Kernel generalization error under isotropic distribution

Week 1

1 Setting

Let us consider a kernel $K(\vec{x}, \vec{y})$ where $\vec{x} \in \mathbb{R}^d$. We will only consider inner-product kernels (or dot-product kernels):

$$K(\vec{x}, \vec{y}) = h(\langle \vec{x}, \vec{y} \rangle)$$

where $h \in \mathcal{C}^\infty$. This way, we should be able to express h as a power series $h(t) = \sum_m^\infty h_m t^m$. [1]

When studying the generalization error of a kernel machine, we are interested in identifying the spectrum of the operator T:

$$Tf(x) = \int d^d\vec{y}\, p(\vec{y}) K(\vec{x}, \vec{y}) f(\vec{y})$$

where $p(\vec{x})$ is the probability distribution of the data. In this first report, we are going to consider a gaussian isotropic case where $p = \mathcal{N}(0, \mathbb{I}_d)$

The operator T is linear, symmetric and positive definite (!), $T : L_2(p) \to L_2(p)$ and can be diagonalized with positive eigenvalues:

$$\int d^d\vec{y}\, p(\vec{y}) K(\vec{x}, \vec{y}) f_\beta(\vec{y}) = \lambda_\beta f_\beta(\vec{x})$$

By Mercer's theorem, this is equivalent to claim that the kernel can be decomposed through:

$$K(\vec{x}, \vec{y}) = \sum_i \lambda_i \varphi_i(\vec{x}) \varphi_i(\vec{y})$$

where φ_β are an orthonormal basis of the $L_2(p)$ space. In this specific case, they can be represented through Hermite polynomials (they are orthonormal with respect to the gaussian measure)

In general, one can show the eigenvalues can be indexed through a multi-index $\beta = (\beta_1, \beta_2, \ldots, \beta_d)$ (TO DO) yielding:

$$K(\vec{x}, \vec{y}) = \sum_\beta \lambda_\beta He_\beta(\vec{x}) He_\beta(\vec{y})$$

where

$$He_\beta(\vec{x}) = \prod_i He_{\beta_i}(x_i)$$

and $He_{\beta_i}(x_i)$ is just the hermite polynomial of order β_i in the variable x_i.

Using an argument of symmetry (the gaussian isotropic measure is invariant under rotation and the same goes for dot-product kernel), we should get a Mercer decomposition of the kind:

$$K(\vec{x}, \vec{y}) = h(\langle \vec{x}, \vec{y} \rangle) = \sum_m^\infty \xi_m \sum_{|\beta|=m} He_\beta(\vec{x}) He_\beta(\vec{y})$$

[1] To be done in the final document: explain why

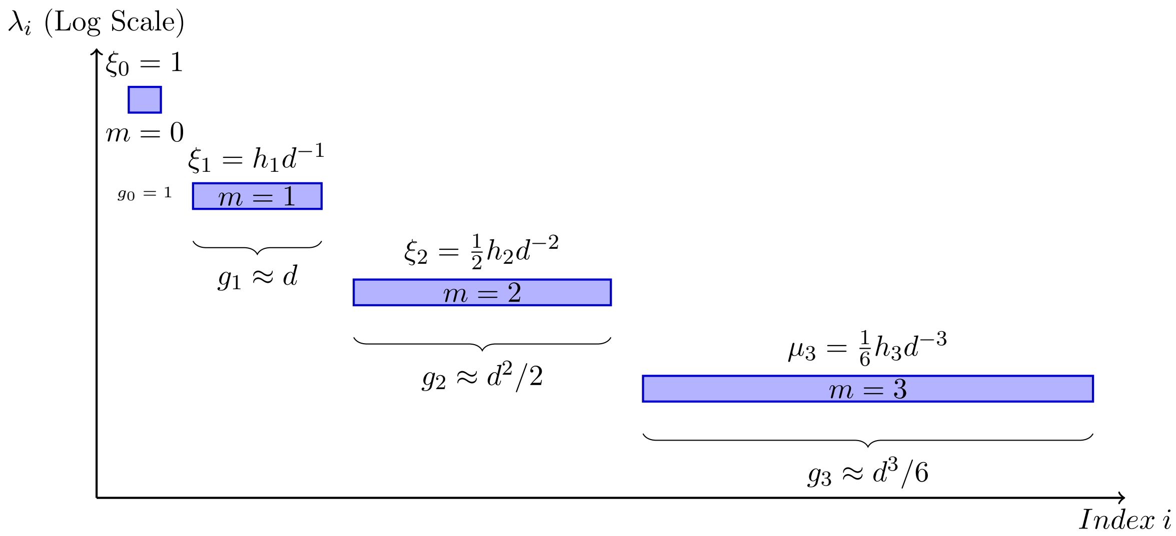

Figure 1: *Kernel spectrum under isotropic gaussian distribution*

The full derivation for the derivation of ξ_m can be found in [Paper 2, Appendix A.1]:

$$\xi_m = h_m m! d^{-m}$$

the degeneracies of a level m is given by the possible ways of arranging $(\beta_1, \beta_2, \ldots, \beta_d)$ such that $|\beta| = \beta_1 + \beta_2 + \cdots + \beta_d = m$. It turns out this is equivalent to:

$$g_m = \binom{d+m-1}{m}$$

when $d \gg 1$, $g_m = \frac{1}{m!} d^m$ (Fig. 1). Note that this is just the number of possible ways to build a vector $\beta = (\beta_1, \ldots, \beta_d)$ such that $|\beta| = m$.

Numerical plots To estimate numerically the spectrum of a general kernel operator:

$$Tf = \int dx' p(x') K(x, x') f(x') = \mathbb{E}[K(x, x') f(x')]_{x' \sim p}$$

we can proceed as follows. Imagine we sample from $p(x)$ and obtain a set of IID values $\{x_1, x_2, x_3, \ldots, x_M\}$. We can then approximate the distribution $p(x)$ as:

$$p(x) \approx \frac{1}{M} \sum_i^M \delta(x' - x_i)$$

Hence:

$$(Tf)(x) \approx \frac{1}{M} \sum_i^M \int dx' \delta(x' - x_i) K(x, x') f(x') = \sum_i^M \frac{1}{M} K(x, x_i) f(x_i)$$

The eigenfunctions $g_k(x)$ with eigenvalue λ_k is such that, $\forall x_j$:

$$(Tg_k)(x_j) = \lambda_k g_k(x_j)$$

Combining:

$$(Tg_k)(x) = \lambda_k g_k(x) \approx \sum_{i=0}^M \frac{1}{M} K(x, x_i) g_k(x_i)$$

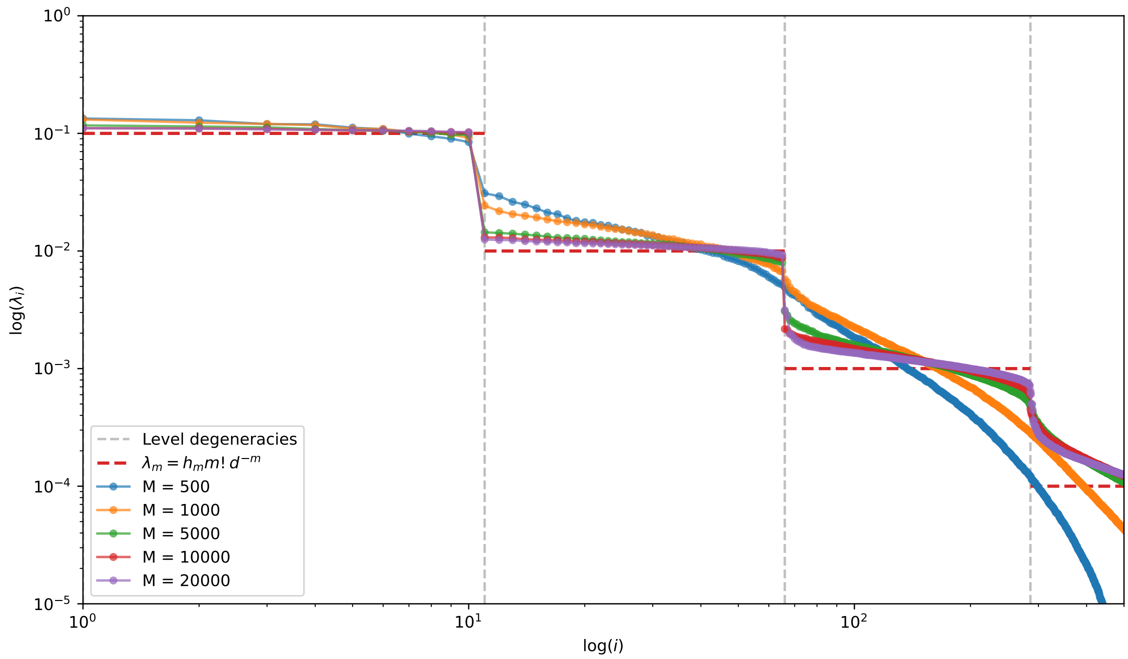

Figure 2: *The red plateau corresponds to the theoretical behavior of the spectrum (here* $d = 10$*). As* M *grows, the curve gets closer and closer to the expected behavior*

If I let $\vec{v} = (g_k(x_1), g_k(x_2), \dots)$, then:

$$\lambda_k v_j = \frac{1}{M} \sum_i K(x_k, x_i) v_i$$

in vectorial form:

$$\lambda_k \vec{v} = \frac{1}{M} K \vec{v}$$

hence I can estimate the eigenvalues λ_k of the operator once I have built the empirical matrix $\frac{1}{M} K(x_i, x_j)$, provided $x_i \sim p(x)$. This is a Montecarlo method essentially (and reproduces the correct spectrum when $M \to \infty$, Fig. 2).

Still, I'm not really convinced this is the best algorithm to perform such a numerical task. As typical in a Monte Carlo approach, the convergence to the target distribution is dictated by $\sqrt{M}$, which is quite slow. A matrix $M \times M$ where $M = 10^5$ has 10^{10} elements, occupying 40 GB in the RAM, so this procedure becomes unfeasible quite soon. I should check and see if there are better algorithms.

2 Kernel error, ridgeless regime $\lambda = 0$

Let's compute the generalization error for a kernel ridge regression problem in the ridgless regime. From Paper1, we know that the kernel error is given once we compute the quantity ν defined through the self-consistency equation:

$$n - \frac{\lambda}{\nu} = Tr(\Sigma(\Sigma + \mathbb{I}\nu)^{-1}) \equiv df(\nu) \tag{1}$$

where Σ is the diagonal matrix of the eigenvalues of $k(\vec{x}, \vec{y})^2$, $\Sigma = Diag(\lambda_1, \dots, \lambda_\infty)$. In this work, we will use the following scaling:

$$n = \alpha d^\kappa$$

[2]Strictly speaking, the eigenvalues are those associated with the kernel integral operator T

that is, $n \sim O(d^\kappa)$ where κ quantifies the sample complexity and α is a constant (of order 1). We will consider a case where $d, n \to \infty$ while α remains finite.

2.1 Computation for ν

Take Eq.1 and substitute $\lambda = 0$ and $n = \alpha d^\kappa$:

$$\alpha = d^{-\kappa} df(\nu) = d^{-\kappa} \sum_i^\infty \frac{\lambda_i}{\lambda_i + \nu} \tag{2}$$

We know that in the isotropic case the eigenvalues are distributed in degenerate shells (call the degeneration g_m for level m), hence:

$$\alpha = d^{-\kappa} \sum_m^\infty g_m \frac{\xi_m}{\xi_m + \nu} = d^{-\kappa} \sum_m^\infty \frac{d^m}{m!} \frac{h_m m! d^{-m}}{h_m m! d^{-m} + \nu} \tag{3}$$

Let's define $h_m m! = \delta_m$. In all usual cases, $\delta_m \sim O(1)$ with respect to m (for polynomials, it eventually becomes zero. For regular functions like the exponential, $\delta_m = 1, \forall m$). The trick here is to work under the thermodynamic limit dividing the sum in three terms:

$$\begin{aligned}
\alpha &= d^{-\kappa} \sum_m^\infty \frac{d^m}{m!} \frac{\delta_m d^{-m}}{\delta_m d^{-m} + \nu} = \\
&= d^{-\kappa} \left(\underbrace{\sum_{m=0}^{\kappa-1} \frac{d^m}{m!} \frac{\delta_m d^{-m}}{\delta_m d^{-m} + \nu}}_{m<\kappa} + \underbrace{\frac{d^\kappa}{\kappa!} \frac{\delta_\kappa d^{-\kappa}}{\delta_\kappa d^{-\kappa} + \nu}}_{m=\kappa} + \underbrace{\sum_{m=\kappa+1}^{\infty} \frac{d^m}{m!} \frac{\delta_m d^{-m}}{\delta_m d^{-m} + \nu}}_{m>\kappa} \right)
\end{aligned} \tag{4}$$

To make things work, we have to make an ansatz. We will guess:

$$\nu = \xi d^{-\kappa}$$

and verify a posteriori this claim. Let's consider the three terms separately:

- **First term**:

$$d^{-\kappa} \sum_{m=0}^{k-1} \frac{d^m}{m!} \frac{\delta_m d^{-m}}{\delta_m d^{-m} + \xi d^{-\kappa}} \tag{5}$$

  Fix a value for $0 \leq m \leq k-1$. Then the eigenvalue term:

$$\frac{\delta_m d^{-m}}{\delta_m d^{-m} + \xi d^{-\kappa}} = \frac{1}{1 + \frac{\xi}{\delta_m} d^{m-k}} \xrightarrow{d \gg 1, m<k} 1 \tag{6}$$

  Since ξ is independent of d by ansatz construction and δ_m is of order 1 with respect to m. The first term then becomes:

$$d^{-\kappa} \sum_{m=0}^{\kappa-1} \frac{d^m}{m!} = \sum_{m=0}^{\kappa-1} \frac{d^{m-\kappa}}{m!} \xrightarrow{d \gg 1} 0 \tag{7}$$

- **Central term**: This is easy, as it evaluates to something of order 1:

$$d^{-\kappa} \frac{d^\kappa}{\kappa!} \frac{\delta_\kappa d^{-\kappa}}{\delta_\kappa d^{-\kappa} + \xi d^{-\kappa}} = \frac{1}{\kappa!} \frac{\delta_k}{\delta_k + \xi} \tag{8}$$

- **Third term**: Again, we write

$$d^{-\kappa}\sum_{m=k+1}^{\infty}\frac{d^m}{m!}\frac{\delta_m d^{-m}}{\delta_m d^{-m}+\xi d^{-\kappa}} \tag{9}$$

Fix a value for $m > \kappa$. Then the eigenvalue term:

$$\frac{\delta_m d^{-m}}{\delta_m d^{-m}+\xi d^{-\kappa}} = \frac{1}{1+\frac{\xi}{\delta_m}d^{m-k}} \approx \frac{\delta_m}{\xi}d^{\kappa-m} \text{ when } d \gg 1 \tag{10}$$

This term alone converges to 0 but it needs to be considered inside the summation:

$$d^{-\kappa}\sum_{m=k+1}^{\infty}\frac{d^m}{m!}\frac{\delta_m d^{-m}}{\delta_m d^{-m}+\xi d^{-\kappa}} \approx d^{-\kappa}\sum_{m=k+1}^{\infty}\frac{d^m}{m!}\frac{\delta_m}{\xi}d^{\kappa-m} = \sum_{m=k+1}^{\infty}\frac{\delta_m}{\xi m!} \tag{11}$$

We will define:

$$l(\kappa) = \sum_{m=k+1}^{\infty}\frac{\delta_m}{m!}$$

and the third term can be rewritten as $\frac{l(\kappa)}{\xi}$[3]

Putting all of this together, we get:

$$\alpha = \frac{1}{\kappa!}\frac{\delta_k}{\delta_k+\xi}+\frac{l(\kappa)}{\xi} \tag{12}$$

Let's solve for ξ:

$$\xi = \xi(\alpha,\kappa) = \frac{1}{2\alpha}\left[\left(\frac{1}{\kappa!}\delta_\kappa + l(\kappa) - \alpha\delta_\kappa\right) + \sqrt{\left(\frac{1}{\kappa!}\delta_\kappa + l(\kappa) - \alpha\delta_\kappa\right)^2 + 4\alpha l(\kappa)\delta_\kappa}\right] \tag{13}$$

which is independent of d, confirming our scaling ansatz. Finally, one has:

$$\nu = \xi(\alpha,\kappa)d^{-\kappa} \tag{14}$$

Eq. 14 clearly shows that ξ does not depend on d, hence our ansatz is justified. Fig. 3 is solid proof that our Eq.14 is the correct one

2.2 Some examples of kernels

When using an exponential kernel, one has $\delta_\kappa = 1, \forall\kappa$ (note that δ_κ is just the derivative of the h function evaluated at 0.)

What happens when we use a polynomial kernel of degree κ_0? This essentialy means $\delta_m = 0\ \forall m \geq \kappa_0$. There can be two cases:

- When $\kappa < \kappa_0$, then $\delta_\kappa \neq 0$ and $l(\kappa) \neq 0$, hence we can safely use Eq.14
- When $\kappa > \kappa_0$, then $\delta_\kappa = 0$ and $l(\kappa) = 0$. This implies $\nu = 0$. This is however problematic Ancora WIP. In teoria $\nu = 0$ non è accettabile perché rende la varianza negativa. A quello che ho capito qui è sbagliata l'ansatz, to be studied

[3] If δ_m is of order 1, then this function should always converge, approfondisci meglio

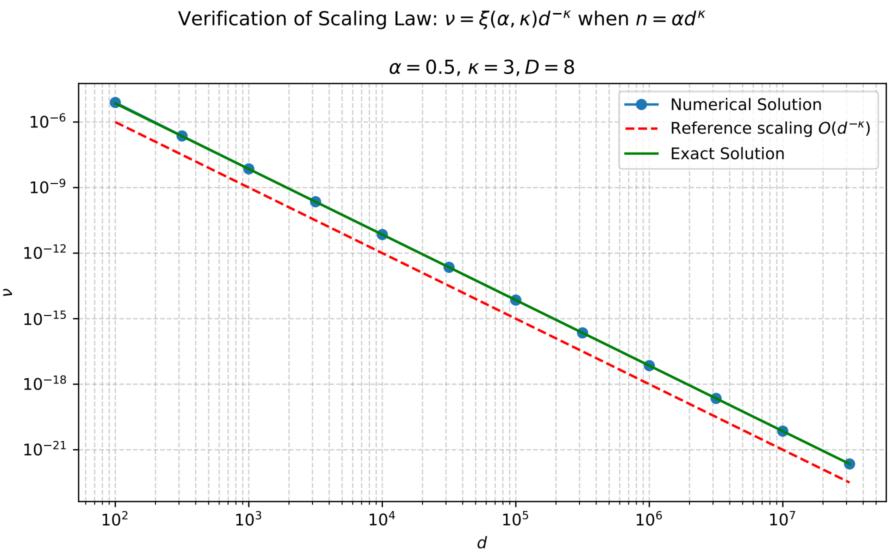

Figure 3: *Behavior of ν against dimension d. The blue dots are obtained through numerical solutions of Eq.1 using the fixed point method to solve the self-consistency equation. The green continuous line represents the curve of Eq.14. The two coincides quite well already when $d \sim$* 1000. *N.B. Quando d cresce, ν va a zero. Forse è meglio se uso $\log \nu$ per evitare valori piccoli*

2.3 Computation for $Var(\alpha, \kappa)$

Again, from Paper 1 (but also Cheng, Montanari) we have:

$$Var(\alpha, \beta, \lambda = 0) = V = \sigma_\epsilon^2 \frac{Tr(\Sigma^2(\Sigma+\nu)^{-2})}{n - Tr(\Sigma^2(\Sigma+\nu)^{-2})} = \frac{\frac{1}{n}Tr(\Sigma^2(\Sigma+\nu)^{-2})}{1 - \frac{1}{n}Tr(\Sigma^2(\Sigma+\nu)^{-2})} = \frac{\tau}{1-\tau} \tag{15}$$

Let's compute τ, the procedure is similar to the one we have seen above.

$$\tau = \frac{1}{n}Tr(\Sigma^2(\Sigma+\nu)^{-2}) = \frac{1}{n}\sum_m g_m \frac{\xi_m^2}{(\xi_m+\nu)^2} \tag{16}$$

As usual, we divide the summation in three pieces and let $n = \alpha d^\kappa$.

$$\tau = \alpha^{-1} d^{-\kappa} \sum_{m=0}^{k-1} g_m \frac{\xi_m^2}{(\xi_m+\nu)^2} + \alpha^{-1} d^{\kappa} g_\kappa \frac{\xi_\kappa^2}{(\xi_\kappa+\nu)^2} + \alpha^{-1} d^{\kappa} \sum_{m=k+1}^{\infty} g_m \frac{\xi_m^2}{(\xi_m+\nu)^2} \tag{17}$$

- **First term**:

$$\begin{aligned}
\alpha^{-1} d^{-\kappa} \sum_{m=0}^{k-1} g_m \frac{\xi_m^2}{(\xi_m+\nu)^2} &= \alpha^{-1} d^{-\kappa} \sum_{m=0}^{k-1} \frac{1}{m!} d^m \frac{\delta_m^2 d^{-2m}}{\xi^2 d^{-2\kappa} + \delta_m^2 d^{-2m} + \delta_m \xi d^{-m-k}} = \\
&= \alpha^{-1} \sum_{m=0}^{k-1} \frac{1}{m!} \frac{\delta_m^2 d^{-\kappa-m}}{\xi^2 d^{-2\kappa} + \delta_m^2 d^{-2m} + \delta_m \xi d^{-m-k}} = \\
&= \alpha^{-1} \sum_{m=0}^{k-1} \frac{1}{m!} \frac{\delta_m^2}{\xi^2 d^{m-\kappa} + \delta_m^2 d^{-m+\kappa} + \delta_m \xi} \underset{d \gg 1}{\approx} \\
&\underset{d \gg 1}{\approx} \sum_{m=0}^{\kappa-1} \frac{1}{\alpha m!} d^{\kappa-m} \xrightarrow{d \gg 1} 0
\end{aligned} \tag{18}$$

Since, again, the terms δ_m, ξ are assumed to be $O(1)$ with respect to d.

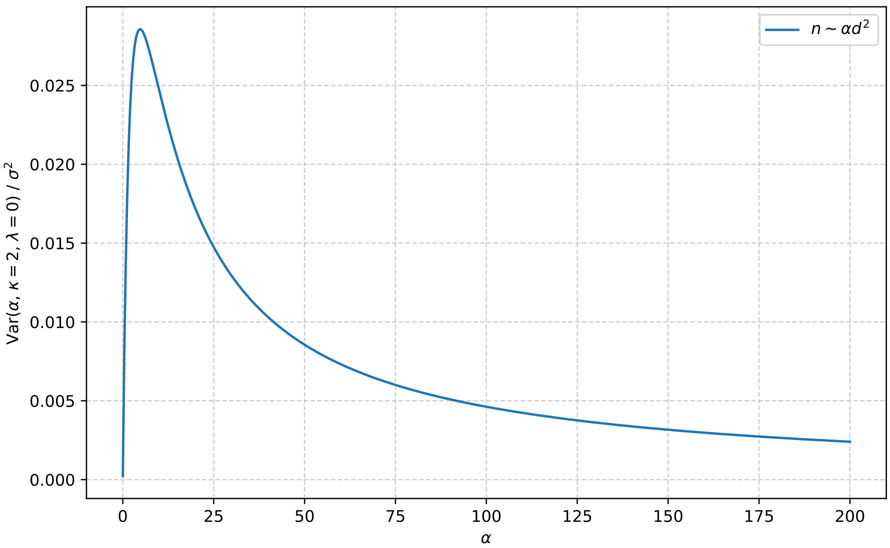

Figure 4: *Variance curve obtained with Eq.22 when using a kernel polynomial of degree* $\kappa_0 = 9$ *and a quadratic scaling regime* $n = \alpha d^2$ *for* d *large. The behavior of the curve is reminiscent of the double descent phenomenon well documented in the machine learning literature. Note that the variance converges to* 0 *when* α *is large (as it should be).*

- **Central term**

$$\alpha^{-1} d^{-\kappa} \frac{1}{\kappa!} d^{\kappa} \frac{\delta_\kappa^2}{(\delta_\kappa + \xi)^2} = \frac{1}{\alpha \kappa!} \frac{\delta_\kappa^2}{(\delta_\kappa + \xi)^2} \tag{19}$$

- **Third term**

$$\begin{aligned} \alpha^{-1} d^{-\kappa} \sum_{m=\kappa+1}^{\infty} g_m \frac{\xi_m^2}{(\xi_m + \nu)^2} &= \alpha^{-1} \sum_{m=\kappa+1}^{\infty} \frac{1}{m!} \frac{\delta_m^2}{\xi^2 d^{m-\kappa} + \delta_m^2 d^{-m+\kappa} + \delta_m \xi} \overset{d \gg 1}{\approx} \\ &\overset{d \gg 1}{\approx} \alpha^{-1} \sum_{m=\kappa+1}^{\infty} \frac{1}{m!} d^{\kappa - m} \xrightarrow{d \gg 1} 0 \end{aligned} \tag{20}$$

So for the variance, the only term that counts is the one of order $m = \kappa$. Finally, we obtain:

$$\tau \approx \frac{1}{\alpha \kappa!} \frac{\delta_\kappa^2}{(\delta_\kappa + \xi(\alpha, \kappa))^2} \tag{21}$$

where $\xi(\alpha, \kappa)$ can be computed as in Eq.14. The full variance will then become:

$$Var(\alpha, \kappa, \lambda = 0) = \sigma_\epsilon^2 \frac{1 - \tau}{\tau} = \sigma_\epsilon^2 \frac{\delta_\kappa^2}{\alpha \kappa! (\delta_\kappa + \xi)^2 - \delta_\kappa^2} \tag{22}$$

Under which condition is the denominator negative? Not trivial

We can now plot the behavior of the variance as α is varied (Fig. 4)

Discorso del multiple descent: il plot in Fig.4 è valido solo per un regime, unendo tutti i regimi uno ottiene una "multiple descent" tipica dei kernel

2.4 Computation for $B(\alpha, \kappa)$

The bias term is given by (again, Paper 1):

$$B(\alpha, \kappa) = \frac{\nu^2 \langle \vec{\theta}, (\Sigma + \nu)^{-2} \vec{\theta} \rangle}{1 - \tau} \tag{23}$$

where $\vec{\theta}$ is the orthonormal decomposition of the target function $f_\star(\vec{x}) \in L_2(\mathcal{N}(0, \mathcal{I}_d))$. In particular, given the generalized d-dimensional Hermite polynomials, we can write:

$$f_\star(\vec{x}) = \sum_{\beta \in \mathbb{N}^d} \theta_\beta He_\beta(\vec{x}) = \sum_{\beta \in \mathbb{N}^d} \theta_\beta \prod_{i=1}^{d} He_{\beta_i}(x_i) \tag{24}$$

In general, the eigendecomposition of $f_\star$ cannot be organized in degenerate levels since we are only assuming $f_\star$ is square-integrable:

$$||f_\star||^2_{L_2} = \sum_{\beta \in \mathbb{N}^d} \theta^2_\beta = 1$$

but $f_\star$ may easily be not rotationally invariant (as was the case for the kernel $k(x, x')$).

Now let's compute the numerator of Eq.23:

$$\nu^2 \langle \vec{\theta}, (\Sigma + \nu)^{-2} \vec{\theta} \rangle = \sum_{\beta \in \mathbb{N}^d} \theta^2_\beta \frac{\nu^2}{(\lambda_\beta + \nu)^2} = \sum_{m=0}^{\infty} \frac{\nu^2}{(\xi_m + \nu)^2} \sum_{|\beta|=m} \theta^2_\beta \tag{25}$$

Let's define:

$$\Theta(m) = \sum_{|\beta|=m} \theta^2_\beta$$

This is the "energy" of the target functions that lies in the m shell. Finally:

$$\begin{aligned}\nu^2 \langle \vec{\theta}, (\Sigma + \nu)^{-2} \vec{\theta} \rangle &= \sum_{m=0}^{\infty} \Theta(m) \frac{\xi^2 d^{-2k}}{(\delta_m d^{-m} + \xi d^{-k})^2} = \sum_{m=0}^{\infty} \Theta(m) \frac{\xi^2 d^{-2k}}{\delta^2_m d^{-2m} + \xi^2 d^{-2\kappa} + 2\delta_m \xi d^{-m-\kappa}} = \\ &= \sum_{m=0}^{\infty} \Theta(m) \frac{\xi^2}{\delta^2_m d^{2(-m+\kappa)} + \xi^2 + 2\delta_m \xi d^{-m+\kappa}}\end{aligned}$$

To solve asymptotically this equation, we will perform the usual trick.

- When $0 \leq m \leq \kappa$, then the terms $d^{-m+\kappa}$ and $d^{2(-m+\kappa)}$ will both diverge to infinity. Consequently, the whole fraction will converge to 0 as $d \to \infty$. The prefactor $\Theta(m)$ is independent of d[4], so the whole term goes to 0 when $m < \kappa$
- When $m = \kappa$, we obtain:

$$\Theta(\kappa) \frac{\xi^2}{(\xi + \delta_\kappa)^2} \tag{26}$$

- When $m > \kappa$, then the terms $d^{-m+\kappa}$ and $d^{2(-m+\kappa)}$ will both converge to 0. We are left with:

$$\sum_{m>k} \Theta(m) \frac{\xi^2}{\delta^2_m d^{2(-m+\kappa)} + \xi^2 + 2\delta_m \xi d^{-m+\kappa}} \approx \sum_{m>k} \Theta(m) \tag{27}$$

  This summation here is convergent, since $||f|| = \sum_m \Theta(m) < \infty$

Finally, we have:

$$\nu^2 \langle \vec{\theta}, (\Sigma + \nu)^{-2} \vec{\theta} \rangle = \frac{\xi^2}{(\xi + \delta_\kappa)^2} \Theta(\kappa) + \sum_{m>k} \Theta(m) \tag{28}$$

This is a nice formula! Combining all together, I get:

$$B(\alpha, \kappa, \vec{\theta}) = \frac{\frac{\xi^2}{(\xi+\delta_\kappa)^2} \Theta(\kappa) + \sum_{m>k} \Theta(m)}{1 - \tau} = \frac{\frac{\xi^2}{(\xi+\delta_\kappa)^2} \Theta(\kappa) + \sum_{m>k} \Theta(m)}{1 - \frac{1}{\alpha\kappa!} \frac{\delta^2_\kappa}{(\delta_\kappa+\xi)^2}} \tag{29}$$

[4] Is it really? Or just O(1)?

When $\alpha \to \infty$ (a lot of data, but still scaling with d^κ, we are "saturating" the κ level but not the $\kappa + 1$), then:

$$Var(\alpha, \kappa) \to 0$$

$$B(\alpha, \kappa, \vec{\theta}) \approx \sum_{m>k} \Theta(m)$$

The last quantity is extremely interesting. It represents the projection of the target function $f_\star$ on the space spanned by Hermite polynomials whose degree is higher than the sample complexity κ. This is the irreducible bias, the error we cannot avoid since we do not have enough data.

TODO:
Asymptotic behavior of ξ, τ as $\alpha \to \infty$. (both goes to 0).
Give a random initialization to the weight θ_β and see what we get as a back-of-the-envelope computations.
Ridge case when $\lambda \neq 0$
Fix the ansatz when $\kappa_0 < \kappa$
And then of course onto the power-law case