Gaussian process regression in the flat limit

Pierre-Olivier Amblard

from a work with Simon Barthelmé, Nico Tremblay and Kostya Usevitch

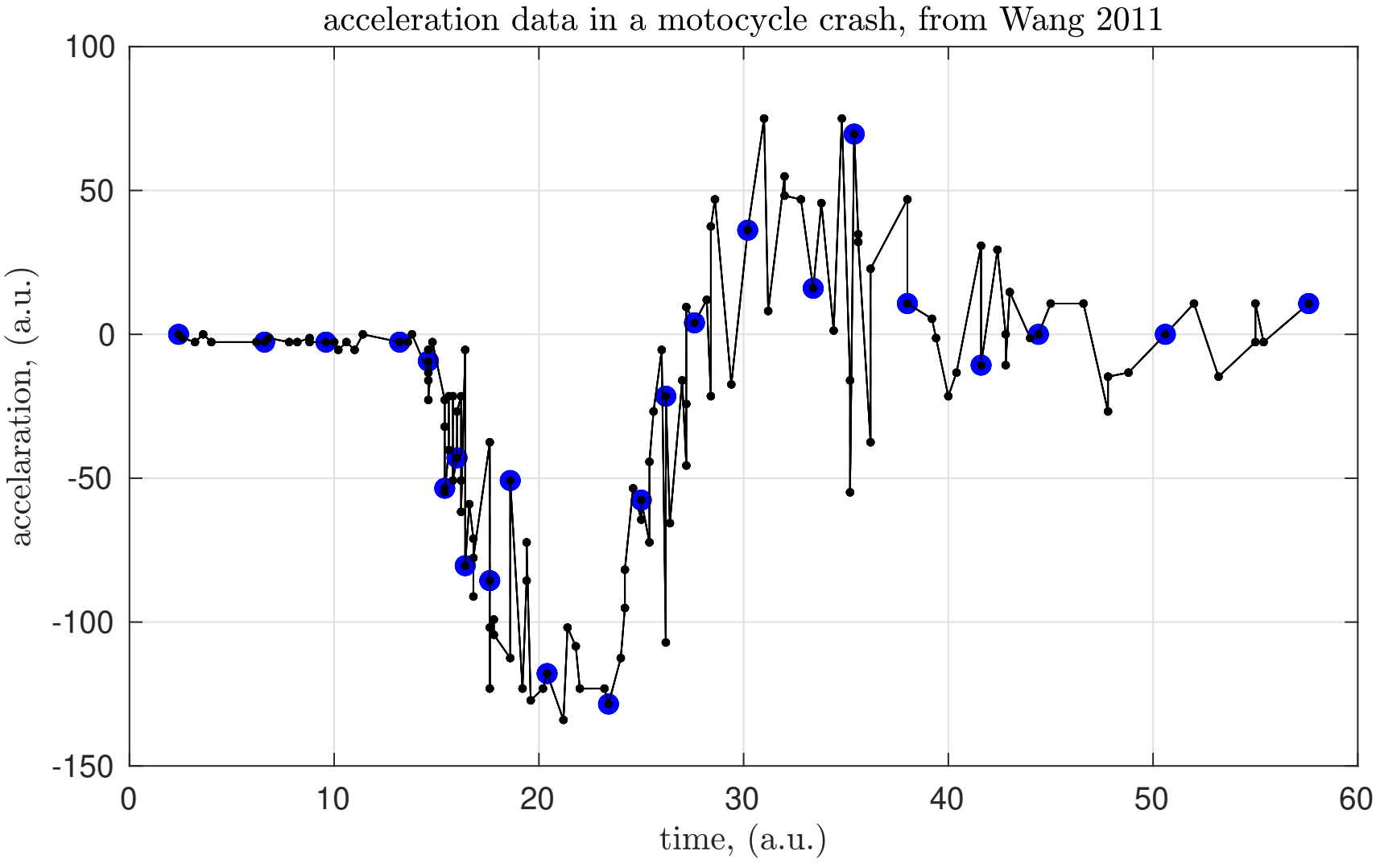
CNRS, GIPSA-lab, U. Grenoble-Alpes, France

W'p "Kernel and sampling methods for design and quantization", Gdr MascoNum 2021

gipsa-lab

Gaussian Process Regression

↪ Find f such that $y_i = f(x_i) + \sigma w_i$ for $i = 1, \ldots, N$



Gaussian Process Regression

$\hookrightarrow$ Find f such that $y_i = f(x_i) + \sigma w_i$ for $i = 1, \ldots, N$

$\hookrightarrow$ In our works, $x_i \in \mathbb{R}^d$, $y_i \in \mathbb{R}$, and w_i a sequence of i.i.d. $\mathcal{N}(0,1)$. BUT here, $d = 1$!

$\hookrightarrow$ In a Bayesian approach, $f : \mathbb{R}^d \to \mathbb{R}$ is given a Gaussian process prior, with zero mean and covariance $k : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$.

$\hookrightarrow$ The *a posteriori* $P(y|\, \boldsymbol{y}, f, \boldsymbol{x}, x)$ at x is Gaussian with

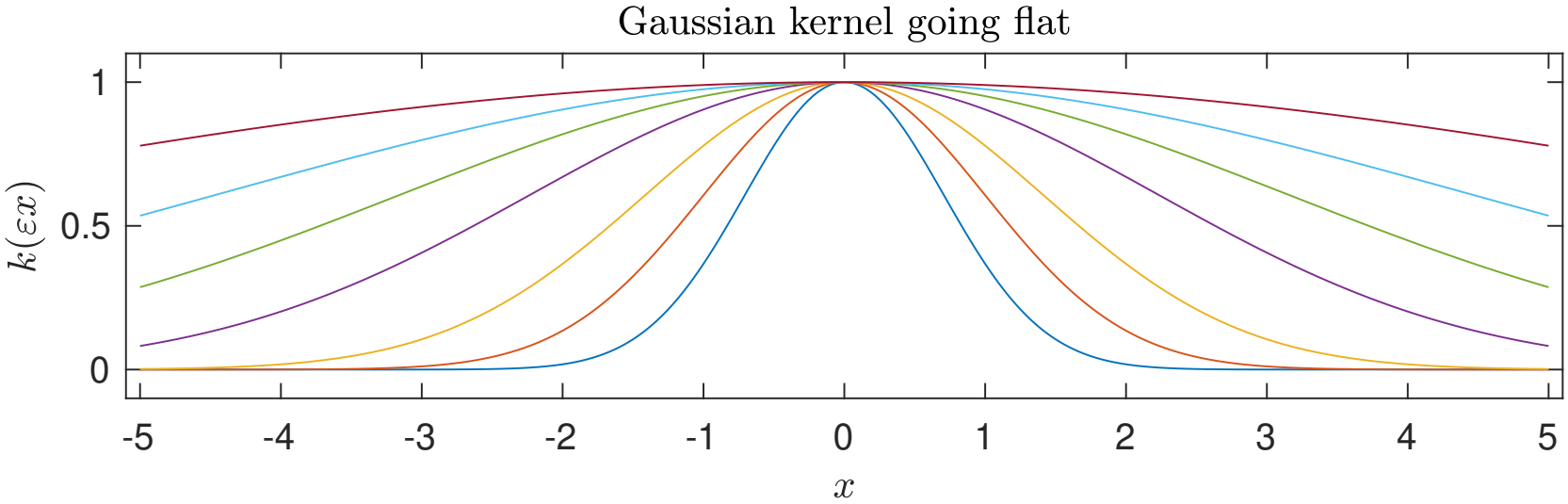
$$\begin{aligned} y^\star(x) = E[y|\boldsymbol{y}, f] &= \sum_i \alpha_i k(x, x_i) = \boldsymbol{\alpha}^\top \boldsymbol{k}_x \text{ with } \boldsymbol{\alpha} = (\boldsymbol{K} + \sigma^2 \boldsymbol{I})^{-1} \boldsymbol{y} \\ \text{Var}[y|\boldsymbol{y}, f] &= k(x,x) - \boldsymbol{k}_x^\top (\boldsymbol{K} + \sigma^2 \boldsymbol{I})^{-1} \boldsymbol{k}_x \end{aligned}$$

$$\boldsymbol{K}_{i,j} = k(x_i, x_j) \qquad \boldsymbol{k}_{x,i} = k(x, x_i)$$

$\hookrightarrow$ $\sigma^2 \longrightarrow 0$: recovering interpolation

Radial Basis Interpolation in the flat limit

- Radial Basis $\sim$ stat./isotropic Gauss. proc. : $k(\boldsymbol{x}, \boldsymbol{y}) = k(\|\boldsymbol{x} - \boldsymbol{y}\|)$.



$\hookrightarrow$ Let ε be the scale parameter, $k(\boldsymbol{x} - \boldsymbol{y}) = k\left(\varepsilon\|\boldsymbol{x} - \boldsymbol{y}\|\right)$.

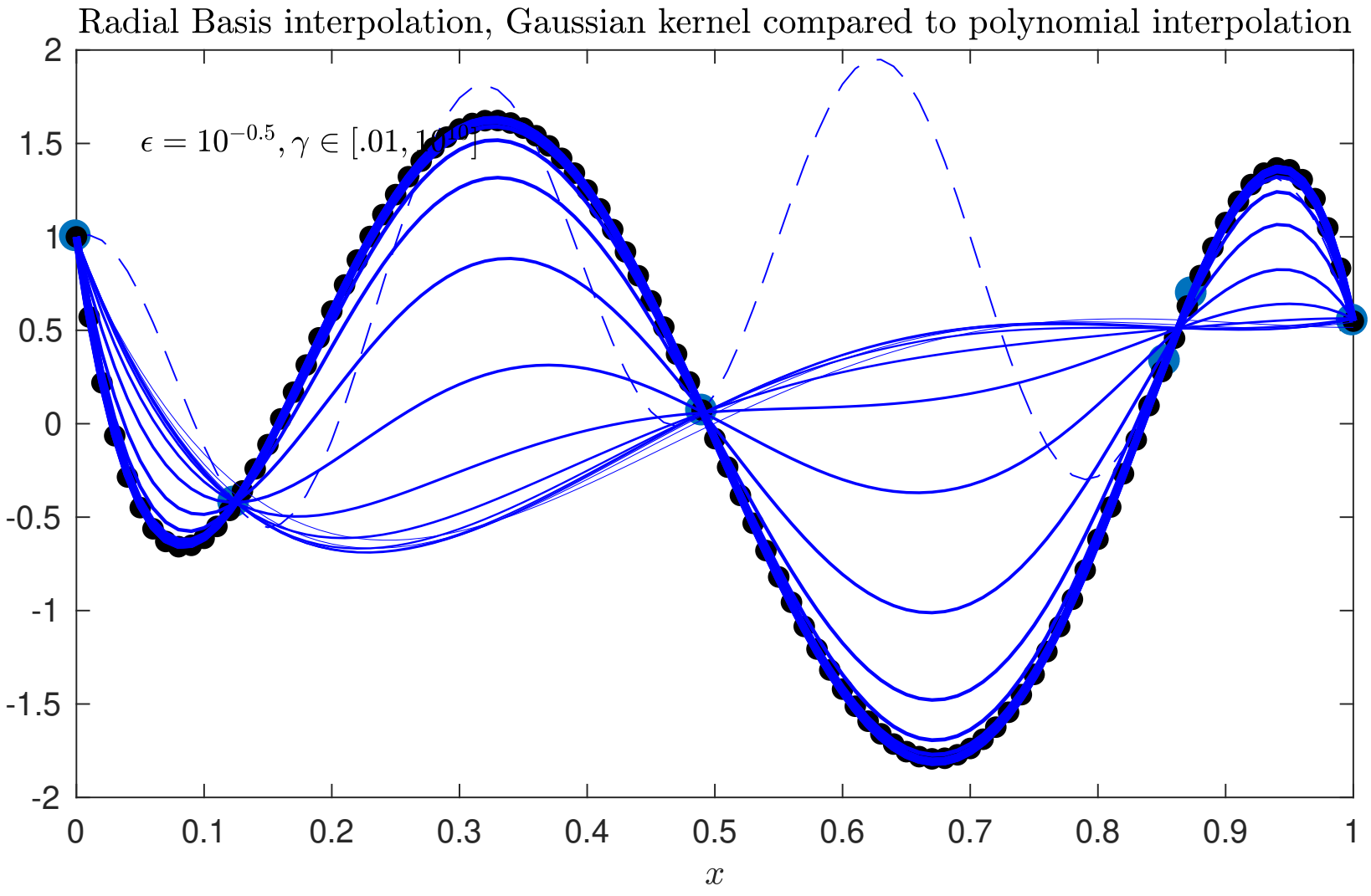
$$y^\star(x) = E[y|\boldsymbol{y}, f] \quad = \quad \sum_i \alpha_i k\left(\varepsilon(\boldsymbol{x} - \boldsymbol{x}_i)\right).$$

Driscoll & Fornberg 2002 proved for $k(\boldsymbol{x} - \boldsymbol{y}) = \exp\left(-\varepsilon^2\|\boldsymbol{x} - \boldsymbol{y}\|^2\right)$

Gaussian RBF interpolation $\overset{\varepsilon \to 0}{\longrightarrow}$ polynomial interpolation

Let's have a look !

RBF interpolation in the flat limit



Radial Basis Interpolation in the flat limit

- Radial Basis $\sim$ stat./isotropic Gauss. proc. : $k(\boldsymbol{x}, \boldsymbol{y}) = k(\|\boldsymbol{x} - \boldsymbol{y}\|)$.

$\hookrightarrow$ Let ε be the scale parameter, $k(\boldsymbol{x} - \boldsymbol{y}) = k\left(\varepsilon\|\boldsymbol{x} - \boldsymbol{y}\|\right)$.

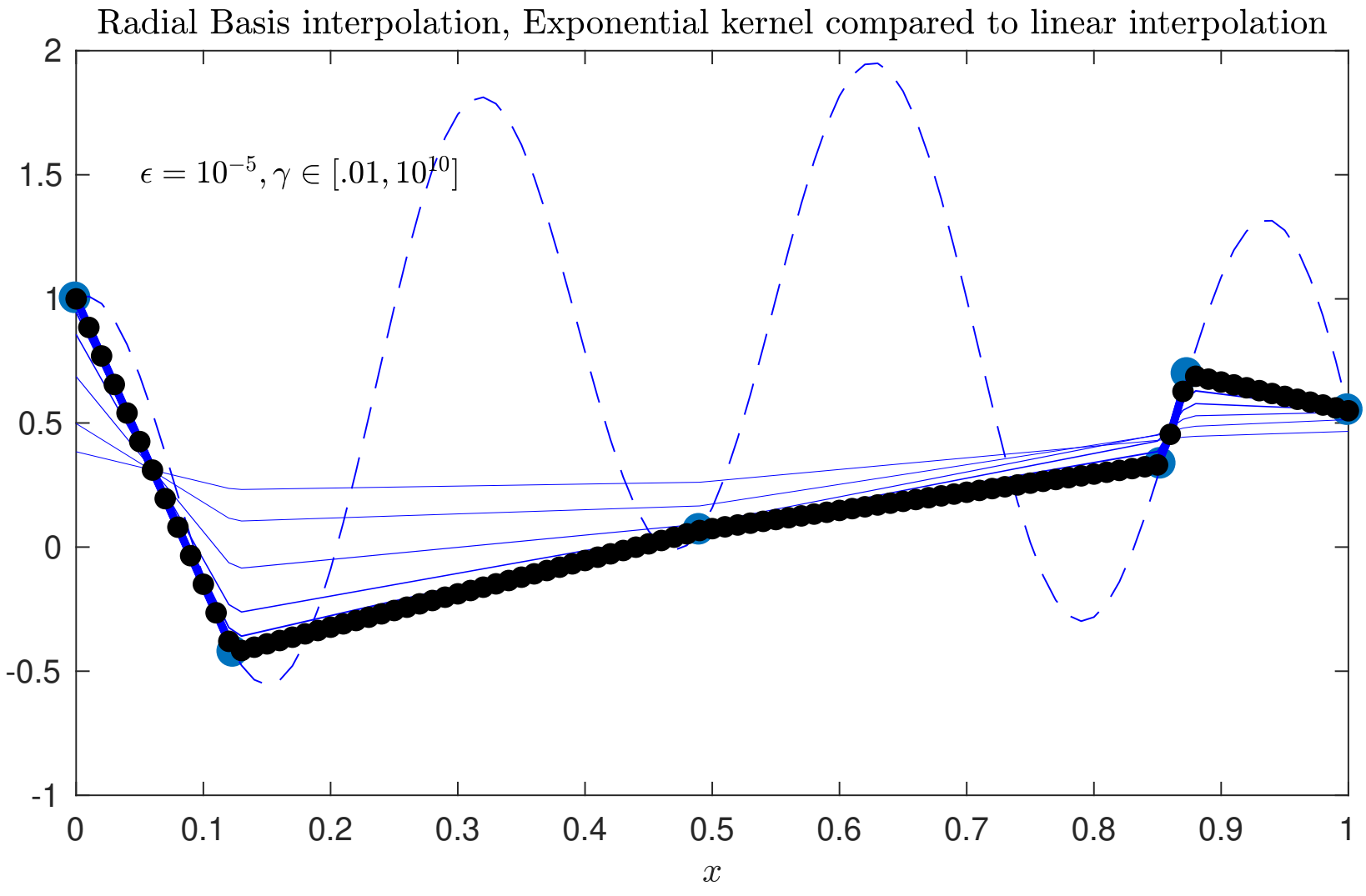
$$y^\star(x) = E[y|\boldsymbol{y}, f] \quad = \quad \sum_i \alpha_i k\left(\varepsilon(\boldsymbol{x} - \boldsymbol{x}_i)\right).$$

Driscoll & Fornberg 2002 proved for $k(\boldsymbol{x} - \boldsymbol{y}) = \exp\left(-\varepsilon^2\|\boldsymbol{x} - \boldsymbol{y}\|^2\right)$

Gaussian RBF interpolation $\overset{\varepsilon\to 0}{\longrightarrow}$ polynomial interpolation

What's up with a harder kernel, *e.g.* $k(\boldsymbol{x} - \boldsymbol{y}) = \exp\left(-\varepsilon\|\boldsymbol{x} - \boldsymbol{y}\|\right)$

RBF interpolation in the flat limit



Radial Basis Interpolation in the flat limit

- Radial Basis $\sim$ stat./isotropic Gauss. proc. : $k(\boldsymbol{x}, \boldsymbol{y}) = k(\|\boldsymbol{x} - \boldsymbol{y}\|)$.

$\hookrightarrow$ Let ε be the scale parameter, $k(\boldsymbol{x} - \boldsymbol{y}) = k\left(\varepsilon\|\boldsymbol{x} - \boldsymbol{y}\|\right)$.

$$y^\star(x) = E[y|\boldsymbol{y}, f] \quad = \quad \sum_i \alpha_i k\left(\varepsilon(\boldsymbol{x} - \boldsymbol{x}_i)\right).$$

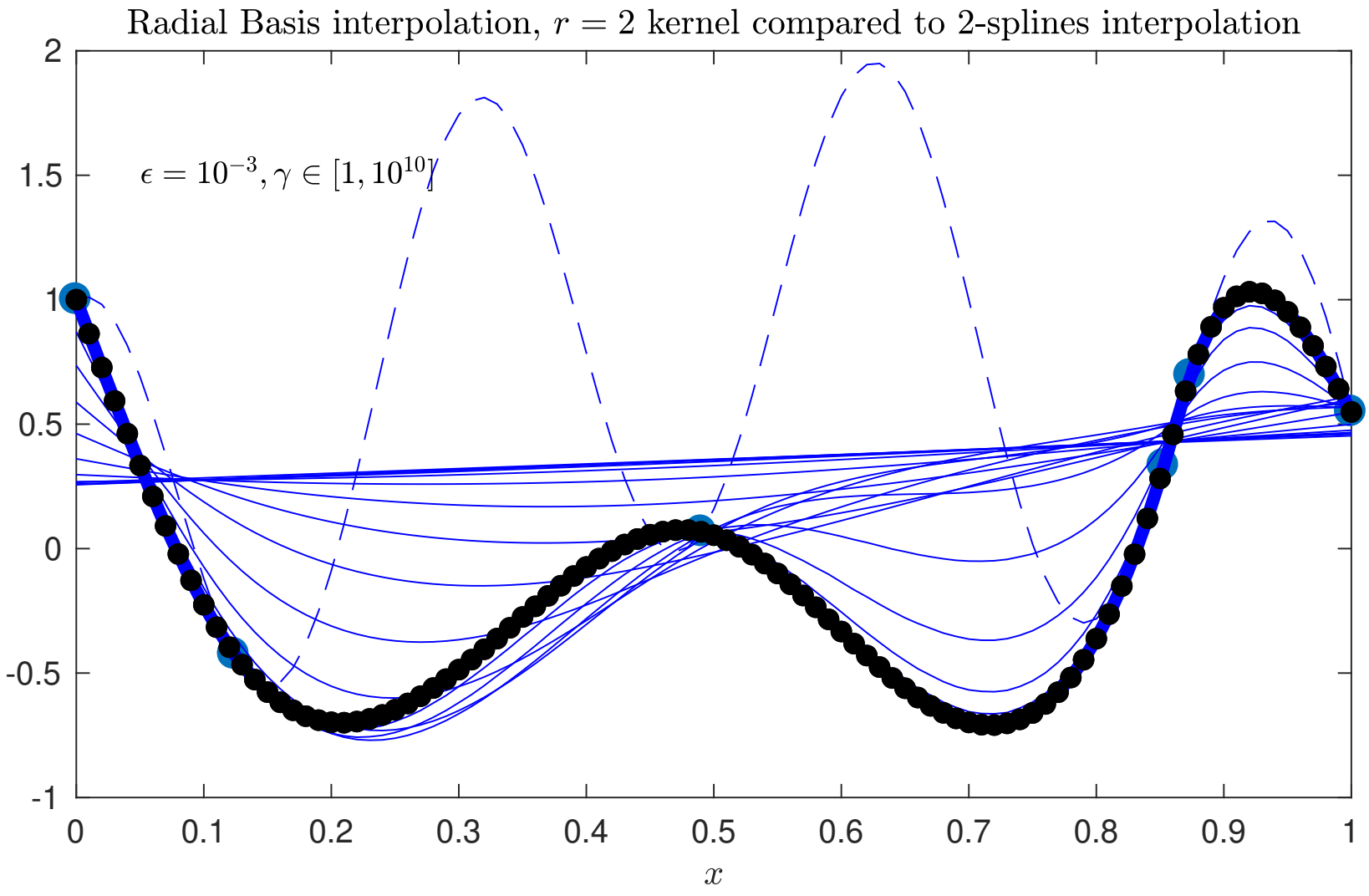
Driscoll & Fornberg 2002 proved for $k(\boldsymbol{x} - \boldsymbol{y}) = \exp\left(-\varepsilon^2\|\boldsymbol{x} - \boldsymbol{y}\|^2\right)$

Gaussian RBF interpolation $\stackrel{\varepsilon\to 0}{\longrightarrow}$ polynomial interpolation

Nice implementation of linear interpolation ! !

Let us smooth it a bit, *e.g.* $k(\boldsymbol{x} - \boldsymbol{y}) = (1 + \varepsilon\|\boldsymbol{x} - \boldsymbol{y}\|)\exp\left(-\varepsilon\|\boldsymbol{x} - \boldsymbol{y}\|\right)$

RBF interpolation in the flat limit



What is hidden ?

Recall : $E[y|\boldsymbol{y}, f] = \boldsymbol{k}_x^\top (\boldsymbol{K} + \sigma^2 \boldsymbol{I})^{-1} \boldsymbol{y}$

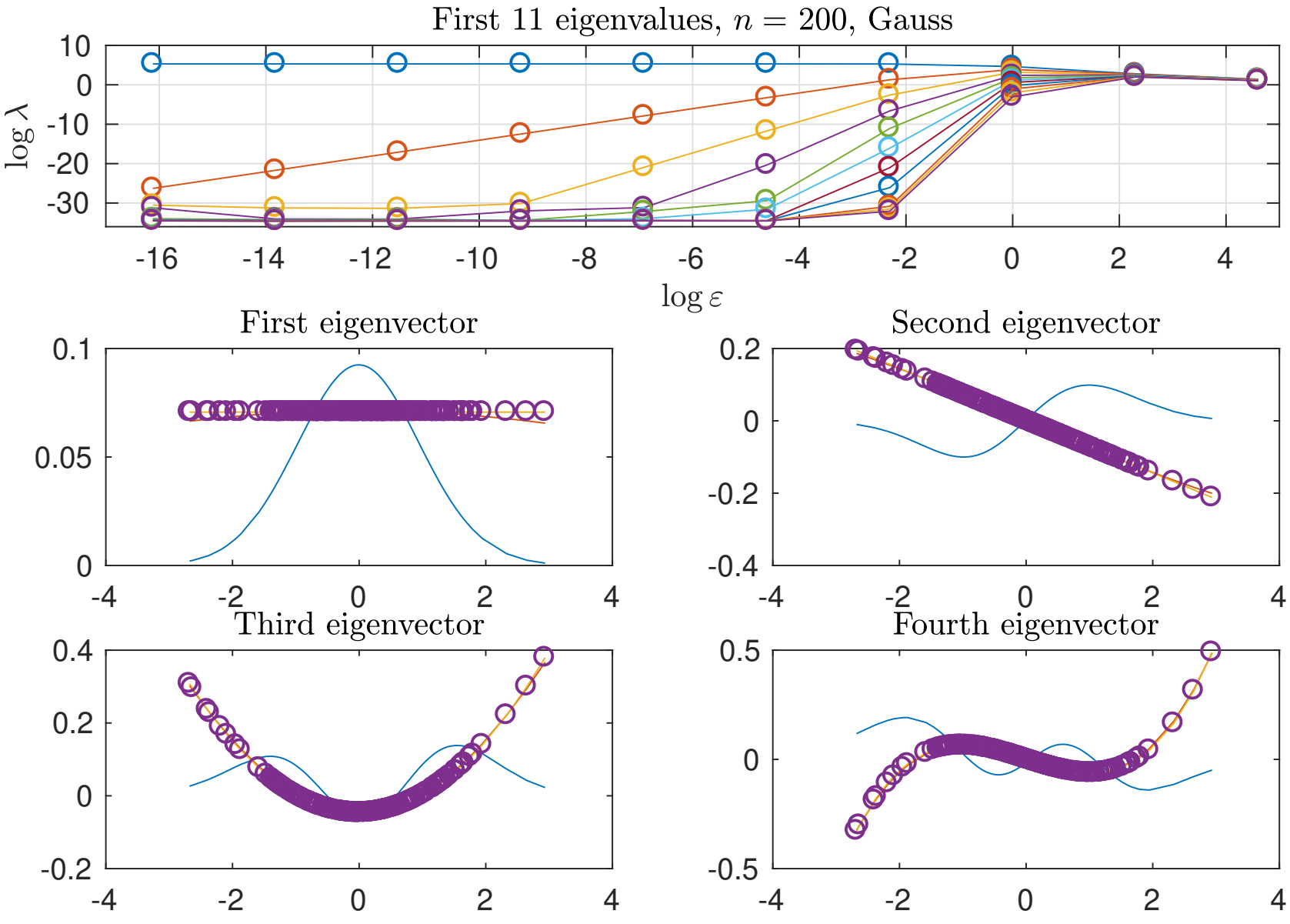
- The kernel is going flat, and thus the Gram matrix goes to a rank 1 matrix !

$\hookrightarrow$ All but one eigenvalues go to zero ! But wait, let us look at the rate !

$\hookrightarrow$ Eigen analysis of the Gram matrix $\boldsymbol{K} = (k(\varepsilon \|x_k - x_l\|))_{k,l}$

$$\boldsymbol{K}\boldsymbol{v}_i(\varepsilon) = \lambda_i(\varepsilon)\boldsymbol{v}_i(\varepsilon)$$

Eigen analysis of the Gram matrix w.r.t. ε, Gauss. k



Smooth kernel : no singularity at zero at any order

- typical example : the Gaussian kernel

$$\exp(-|x|^2) = f_0 + f_2|x|^2 + f_4|x|^4 + \dots$$

$\hookrightarrow$ for scalar inputs, as $\varepsilon \to 0$,

$$\lambda_i(\varepsilon) = \varepsilon^{2i}(\widetilde{\lambda}_i + O(\epsilon)), \forall i = 0, \dots, n-1$$

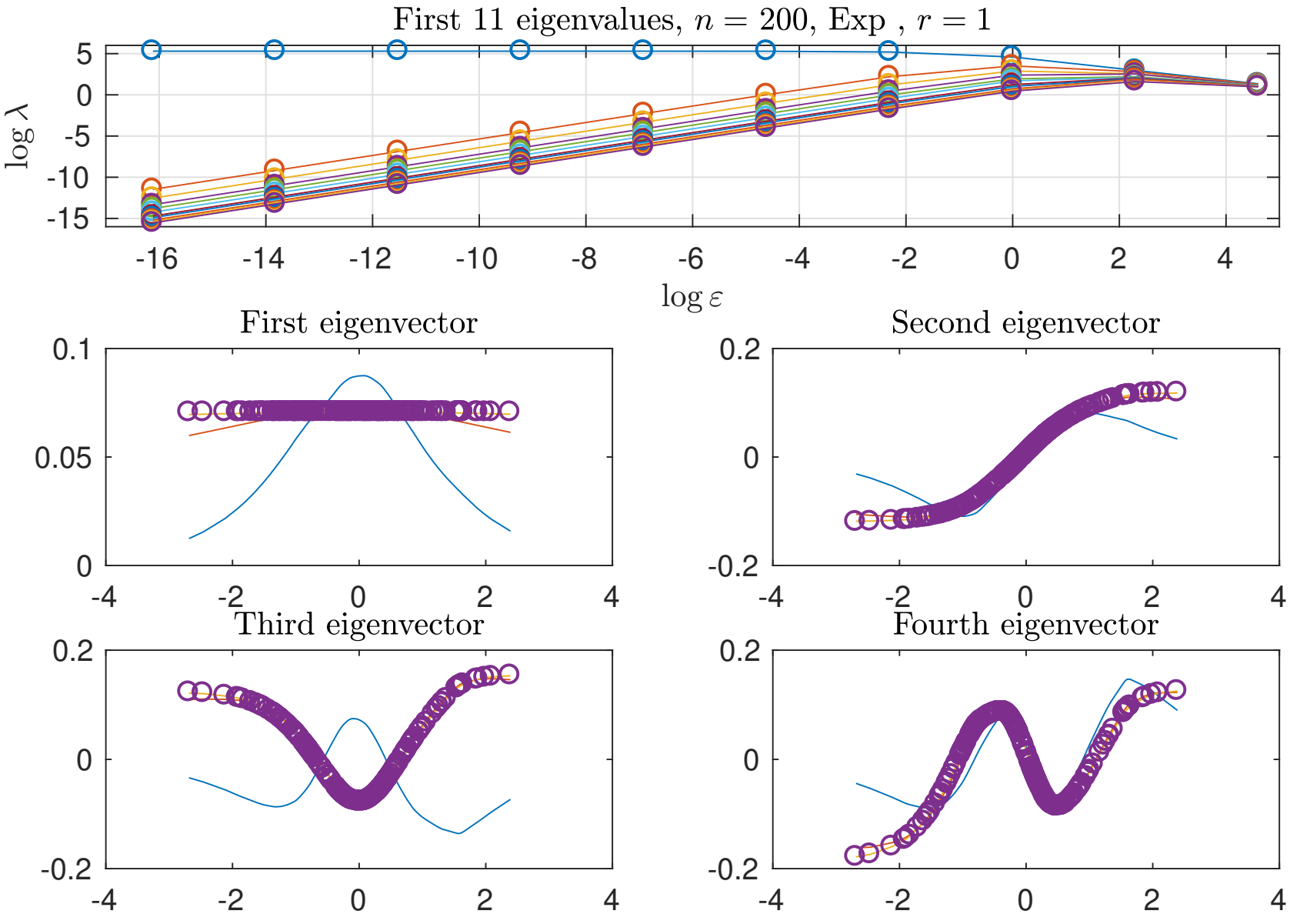
$$\boldsymbol{v}_i = \text{e.}\overrightarrow{v}. \begin{pmatrix} 1 & x_1 & x_1^2 & \dots & x_1^{n-1} \\ 1 & x_2 & x_2^2 & \dots & x_2^{n-1} \\ \vdots & \vdots & \vdots & \dots & \vdots \\ 1 & x_n & x_n^2 & \dots & x_n^{n-1} \end{pmatrix}$$

i.e. orthogonal polynomials.

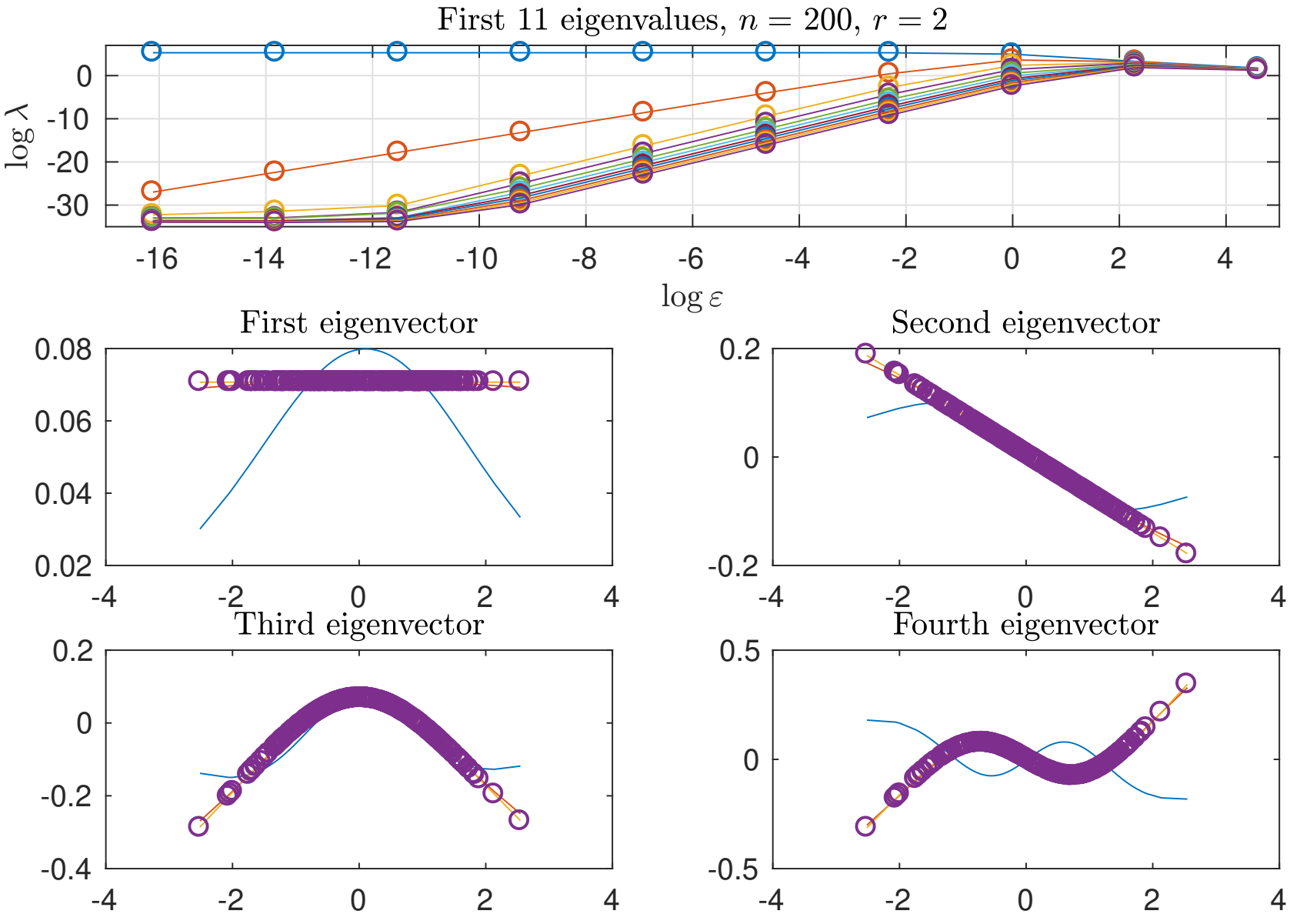
But what if :

$$k(|x|) = f_0 + f_2|x|^2 + f_4|x|^4 + \dots + f_{2r-1}|x|^{2r-1} + f_{2r}|x|^{2r} + \dots$$

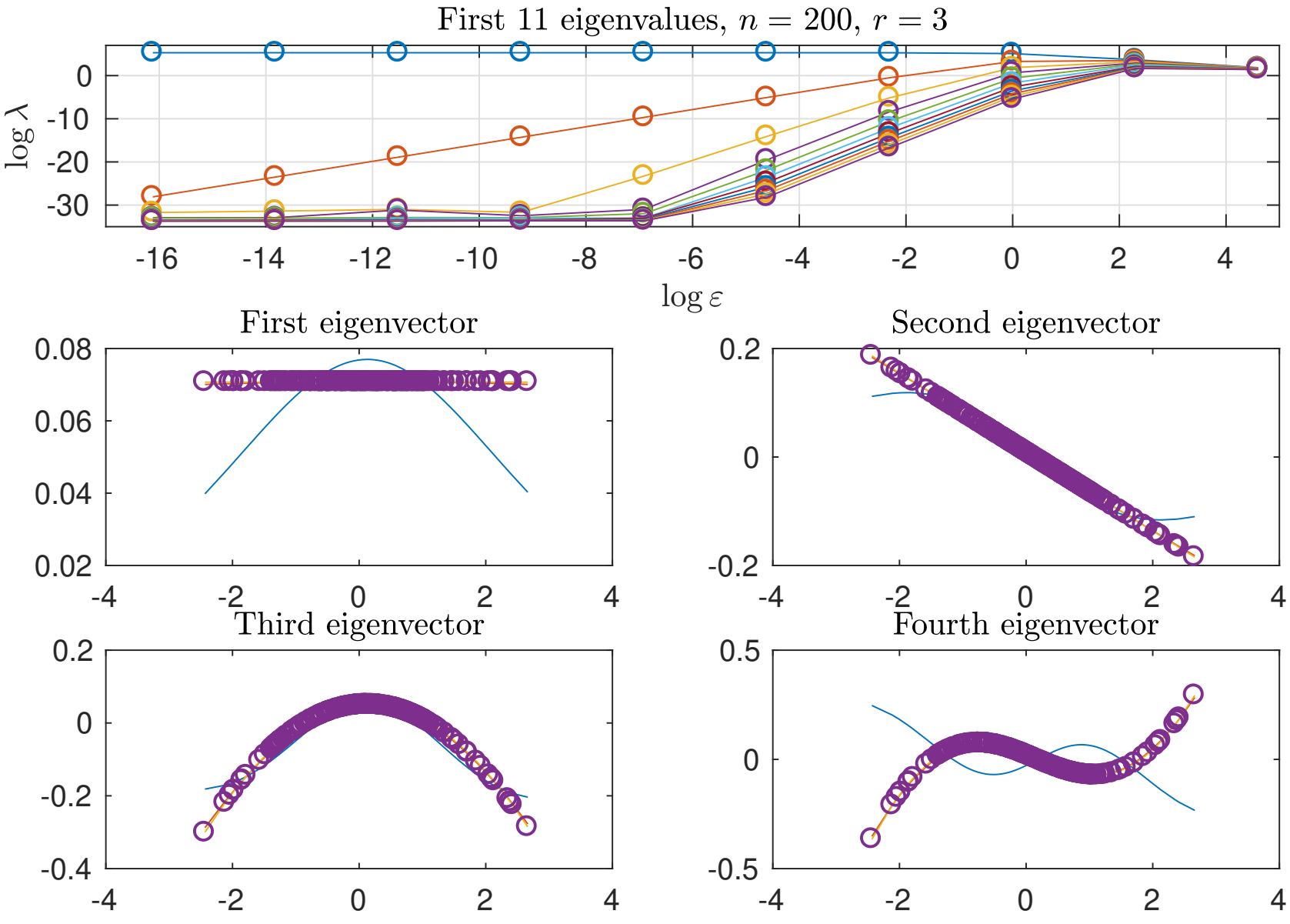
Eigen analysis of the Gram matrix w.r.t. ε, $r = 1$



Eigen analysis of the Gram matrix w.r.t. ε, $r = 2$



Eigen analysis of the Gram matrix w.r.t. ε, $r = 3$



Non smooth kernel : singularity at zero somewhere

$$k(|x|) = f_0 + f_2|x|^{\,2} + f_4|x|^{\,4} + \ldots + f_{2r-1}|x|^{2r-1} + f_{2r}|x|^{2r} + \ldots$$

$\hookrightarrow$ for scalar inputs, as $\varepsilon \to 0$,

$$\begin{aligned}
\lambda_i(\varepsilon) &= \varepsilon^{2i}(\widetilde{\lambda}_i + O(\epsilon)), \forall i = 0, \ldots, r-1 \\
\lambda_i(\varepsilon) &= \varepsilon^{2r-1}(\widetilde{\lambda}_i + O(\epsilon)), \forall i = r, \ldots, n
\end{aligned}$$

$$\begin{aligned}
\boldsymbol{v}_i &= \text{e.}\overrightarrow{v.} \begin{pmatrix} 1 & x_1 & x_1^2 & \ldots & x_1^{r-1} \\ 1 & x_2 & x_2^2 & \ldots & x_2^{r-1} \\ \vdots & \vdots & \vdots & \ldots & \vdots \\ 1 & x_n & x_n^2 & \ldots & x_n^{r-1} \end{pmatrix} \forall i = 0, \ldots, r-1 \\
\boldsymbol{v}_i &= \text{linked to } (\|x_k - x_l\|^{2r-1})_{k,l}, \quad \forall i = r, \ldots, n
\end{aligned}$$

Back to GPR, degrees of freedom

$$y^\star(x) = E[y|\boldsymbol{y}, f] = \sum_i \alpha_i k(x, x_i) = \boldsymbol{\alpha}^\top \boldsymbol{k}_x^\top (\boldsymbol{K} + \sigma^2 \boldsymbol{I})^{-1} \boldsymbol{y}$$

$$\mathrm{Var}[y|\boldsymbol{y}, f] = k(x, x) - \boldsymbol{k}_x^\top (\boldsymbol{K} + \sigma^2 \boldsymbol{I})^{-1} \boldsymbol{k}_x$$

Prediction mean and variance at x given $\boldsymbol{x}$ can be computed from the smoother matrix for $(\boldsymbol{x}, x)$

- For linear smoothers $y^\star(\boldsymbol{x}) = \boldsymbol{M}\boldsymbol{y}$, d.o.f. : $\mathrm{dof} = \mathrm{Tr}(\boldsymbol{M})$ (Hastie/Tibs./Fried, TESL)

all info. in the smoothing matrix, here : $\boldsymbol{M}(\varepsilon) = \boldsymbol{K}(\boldsymbol{K} + \sigma^2 \boldsymbol{I})^{-1}$!

$\hookrightarrow$ For GPR

$$dof(\varepsilon) = \mathrm{Tr}(\boldsymbol{M}) = \sum_{i=1}^{n} \frac{\lambda_i(\varepsilon)}{\lambda_i(\varepsilon) + \sigma^2} \approx \sum_{i=1}^{n} \frac{\widetilde{\lambda}_i \varepsilon^{e_i}}{\widetilde{\lambda}_i \varepsilon^{e_i} + \sigma^2} \xrightarrow{\varepsilon \to 0} \frac{\widetilde{\lambda_1}}{\widetilde{\lambda_1} + \sigma^2} \leq 1!$$

A proj. (*e.g. poly. reg.*) onto any $q \leq n$ dimensional space : $\mathrm{dof} = q$! !

$\hookrightarrow$ We must add another parameter to control the d.o.f.

Rescaling

We didn't control the amplitude !
$\hookrightarrow$ Let the kernel be $\gamma(\varepsilon)k(\varepsilon\| \boldsymbol{x} - \boldsymbol{y}\|)$.

The d.o.f. writes

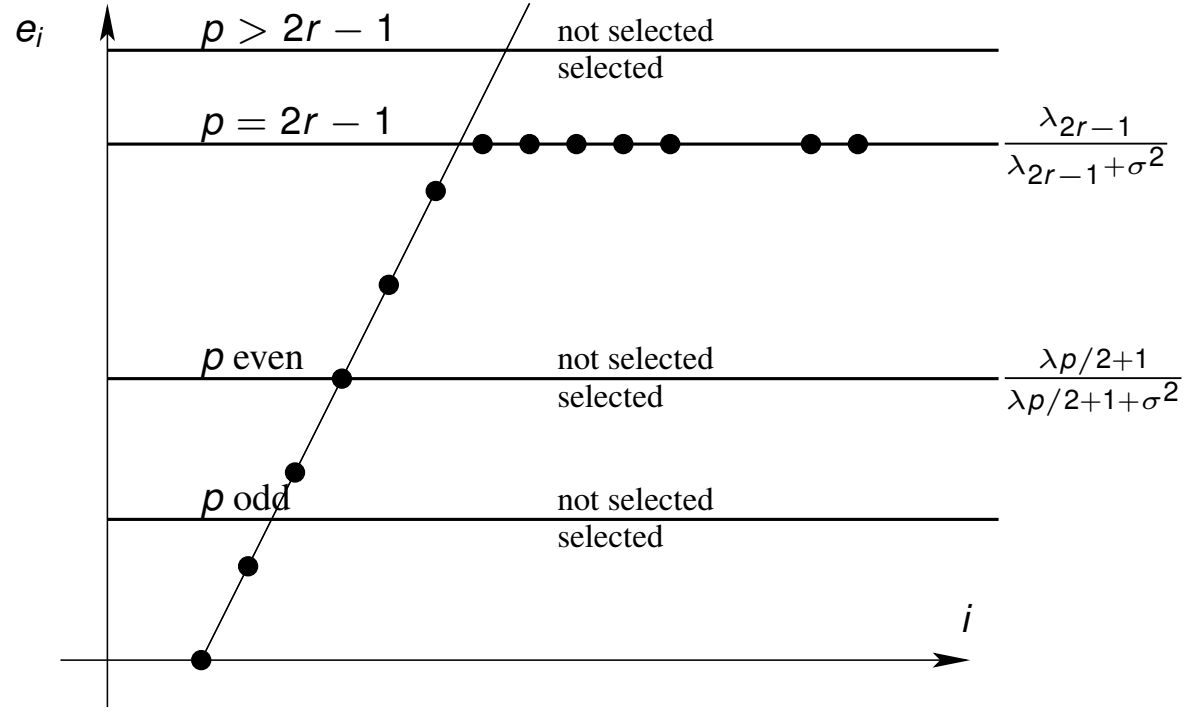
$$\mathrm{Tr}(\boldsymbol{M}) = \sum_{i=1}^{n} \frac{\gamma(\varepsilon)\lambda_i(\varepsilon)}{\gamma(\varepsilon)\lambda_i(\varepsilon) + \sigma^2} \approx \sum_{i=1}^{n} \frac{\widetilde{\lambda}_i}{\widetilde{\lambda}_i + \gamma(\varepsilon)^{-1}\varepsilon^{-e_i}\sigma^2}$$

Natural choice $\gamma(\varepsilon) \propto \varepsilon^{-p}$: p controls the d.o.f. through ε^{p-e_i}

$$y^\star(\boldsymbol{x}) = \sum_{i=1}^{n} \frac{\widetilde{\lambda}_i}{\widetilde{\lambda}_i + \gamma(\varepsilon)^{-1}\varepsilon^{-e_i}\sigma^2} \boldsymbol{v}_i(\varepsilon)\boldsymbol{v}_i^\top(\varepsilon)\boldsymbol{y}$$

$$\frac{\widetilde{\lambda}_i}{\widetilde{\lambda}_i + \varepsilon^{p-e_i}\sigma^2} \longrightarrow \begin{cases} 1 : \text{eigen vector } i \text{ is kept} \\ \text{a constant} : \text{eigen vector } i \text{ is penalized} \\ 0 : \text{eigen vector } i \text{ is left} \end{cases}$$

$\lambda/(\lambda+\sigma^2)$ as a filter



$$\frac{\widetilde{\lambda}_i}{\widetilde{\lambda}_i+\varepsilon^{p-2(i-1)}\sigma^2} \longrightarrow \begin{cases} 1 : \text{the eigen vector is kept} \\ \text{a constant} : \text{the eigen vector is penalized} \\ 0 : \text{the eigen vector is left} \end{cases}$$

$$\frac{\widetilde{\lambda}_i}{\widetilde{\lambda}_i+\varepsilon^{p-(2r-1)}\sigma^2} \longrightarrow \begin{cases} 1 : \text{the eigen vectors are kept} \\ \text{a constant} : \text{the eigen vectors are penalized} \\ 0 : \text{the eigen vectors are left} \end{cases}$$

Theorem

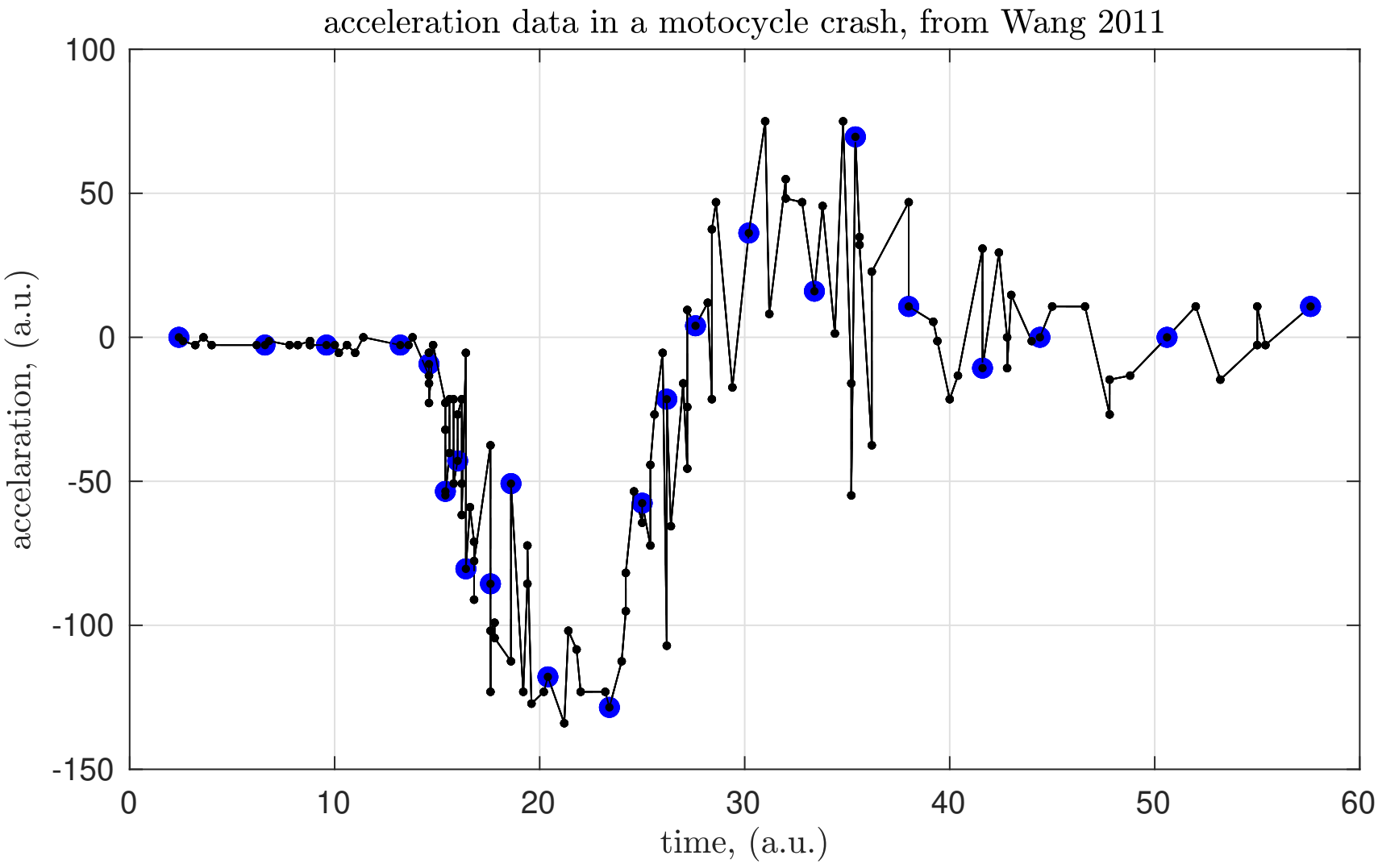
For the kernel $\varepsilon^{-p} k(\varepsilon \|\boldsymbol{x} - \boldsymbol{y}\|)$ with regularity r, the associated Gaussian process regression converges as $\varepsilon \to 0$ toward :

- interpolation if $p > 2r - 1$;
- smoothing splines of order r is $p = 2r - 1$;
- penalized polynomial regression if $p < 2r - 1$ and even ;
- polynomial regression of order $\lfloor p/2 + 1 \rfloor$ if $p < 2r - 1$ and odd.

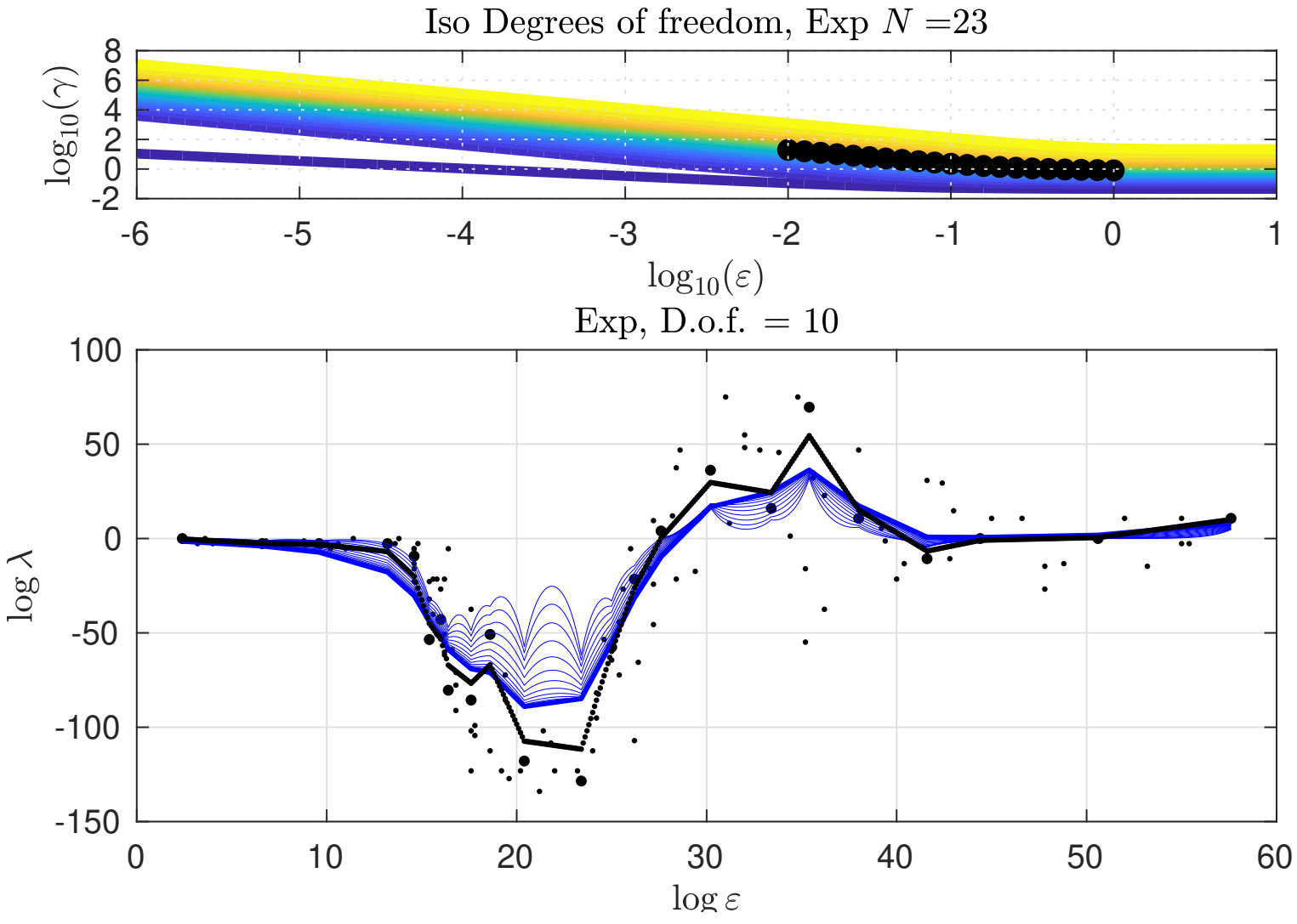
Same kind of results in the multivariate case ! and stuffs like :

as $\varepsilon \to 0$, GPR for $\varepsilon^{-p} \exp(\varepsilon \|\boldsymbol{x} - \boldsymbol{y}\|^2)$ behaves as GPR with $(\boldsymbol{x}^\top \boldsymbol{y})^{p/2-1}$ (poly. reg., penalized or not)

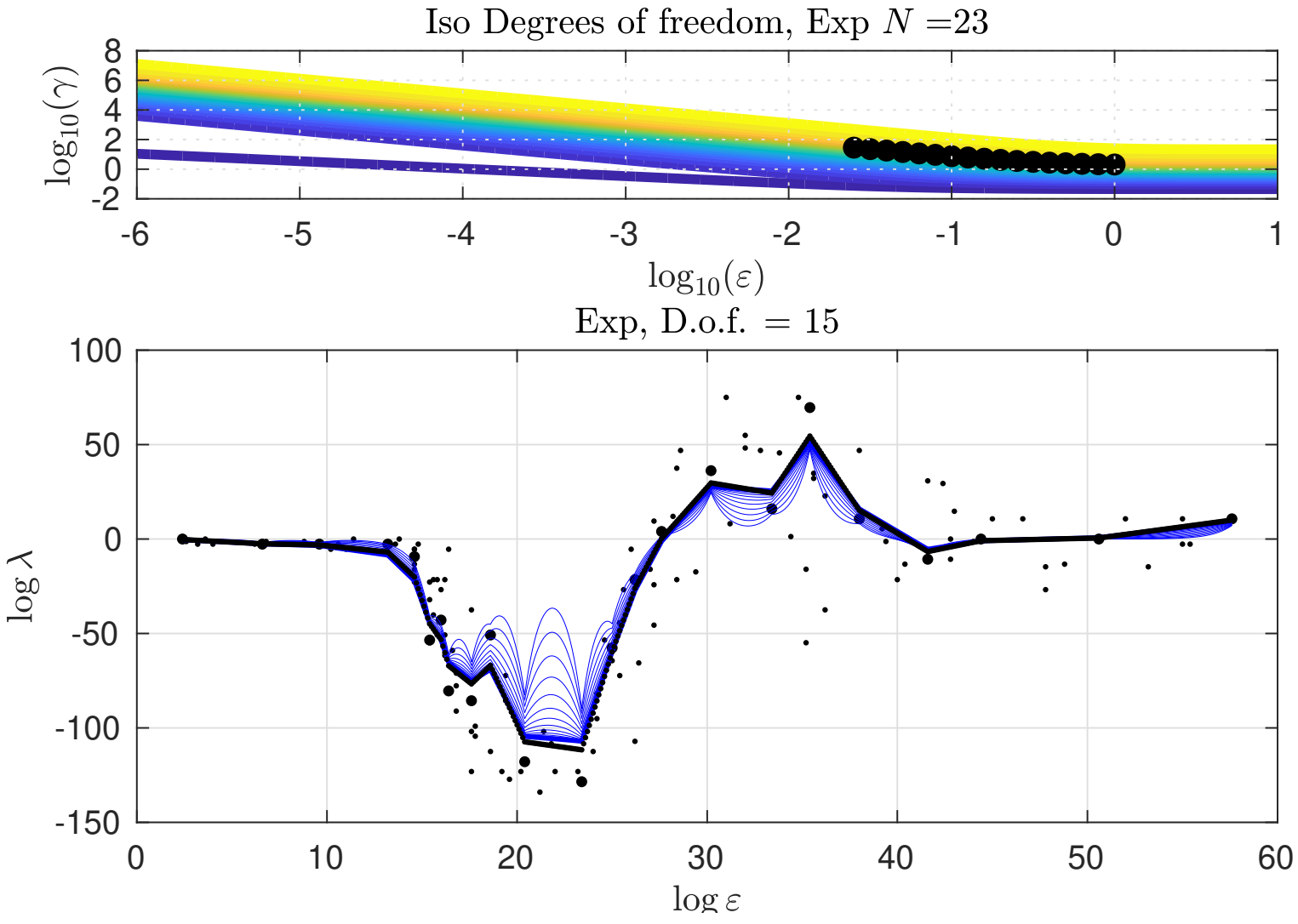
Illustrations



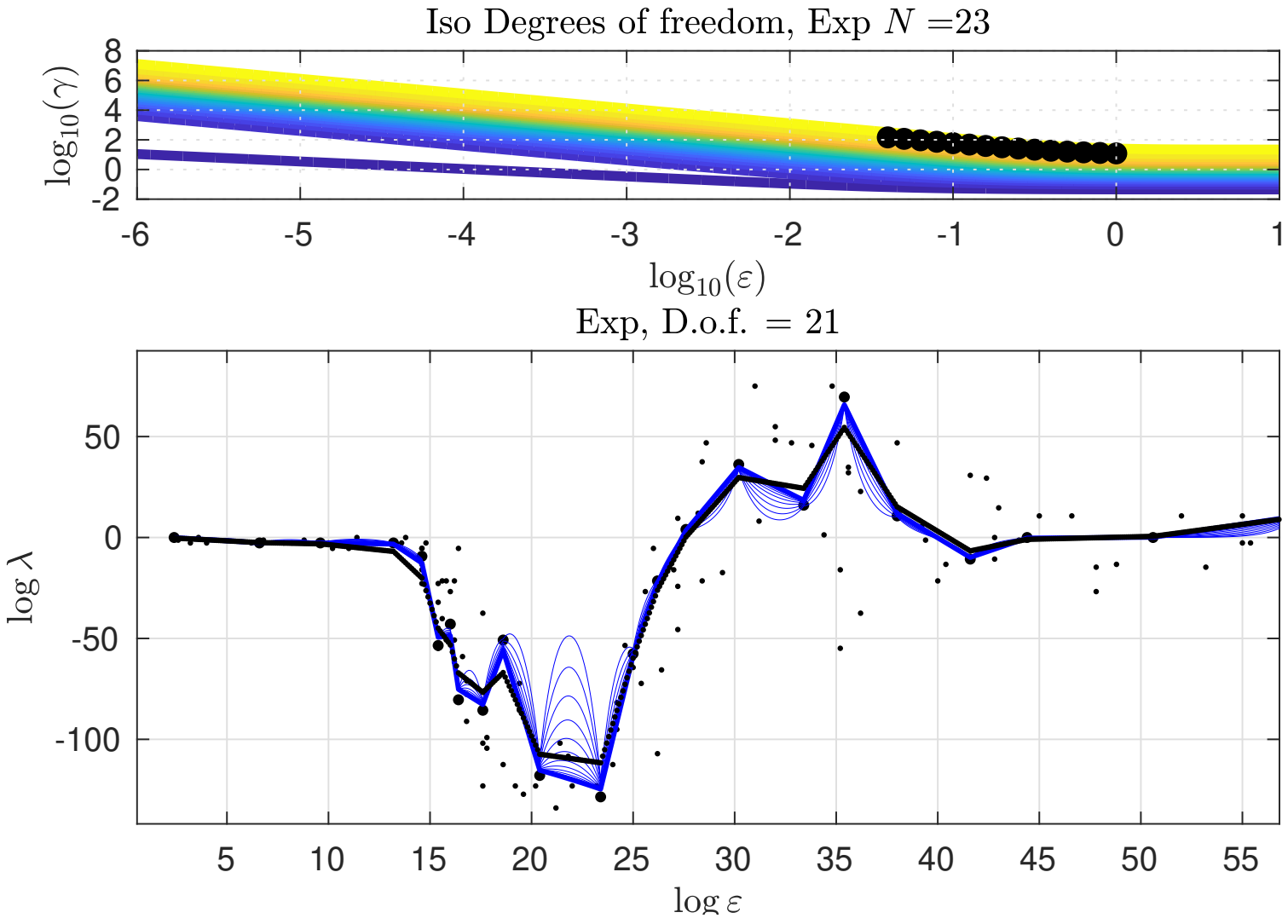
Exponential kernel, $r = 1$



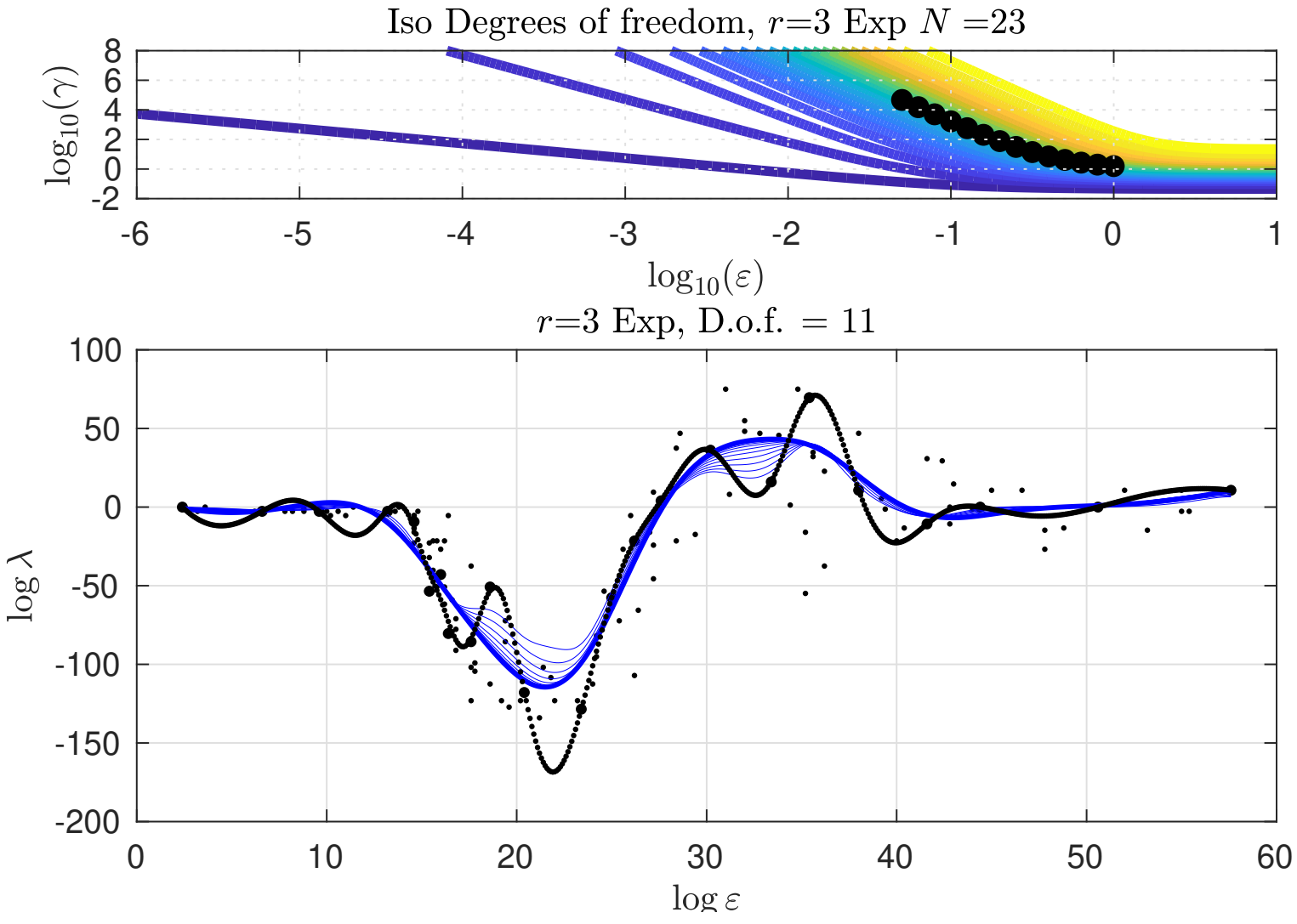
Exponential kernel, $r = 1$



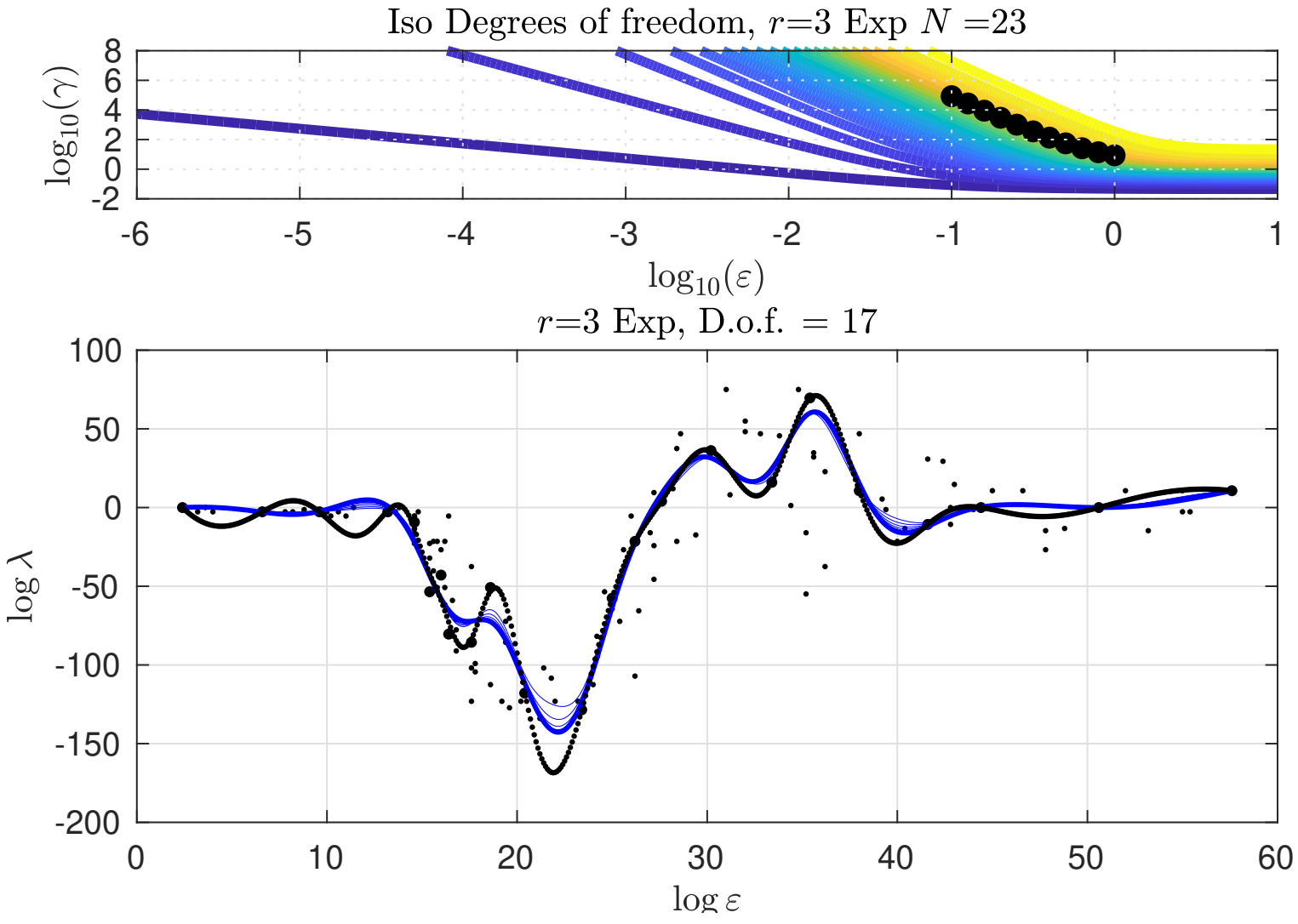
Exponential kernel, $r = 1$



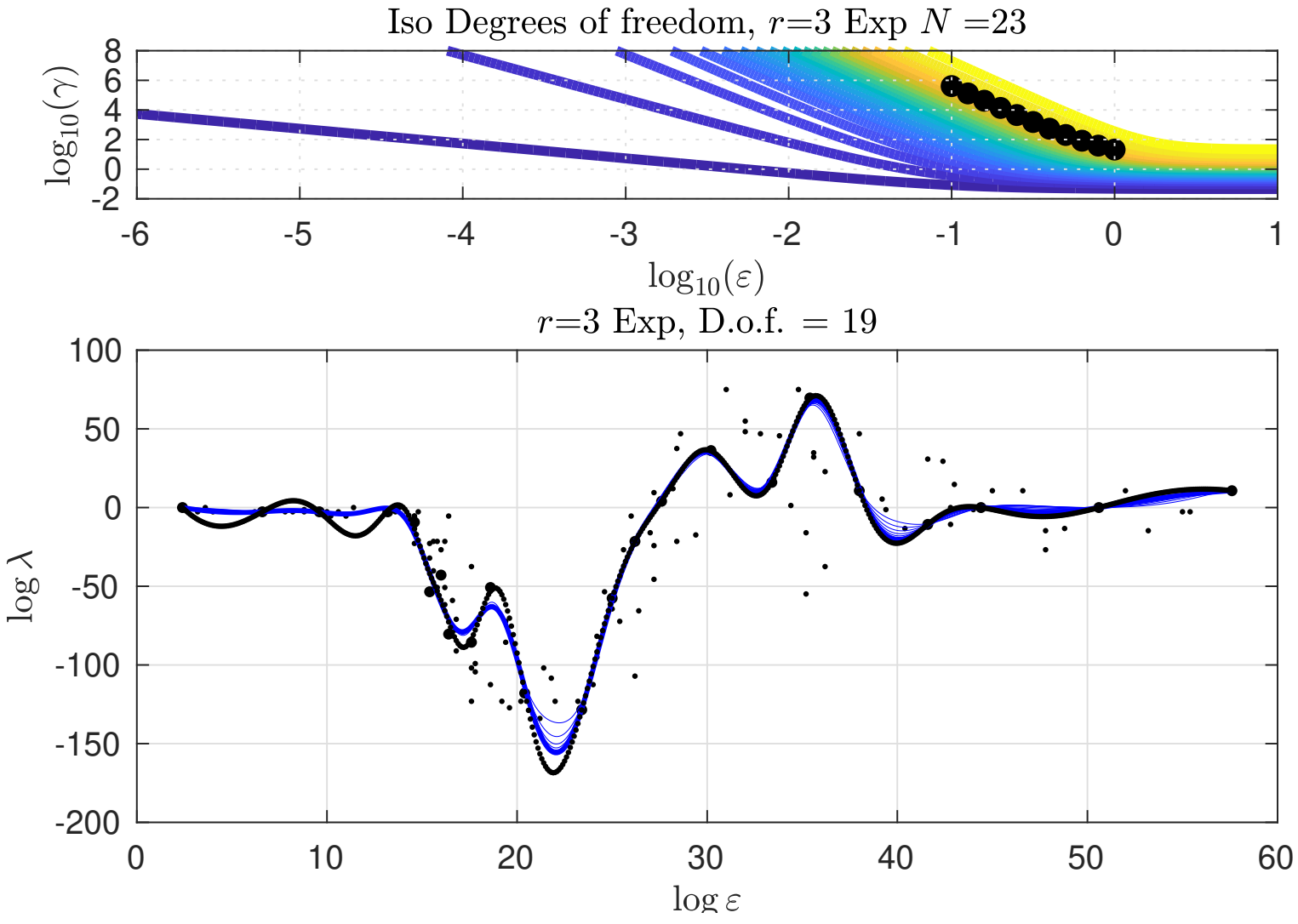
Exponential kernel, $r = 3$



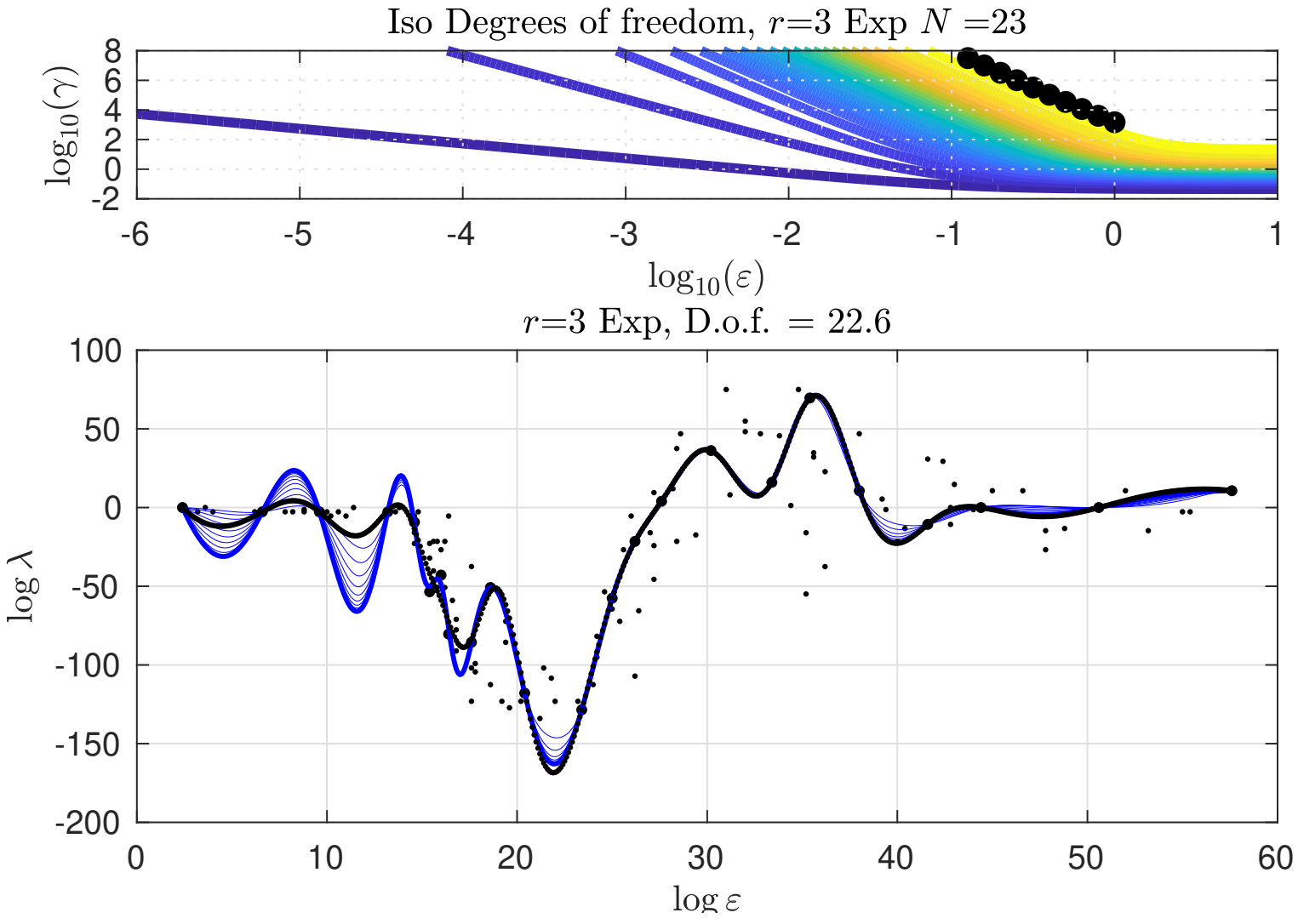
Exponential kernel, $r = 3$



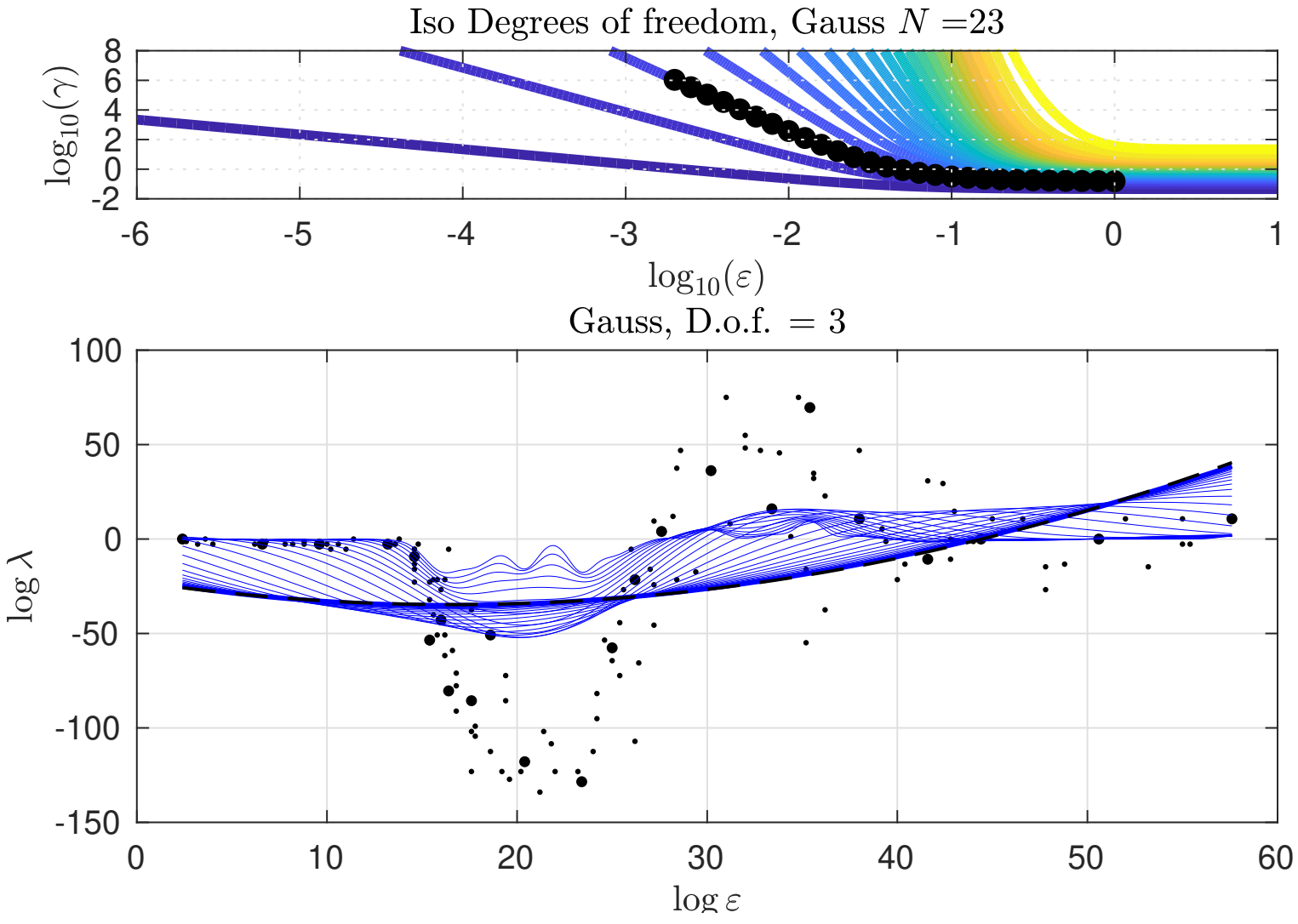
Exponential kernel, $r = 3$



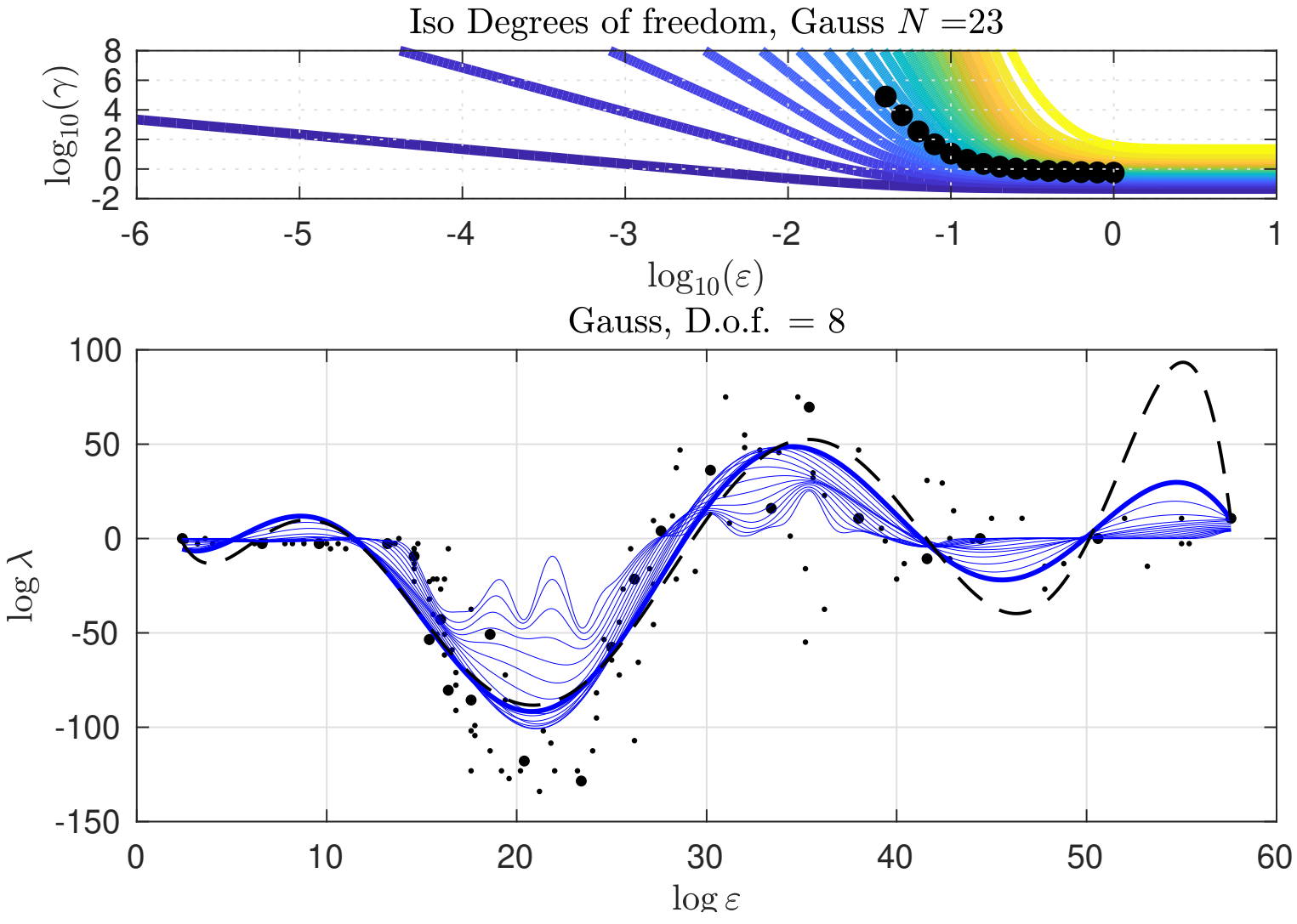
Exponential kernel, $r = 3$



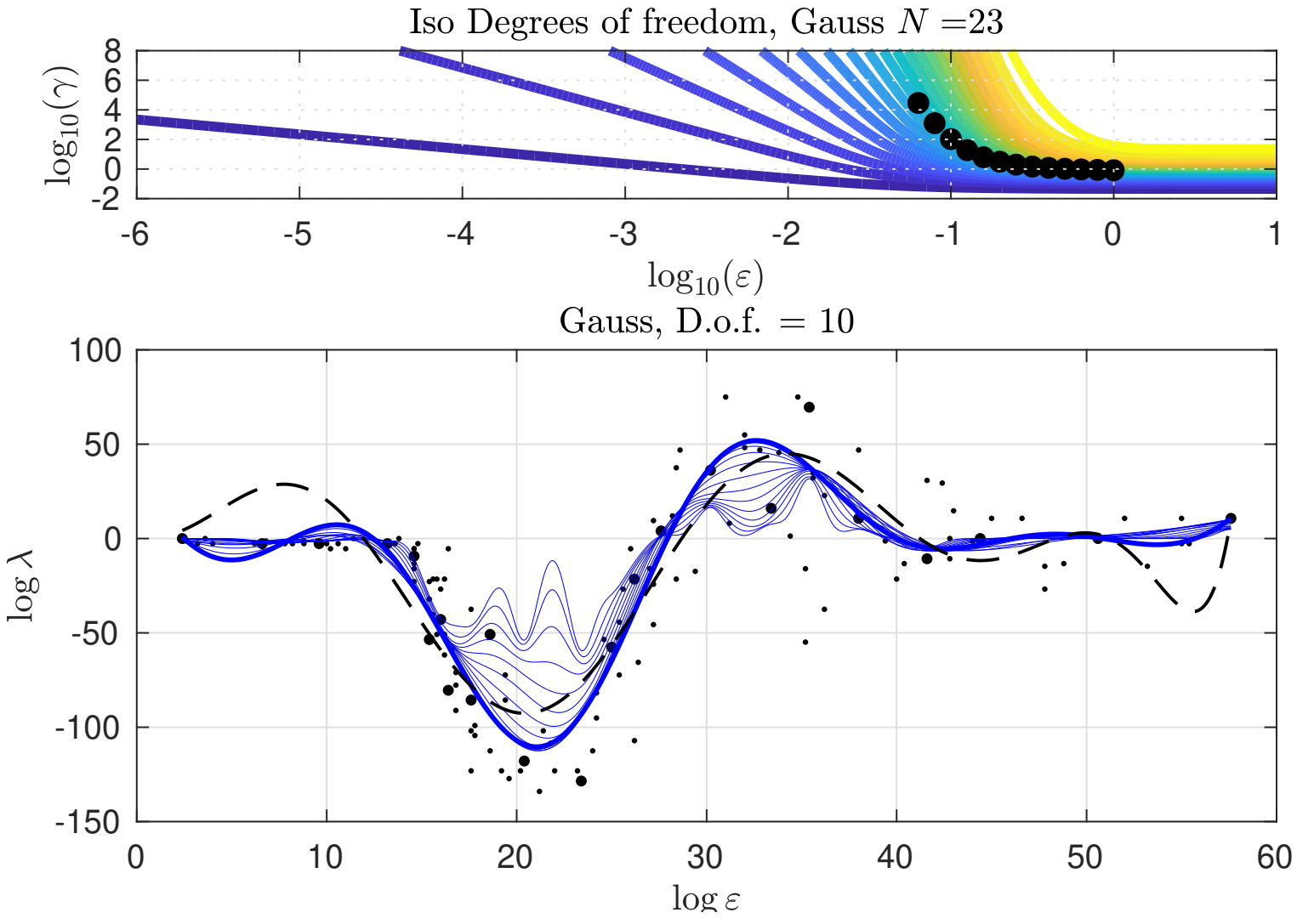
Gaussian kernel



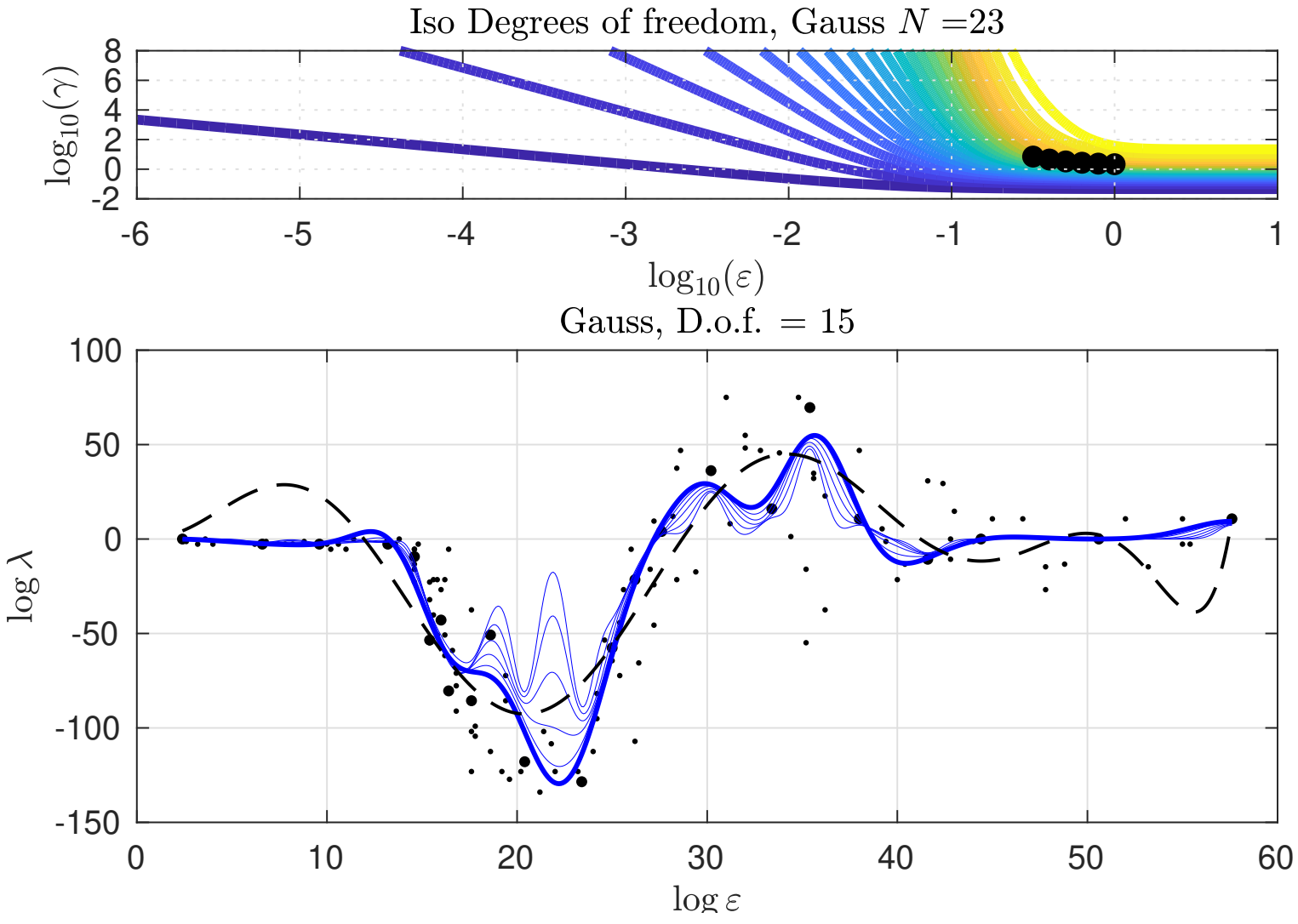
Gaussian kernel



Gaussian kernel



Gaussian kernel



To conclude

$\hookrightarrow$ Deep links between GP & splines/orthogonal polynomials in the flat limit

$\hookrightarrow$ Flat limit may be obtain quite rapidly

$\hookrightarrow$ Flat limits attractive (*e.g.* polyharmonic splines) since less hyperparameter to control

$\hookrightarrow$ We've studied the flat limit in DPP's !

More in :

Bathelmé&Usevitch, SIMAX 2020 and our ArXiv papers currently submitted (or to be)