SUPPLEMENTARY MATERIAL FOR : "REGRESSION FUNCTION ESTIMATION AS A PARTLY INVERSE PROBLEM"

F. COMTE[(1)] AND V. GENON-CATALOT[(2)]

Abstract. This paper is about nonparametric regression function estimation. Our estimator is a one step projection estimator obtained by least-squares contrast minimization. The specificity of our work is to consider a new model selection procedure including a cutoff for the underlying matrix inversion, and to provide theoretical risk bounds that apply to non compactly supported bases, a case which was specifically excluded of most previous results. Upper and lower bounds for resulting rates are provided.

MSC2010 *Subject classifications.* 62G08 - 62M05
Key words and phrases. Hermite basis. Laguerre basis. Model selection. Non parametric estimation. Regression function.

To ease the reading and illustrate the method, we add in this supplementary material theoretical toolds used in the paper and simulation results.

Appendix A. Theoretical tools

A proof of the following theorem can be found in Stewart and Sun (1990).

Theorem A.1. *Let* $\mathbf{A}$, $\mathbf{B}$ *be* $(m \times m)$ *matrices. If* $\mathbf{A}$ *is invertible and* $\|\mathbf{A}^{-1}\mathbf{B}\|_{\text{op}} < 1$, *then* $\tilde{\mathbf{A}} := \mathbf{A} + \mathbf{B}$ *is invertible and it holds*

$$\|\tilde{\mathbf{A}}^{-1} - \mathbf{A}^{-1}\|_{\text{op}} \leq \frac{\|\mathbf{B}\|_{\text{op}}\|\mathbf{A}^{-1}\|_{\text{op}}^2}{1 - \|\mathbf{A}^{-1}\mathbf{B}\|_{\text{op}}}$$

Theorem A.2 (Bernstein Matrix inequality)**.** *Consider a finite sequence* $\{\mathbf{S}_k\}$ *of independent, random matrices with common dimension* $d_1 \times d_2$. *Assume that*

$$\mathbb{E}\mathbf{S}_k = 0 \quad \text{and} \quad \|\mathbf{S}_k\|_{\text{op}} \leq L \quad \textit{for each index } k.$$

Introduce the random matrix $\mathbf{Z} = \sum_k \mathbf{S}_k$. *Let* $\nu(\mathbf{Z})$ *be the the variance statistic of the sum:* $\nu(\mathbf{Z}) = \max\{\lambda_{\max}(\mathbb{E}[\mathbf{Z}'\mathbf{Z}]), \lambda_{\max}(\mathbb{E}[\mathbf{Z}\mathbf{Z}'])\}$. *Then*

$$\mathbb{E}\|\mathbf{Z}\|_{\text{op}} \leq \sqrt{2\nu(\mathbf{Z})\log(d_1+d_2)} + \frac{1}{3}L\log(d_1+d_2).$$

Furthermore, for all $t \geq 0$

$$\mathbb{P}\left[\|\mathbf{Z}\|_{\text{op}} \geq t\right] \leq (d_1+d_2)\exp\left(-\frac{t^2/2}{\nu(\mathbf{Z}) + Lt/3}\right).$$

A proof can be found in Tropp (2012) or Tropp (2015).
We recall the Talagrand concentration inequality given in Klein and Rio (2005).

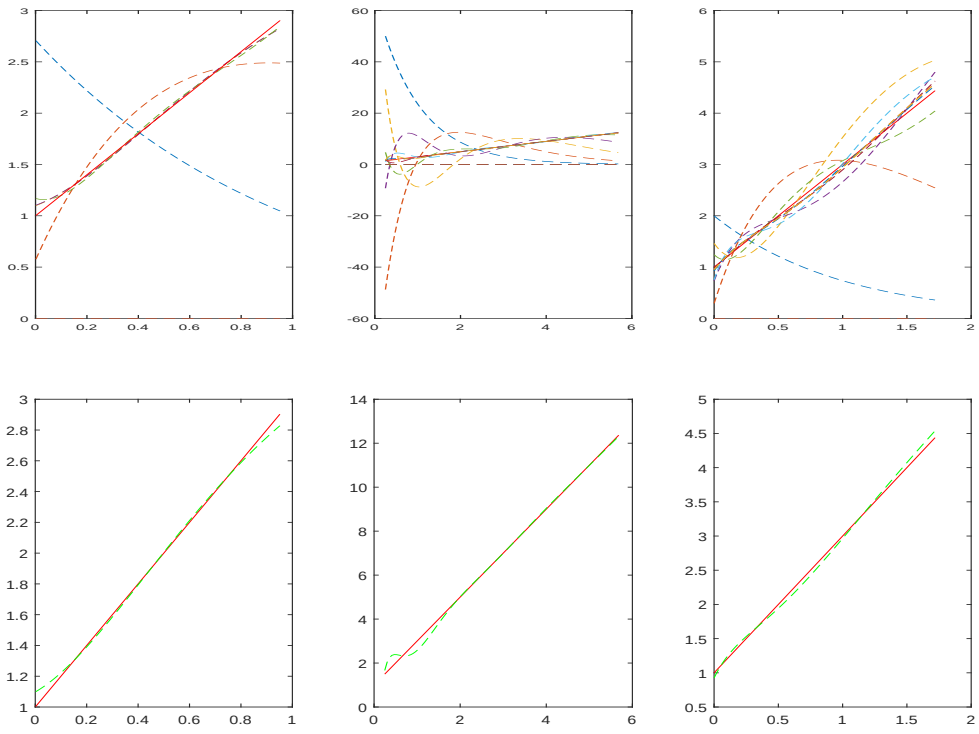
(1): Université Paris Descartes, Laboratoire MAP5, email: fabienne.comte@parisdescartes.fr.
(2): Université Paris Descartes, Laboratoire MAP5, email: valentine.genon-catalot@parisdescartes.fr.

Theorem A.3. *Consider $n \in \mathbb{N}^*$, $\mathcal{F}$ a class at most countable of measurable functions, and $(X_i)_{i\in\{1,\dots,n\}}$ a family of real independent random variables. Define, for $f \in \mathcal{F}$, $\nu_n(f) = (1/n)\sum_{i=1}^n (f(X_i) - \mathbb{E}[f(X_i)])$, and assume that there are three positive constants M, H and v such that $\sup_{f\in\mathcal{F}} \|f\|_\infty \le M$, $\mathbb{E}[\sup_{f\in\mathcal{F}} |\nu_n(f)|] \le H$, and $\sup_{f\in\mathcal{F}} (1/n)\sum_{i=1}^n \mathrm{Var}(f(X_i)) \le v$. Then for all $\alpha > 0$,*

$$\mathbb{E}\left[\left(\sup_{f\in\mathcal{F}} |\nu_n(f)|^2 - 2(1+2\alpha)H^2\right)_+\right] \le \frac{4}{b}\left(\frac{v}{n}e^{-b\alpha\frac{nH^2}{v}} + \frac{49M^2}{bC^2(\alpha)n^2}e^{-\frac{\sqrt{2}bC(\alpha)\sqrt{\alpha}}{7}\frac{nH}{M}}\right)$$

with $C(\alpha) = (\sqrt{1+\alpha} - 1) \wedge 1$, and $b = \frac{1}{6}$.

By density arguments, this result can be extended to the case where $\mathcal{F}$ is a unit ball of a linear normed space, after checking that $f \to \nu_n(f)$ is continuous and $\mathcal{F}$ contains a countable dense family.



$X \sim \mathcal{U}([0,1])$ $X \sim \mathcal{N}(4,1)$ $X \sim f_3$

FIGURE 1. First line: beam of the proposals $\hat{f}_m$ for $m = 1$ to $m_{\max}$ in the Laguerre basis. Second line: the estimator as selected by the procedure, $\hat{f}_{\hat{m}}$. Function $b(x) = 2x+1$, $n = 1000$, density $f_k(x) = (k-1)/(1+x)^k \mathbf{1}_{x\ge 0}$.

Appendix B. Numerical illustrations

In this section, numerical illustrations of our method are presented. The estimation procedure is implemented for the Laguerre (Figures 1 to 4) and the Hermite basis (Figure 5). The $(\varepsilon_i)_{1\le i\le n}$ are generated as an i.i.d. sample of Gaussian $\mathcal{N}(0,\sigma^2)$ with $\sigma = 0.5$. Then, we choose different functions $b(.)$ (bounded or not) and different types of distribution of the design $(X_i)_{1\le i\le n}$. Typically, a linear function $x \mapsto 2x+1$ is experimented without the information of its linearity, which allows to test moment conditions; on the contrary, $x \mapsto 4x/(1+x^2)$ is bounded and should be easier to reconstruct. For the design density, we consider standard uniform or Gaussian cases, and also different heavy tailed distributions.

As usual in model selection methods, the constant κ is calibrated by preliminary simulation experiments, and we took $\kappa = 4$, see comments after Theorem 4.1.

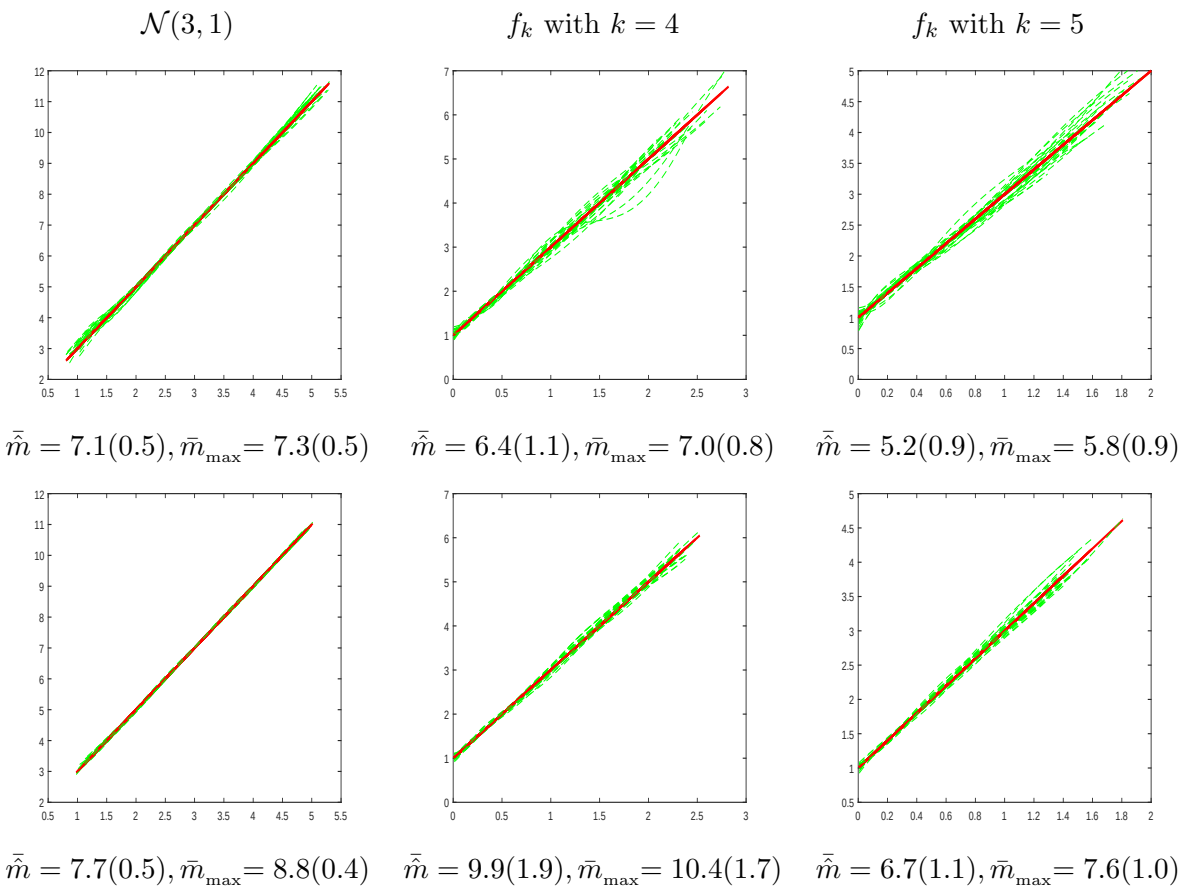
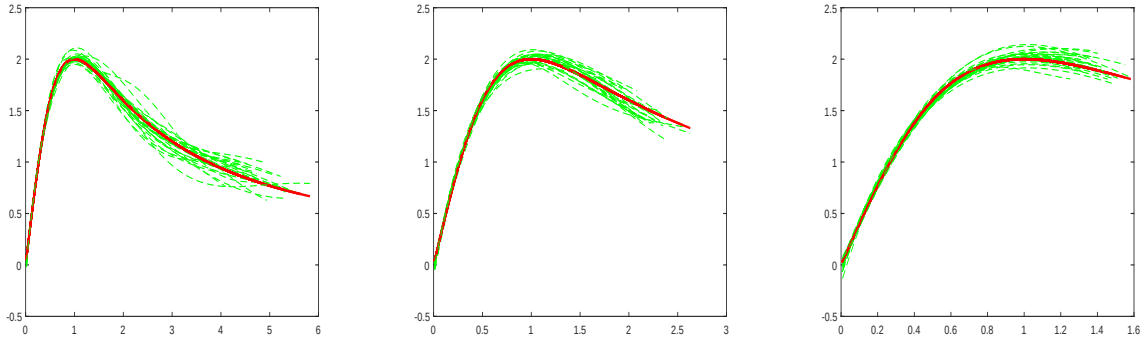


Figure 2. 25 estimated curves in Laguerre basis (dotted -green/grey), the true in bold (red), first line: $n = 250$, last line: $n = 1000$, $b(x) = 2x+1$ and different laws for the design, $f_k(x) = (k-1)/(1+x)^k \mathbf{1}_{x\ge 0}$.

In Figure 1, we plot in the first line the collection of estimators in the Laguerre basis, among which the algorithm makes the selection. The number of computed estimators is different from one example to another, as the collection of models $\widehat{\mathcal{M}}_n$ is random and

depends on $\|\widehat{\Psi}_m^{-1}\|_{\text{op}}$. In the practical implementation, the collection $\widehat{\mathcal{M}}_n$ may be small. Therefore, we have considered the (random) maximum value $m_{\max}$ such that $\|\widehat{\Psi}_m^{-1}\|_{\text{op}} \leq \sqrt{n}$, which is slightly larger than in the theory. But inversion of the matrix $\widehat{\Psi}_m$ remains possible in such cases. Surprisingly, we can see that very few estimators are sometimes computed (see the example of uniform distribution on the right). They are also very different from one dimension to another. The second line presents the final estimator, selected by the procedure. In the example of Figure 1, the curve is linear, and is perfectly estimated, although its particular form is unknown and was not *a priori* easy to obtain with the Laguerre basis.



$k=3$ $\bar{\hat{m}} = 5.6(1.0), \bar{m}_{\max} = 15.6(0.7)$ | $k=4$ $\bar{\hat{m}} = 4.6(1.1), \bar{m}_{\max} = 10,1(1.6)$ | $k=5$ $\bar{\hat{m}} = 3.5(1.3), \bar{m}_{\max} = 7.2(1.0)$

FIGURE 3. 25 estimated curves in the Laguerre basis (dotted -green/grey), the true in bold (red), $n = 1000$, density $f_k(x) = (k-1)/(1+x)^k \mathbf{1}_{x\geq 0}$ for $k = 3, 4$ and 5, $b(x) = 4x/(1+x^2)\mathbf{1}_{x\geq 0}$.

In Figures 2 to 5, we present beams of 25 estimators computed either in the Laguerre basis (Figures 2, 3, 4), or in the Hemite basis (Figure 5). The beams give information about the variability of the estimation procedure.

Below each plot, we give the density of the design and the value of $\bar{\hat{m}}$ which is the mean of the selected dimensions for the 25 estimators represented on the figure, with standard deviation in parenthesis. It is associated with the value of $\bar{m}_{\max}$ which is the mean of the maximal dimension for which the estimator is computed, with standard deviation in parenthesis. We can see that the maximal dimension is rather small (less than ten models are compared for selection, in general) but an adequate choice seems always to exist in this small collection. This means that the squared-bias variance compromise in the restricted set $\mathcal{M}_n$ has good performance and that the non compact Laguerre and Hermite bases are very interesting and simple estimation tools. Indeed, the method is very fast and its low complexity has an important practical interest.

Figure 2 is complementary of Figure 1 and considers the same linear regression function with similar distributions for X. The interest of the linear case is also to illustrate the sharpness of the moment conditions: indeed the condition $\mathbb{E}[b^2(X_1)] < +\infty$ for X with density $f_k(x) = (k-1)/(1+x)^k \mathbf{1}_{x\geq 0}$ is satisfied for $k > 3$ and the condition $\mathbb{E}[b^4(X_1)] < +\infty$ holds for $k > 5$. We checked, in the case of linear $b(.)$, that the method does not

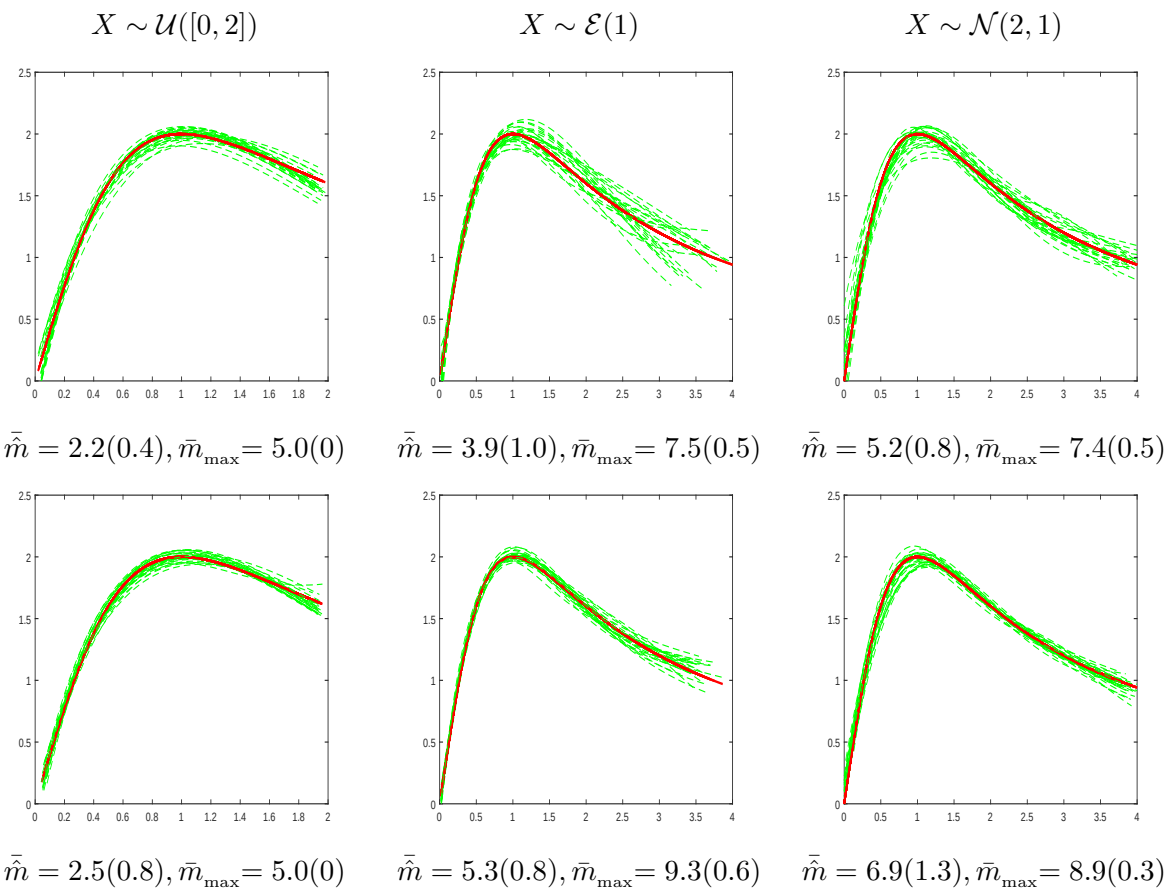


FIGURE 4. 25 estimated curves in Laguerre basis (dotted -green/grey), the true in bold (red), first line: $n = 250$, last line: $n = 1000$, $b(x) = 4x/(1+x^2)\mathbf{1}_{x\geq 0}$ and different laws for the design.

work for $k = 2, 3$, but the last two plots of Figure 2 show that it works rather well for $k = 4, 5$. The minimal theoretical condition may thus be weakened from $\mathbb{E}[b^4(X_1)] < +\infty$ to $\mathbb{E}[b^2(X_1)] < +\infty$.

Figure 2 allows also to compare results between two sample sizes, $n = 250$ and $n = 1000$: the improvement is obvious but the estimation remains correct for $n = 250$.

Figures 3 and 4 present the results for the function $b(x) = 4x/(1+x^2)\mathbf{1}_{x\geq 0}$ and different distributions for X, heavy tailed or not. The beams are more concentrated around the true function in Figure 4 for non heavy tailed distribution.

The estimation with the Hermite basis has similar behaviour, as can be seen in Figure 5, and a comparison between $n = 100$ and $n = 1000$ is provided here. The regression function is unbounded, but a non heavy-tailed density for X is used: this makes the problem obviously easier, and the results excellent even for $n = 100$.

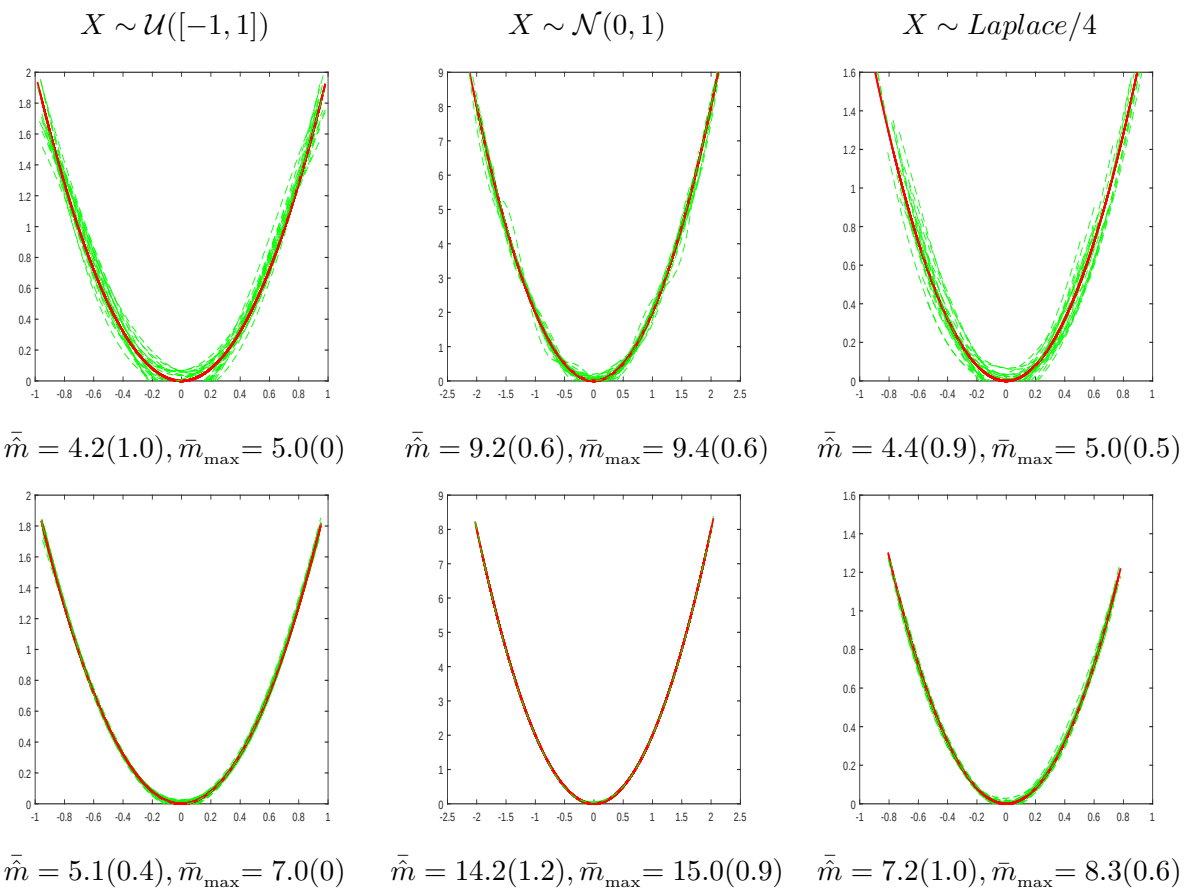


FIGURE 5. 25 estimated curves in Hermite basis (dotted -green/grey), the true in bold (red), first line: $n = 100$, last line: $n = 1000$, $b(x) = 2x^2$ and different laws for the design.

REFERENCES

[Klein and Rio, 2005] Klein, T. and Rio, E. (2005) Concentration around the mean for maxima of empirical processes. *Ann. Probab.* **33**, no. 3, 1060-1077.

[Stewart and Sun, 1990] Stewart, G. W. and Sun, J.-G. (1990). *Matrix perturbation theory.* Boston etc.: Academic Press, Inc.

[Tropp, 2012] Tropp, J. A. (2012). User-friendly tail bounds for sums of random matrices. *Found. Comput. Math.*, 12(4):389–434.

[Tropp, 2015] Tropp, J. A. (2015). An introduction to matrix concentration inequalities. *Found. Trends Mach. Learn.*, 8(1-2):1–230.