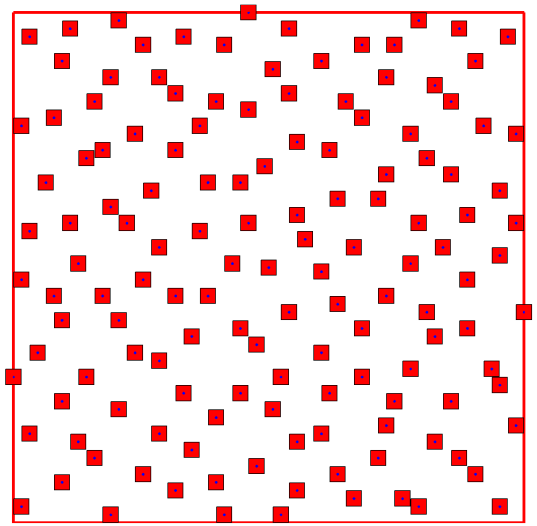
$K =$ tensor product of Matérn 3/2:

$K_{3/2,\theta}(x, x') = (1 + \sqrt{3}\theta|x - x'|)\exp(-\sqrt{3}\theta|x - x'|)$, $\theta = 10$ ($n_{\max}^{1/d} \simeq 11.8$)

scrambled Sobol' LDS $\mathbf{X}_{140}^{sS}$



Vertex Direction $\mathbf{X}_{140}^{VD}$



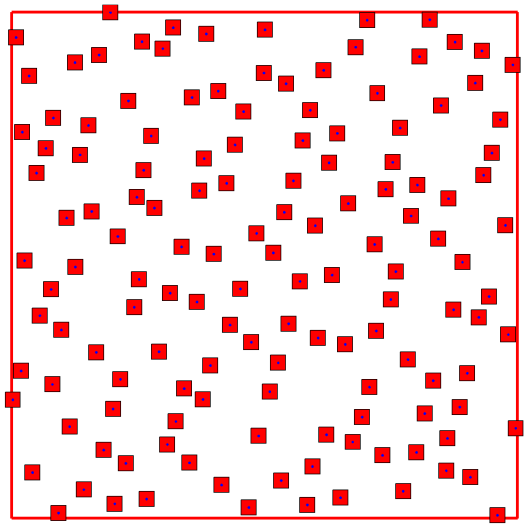
Example: $d = 2$, $\mathscr{X} = [0,1]^2$, $\mathscr{X}_Q =$ regular grid 64×64

$n_{\max} = 140$, $\mathbf{X}_1 = (0.5, 0.5)$, $\xi_{(1)} = \delta_{\mathbf{x}_1}$

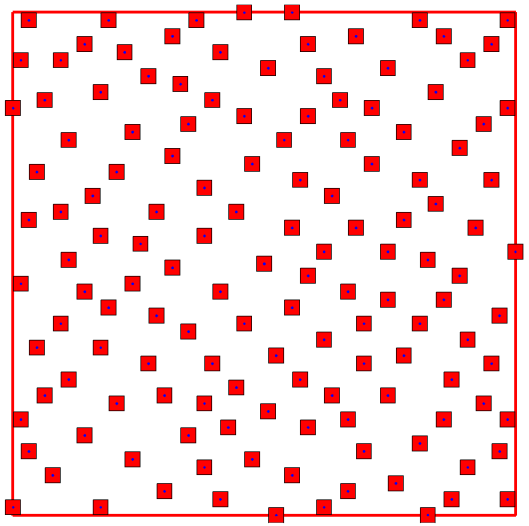
$K =$ tensor product of Matérn 3/2:

$K_{3/2,\theta}(x, x') = (1 + \sqrt{3}\theta|x - x'|)\exp(-\sqrt{3}\theta|x - x'|)$, $\theta = 10$ ($n_{\max}^{1/d} \simeq 11.8$)

scrambled Sobol' LDS $\mathbf{X}_{140}^{sS}$



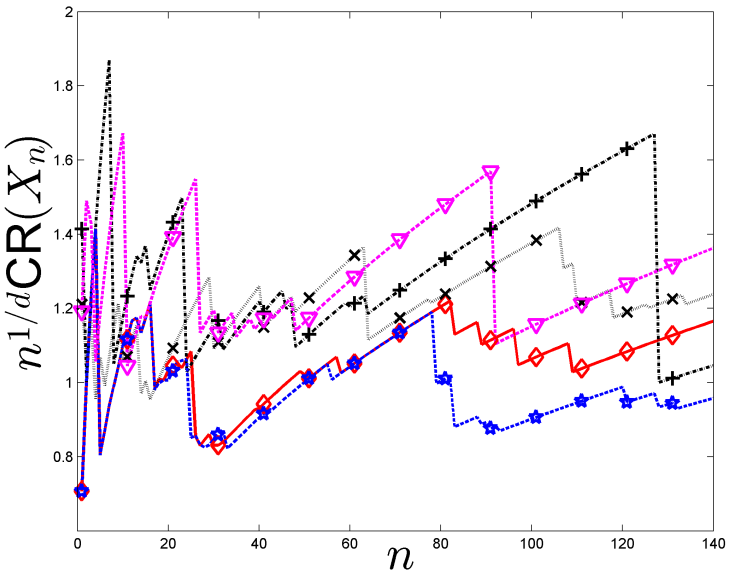
Minimum-Norm $\mathbf{X}_{140}^{MN}$



Space-filling performance

$$n^{1/d}\,\mathrm{CR}(\mathbf{X}_n)$$

(smaller is better)

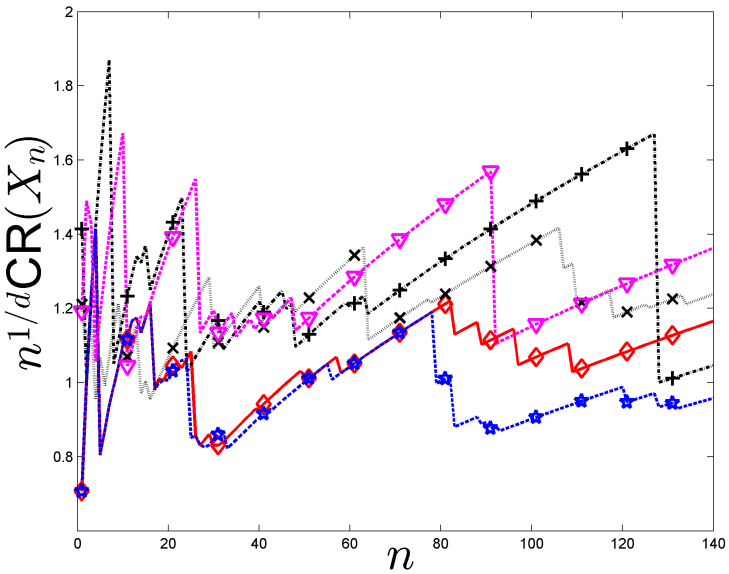


$\mathbf{X}_n^{S}$, $\mathbf{X}_n^{sS}$, $\mathbf{X}_n^{EL}$, $\mathbf{X}_n^{VD}$, $\mathbf{X}_n^{MN}$

Space-filling performance

$n^{1/d}\,\mathrm{CR}(\mathbf{X}_n)$

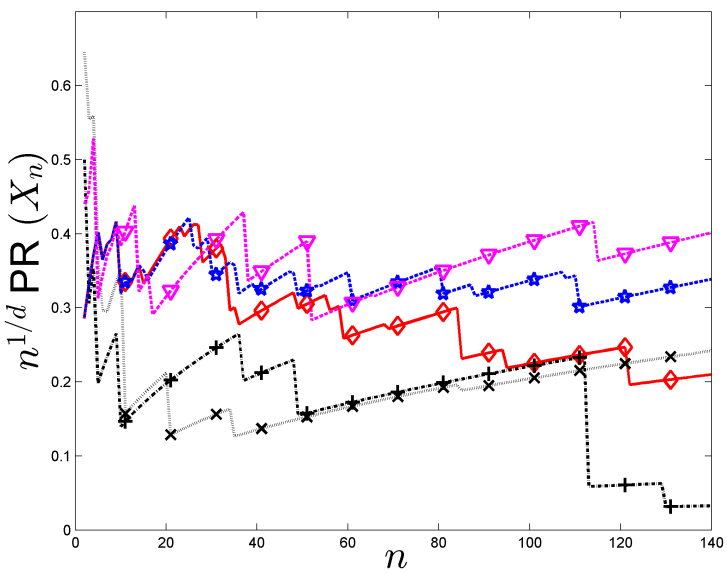
(smaller is better)



$\mathbf{X}_n^{S}$, $\mathbf{X}_n^{sS}$, $\mathbf{X}_n^{EL}$, $\mathbf{X}_n^{VD}$, $\mathbf{X}_n^{MN}$

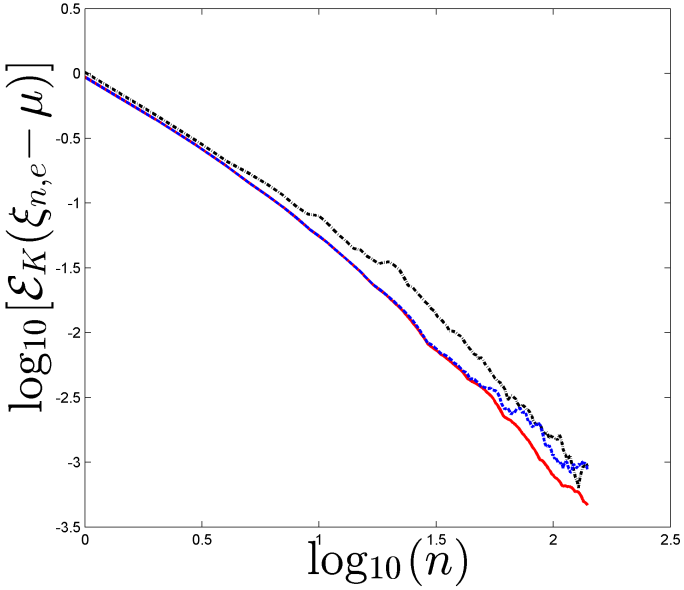
$n^{1/d}\,\mathrm{PR}(\mathbf{X}_n)$

(larger is better)

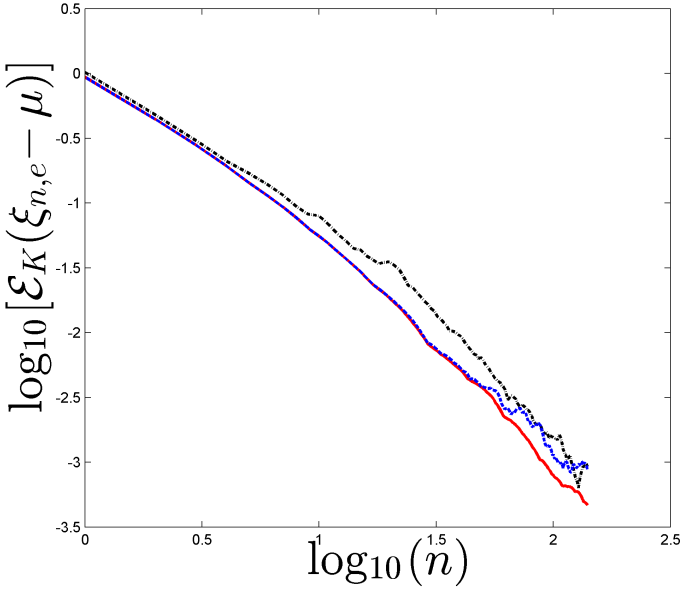


$\mathbf{X}_n^{S}$, $\mathbf{X}_n^{sS}$, $\mathbf{X}_n^{EL}$, $\mathbf{X}_n^{VD}$, $\mathbf{X}_n^{MN}$

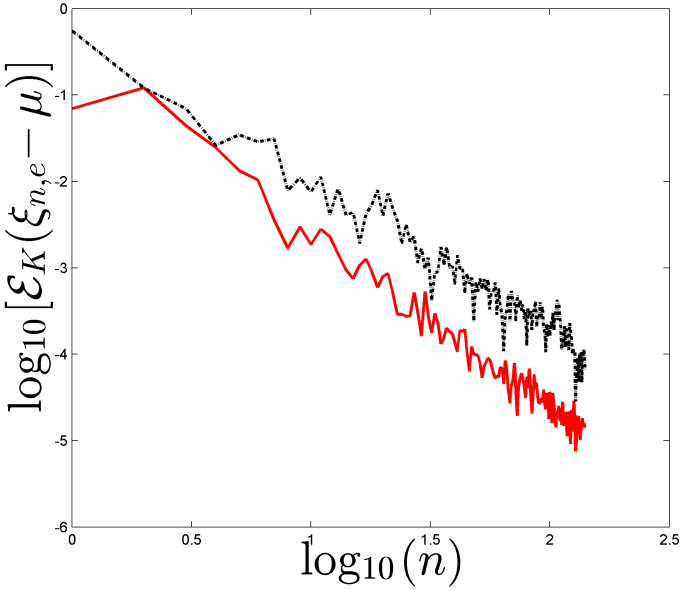
Decrease of $\mathscr{E}_{K_{3/2,10}}(\xi_n - \mu)$
$(\mathbf{X}_n^S, \mathbf{X}_n^{VD}, \mathbf{X}_n^{MN})$



Decrease of $\mathscr{E}_{K_{3/2,10}}(\xi_n - \mu)$
$(\mathbf{x}_n^S, \mathbf{x}_n^{VD}, \mathbf{x}_n^{MN})$



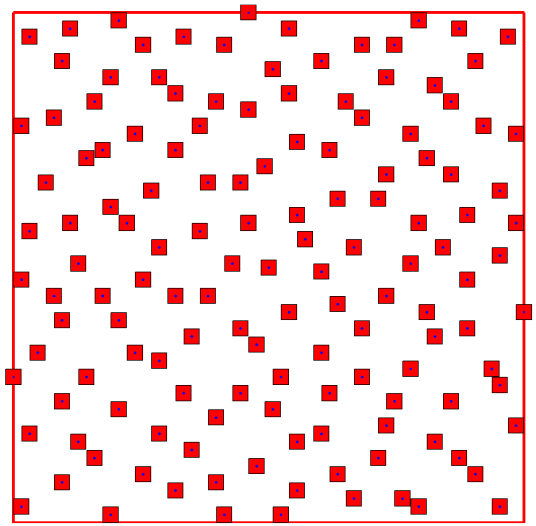
Decrease of $\mathscr{E}_{K_{3/2,1}}(\xi_n - \mu)$
$(\mathbf{x}_n^S, \mathbf{x}_n^{VD})$



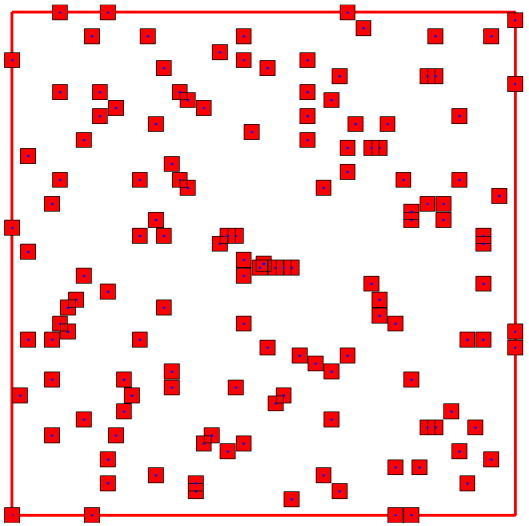
For a large enough correlation length the rate of decrease can reach $\mathcal{O}(1/n^2)$ (no proof available)

How meaningful is the rate of convergence?

"kernel herding" with $K_{3/2,10}$
$\mathbf{X}_{140}^{VD}$ (rate $\lesssim 1/k$)



"kernel herding" with $K_{3/2,1}$
$\mathbf{X}_{140}^{VD}$ (rate $\lesssim 1/k^2$)



A faster rate does not mean that points are more evenly distributed!

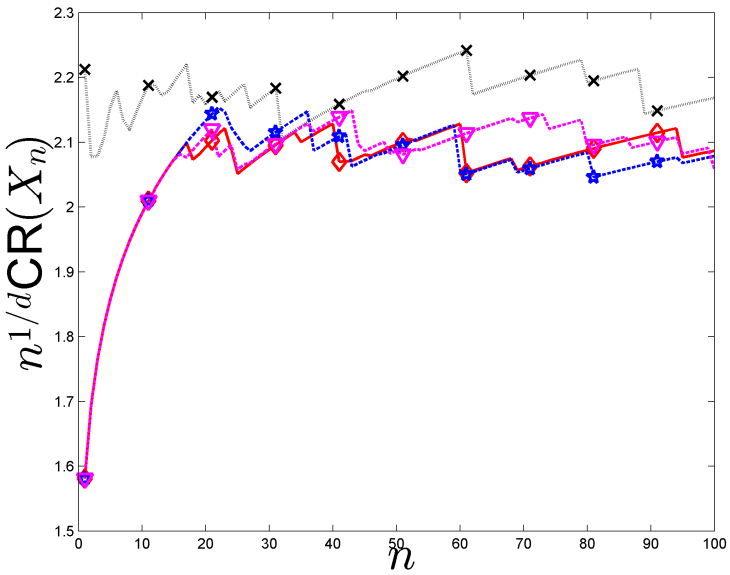
Example: $d = 10$, $\mathscr{X}_Q = 2^{12}$ points of scrambled Sobol' in $\mathscr{X} = [0,1]^2$

$n_{\max} = 140$, $\mathbf{X}_1 = \mathbf{1}_d/2$, $\xi_{(1)} = \delta_{\mathbf{X}_1}$

K = tensor product of Matérn 3/2:

$K_{3/2,\theta}(x,x') = (1+\sqrt{3}\theta|x-x'|)\exp(-\sqrt{3}\theta|x-x'|)$, $\theta = n_{\max}^{1/d}$

$\mathbf{X}_n^S$, $\mathbf{X}_n^{VD}$, $\mathbf{X}_n^{MN}$, $\mathbf{X}_n^{MN-\log}$



$[K_{\log}(x_i, x_i') = -\log|x_i - x_i'|]$

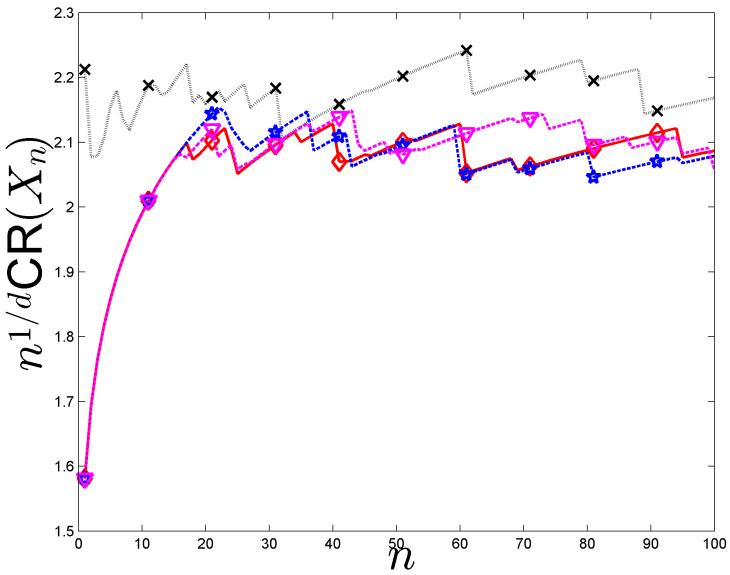
Example: $d = 10$, $\mathscr{X}_Q = 2^{12}$ points of scrambled Sobol' in $\mathscr{X} = [0,1]^2$

$n_{\max} = 140$, $\mathbf{X}_1 = \mathbf{1}_d/2$, $\xi_{(1)} = \delta_{\mathbf{X}_1}$

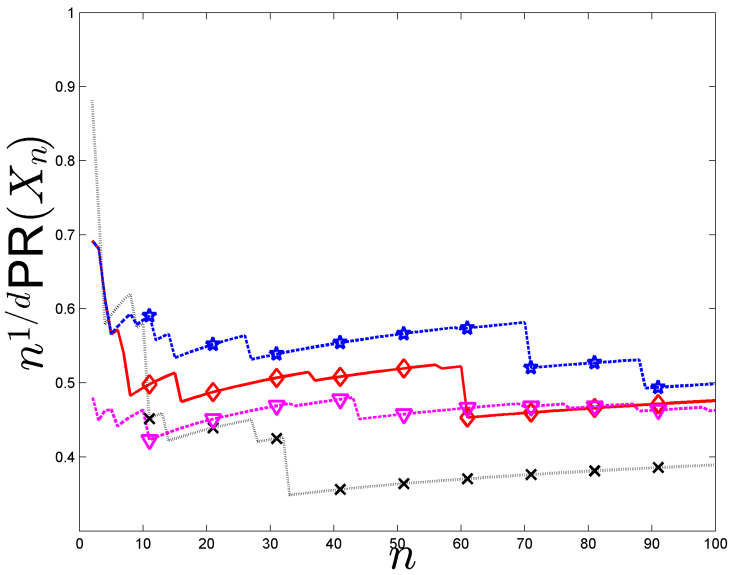
K = tensor product of Matérn 3/2:

$K_{3/2,\theta}(x,x') = (1+\sqrt{3}\theta|x-x'|)\exp(-\sqrt{3}\theta|x-x'|)$, $\theta = n_{\max}^{1/d}$

$\mathbf{X}_n^S$, $\mathbf{X}_n^{VD}$, $\mathbf{X}_n^{MN}$, $\mathbf{X}_n^{MN-\log}$



$\mathbf{X}_n^S$, $\mathbf{X}_n^{VD}$, $\mathbf{X}_n^{MN}$, $\mathbf{X}_n^{MN-\log}$



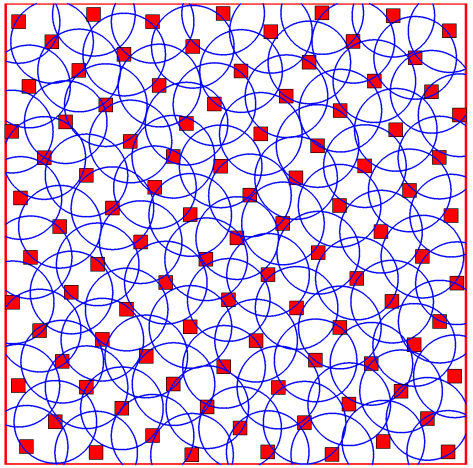
$[K_{\log}(x_i, x_i') = -\log|x_i - x_i'|]$

3/ n fixed: minimise $s_n^2 = \mathscr{E}_K(\hat{\xi}_n - \mu) = \frac{1}{\mathbf{1}_n^\top \widetilde{\mathbf{K}}_n^{-1} \mathbf{1}_n}$ w.r.t. $\mathbf{X}_n$

a) First (initialisation), use a greedy algorithm, stop at n $\boxed{\rightarrow \mathbf{X}_n^{VD}}$

b) Then, maximise $\mathbf{1}_n^\top \widetilde{\mathbf{K}}_n^{-1} \mathbf{1}_n$ viewed as a function of $\mathbf{X}_n$ $\boxed{\rightarrow \mathbf{X}_n^{VDopt}}$

a) Initialisation at $\mathbf{X}_{100}^{VD}$ generated by Vertex-Direction method with $\mathrm{CR}(\mathbf{X}_{100}^{VD}) \simeq 0.0925$, $2\,\mathrm{PR}(\mathbf{X}_{100}^{VD}) \simeq 0.0523$



$\rightarrow$ b) $\mathrm{CR}(\mathbf{X}_{100}^{VDopt}) \simeq 0.0897$, $2\,\mathrm{PR}(\mathbf{X}_{100}^{VDopt}) \simeq 0.0738$

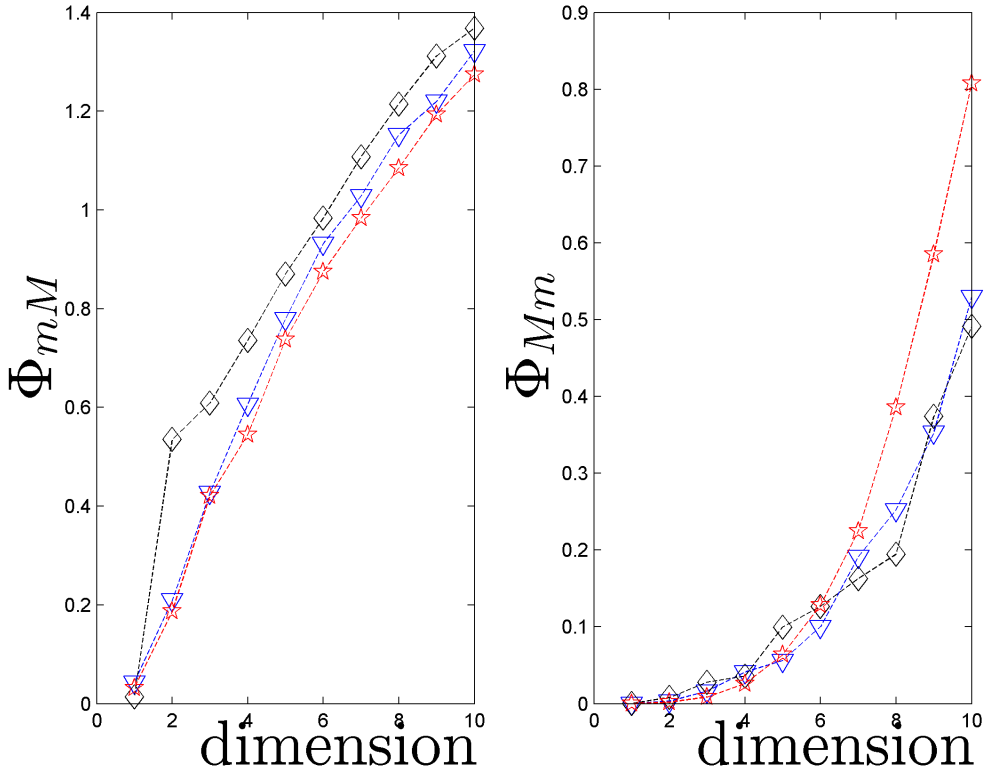
Example: $d = 10$, $\mathscr{X} = [0,1]^{10}$, $n_{\max} = 100$

$K = \bigotimes K_{3/2,\theta}(x,x')$, $\theta = \lfloor n_{\max}^{1/d} \rfloor = 1$

a) VD $\mathbf{X}_n^{VD}$, $n = 1, \ldots, 100$, then b) $\mathbf{X}_{100}^{VDopt}$

$\max \mathrm{CR} = \max \Phi_{mM}$ and $2 \min \mathrm{PR} = \min \Phi_{Mm}$ on projections

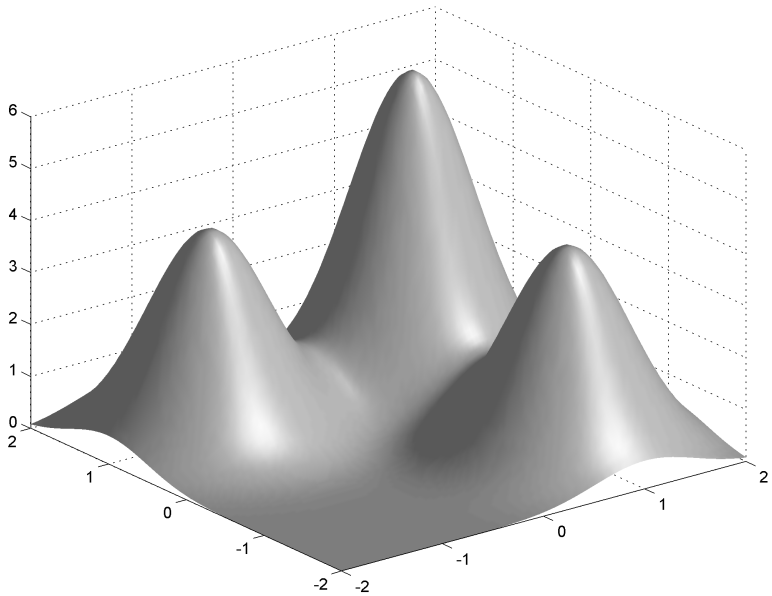
$\diamondsuit$ $\mathbf{X}_{100}^{S}$, $\triangledown$: $\mathbf{X}_{100}^{VD}$, $\bigstar$: $\mathbf{X}_{100}^{VDopt}$

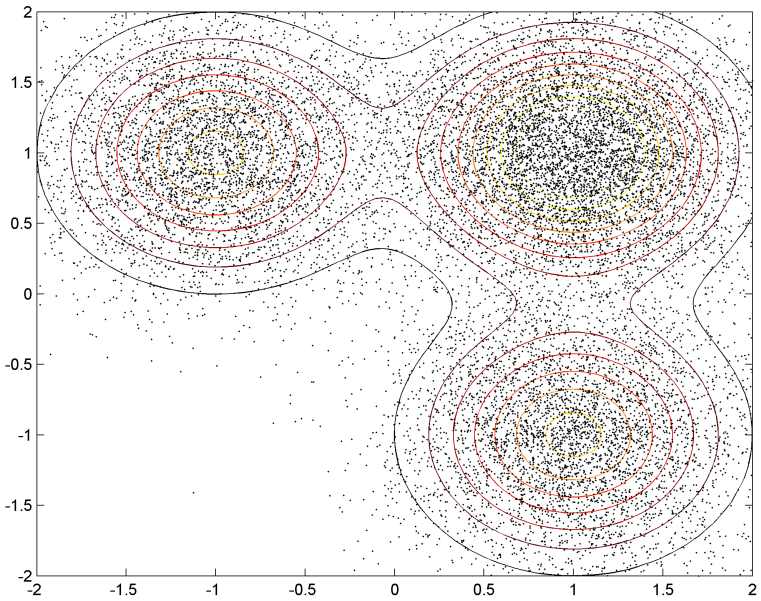


Kernel herding is not restricted to μ being uniform:

Example: Gaussian mixture $\mu = \sum_{j=1}^{3} \beta_j \, \mu_{\mathcal{N}}(\mathbf{a}_j, \sigma_j)$, $Q = 2^{14} = 16\,384$

(for $K_\theta(\mathbf{x}, \mathbf{x}') = \exp -(\theta \, \|\mathbf{x} - \mathbf{x}'\|^2)$, we know $P_\mu(\cdot)$ and $\mathscr{E}_K(\mu)$)

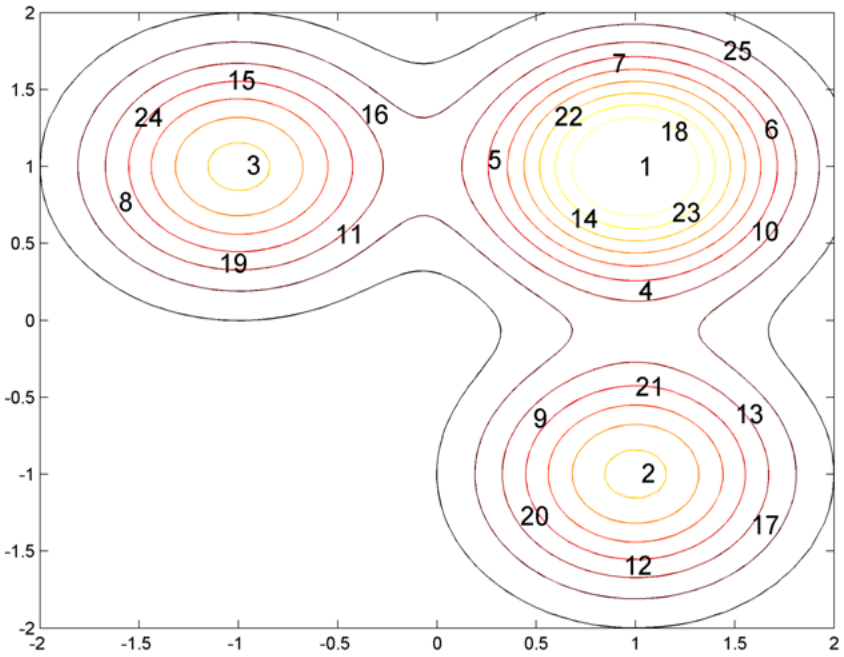




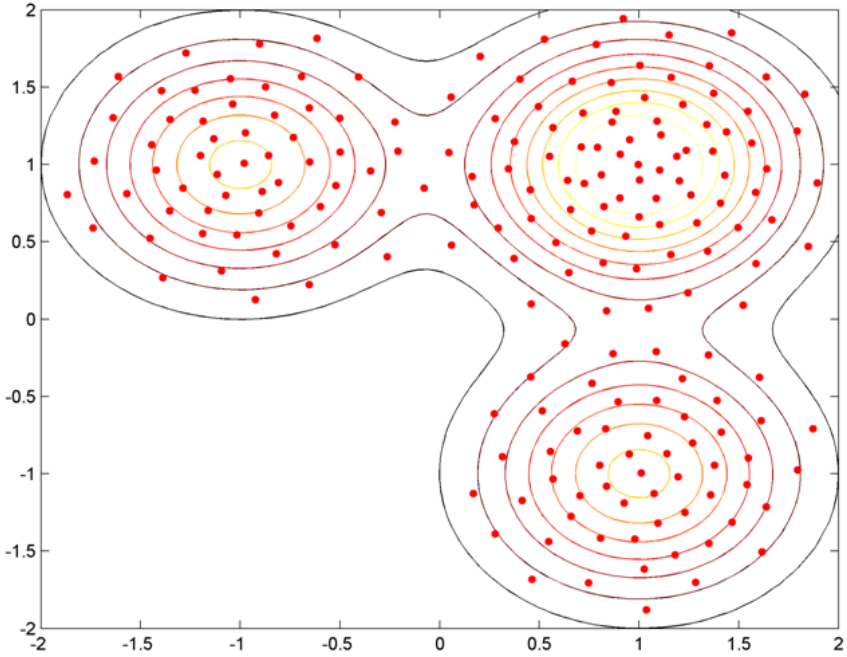
Kernel herding is not restricted to μ being uniform:

Example: Gaussian mixture $\mu = \sum_{j=1}^{3} \beta_j \, \mu_{\mathcal{N}}(\mathbf{a}_j, \sigma_j)$, $Q = 2^{14} = 16\,384$

(for $K_\theta(\mathbf{x}, \mathbf{x}') = \exp -(\theta \, \|\mathbf{x} - \mathbf{x}'\|^2)$, we know $P_\mu(\cdot)$ and $\mathscr{E}_K(\mu)$)



$n = 25$

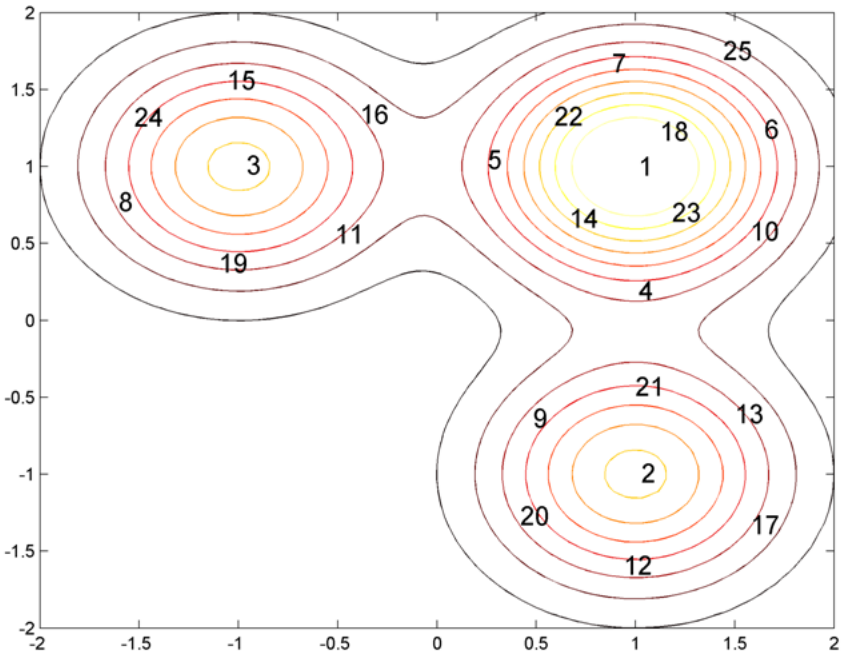


$n = 200$

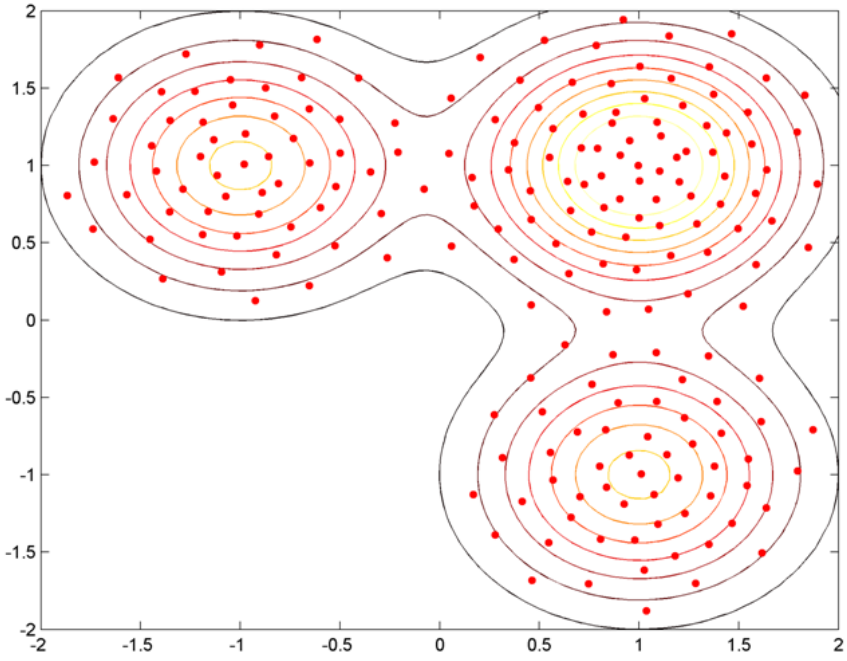
Kernel herding is not restricted to μ being uniform:

Example: Gaussian mixture $\mu = \sum_{j=1}^{3} \beta_j \, \mu_{\mathcal{N}}(\mathbf{a}_j, \sigma_j)$, $Q = 2^{14} = 16\,384$

(for $K_\theta(\mathbf{x}, \mathbf{x}') = \exp -(\theta \, \|\mathbf{x} - \mathbf{x}'\|^2)$, we know $P_\mu(\cdot)$ and $\mathscr{E}_K(\mu)$)



$n = 25$



$n = 200$

- Comparison between kernel herding, greedy MMD minimisation and Sequential Bayesian Quadrature (P., 2021)
- Extension to Stein discrepancy (Teymur et al., 2020) (K'_μ such that $P_\mu(\cdot) \equiv 0$ and $\mathscr{E}_{K'_\mu}(\mu) = 0$ without knowing the normalising constant in μ)
- Singular kernels (via completely monotone functions) (P. & Zhigljavsky, 2021)

4 Conclusions

- Maximum Mean Discrepancy connects Bayesian quadrature, potential theory, design for parameter estimation, space filling design
  - For many kernels K: metric for the weak cv. of prob. measures, easier to compute than optimal transport (= Wasserstein)
  - (design for) Bayesian integration $\Leftrightarrow$ (design for) parameter estimation in a model with modified covariance

4 Conclusions

- Maximum Mean Discrepancy connects Bayesian quadrature, potential theory, design for parameter estimation, space filling design
  - For many kernels K: metric for the weak cv. of prob. measures, easier to compute than optimal transport (= Wasserstein)
  - (design for) Bayesian integration $\Leftrightarrow$ (design for) parameter estimation in a model with modified covariance
  - $\rightarrow$ minimising $s_n^2 = \frac{1}{\mathbf{1}_n^\top \widetilde{\mathsf{K}}_n^{-1} \mathbf{1}_n}$ as an alternative to minimising $\mathrm{IMSPE}(\mathbf{X}_n) = \int_{\mathscr{X}} \rho_n^2(\mathbf{x}) \, \mathrm{d}\mu(\mathbf{x})$

4 Conclusions

- Maximum Mean Discrepancy connects Bayesian quadrature, potential theory, design for parameter estimation, space filling design
  - For many kernels K: metric for the weak cv. of prob. measures, easier to compute than optimal transport (= Wasserstein)
  - (design for) Bayesian integration $\Leftrightarrow$ (design for) parameter estimation in a model with modified covariance
  - $\rightarrow$ minimising $s_n^2 = \frac{1}{\mathbf{1}_n^\top \widetilde{\mathsf{K}}_n^{-1} \mathbf{1}_n}$ as an alternative to minimising $\mathrm{IMSPE}(\mathbf{X}_n) = \int_{\mathscr{X}} \rho_n^2(\mathbf{x}) \, \mathrm{d}\mu(\mathbf{x})$
- Optimisation algorithm for n fixed:
  Any nonlinear programming method (unconstrained) can be used
  (a Maj.-Min. approach is used in (Mak and Joseph, 2017, 2018))

A few more general questions and open issues:

- Space-filling performance? (discrepancy $\not\Leftrightarrow$ dispersion)
  Efficiency bounds for CR and PR?
- Which K? (wide choice, but does not seem to be crucial)

A few more general questions and open issues:

- Space-filling performance? (discrepancy $\not\Leftrightarrow$ dispersion)
  Efficiency bounds for CR and PR?
- Which K? (wide choice, but does not seem to be crucial)
- Steepest descent or Greedy optimisation?
  $\rightarrow$ To construct a sequence of embedded designs with small $\mathrm{CR}(\mathbf{X}_n)$, is kernel herding ($=$ Frank-Wolfe applied to MMD minimisation) preferable to methods exploiting submodularity, such as (Nogales Gómez et al., 2021)?

A few more general questions and open issues:

- Space-filling performance? (discrepancy $\not\Leftrightarrow$ dispersion)
  Efficiency bounds for CR and PR?
- Which K? (wide choice, but does not seem to be crucial)
- Steepest descent or Greedy optimisation?
  $\rightarrow$ To construct a sequence of embedded designs with small $\mathrm{CR}(\mathbf{X}_n)$, is kernel herding ($=$ Frank-Wolfe applied to MMD minimisation) preferable to methods exploiting submodularity, such as (Nogales Gómez et al., 2021)?

Thank you for your attention !

References I

Bach, F., Lacoste-Julien, S., Obozinski, G., 2012. On the equivalence between herding and conditional gradient algorithms. In: Proc. 29th Annual International Conference on Machine Learning. pp. 1355–1362.

Damelin, S., Hickernell, F., Ragozin, D., Zeng, X., 2010. On energy, discrepancy and group invariant measures on measurable subsets of Euclidean space. J. Fourier Anal. Appl. 16, 813–839.

Grenander, U., 1950. Stochastic processes and statistical inference. Arkiv för Matematik 1 (3), 195–277.

Hájek, J., 1956. Linear estimation of the mean value of a stationary random process with convex correlation function. Czechoslovak Mathematical Journal 6 (81), 94–117.

Hickernell, F., 1998. A generalized discrepancy and quadrature error bound. Mathematics of Computation 67 (221), 299–322.

Mak, S., Joseph, V., 2017. Projected support points, with application to optimal MCMC reduction. arXiv preprint arXiv:1708.06897.

Mak, S., Joseph, V., 2018. Support points. Annals of Statistics 46 (6A), 2562–2592.

Näther, W., 1985. Effective Observation of Random Fields. Vol. 72. Teubner-Texte zur Mathematik, Leipzig.

References II

Nogales Gómez, A., Pronzato, L., Rendas, M.-J., 2021. Incremental space-filling design based on coverings and spacings: improving upon low discrepancy sequences. Journal of Statistical Theory and Practice (to appear).

O'Hagan, A., 1991. Bayes-Hermite quadrature. Journal of Statistical Planning and Inference 29 (3), 245–260.

Pronzato, L., 2021. Performance analysis of greedy algorithms for minimising a maximum mean discrepancy. hal-03114891, arXiv:2101.07564.

Pronzato, L., Zhigljavsky, A., 2019. Measures minimizing regularized dispersion. Journal of Scientific Computing 78 (3), 1550–1570.

Pronzato, L., Zhigljavsky, A., 2020. Bayesian quadrature, energy minimization and space-filling design. SIAM/ASA J. Uncertainty Quantification 8 (3), 959–1011.

Pronzato, L., Zhigljavsky, A., 2021. Minimum-energy measures for singular kernels. Journal of Computational and Applied Mathematics 382, (16 pages, to appear) hal-02495643.

Sejdinovic, S., Sriperumbudur, B., Gretton, A., Fukumizu, K., 2013. Equivalence of distance-based and RKHS-based statistics in hypothesis testing. The Annals of Statistics 41 (5), 2263–2291.

Sriperumbudur, B., Gretton, A., Fukumizu, K., Schölkopf, B., Lanckriet, G., 2010. Hilbert space embeddings and metrics on probability measures. Journal of Machine Learning Research 11, 1517–1561.

References III

Szabó, Z., Sriperumbudur, B., 2018. Characteristic and universal tensor product kernels. Journal of Machine Learning Research 18, 1–29.

Székely, G., Rizzo, M., 2013. Energy statistics: A class of statistics based on distances. Journal of Statistical Planning and Inference 143 (8), 1249–1272.

Teymur, O., Gorham, J., Riabiz, M., Oates, C., 2020. Optimal quantisation of probability measures using maximum mean discrepancy. arXiv preprint arXiv:2010.07064v1.