Kernel matrices in the flat limit

Simon Barthelmé
Konstantin Usevich
Nicolas Tremblay
P-O Amblard

23rd September 2022

Studying kernel methods

- Kernel methods (GPs, SVMs, Spectral clustering, etc.) are central to modern computational statistics
- Theoretical studies often focus on large-n asymptotics, relating kernel matrices to linear operators on function spaces
- Here I'll describe a fixed-sample limit that lets us relate kernel methods to (multivariate) polynomials and splines
- Application to GP regression

Reminder: kernel matrices

- We are given a set of n locations (nodes) $x_1 \dots x_n \in \mathbb{R}^d$
- Kernel matrix K is a $n \times n$ matrix with entries:

$$K_{ij} = k(x_i, x_j)$$

- $k(x, y)$ is a positive definite function
- Here we look at *stationary*, RBF kernels, $k(x, y)$ is a function of $\|x - y\|$

Two examples of kernel functions

- The squared-exponential kernel function:

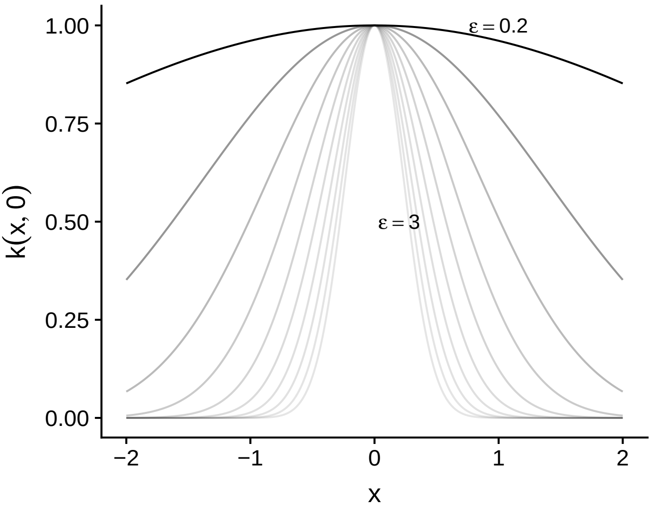
$$k(x, y) = \exp\left(-(\epsilon \|x - y\|)^2\right)$$

- The exponential kernel function (special case of Matern):

$$k(x, y) = \exp\left(-\epsilon \|x - y\|\right)$$

- Note that ϵ plays the role of an inverse scale parameter
- These two kernels turn out to have widely different behaviour in the limit we're interested in!

The flat limit (kernel function)



Wait, does this make any sense?

- As $\epsilon \to 0$, the kernel matrix turns into a constant matrix. Surely this limit is completely trivial?
- It's not, but it's not easy to see why
- One hint is given by Driscoll & Fornberg's ground-breaking result on Radial Basis Function interpolation in the flat limit (2002)

Radial Basis Function interpolation

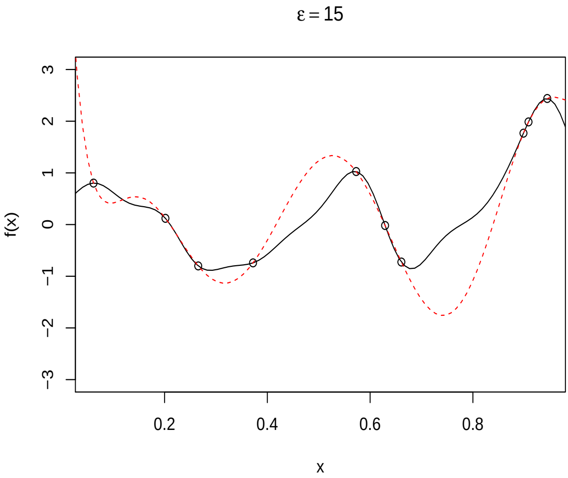
- The goal is to interpolate the value of a function $f(x)$ given measurements at $x_1, \ldots, x_n$.
- In RBF interpolation the interpolant $\tilde{f}$ is constructed from a kernel function:

$$\tilde{f}(x) = \sum_{i=1}^{n} \alpha_i k(x_i, x)$$

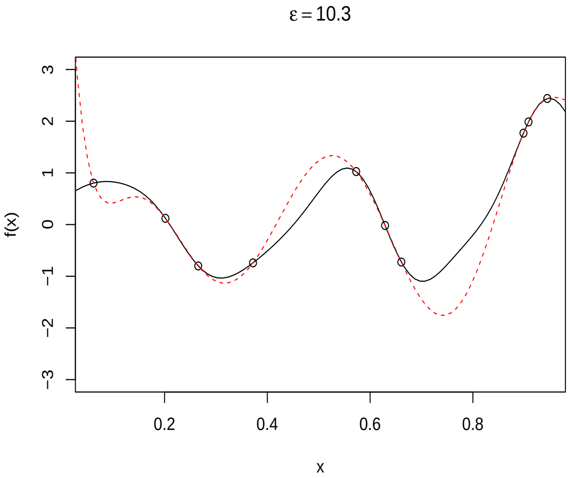
- RBF interpolation = noiseless limit of GP regression
- In polynomial interpolation,

$$\tilde{f}(x) = \sum_{j=0}^{n-1} \beta_j x^j$$

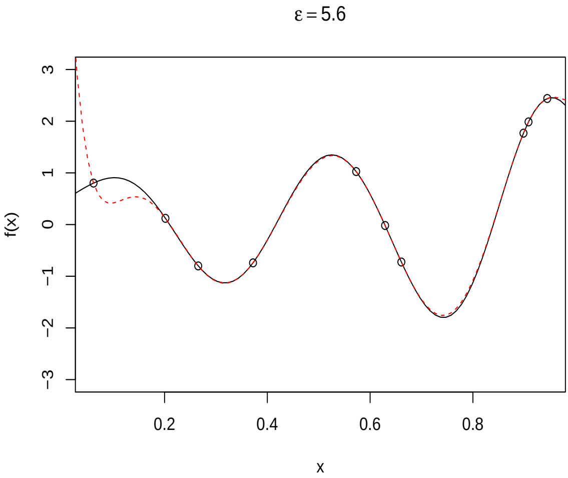
RBF interpolation in the flat limit



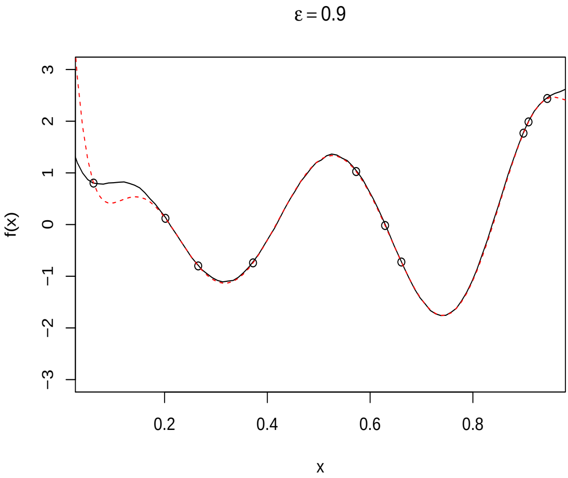
RBF interpolation in the flat limit



RBF interpolation in the flat limit



RBF interpolation in the flat limit



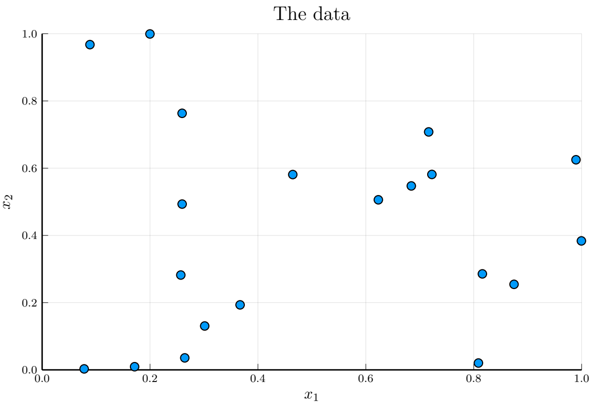
RBF interpolation in the flat limit

- In red: a polynomial interpolant of degree $n-1$
- Driscoll & Fornberg showed that the Gaussian RBF interpolant tends to the polynomial interpolant as $\epsilon \to 0$
- This result should be very surprising: the basis functions become flat, but the interpolant stays well-defined in the limit
- Consequently: (a) there's something non-trivial going on with the flat limit, and (b) it's something to do with polynomials

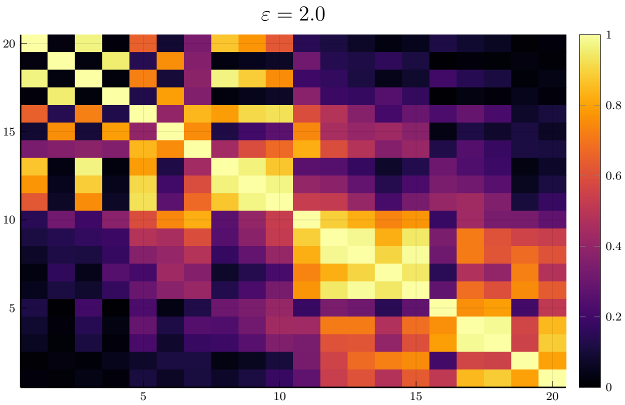
Hint 2: Empirical behaviour of eigenvalues

- Another hint that something interesting is going on comes from empirical observations
- In the next few slides we will see the typical behaviour of the eigenvalues of a kernel matrix (with Gaussian kernel)
- Kernel matrix for 20 points, drawn randomly from unit square

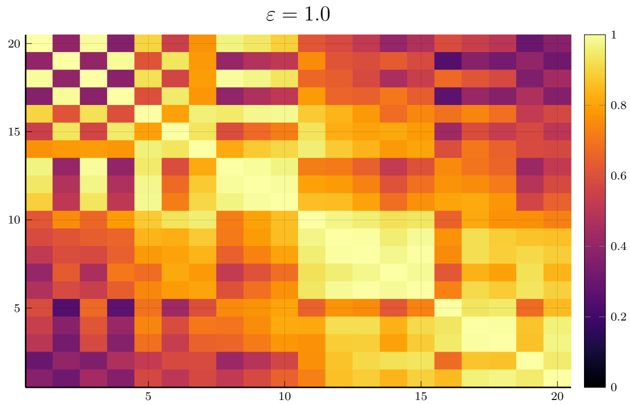
Empirical behaviour of eigenvalues (Gaussian kernel)



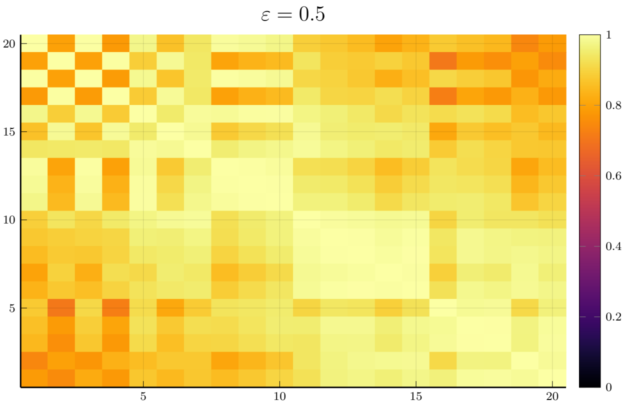
Empirical behaviour of eigenvalues (Gaussian kernel)



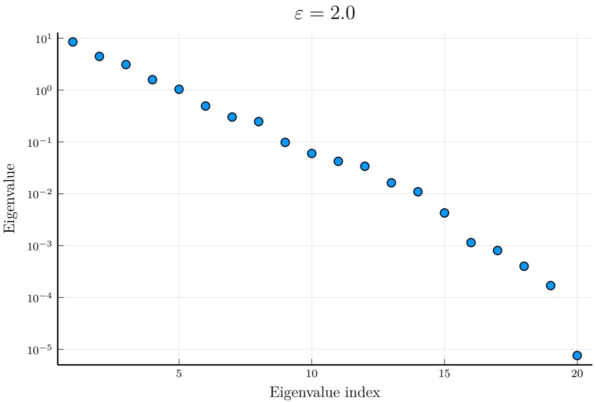
Empirical behaviour of eigenvalues (Gaussian kernel)



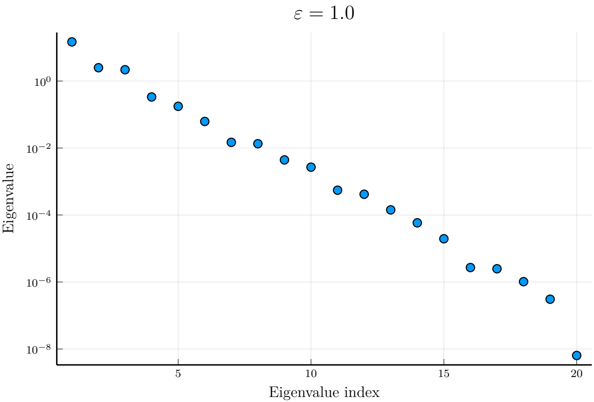
Empirical behaviour of eigenvalues (Gaussian kernel)



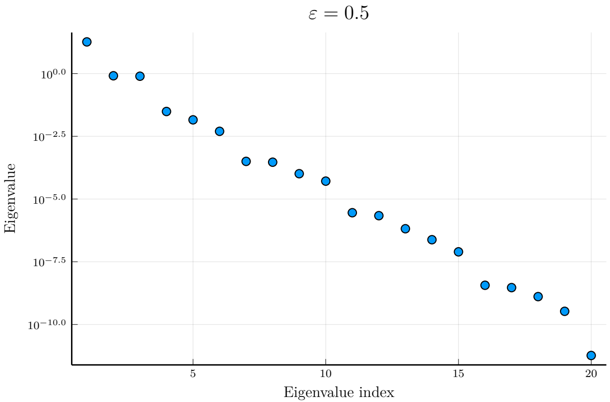
Empirical behaviour of eigenvalues (Gaussian kernel)



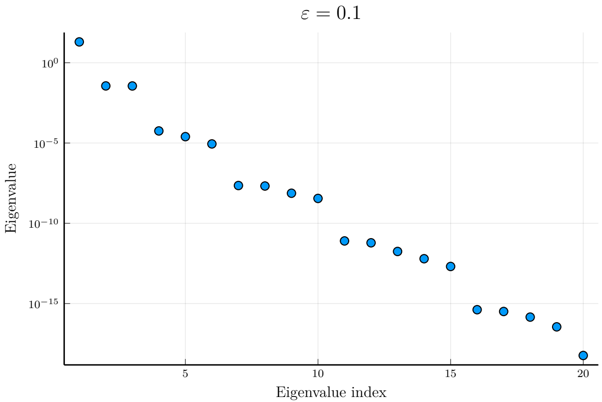
Empirical behaviour of eigenvalues (Gaussian kernel)



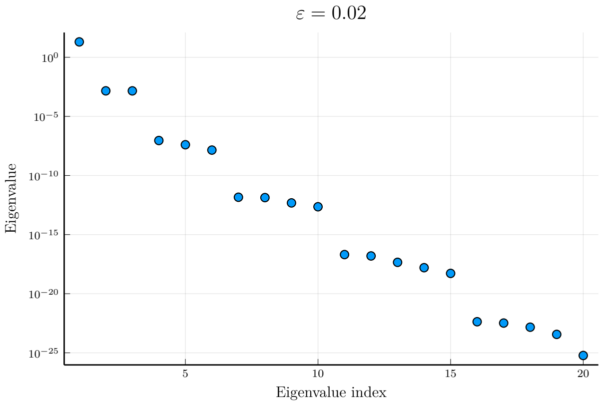
Empirical behaviour of eigenvalues (Gaussian kernel)



Empirical behaviour of eigenvalues (Gaussian kernel)



Empirical behaviour of eigenvalues (Gaussian kernel)



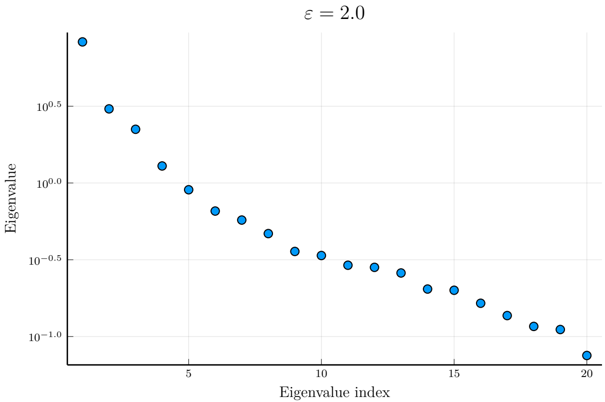
The flat limit

- The "slanted staircase" pattern appears because the first block of eigenvalues has order $\mathcal{O}(1)$, the second $\mathcal{O}(\varepsilon^2)$, the third $\mathcal{O}(\varepsilon^4)$, etc.
- First proved (in passing) by Schaback (2005)
- Hints at strong structure in the spectral behaviour of kernel matrices in the flat limit...

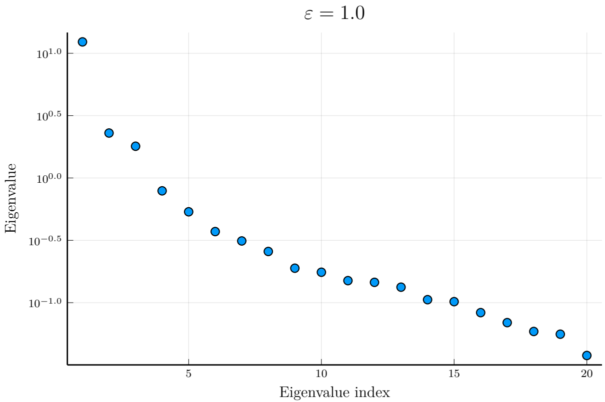
Objective

- Our objective was to characterise the eigenvectors and eigenvalues of kernel matrices in the flat limit
- We also have determinants, regularised inverses, and some other things but eigenvectors and values are the most helpful when trying to understand the behaviour of kernel methods
- Our results cover both infinitely smooth kernels (e.g. Gaussian) but also finitely smooth kernels (nearly everything else)
- Behaviour is quite different!

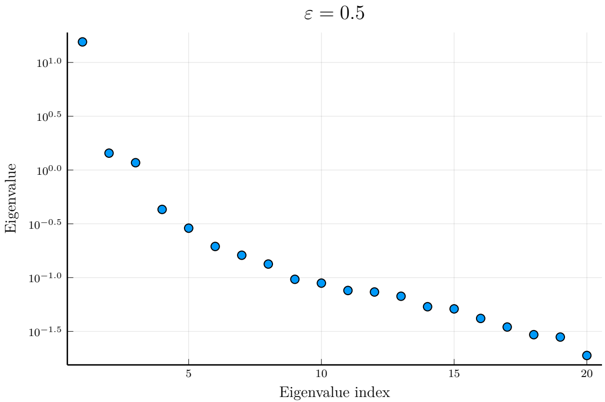
Empirical behaviour of eigenvalues (Exp. kernel)



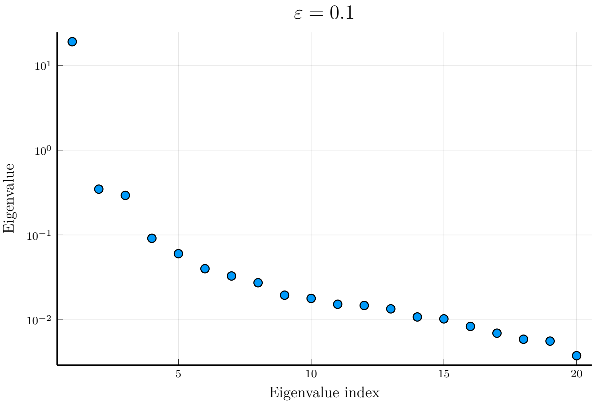
Empirical behaviour of eigenvalues (Exp. kernel)



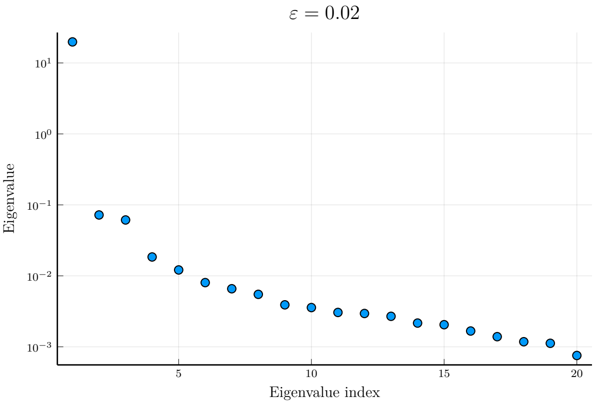
Empirical behaviour of eigenvalues (Exp. kernel)



Empirical behaviour of eigenvalues (Exp. kernel)



Empirical behaviour of eigenvalues (Exp. kernel)



The flat limit (Exp. kernel)

- In the case of the exponential kernel, there are only two blocks of eigenvalues.
- One is of order $\mathcal{O}(1)$, the second $\mathcal{O}(\varepsilon)$
- Why such different behaviours?
- Has to do with *regularity* of kernel function
- Note that $\exp(-\varepsilon \|x - y\|)$ is not differentiable at $x = y$.

Kernel regularity is what matters

- The following kernel has three distinct blocks:

$$k(x, y) = (1 + \varepsilon \|x - y\|) \exp(-\varepsilon \ \|x - y\|)$$

- Can also give you kernels with four blocks, five, etc.
- Our results show: how many blocks = 1 + how many times kernel differentiable at $x = y$
- It's the most important parameter that distinguishes kernels in the flat limit!

Main result: eigenvalues/vectors of kernel matrices in the flat limit

- Our main result is an asymptotic expansion of the eigenvalues and eigenprojectors of $\mathrm{K}(\epsilon)$ as $\epsilon \to 0$
- Some subtleties and corner cases, will try to keep it simple
- *From now on all results are in $d = 1$.* $d > 1$ is similar but needs a lot more notation.

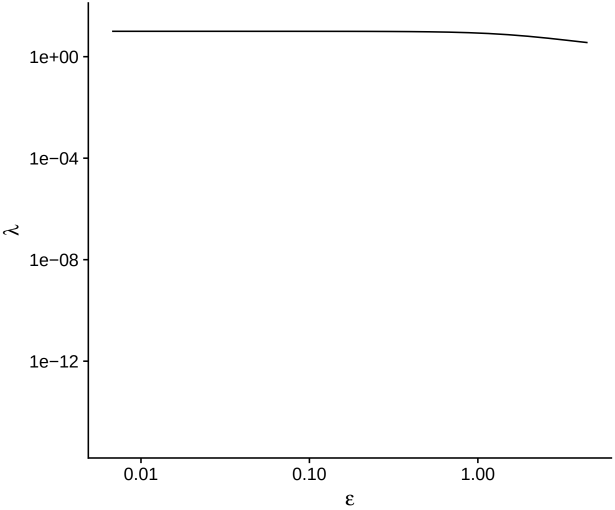
Main result: eigenvalues, completely smooth case

- Infinitely smooth kernel, e.g. squared-exponential, $d = 1$
- Each eigenvalue $\lambda_i(\varepsilon)$ can be written

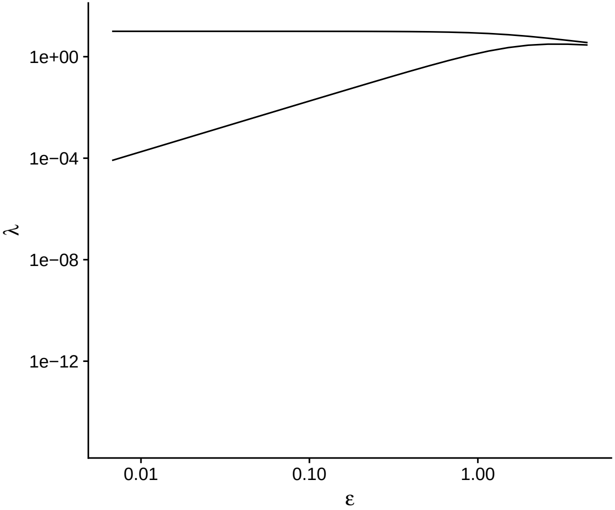
$$\lambda_i(\varepsilon) = \varepsilon^{2(i-1)}(\tilde{\lambda}_i + \mathcal{O}(\varepsilon))$$

- This means $\lambda_1(\varepsilon) = \mathcal{O}(1), \lambda_2(\varepsilon) = \mathcal{O}(\varepsilon^2), \lambda_3(\varepsilon) = \mathcal{O}(\varepsilon^6), \ldots$
- In addition we have a simple closed-form expression for $\tilde{\lambda}_i$

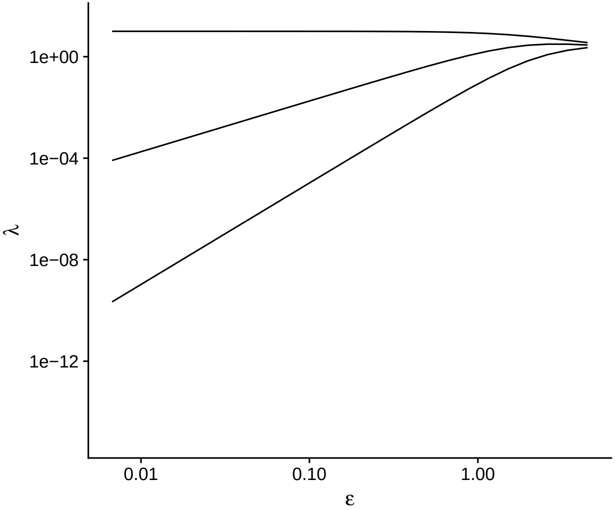
Main result: eigenvalues



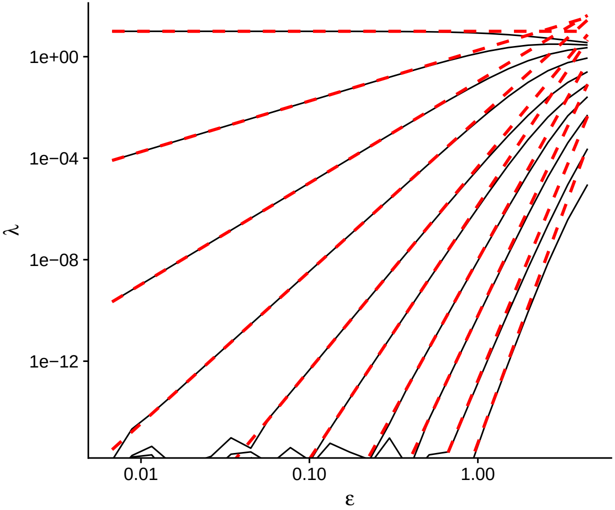
Main result: eigenvalues



Main result: eigenvalues



Main result: eigenvalues



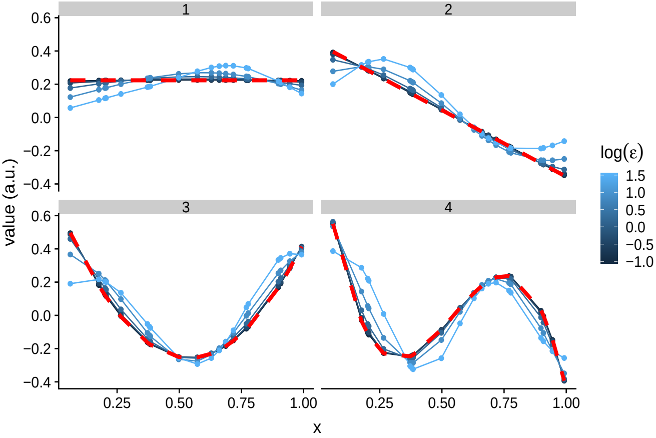
Interpretation

- As $\epsilon \to 0$, every eigenvalue goes to 0 except for the top one.
- λ_{j+1} goes to 0 faster than λ_j, so that every *eigengap* increases

Main result: eigenvectors

- If the kernel is completely smooth, the eigenvectors are *orthogonal polynomials*
- Specifically: let $\boldsymbol{v}_j = \begin{pmatrix} x_1^j \\ x_2^j \\ \vdots \\ x_n^j \end{pmatrix}$
- Apply the Gram-Schmidt process to $\boldsymbol{v}_0, \boldsymbol{v}_1, \ldots$ to get $\boldsymbol{q}_0, \boldsymbol{q}_1, \ldots$
- Then the j-th eigenvector of K_ε goes to $\boldsymbol{q}_{j-1}$ as $\varepsilon \to 0$.

Main result: eigenvectors



The finite-smoothness case

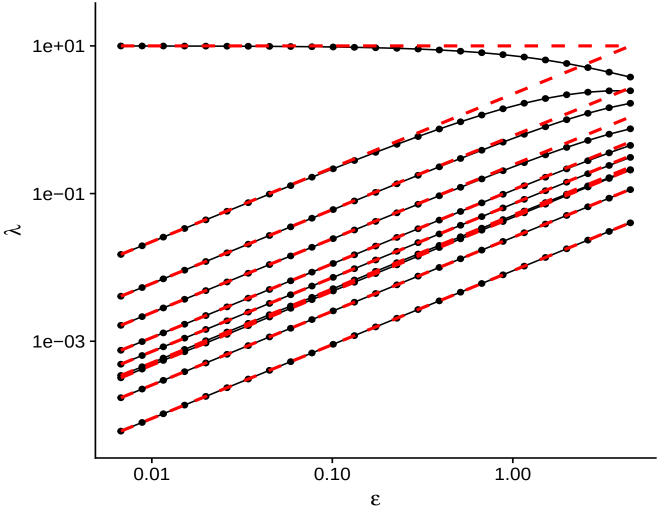
Whole story too long to tell! Brief summary for the *exponential* kernel.

- There is one eigenvalue of constant order, and $n-1$ eigenvalues of order ϵ.
- The limiting eigenvectors are the constant vector (for λ_1), and the $n-1$ non-null eigenvectors of

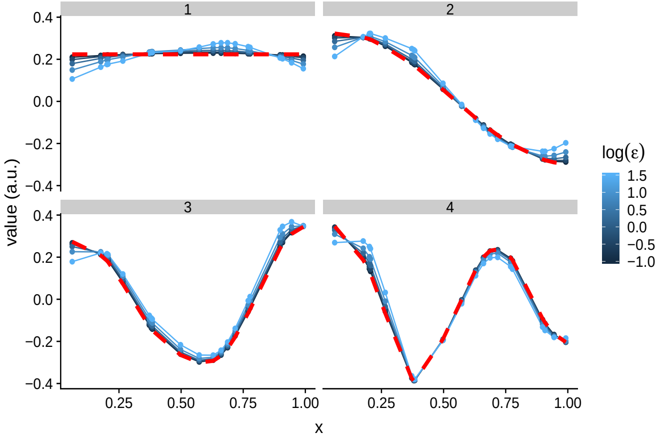
$$(\mathsf{I}-(1/n)\mathbf{1}\mathbf{1}^t)\mathsf{D}_{(1)}(\mathsf{I}-(1/n)\mathbf{1}\mathbf{1}^t)$$

  where $\mathsf{D}_{(1)}(i,j)=|x_i-x_j|$.
- These turn out to be (piecewise linear) splines!

Exponential kernel: eigenvalues



Exponential kernel: eigenvectors



Summary so far

- The eigenvectors and eigenvalues of kernel matrices in the flat limit exhibit interesting patterns, *even though the limit is just a constant matrix*
- Hallmark of a *singular perturbation problem*: the limit of the eigenvectors $\neq$ the eigenvectors of the limit. Makes it interesting but difficult (see Kato's book)
- Limit depends mostly on *regularity* of kernel function: either polynomials or splines appear at different orders
- See Barthelme & Usevich (2020) for complete story
- Now for an application

GP regression/Kernel Ridge Regression

- GP regression = Kernel Ridge Regression (to some extent)
- Pick a kernel function $k(x, y)$, then build a Hilbert space $\mathcal{H}$ of functions with $k(x, y)$ as a reproducing kernel.
- Look for a function in $\mathcal{H}$ for a good fit to the data whose norm isn't too high

$$\hat{f} = \underset{f \in \mathcal{H}}{\operatorname{argmin}} \sum_{i=1}^{n} (y_i - f(x_i))^2 + \gamma^{-1} \|f\|_{\mathcal{H}}^2 \tag{1}$$

- Here $\|f\|_{\mathcal{H}}^2$ is the norm in the RKHS induced by the kernel function $k(x, y)$

GP regression/Kernel Ridge Regression

- (Representer theorem, Schölkopf & Smola 2002) Solution is just:

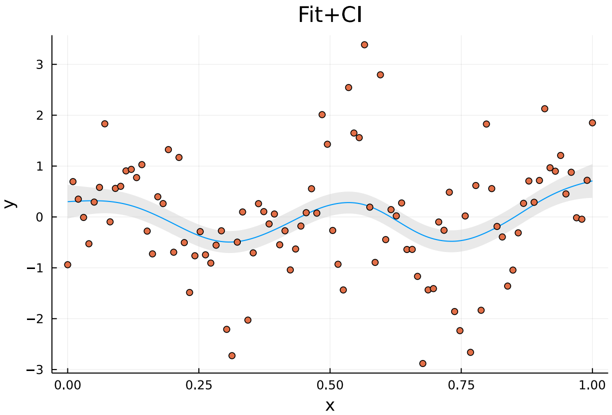
$$\hat{f}(x) = \sum_{i=1}^{n} \alpha_i k(x, x_i)$$

- The weights are just

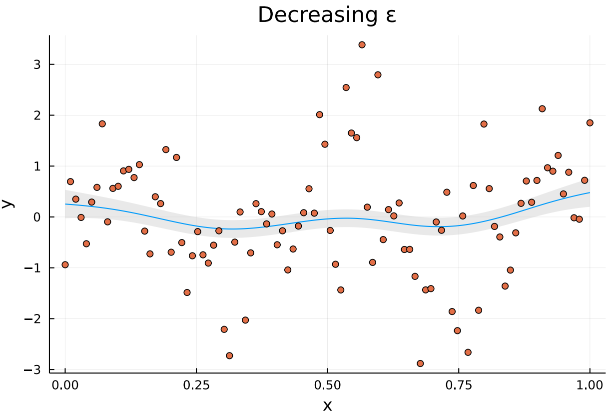
$$\boldsymbol{\alpha} = (\mathsf{K} + \gamma^{-1}\mathsf{I})^{-1}\mathsf{y}$$

  where K is the kernel matrix $K_{ij} = k(x_i, x_j)$.
- Notice γ controls regularisation here: another hyperparameter to set.

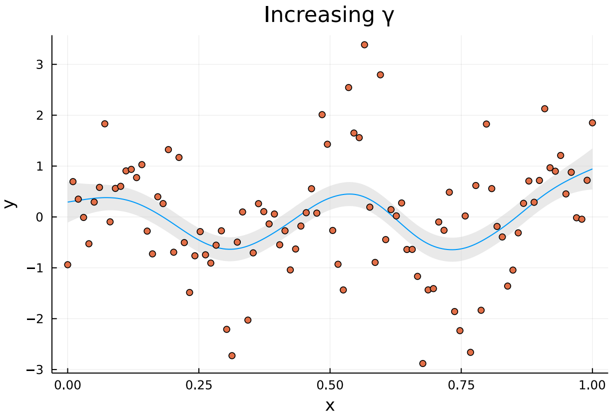
GP regression in a nutshell



GP regression in a nutshell



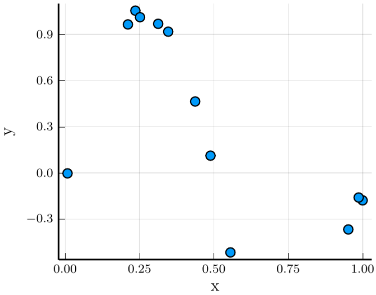
GP regression in a nutshell

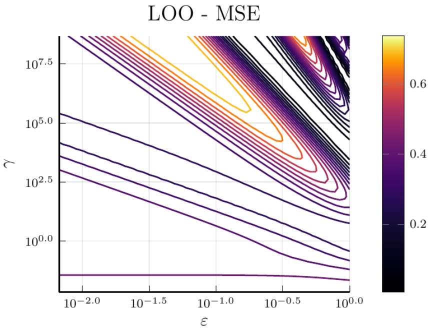


Handling hyperparameters

- There are (at least) two hyperparameters in a typical GP regression problem
- ε sets “width” of the kernel function
- γ sets the amount of regularisation: high γ, low regularisation
- Typically γ and ε are picked using a hyperparameter selection technique (e.g. leave-one-out)

Cross-validation example





Taking the flat limit the correct way

- Leaving γ fixed as $\varepsilon \to 0$ gives a trivial limit: the limiting GP fit is flat.
- As we lower ε, we need to increase γ
- Take the limit along the contours of the hyperparameter selection criterion

Informal result

- Formal result requires a lot of notation; sorry!
- Informally: in the flat limit, a GP model with inf-smooth kernel behaves like a polynomial regression with the same d.f.
- Both the fit (pointwise predictive means) *and* variances converge

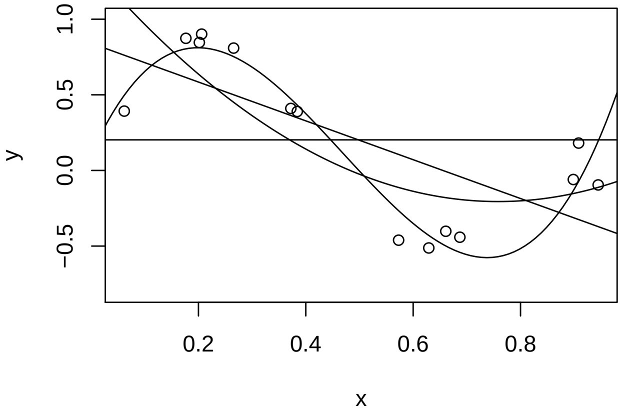
Polynomial regression

- Polynomial regression is just the following problem:

$$\tilde{f}_m = \operatorname*{argmin}_{f \in \mathcal{P}_m} \sum_{i=1}^{n} (y_i - f(x_i))^2$$

  where $\mathcal{P}_m$ is the space of polynomial functions of degree m.
- Single hyperparameter: degree m. Higher m, wigglier fit.
- NB: polynomial interpolation appears as a special case when there are sufficient degrees of freedom ($m = n - 1$).

GP regression → polynomial regression



The finitely-smooth case

- Convergence to poly. regression holds for sufficiently smooth kernels like the Gaussian.
- For exponential (and other Màtern) kernels the story is a bit more complicated
- Roughly: depending on the degrees of freedom and the kernel, the GP model converges to either a polynomial regression, or to a polyharmonic splines.
- Polyharmonic splines = multivariate extension of univariate splines, introduced by Duchon (1977)

Conclusion

- In the flat limit, GP regression tends to a polynomial regression or polyharmonic spline regression depending on smoothness of kernel
- Can also derive flat limit behaviour of other kernel methods, like DPPs (Barthelme et al, 2022), or kernel independence tests (Amblard et al, forthcoming)
- More at arxiv:2201.01074

Conclusion (practical aspects)

- As a practical approximation, the flat limit works sometimes very well and sometimes poorly: it seems to depend on geometry, in a way we don't understand yet.
- It always applies to diagonal blocks in kernel matrices for points that are close together: should be able to use far-field approximations for off-diagonal blocks, we are investigating that.
- It makes ε-free models like polyharmonic splines quite attractive: they occur as a limit of standard GPs, but there's one fewer hyperparameter to worry about!
- NB: Polyharmonic (AKA Duchon) splines are already implemented in *mgcv*, a standard package for GAM fitting by Simon Wood.

References (1)

- Original work on eigenvalues/eigenvectors is in: Barthelmé, S., & Usevich, K. (2021). Spectral properties of kernel matrices in the flat limit. SIAM Journal on Matrix Analysis and Applications, 42(1), 17-57.
- We have applied the results to DPPs, GPs, and kernel independence tests in subsequent papers.

References (2)