Concentration of probability measures explains cut-off phenomena in model selection: illustration in a simple setting

Pascal Massart
Université Paris-Saclay, Orsay

Lecture 1

1. Overview

Concentration of product probability measures

Michel Talagrand received the Abel Prize in 2024. His scientific output is enormous, and it would be futile for anyone to attempt to cover it all or even summarize it.

This remains true even if we focus on his work on product probability measures in the 1990s that has had an impact in statistics and machine learning (among other domains). Our goal here is much more modest.

Instead we propose to solve a kind of « guided exercise » that will illustrate the power of this connection, following the lines of our recent paper with V. Rivoirard.

Why should we care about concentration of product probability measures ?

- Functions of independent random variables include suprema of empirical processes
- Non asymptotic sub-Gaussian probability bounds for suprema of empirical processes can be used as non asymptotic substitutes of the central limit theorem to study the behavior of empirical risk minimizers.
- This was the main reason for which these concentration results became popular in statistics and machine learning.

This is the global picture. Let us now turn to a more specific issue which illustrate the importance of getting exponential probability bounds.

Estimator (and model) selection

A versatile approach to functional estimation consists of considering some (possibly huge) collection of estimators $\{\hat{f}_m\}_{m\in\mathcal{M}}$ of some target f and define some genuine selection rule $\hat{m}$ from the data. This includes the case of a collection of empirical risk minimizers (model selection).

The statistical performance of such a procedure can be evaluated through some given loss function ℓ . One would like the risk $E\ell(f,\hat{f}_{\hat{m}})$ to be as close as possible to the oracle benchmark

$$\inf_{m\in\mathcal{M}} E\ell(f,\hat{f}_m)$$

In many cases the selection procedure involves some hyperparemeter λ and most of the positive results ensuring that the selected estimator behaves approximately like an oracle are proved under the constraint that λ is larger than some quantity which is more or less precisely known. The choice of λ is left to the user…

Main point in what follows

Negative results ensuring that below some critical value the procedure breaks down can be very helpful. The existence of such cut-off phenomena fully can be proved via concentration

Let us now move on to the specific simple context in which we will illustrate this point (our guided exercice!).

2. The classical linear regression framework

In the Euclidean space $\mathbb{R}^n$, one observes the random vector

$$Y = f + \sigma\epsilon$$

where the variables $\epsilon_j, 1 \leq j \leq n$ are i.i.d. centered and normalized rv's . σ is is the level of noise which is assumed to be known. f is the unknown mean vector to be estimated.

If we go back to statistics in the fifties the linear model assumption would specify that the unknown mean vector belongs to some linear subspace S of the Euclidean space $\mathbb{R}^n$

Once this constraint is given, f can merely be estimated by the least squares estimator (LSE) defined by

$$\hat{f} = \arg\min_{g \in S} \| Y - g \|^2$$

In the seventies many efforts have been developed to propose model selection criteria in order to relax the constraint that the mean vector belongs to a linear model which is given in advance. In the spirit of the works by Birgé and M. we have in mind that models as well as their collection are allowed to depend on n

3. Model selection

Model selection proceeds in two steps:

- Take some finite collection of linear models $(S_m)_{m \in \mathcal{M}}$ (which are merely subsets of $\mathbb{R}^n$). To each model S_m corresponds the LSE $\hat{f}_m$ defined on it.
- Use the data to select a value $\hat{m}$ in $\mathcal{M}$, the selected model being $S_{\hat{m}}$ and the corresponding estimator $\hat{f}_{\hat{m}}$.

According to the non asymptotic point of view the quality of a model selection procedure is measured by the quadratic risk of the resulting estimator $E_f \| \hat{f}_{\hat{m}} - f \|^2$.

In other words one ideally would like the selected estimator to behave like an *oracle*, i.e. minimize the quadratic risk of $\hat{f}_m$, which by Pythagoras' identity can be written as

$$E_f\|\hat{f}_m - f\|^2 = \|f - f_m\|^2 + E_f\|\hat{f}_m - f_m\|^2$$

where f_m denotes the orthogonal projection $\Pi_m f$ of f onto S_m

It is interesting to analyze the random term of this expression

$$\|\hat{f}_m - f_m\|^2 = \sigma^2\|\Pi_m \epsilon\|^2$$

Taking some orthonormal basis $\{\phi_j^{(m)}, 1 \leq j \leq D_m\}$ of S_m we see that

$$\chi_m^2 = \| \Pi_m(\epsilon) \|^2 = \sum_{1 \leq j \leq D_m} \langle \epsilon, \phi_j^{(m)} \rangle^2$$

so that the random term is a chi-square type statistics. In particular the quadratic risk can be computed as

$$E_f \|\hat{f}_m - f\|^2 = \|f - f_m\|^2 + \sigma^2 D_m$$

Mallows' tricky idea consists of noticing that minimizing the quadratic risk is equivalent to minimizing this risk minus the constant term $\|f\|^2$, i.e.

Unbiased risk estimation and Mallows' heuristics

$$E_f\|\hat{f}_m - f\|^2 - \|f\|^2 = -\|f_m\|^2 + \sigma^2 D_m \qquad (*)$$

just, by using Pythagoras' identity again.

By similar arguments, we also have

$$E_f\|\hat{f}_m\|^2 - \|f_m\|^2 = \sigma^2 D_m$$

So that the quantity

$$\mathrm{crit}(m) = -\|\hat{f}_m\|^2 + 2\sigma^2 D_m$$

turns out to be an *unbiased* estimator of (*)

This is precisely Mallows' criterion!

Mallows' heuristics relies on the belief that what you see in expectation is what you see on the data.

Justifying or correcting this belief has been the main motivation for introducing concentration inequalities in our long-standing collaboration with Lucien Birgé on this topic, starting 30 years ago.

The key is to specify how close is $\|\hat{f}_m\|^2$ to its expectation $\|f_m\|^2 + \sigma^2 D_m$, *uniformly* with respect to $m \in \mathcal{M}$.

Essentially, this amounts to understand the behavior of

$$\chi_m^2 = \| \Pi_m(\epsilon) \|^2 = \sum_{1 \le j \le D_m} \langle \epsilon, \phi_j^{(m)} \rangle^2$$

Link with concentration issues

The idea is to rewrite the square root of the chi-square type statistics as

$$\chi_m = \sup_{b \in S_m, \|b\| \leq 1} \langle b, \epsilon \rangle$$

This makes clear that χ_m is a 1-Lipschitz function of ϵ.
In Birgé and M. (2000) this fact has been used in the Gaussian case since if ϵ is standard normal on $\mathbb{R}^n$, it is known since 75' that for a 1-Lipschitz function ζ

$$P\{\zeta(\epsilon) \geq M + t\} \leq e^{-t^2/2}$$

where M denotes either the mean or the median of $\zeta(\epsilon)$

4. Suprema of Rademacher processes

Assume now that ϵ is a Rademacher random vector. This is more tricky since it is known that Lipschitz is not enough to warrant concentration. Fortunately, the function of interest is also convex, so that if we consider

$$\zeta(x) = \sup_{b \in B} \langle b, x \rangle$$

where B is some closed subset of the unit Euclidean ball of $\mathbb{R}^n$,we can rewrite it as $\zeta(x) = \langle b^*(x), x \rangle$ and notice that for all $x, y \in [-1,1]^n$,

$$\zeta(x) - \zeta(y) \leq \langle b^*(x), x \rangle - \langle b^*(x), y \rangle$$

$$\leq 2 \sum_{i=1}^{n} |b_i^*(x)| \mathbf{1}_{x_i \neq y_i}$$

We see that the function ζ obeys to $\mathscr{C}_v$

Weak bounded differences condition

$$\zeta(x_1, \ldots, x_n) - \zeta(y_1, \ldots, y_n) \leq \sum_{i=1}^{n} c_i(x) \mathbf{1}_{x_i \neq y_i}$$

for all $x_1, \ldots, x_n$ and $y_1, \ldots, y_n$.
where the c_i's are non negative measurable function
satisfying $\| \sum_{i=1}^{n} c_i^2 \|_\infty \leq v$

with $v = 4$ in the Rademacher case above. Note that the Rademacher assumption is not essential here. If the errors are absolutely bounded by M, then the conclusion would be the same but with a different value for v ($v = 4M^2$).

Mc Diarmid's bounded difference condition (1978) involves $\sum_{i=1}^{n} \|c_i^2\|_\infty$ instead of $\| \sum_{i=1}^{n} c_i^2 \|_\infty$. It took time before one realizes that it was possible to prove sub-Gaussian concentration bounds under the weaker condition $\mathscr{C}_v$.

Talagrand has introduced a clever tool to deal with such functions: the convex distance !

As we shall see Talagrand's approach naturally leads to concentration around the median. If one wants to save constants it is easier to use the transportation which leads to concentration around the mean through some nice variational argument.

Let's see how it all works!

Talagrand's convex distance

M. Talagrand's approach to the concentration of product probability measures is inspired by geometry.
Basically his approach is a cleverly modified version of the isoperimetric approach introduced by Vitali Milman and which goes as follows.

The isoperimetric approach

If $(\mathcal{X}, d)$ and μ is a Borel probability measure, the concentration rate function γ of μ is defined as

$$\gamma(t) = \sup_{A, \mu(A) \geq 1/2} \mu\{d(\,.\,, A) \geq t\}$$

Note that if μ is the standard normal distribution on a Euclidean space, then Borell's theorem ensures that

$$\gamma(t) \leq \int_t^\infty \exp(-u^2/2) du \leq \frac{1}{2} e^{-t^2/2}$$

Now, if f is a 1-Lipschitz function and x is a point such that $d(x, \{f \leq s\}) < t$, then, there exists some point y such that $f(y) \leq s$ and $d(x, y) < t$ and therefore

$$f(x) \leq f(y) + d(x, y) < s + t.$$

In other words the level sets of a 1-Lipschitz function have the property that

$$\{f \geq s + t\} \subseteq \{d(\,.\,, \{f \leq s\}) \geq t\}$$

which directly implies that choosing s as a median M of f under μ, $\mu\{f - M \geq t\} \leq \gamma(t)$ and changing f into $-f$ leads to a similar result on the left tail. Finally

$$\mu\{\,|f - M| \geq t\} \leq 2\gamma(t)$$

which establishes the connection between the concentration of measure and the concentration of 1-Lipschitz functions around their median.

Coming back to M. Talagrand's approach to the concentration of product probability measures, let us now introduce his concept of « convex distance » d_T.

Definition

Let $\mathcal{X}^n$ be some product space and denote by $\mathbf{B}_n^+$ the set $\{\alpha \in \mathbb{R}^n \,|\, \sum_{i=1}^n \alpha_i^2 \leq 1 \text{ and } \alpha_i \geq 0, \forall 1 \leq i \leq n\}$. Given some point $x \in \mathcal{X}^n$ and some subset A of $\mathcal{X}^n$ one defines

$$d_T(x, A) = \sup_{\alpha \in \mathbf{B}_n^+} \inf_{y \in A} \sum_{i=1}^n \alpha_i \mathbf{1}_{x_i \neq y_i}$$

This « distance » is especially well designed to deal with functions which satisfy the weak bounded differences condition above. Indeed assume that f is such a function

$$f(x) - f(y) \leq \sum_{i=1}^{n} c_i(x) \mathbf{1}_{x_i \neq y_i}$$

Suppose also that $\| \sum_{i=1}^{n} c_i^2 \|_\infty \leq 1$.

Choosing A to be a level set $\{f \leq s\}$ and taking some point x, if $d_T(x, A) < t$, since $c(x) \in \mathbf{B}_n^+$, this means that

$$\inf_{y \in A} \sum_{i=1}^{n} c_i(x) \mathbf{1}_{x_i \neq y_i} \leq d_T(x, A) < t$$

So that there exists some point y such that $f(y) \leq s$, for which

$$\sum_{i=1}^{n} c_i(x) \mathbf{1}_{x_i \neq y_i} < t$$

and the above condition on f implies that $f(x) < s + t$.

In other words we have proved that under the weak bounded difference condition, the level sets of a function have the property (*) that

$$\{f \geq s + t\} \subseteq \{d_T(\,.\,, \{f \leq s\}) \geq t\}$$

provided that $\| \sum_{i=1}^{n} c_i^2 \|_\infty \leq 1$

Here, everything works as if Talagrand's convex distance were an actual distance and the function f were 1-Lipschitz with respect to this distance.

If one copy/paste Milman's isoperimetric approach, ultimately, one has to switch to $-f$. Obviously this is not possible to do that here because the weak bounded difference condition is not symmetric. One needs a new idea…

Talagrand has the solution! At this step: use a stronger version of the concentration rate.

If X is a random variable taking its values in $\mathcal{X}^n$
the quantity introduced by Talagrand is

$$\theta(t) = \sup_A P(X \in A) P(d_T(X, A) \geq t)$$

By property (*), it is true that for all s and all positive t

$$P(f(X) \leq s) P(f(X) \geq s + t) \leq \theta(t).$$

Now, taking M to be a median of $f(X)$ and choosing either $s = M$ or $s = M - t$ leads to

$$\boxed{P(f(X) \geq M + t) \vee P(f(X) \leq M - t) \leq 2\theta(t)}$$

In other words, if one is able to handle $\theta(t)$, the concentration of $f(X)$ around its median follows. One of the beautiful results established by Talagrand is that $\theta(t) \leq \exp(-t^2/4)$ whenever the variables $X_1, X_2, \ldots, X_n$ are independent.

Talagrand's convex distance inequality

Let $X_1, X_2, \ldots, X_n$ be independent random variables taking there values in $\mathcal{X}^n$. For all measurable set $A \in \mathcal{X}^n$ the following inequality holds for all positive t

$$P(X \in A)P(d_T(X, A) \geq t) \leq \exp(-t^2/4)$$

This result has an immediate corollary.

Corollary

Let $Z = f(X_1, \ldots, X_n)$ be a measurable function of independent random variables. Assume that the function f satisfies to the weak bounded differences condition above and suppose furthermore that $\| \sum_{i=1}^{n} c_i^2 \|_\infty \leq v$. Then

$$P(Z \geq M + t) \vee P(Z \leq M - t) \leq 2 \exp(-\frac{t^2}{4v})$$

Concentration around the mean through transportation

The transportation approach to concentration starts with the following idea. Let us consider some coupling $\mathbf{P} \in \mathscr{P}(P^n, Q)$ between P^n and Q.
We start with the identity

$$E_Q(\zeta) - E_{P^n}(\zeta) = E_{\mathbf{P}}(\zeta(Y) - \zeta(X))$$

By the weak bounded differences condition $\mathscr{C}_v$ and Cauchy-Schwarz we can write

$$E_{\mathbf{P}}(\zeta(Y) - \zeta(X) \,|\, Y) \leq \sum_{i=1}^{n} c_i(Y) \mathbf{P}(X_i \neq Y_i \,|\, Y)$$

$$\leq \sqrt{\sum_{i=1}^{n} c_i^2(Y)} \sqrt{\sum_{i=1}^{n} \mathbf{P}^2(X_i \neq Y_i \,|\, Y)}$$

And by Cauchy-Schwarz inequality again

Since $\| \sum_{i=1}^{n} c_i^2 \|_\infty \leq v$ we finally derive that

$$E_Q(\zeta) - E_{P^n}(\zeta) \leq \sqrt{v \inf_{\mathbf{P} \in \mathscr{P}(P,Q)} \sum_{i=1}^{n} E_{\mathbf{P}}\left(\mathbf{P}^2(X_i \neq Y_i \,|\, Y)\right)}$$

Now it remains to use the following remarkable coupling inequality due to K. Marton

Marton's coupling inequality

Let $P^n = \mu_1 \otimes \mu_2 \otimes \ldots \otimes \mu_n$ be some product probability measure on some product space $\mathcal{X}^n$ and Q be a probability distribution on $\mathcal{X}^n$ such that $Q \ll P^n$ then

$$\min_{\mathbf{P} \in \mathcal{P}(P,Q)} E_{\mathbf{P}}\Big(\sum_{i=1}^{n} \mathbf{P}^2(X_i \neq Y_i \,|\, X) \Big) \leq 2D(Q||P^n)$$

and

$$\min_{\mathbf{P} \in \mathcal{P}(P,Q)} E_{\mathbf{P}}\Big(\sum_{i=1}^{n} \mathbf{P}^2(X_i \neq Y_i \,|\, Y) \Big) \leq 2D(Q||P^n)$$

If we use this coupling result we can conclude that

$$E_Q(f) - E_{P^n}(f) \leq \sqrt{2vD(Q||P^n)}$$

And the same inequality holds for $-f$ just conditioning by X instead of Y. What about concentration then?

From transportation to concentration

The key is the following lemma which tells us that the preceding inequality is a way of « encoding » sub-gaussianity

Lemma

Let Z be some integrable random variable. Given $v > 0$, the two following assertions are equivalent.

1) For all $\lambda \in \mathbb{R}^+$, $\log E_P(e^{\lambda(Z-E_PZ)}) \leq \frac{\lambda^2}{2v}$

2) For all $Q \ll P$, $E_Q(Z) - E_P(Z) \leq \sqrt{2vD(Q\|P)}$

Proof?

It is a direct consequence of the

Variational formula for entropy

Let Y be some real valued random variable,

$$\log E_P\left(e^Y\right) = \sup_{Q \ll P} E_Q(Y) - D(Q||P)$$

Indeed, given positive numbers a and v let us start from the following elementary formula:

$$\inf_{\lambda>0} \left(\frac{a}{\lambda} + \frac{\lambda v}{2}\right) = \sqrt{2av}$$

Use this formula with $a = D(Q\|P)$ allows to rewrite assertion 2) as

$$E_Q Z - E_P Z \leq \frac{D(Q\|P)}{\lambda} + \frac{\lambda v}{2}$$

For all positive λ and all $Q \ll P$ or equivalently

$$\sup_{Q \ll P} \lambda(E_Q(Z) - E_P(Z)) - D(Q\|P) \leq \lambda^2 v/2$$

which is exactly equivalent to 1) because of the variational formula ■

Conclusion: a function $Z = \zeta(X_1, \ldots, X_n)$ of independent variables with ζ satisfying $\mathscr{C}_v$ is sub-Gaussian with variance factor v

This means that for all λ

$$\log E(e^{\lambda(Z-EZ)}) \leq \frac{\lambda^2}{2v}$$

And of course by Chernoff's inequality sub-Gaussian tail bounds follow

$$P\{Z - EZ \geq t\} \vee P\{EZ - Z \geq t\} \leq e^{-\frac{t^2}{2v}}$$

Ultimately these tail bounds are the same as in the Gaussian case for a $\sqrt{v}$ -Lipschitz function.

Bonus (teasing for Anna's talk): A proof of Talagrand's convex distance inequality can be derived from this result just because $d_T(\,.\,, A)$ itself satisfies to the weak bounded differences condition $\mathscr{C}_1$.

Lecture 2

Back to the model selection issue

In the Euclidean space $\mathbb{R}^n$, one observes the random vector

$$Y = f + \sigma\epsilon$$

where the variables $\epsilon_j, 1 \leq j \leq n$ are i.i.d. centered and normalized rv's . σ is is the level of noise which is assumed to be known. f is the unknown mean vector to be estimated.

We consider some collection $(S_m)_{m\in\mathcal{M}}$ of linear subspaces of $\mathbb{R}^n$). To each model S_m corresponds the LSE $\hat{f}_m$ defined on it, in other words $\hat{f}_m$ is merely the orthogonal projection of Y onto S_m. Denoting by f_m the orthogonal projection of f onto S_m, the quality of model S_m is reflected by the quadratic risk of $\hat{f}_m$

$$E_f\|f - \hat{f}_m\|^2$$

Unbiased risk estimation principle

Here the quadratic risk is explicitly computable

$$E_f\|\hat{f}_m - f\|^2 - \|f\|^2 = -\|f_m\|^2 + \sigma^2 D_m \qquad (*)$$

and Mallow's criterion

$$\mathrm{crit}(m) = -\|\hat{f}_m\|^2 + 2\sigma^2 D_m$$

is merely an *unbiased* estimator of $(*)$

This analysis is performed in expectation, for each given model. What happens if $\mathscr{M}$ is huge?

Is it correct in any case? How to correct it if it breaks down?

The key is to specify how close is $\|\hat{f}_m\|^2$ to its expectation $\|f_m\|^2 + \sigma^2 D_m$, *uniformly* with respect to $m \in \mathcal{M}$.

Essentially, this amounts to understand the behavior of

$$\chi_m^2 = \| \Pi_m(\epsilon) \|^2 = \sum_{1 \le j \le D_m} \langle \epsilon, \phi_j^{(m)} \rangle^2$$

This a task that one can perform in different ways in the Gaussian case since this quantity follows a chi-square distribution.
In the non-Gaussian case it is no longer a sum of independent variables and this is why concentration tools are interesting here.

Back to suprema of Rademacher processes

This quantity of interest in our statistical problem can be written as

$$\chi_m = \sup_{b \in S_m, \|b\| \leq 1} \langle b, \epsilon \rangle$$

In this case $E(\chi_m^2) = D_m$ and a Poincaré type inequality can be used to show that $\mathrm{Var}(\chi_m) \leq 2$. Combining this with the preceding tail bounds leads to the following controls. Except on a set with probability less than e^{-x}

$$\chi_m \leq \sqrt{D_m} + 2\sqrt{2x}$$

And similarly

$$\chi_m \geq \sqrt{(D_m - 2)_+} - 2\sqrt{2x}$$

The needed upper and lower tail bounds are on the shelf!

5. Cut-off phenomena for penalized model selection

Coming back to the model selection issue, one can prove two complementary results that help the understanding of penalized least-squares criteria in a sharp way.

Those results will have the same flavor as those proved 25 years ago by Birgé and M. in the Gaussian case except that we have relaxed here the Gaussian assumption.

Upper tails in action

Model selection Theorem 1

Let $(x_m)_{m\in\mathcal{M}}$ be a family of non negative numbers such that

$$\sum_{m\in\mathcal{M}} \exp(-x_m) = \Sigma < \infty$$

Let $K > 1$ be given and assume that $\mathrm{pen}(m) \geq K\sigma^2(\sqrt{D_m} + 2\sqrt{2x_m})^2$ for all $m \in \mathcal{M}$. Let $\hat{m}$ minimizing the penalized least-squares criterion

$$\mathrm{crit}(m) = -\|\hat{f}_m\|^2 + \mathrm{pen}(m)$$

over $m \in \mathcal{M}$. The corresponding penalized least-squares estimator $\hat{f}_{\hat{m}}$ satisfies to the following risk bound

$$E_f\|\hat{f}_{\hat{m}} - f)\|^2 \leq C(K)\Big(\inf_{m\in\mathcal{M}}(\|f - f_m\|^2 + \mathrm{pen}(m)) + (1+\Sigma)\sigma^2\Big)$$

where $C(K)$ depends only on K.

Penalized least squares you say?

One could be surprised to see a non-positive quantity appear in the « least-squares » penalized criterion. It is just an artifact, because the identity

$$\|Y - \hat{f}_m\|^2 - \|Y\|^2 = -\|\hat{f}_m\|^2$$

ensures that substituting $\|Y - \hat{f}_m\|^2$ to $-\|\hat{f}_m\|^2$ in the definition of the criterion is painless.

Choice of the weights and link with the oracle

As in the original work of Birgé and M. , the weights have some Bayesian flavor since it plays the role of a prior finite measure on the list of models. But one can choose them in a way that enlightens the price to pay for redundancy of models with the same dimension.

Choice of the weights

Typical choice $x_m = x(D_m)$

Then

$$\sum_{m \in \mathcal{M}} e^{-x_m} = \sum_{D \geq 1} |\{m / D_m = D\}| \, e^{-x(D)}$$

choosing $x(D) = \alpha D + \beta_g |\{m / D_m = D\}|$

$$\sum_{m \in \mathcal{M}} e^{-x_m} \leq \frac{e^{-\alpha}}{1 - e^{-\alpha}} = \frac{1}{e^{\alpha} - 1}$$

Take as a penalty

$$pen(m) = K\,\sigma^2\left(\sqrt{D_m} + 2\sqrt{2x(D_m)}\right)^2$$

Risk bound

$$\mathbb{E}\,\|f - f_{\hat{m}}\|^2 \le C'(K) \inf_{D \ge 1}\left(b_D^2(f) + \sigma^2(D + x(D))\right)$$

$b_D^2(f)$: $\inf_{D_m = D} \|f - f_m\|^2$

gain with redundancy

$\sigma^2(D + x(D))$: price to pay for redundancy

In particular if $\kappa(\mathcal{D}) \leq L \mathcal{D}$

$$\mathbb{E}\|f - \hat{f}_{\hat{m}}\|^2 \leq C(K, L) \inf_{m \in \mathcal{M}} \mathbb{E}\|f - \hat{f}_m\|^2$$

oracle risk

This is typically the case when there is at most one model per dimension.

Sketch of proof of Theorem 1

We start from

$$\|y - \hat{f}_{\hat{m}}\|^2 + \text{pen}(\hat{m}) \leq \|y - \hat{f}_m\|^2 + \text{pen}(m)$$

$$\leq \|y - f_m\|^2 + \text{pen}(m)$$

using $y = f + \sigma\varepsilon$ leads to

$$\|y - g\|^2 = \|f - g\|^2 + 2\langle f - g, \varepsilon \rangle + \sigma^2 \|\varepsilon\|^2$$

And finally

$$\left[\|f - \hat{f}_{\hat{m}}\|^2 \leq \|f - f_m\|^2 + \text{pen}(m) + 2\sigma \langle \hat{f}_{\hat{m}} - f_m \rangle - \text{pen}(\hat{m}) \right] \quad (*)$$

Time to work!

Let us introduce

$$\forall_{m,m'} = \sup_{g \in S_{m'}} \frac{\langle g - f_m, \varepsilon \rangle}{\|f - f_m\| + \|g - f\|}$$

By definition

$$2\sigma \langle \hat{f}_{\hat{m}} - f_m, \varepsilon \rangle$$

$$\leq 2\sigma \left(\|f - f_m\| + \|\hat{f}_{\hat{m}} - f\| \right) \forall_{m,\hat{m}}$$

and using repeatedly $2ab \leq a^2 + b^2$

(*) leads to

$$\gamma \|f - \hat{f}_m\|^2 \leq \gamma' \|f - f_m\|^2 + \text{pen}(m) + \left(\frac{1+\eta}{1-\eta} \right) \sigma^2 \forall_{m,m'}^2 - \text{pen}(m')$$

(**)

It remains to control $\chi^2_{m,m'}$.
If $m = m'$ and $f \in S_m$ this quantity is exactly a pseudo chi-square. But still we deal we a supremum of Rademacher process and adapting the arguments that we used before still provides something neat:

$$\chi_{m,m'} \leq 1 + \sqrt{D_{m'}} + 2\sqrt{2x}$$

except on a set of probability less than e^{-x}.

To obtain the final result it remains to use a union bound and some tedious algebra starting from (**)

Choosing γ adequately (as a function of K) leads to

$$\gamma \|f - \hat{f}_m\|^2 \leq \gamma^{-1} \|f - f_m\|^2 + \mathrm{pen}(m) + \frac{(1+\gamma)^2}{\gamma(1-\gamma)} \sigma^2 \left(1 + 2\sqrt{2x}\right)^2$$

except on a set of probability $\leq \sum e^{-x}$

Integrating this bound completes the proof ∎

If we want to analyze the sharpness of the constraint $K > 1$ in the preceding Theorem, one has to take some asymptotic point of view in which we explicitly allow the collection of models $\mathscr{M}_N$ to depend on N and let N (and therefore n) tends to infinity. More precisely we assume that for some model S_{m_N} with dimension N one has $S_m \subseteq S_{m_N}$. If the number of models per dimension is sub-exponential and if we take a penalty of the form

$$\text{pen}(m) = K\sigma^2 D_m$$

the preceding theorem gives a positive result as soon as $K > 1$. What about $K < 1$?

The answer is that the criterion explodes!

Lower tails in action

Model selection Theorem 2

Assume that $N^{-1}\log|\mathscr{M}_N| \to 0$. Let $K < 1$ be given.
Let $\hat{m}$ minimizing the penalized least-squares criterion

$$-\|\hat{f}_m\|^2 + K\sigma^2 D_m$$

over $m \in \mathscr{M}_N$. For any positive δ, there exists some $N_0(K,\delta)$ (which depends neither on σ nor on f) such that whenever $N \geq N_0(K,\delta)$,

$$P\{D_{\hat{m}} \geq N/2\} \geq 1-\delta$$

Meanwhile, if N is large enough

$$E_f\|\hat{f}_{\hat{m}} - f\|^2 \geq \|f - f_{m_N}\|^2 + \sigma^2 N/4$$

Sketch of proof of Theorem 2

To make our life easier assume that $f=0$. In this pure noise situation the criterion can be written as:

$$\sigma^{-2}(\mathrm{crit}(m_N) - \mathrm{crit}(m_1)) = -(\mathcal{K}^2_{m_N} - \mathcal{K}^2_{m}) + K(N - D_m)$$

The point now is that the negative term in this expression has expectation $= -(N - D_m)$

Therefore, using the lower tail part of the concentration inequality $\chi^2_{m_N} - \chi^2_m$ stays above

$$(1-\eta)(N - D_m)$$

for all models m with $D_m \leq \frac{N}{2}$ with probability $\geq 1-\delta$.

So that there is room to commute the computation in expectation into a probability bound.

$$\Rightarrow \sigma^{-2}(\text{crit}(m_N) - \text{crit}(m)) \leq \left(\frac{1-\chi^2}{2}\right)(N - D_m)$$

with probability $\geq 1-\delta$ $\forall m$ with $D_m \leq \frac{N}{2}$.

Illustration

Let us consider the simplest situation for which there is only one model per dimension. The typical situation of this kind is as follows. Take some orthonormal system $\{\phi_j, 1 \leq j \leq N\}$ and take as a collection of models $(S_D)_{1 \leq D \leq N}$, where S_D is spanned by $\{\phi_j, 1 \leq j \leq D\}$. Combining the two Theorems above tells us that for the criterion

$$-\|\hat{f}_D\|^2 + K\sigma^2 D$$

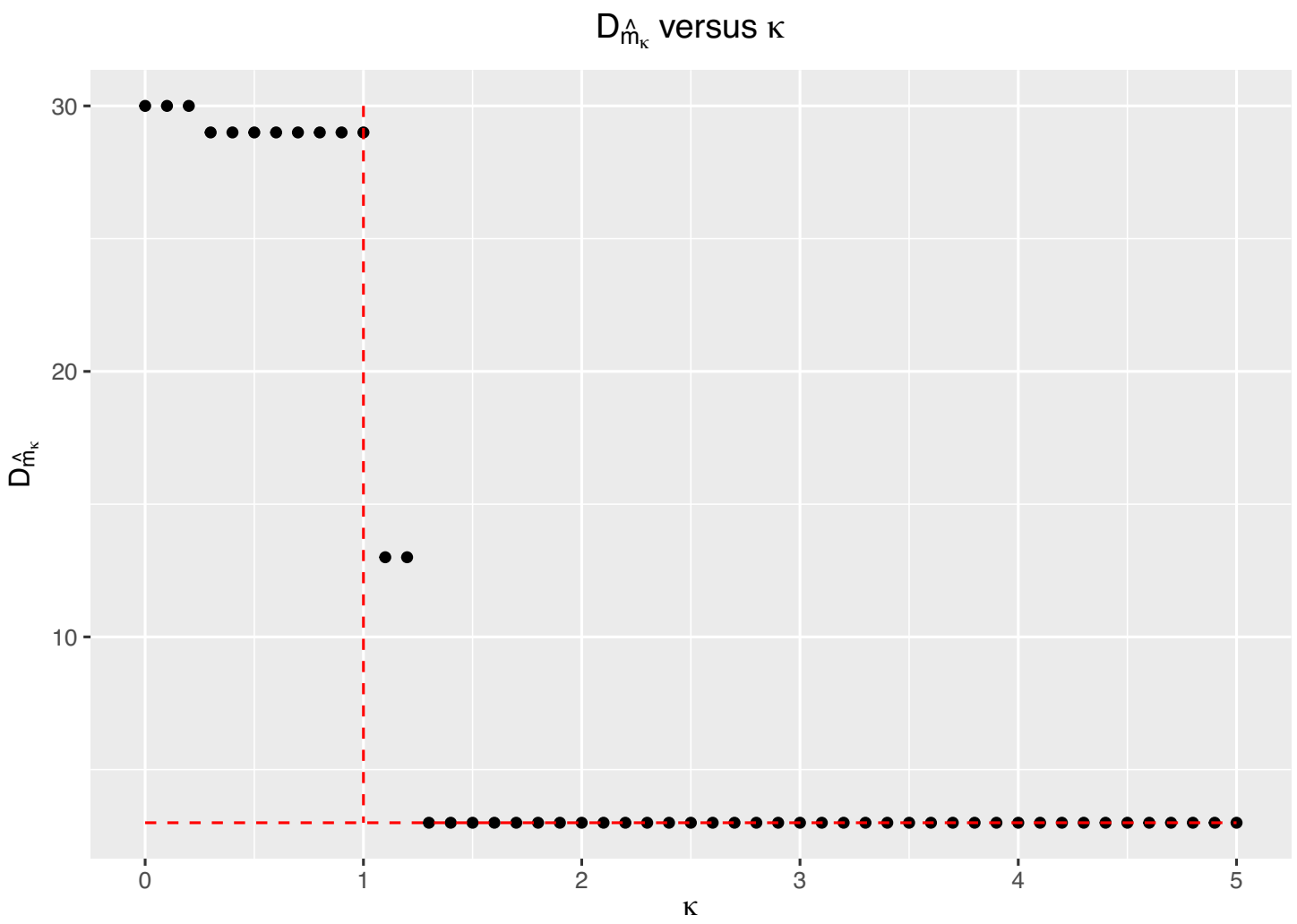
there is some cut-off at the critical value $K = 1$

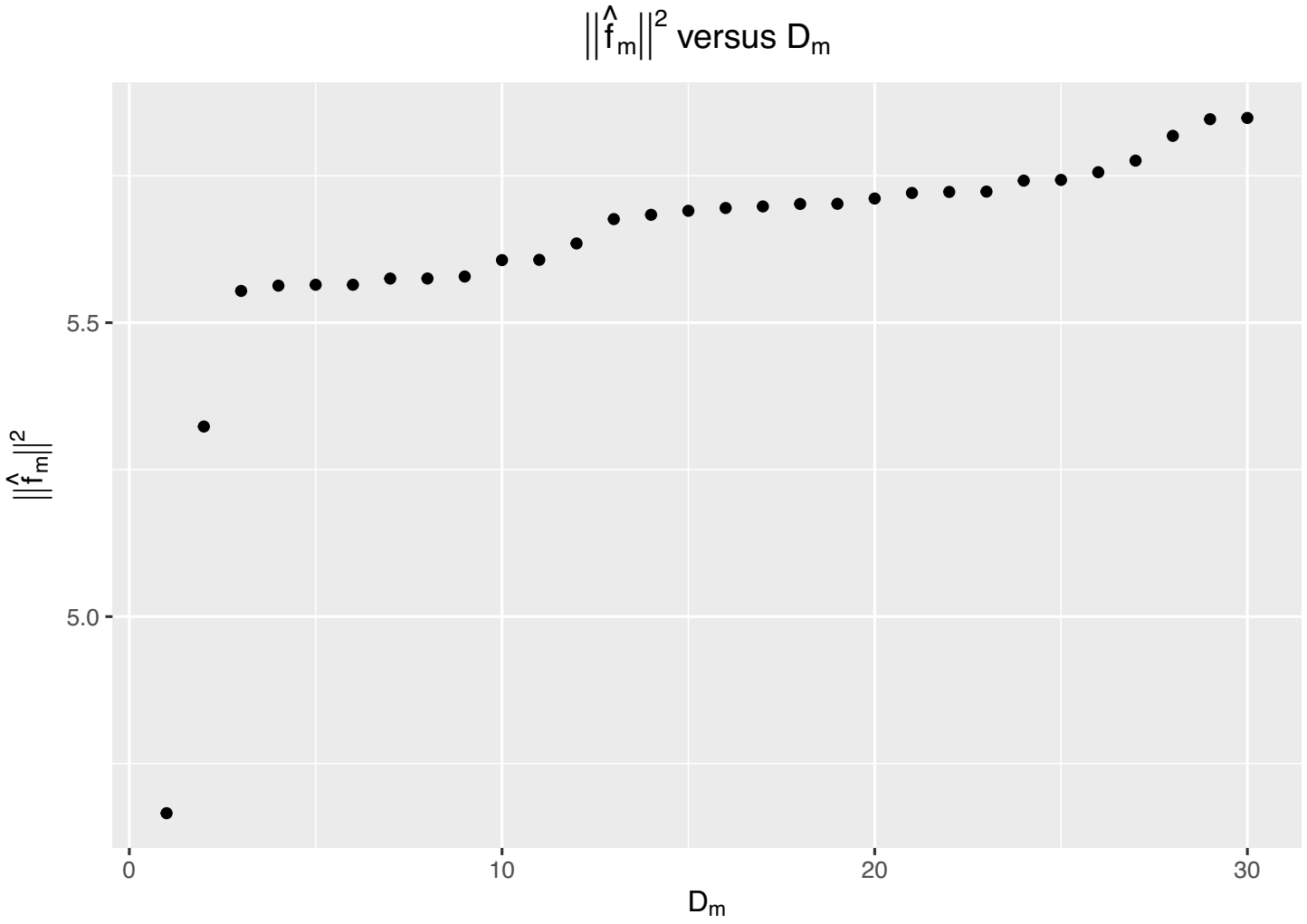
- Above this value the criterion chooses a sensitive model (it can be proved in this case that $K = 2$ is asymptotically the best value)
- Below this value the criterion explodes in the sense that that it chooses large dimensional models with high probability.

Connection with non parametric estimation

This framework is especially meaningful in the situation where one wants to estimate some function f on $[0,1]$ from noisy observations $Y_i = f(i/n) + \sigma\epsilon_i,\ 1 \leq i \leq n$. In this case the Fourier basis can be used to build the above nested family of models and having in mind that n is large makes sense.
Choosing a good model here means approximating the function f from the data by a convenient trigonometric polynomial.

The cut-off phenomenon can be exploited to perform fully automatic model selection without knowing the level of noise, just by identifying the minimal penalty from the data and multiplying it by 2 to perform the final model selection.


$D_{\hat{m}_\kappa}$ versus κ
30
20
10
$D_{\hat{m}_\kappa}$
0
1
2
3
4
5
κ


$\|\hat{f}_m\|^2$ versus D_m
$\|\hat{f}_m\|^2$
5.5
5.0
0
10
20
30
D_m