Semester projet report

Ylann Rouzaire

June 3, 2020

This semester project can be separated in two distinct parts : the investigation of the dynamics of learning in a Teacher/Student Kernel Regression framework on one hand and the investigation of the behaviour of a noiseless hard-margin SVM classifier in presence of a gap at the interface between the 2 classes.

Each work will constitute a section of this report, organised as follows : in a first part, I shall explain the mathematical setting of the framework. The second part is the actual code and its explanation, available publicly on GitHub. The README.md if self-contained so it is not necessary to duplicate the information. Then come the figures and main numerical results. The last part sums up the analytical developments aiming at proving the scaling laws observed numerically.

This project was supervised by Matthieu Wyart and Stefano Spigler.

Contents

1 Teacher/Student Kernel Regression Dynamics

1.1 Settings

Kernel : A kernel is a function from $\mathbf{R}^d$ x $\mathbf{R}^d$ to $\mathbf{R}$. We will only deal with isotropic translation-invariant kernel $k(x, x') = k(||x - x'||)$ of the Matérn family :

$$f_\nu(h) = \frac{2^{1-\nu}}{\Gamma(\nu)} \left(\sqrt{2\nu} \frac{h}{\rho} \right)^\nu K_\nu \left(\sqrt{2\nu} \frac{h}{\rho} \right), \tag{1}$$

where Γ is the Euler Gamma function, K_ν is the modified Bessel function of the second kind, $\rho = 1$ is the characteristic length scale and $\nu > 0$ is the smoothness parameter (the larger the smoother, $\nu = \infty \leftrightarrow$ Gaussian kernel).

Teacher/Student framework : The idea is the following : the Teacher generates data according to its own "law" K_T. A subset of size P of that data called the *training set* is passed to the Student. From it and thanks to its "law" K_S, the Student has to infer the underlying "law" K_T. Its performance is measured by the test error ϵ , defined as a quadratic loss between the prediction of the Student and true values at several *new/unseen* points generated by the Teacher.

$$\epsilon = \epsilon\left(P, \nu_T, \nu_S\right) = \mathrm{E}_T \, \frac{1}{P_{test}} \sum_{\mu=1}^{P_{test}} |\hat{Z}^S(x_\mu) - Z^T_{test}(x_\mu)|^2 \tag{2}$$

where E_T is the expectation over the Teacher random process.

In this project, Teacher kernels are chosen among the Matérn family (1) and the Student kernel will always be the Laplace kernel :

$$\text{Laplace}(x, x') = \text{Matérn}\left[\nu = 1/2\right](x, x') = \exp(-||x - x'||) \tag{3}$$

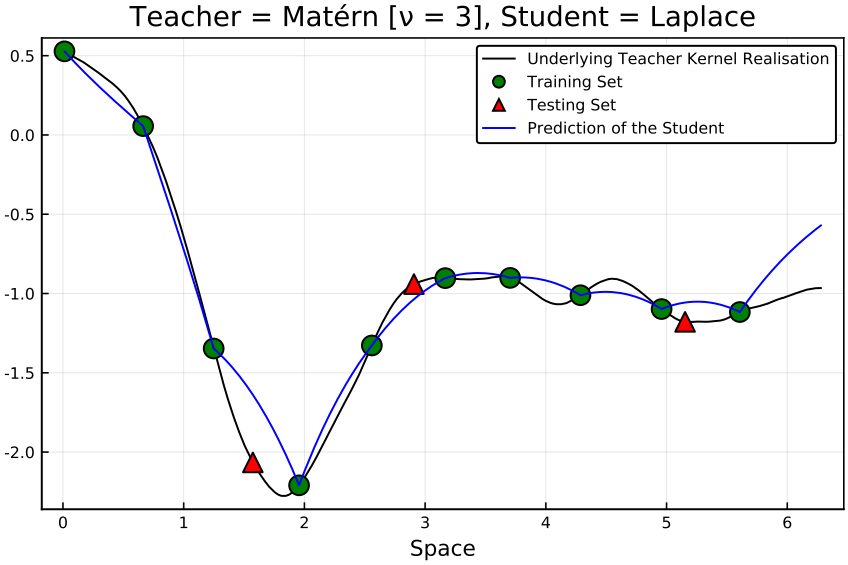


Figure 1: Illustration of the Teacher/Student framework in the context of Matérn kernel Regression. The test error ϵ is the sum over all testing set of the square distance between the red triangles and value of the blue curve at the same x coordinate.

Generation of the data by the Teacher kernel :

$$Z^T \sim \mathcal{N}(0, K_T) \tag{4}$$

so that

$$\mathrm{E}\ Z^T(x) = 0 \text{ and } \mathrm{E}\ Z^T(x)Z^T(x') = K_T(x, x') \tag{5}$$

Prediction of the Student kernel :

$$\hat{Z}^S(x_0)(t) = \sum_{\mu=1}^{P} a_\mu(t) K_S(x_\mu, x_0) = a(t) \cdot k_S(x_0) \tag{6}$$

where $a(t)$ is a vector of weights of size P changing over time during the learning procedure. The different optimisation routines implemented in the code are described in the README.md file of the project.

1.2 Features of the code

The Julia code can be found here : `github.com/Rouzaire/Kernel_Regression`.
The README.md file is self-explanatory and contains general information on the code and its structure. The code itself is commented and adds more precise details.

1.3 Results and Figures

The first checkup of the code was to recover the scaling for ϵ versus P found in [1] :

$$\epsilon \sim p^{-\beta}, \text{ with } \beta = \frac{1}{d} \min(\alpha_T - d, 2\alpha_S) \tag{7}$$

My main contribution at this stage is that my code is faster than what was previously coded.

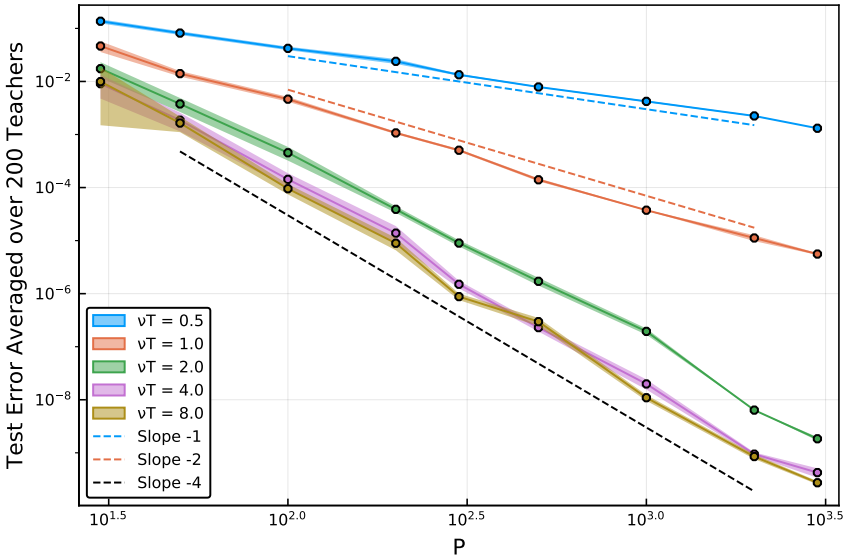
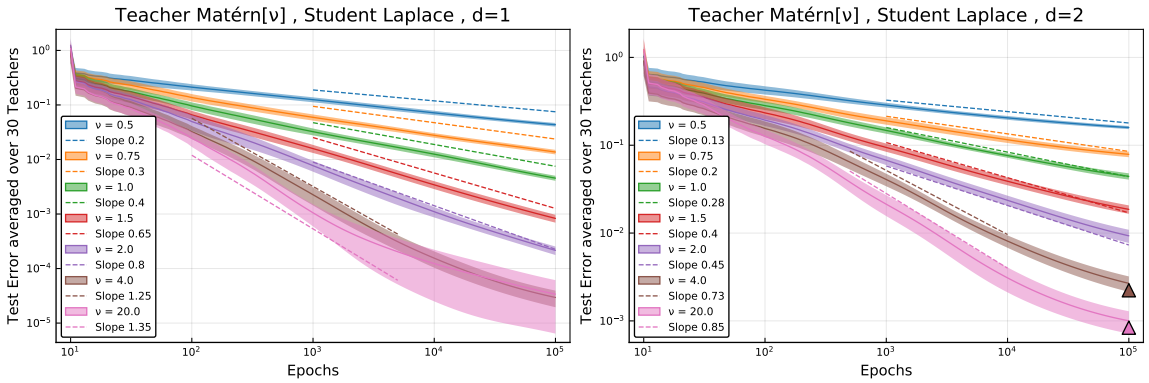


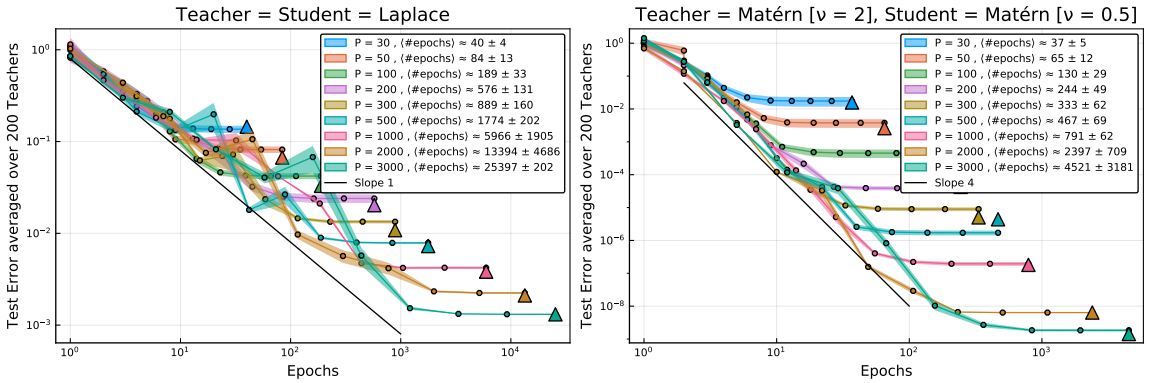
Figure 2: Learning curves for several Teacher kernels, here for $d = 1$. Student kernel is fixed to Laplace, cf. eq (3). One recovers the scaling predicted in [1].

Then I could investigate the dynamics of the test error during the optimization procedure. The conclusion is that for the Gradient Descent algorithm, the dynamics is very monotonic and exhibits no overfit : the test error is always decreasing and its limit is the exact test error plotted with triangles (only in Fig. (3b) for readability) :



(a) (b)

Figure 3: Optimization : Gradient Descent.



(a) (b)

Figure 4: Optimization : Conjugate Gradient Descent.

One can notice that for intermediate time, both for GD and CGD, the test error decays in a powerlaw. This is part of the analytic work in progress.

1.4 Analytical developments

In [3], the authors suggest that the test error can be decomposed in a sum over the modes ρ of the target function of error per mode. They provide with an expression for the error per mode : NB : $\Lambda \geq 0$ is the rigde parameter.

$$E = \sum_\rho E_\rho, \text{ where } E_\rho(p) = \frac{\langle \bar{\omega_\rho^2} \rangle}{\lambda_\rho} \left(\frac{1}{\lambda_\rho} + \frac{p}{\Lambda + t(p)} \right)^{-2} \left(1 - \frac{p\,\gamma(p)}{(\Lambda + t(p))^2} \right)^{-1} \tag{8}$$

$$t(p) = \sum_\rho \left(\frac{1}{\lambda_\rho} + \frac{p}{\Lambda + t(p)} \right)^{-1} \tag{9}$$

$$\gamma(p) = \sum_\rho \left(1 - \frac{p\,\gamma(p)}{(\Lambda + t(p))^2} \right)^{-2} \tag{10}$$

Note that these equations are the exact errors, there is no notion of dynamics here. We wanted to check whether for large P, these equations boiled down to the scalings relations already found in [1] for the ridgeless case. The answer is yes and here are some intermediate results :

Without rigde : $\Lambda = 0$, for $p \gg 1$

- $t(p) \sim p^{-s} \quad s = \frac{2-\theta}{\theta-1}$
- $\gamma(p) \sim p^{-r} \quad r = \frac{3-\theta}{\theta-1}$
- $E(p) \sim p^{-\beta} \quad \beta = \frac{1}{d} \min(\alpha_T - d, 2\,\alpha_S)$

With rigde : $\Lambda \sim \mathcal{O}(1)$, for $p \gg 1$

- $t(p) \sim p^{-s} \quad s = 2 - \theta$
- $\gamma(p) \sim p^{-r} \quad r = 3 - \theta$
- $E(p) \sim p^{-\beta} \quad \beta = \frac{1}{\alpha_S} \min(\alpha_T - d, 2\,\alpha_S)$

NB : $\alpha = d + 2\,\nu > d$ so if the data is not noisy, a regression with ridge will never be optimal/better than a ridgeless regression.

2 Kernel Classification with a gap at the interface

2.1 Settings

We investigate the learning curves of a supervised *2-class hard-margin kernel classification* task when the data present many invariant. In this project, the data labels will only depend on the first component x_1 of $\vec{x} \in \mathrm{R}^d$, all other dimensions are irrelevant for labelling :

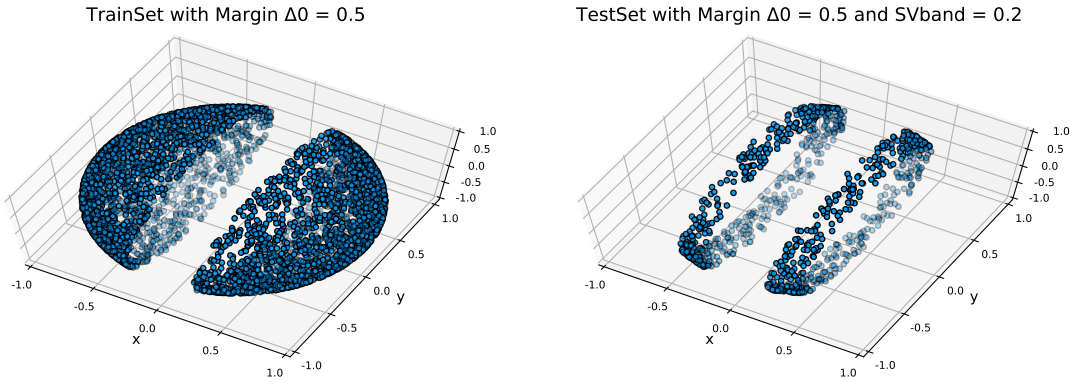
$$y(\vec{x}) = \text{sgn } x_1 \tag{11}$$

Generation of the data :
The data is generated uniformly on the unit hypersphere of dimension d [1] by normalizing (to 1) points from a multivariate random normal distribution $\mathcal{N}(0, I_{d+1})$ in d+1 dimensions.

This project studies the consequences of a *gap* of size $\Delta_0 \geq 0$ at the interface between labels. There is no data (neither training nor testing) in the gap. Intuitively, since the more distant from the interface, the easier to classify, one could expect less misclassification errors. Indeed, one shall see that the learning curve. decays as an exponential and not as a power law anymore when there is a gap between labels : see Fig. (**??**).

This very fact that the error $\epsilon\,(P)$ decays exponentially fast in presence of a gap $\Delta_0 > 0$ means that the testing set has to be at least bigger than $1/\epsilon$ to record at least one error. This quickly becomes numerically untractable in terms of memory and CPU resources. This is why the code implements and *Importance Sampling* (IS) algorithm, at least conceptually. The idea is that if a test point is far from the interface, it will for sure be correctly classified : it is therefore useless to compute that prediction. Thus, the testing set only contains data points in the vicinity of the gap. The resulting test error will then be weighted by the probability of falling into that area compared to all the surface of the unit hypersphere (minus the Δ_0 zone). This last step relies on the assumption that all the possibly misclassified points are contained in the testing set : the choice of the distance (called SVband) to the gap is therefore of capital importance.

[1] For clarity : $d = 1$ means the unit circle and $d = 2$ means the usual sphere embedded in the natural 3 dimensions.



(a) Example of Training Set for $d = 2$.

(b) Example of Testing Set for $d = 2$.

Figure 5: Data is uniformly distributed around the unit hypersphere.

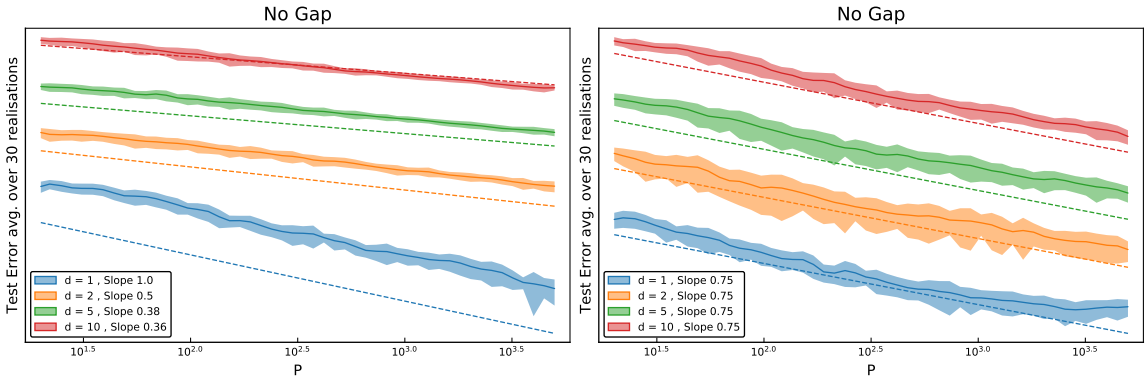
2.2 Features of the code

The Julia code can be found here : github.com/Rouzaire/Kernel_Classification.
The README.md file is self-explanatory and contains general information on the code and its structure. The code itself is commented and adds more precise details.

2.3 Results and Figures

Left panels : Laplace kernel $\xi = 1$, $k(h) = \exp(-h/\sigma)$
Right panels : Gaussian kernel $\xi = 2$, $k(h) = \exp(-\frac{h^2}{2\sigma^2})$
Unless stated otherwise, $\sigma = 100$ so that the kernel is evaluated only close to the origin (simpler to work out analytically thanks to Taylor expansions).

WARNING : The theory developed in [4] does not apply to Gaussian kernels $\xi = 2$. Thus, the fits of the right panels are "visual" fits and are not (yet) arising from any theoretical foundation.



(a) Laplace

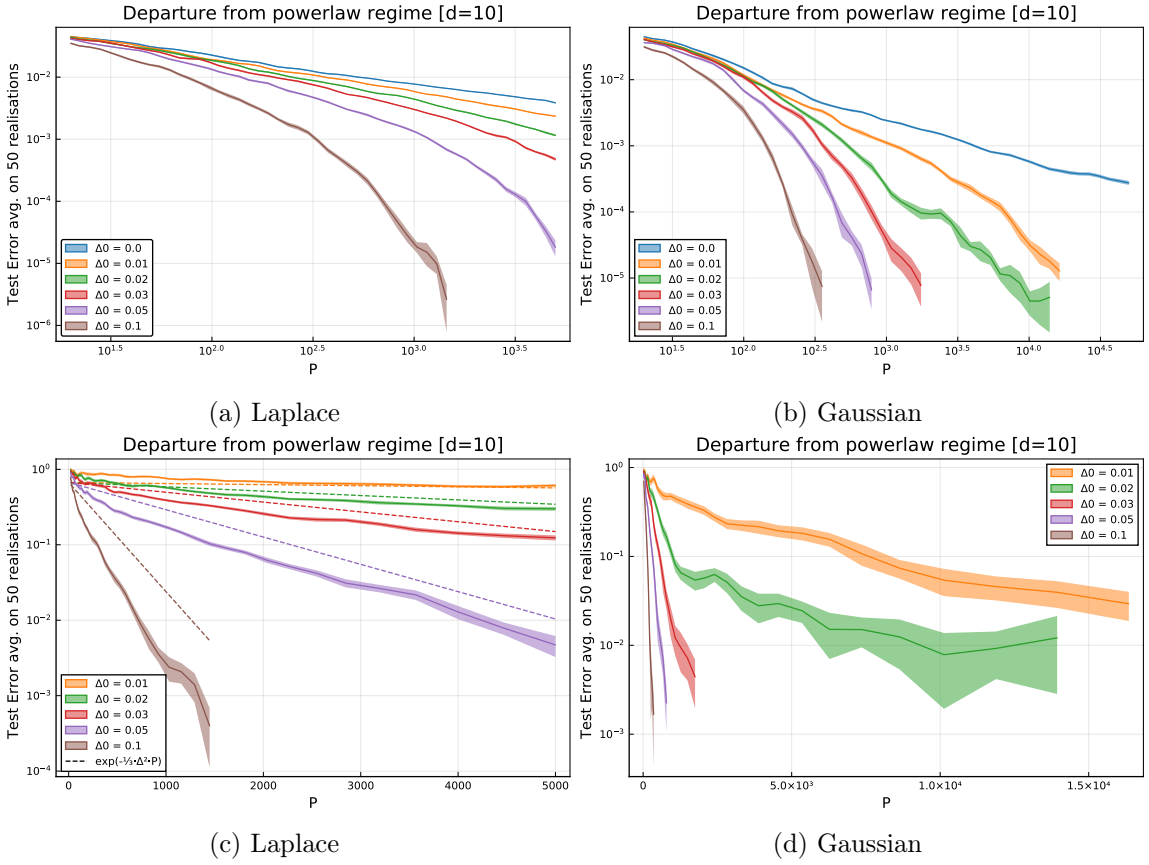
(b) Gaussian

Figure 6: Learning curves, No Gap Setup. NB : curves have artificially been shifted up for readability reasons so yaxis values become meaningless.

The test error of Fig. (6) decreases in a power law manner : $\epsilon(p) \sim p^{-\beta}$
For Laplace kernel, the theory developed in [4] applies and is in perfect accordance : one recover the correct exponent

$$\beta = \frac{d + \xi - 1}{3d - 3 + \xi} \tag{12}$$

and therefore the regression is subject to the so-called *curse of dimensionality*.
On the contrary, it appears that the Gaussian kernel exhibits a more or less constant exponent β that is numerically found to be close to $3/4$. This could motivate further theoretical work.



(a) Laplace (b) Gaussian

(c) Laplace (d) Gaussian

Figure 7: Departure from powerlaw regime, here $d = 10$. Top panels : LogLog scale exhibits a misxture of powerlaw and exponential behaviour. Bottom panels : LogLin scale. Curves are divided by the blue curve of top panels ($\Delta_0 = 0$) to isolate the exponential component.

Left panels of Fig. (7) suggest the following hypothesis (for Laplace kernels) :

$$\epsilon(p) \sim p^{-\beta} \ \exp\left\{ \frac{1}{2} f(d, \xi) \left(\frac{\Delta_0}{\gamma} \right)^2 p \right\} \tag{13}$$

2.4 Analytical developments

Work in progress

References

[1] Spigler, Geiger and Wyart
Learning Curves of Kernel Methods, empirical data vs. Teacher/Student paradigm

[2] Jacot, Gabriel and Hongler
Neural Tangent Kernel Convergence and Generalization in Neural Networks

[3] Bordelon, Canatar and Pehlevan
Spectrum dependent learning curves in kernel regression and wide neural networks

[4] Paccolat, Spigler and Wyart
How isotropic kernels learn simple invariants

[5] Pillaud-Vivien, Rudi and Bach
Exponential Convergence of Testing Error for Stochastic Gradient Methods

[6] Caponnetto and De Vito
Optimal Rates for the Regularized Least-Squares Algorithm

[7] Bartlett and Mendelson
Rademacher and Gaussian Complexities