Kernel change-point detection

Alain Celisse

[1]UMR 8524 CNRS - Université Lille 1

[2]Modal INRIA team-project

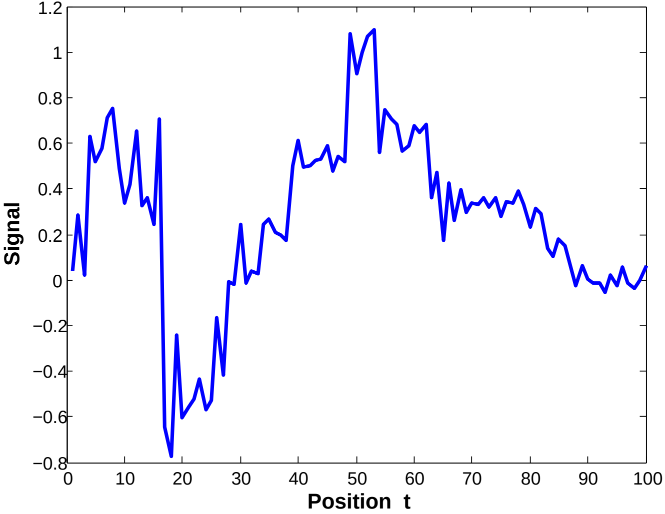
[3]SSB Group, Paris

joint work with Sylvain Arlot and Zaïd Harchaoui

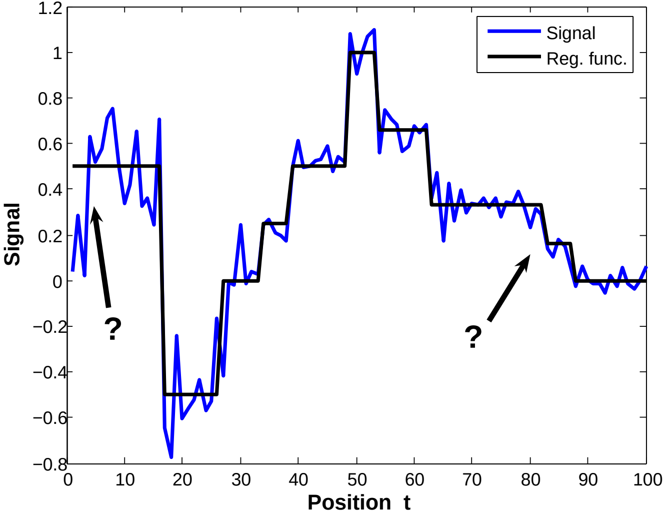
Workshop: "Recent advances in changepoints analysis"

Warwick University, March 28, 2012

1-D signal (example)

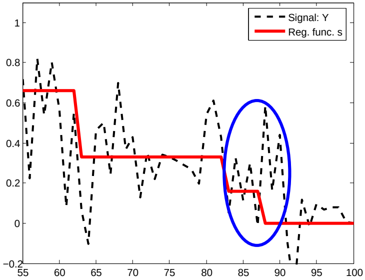


1-D signal (example): Find abrupt changes in the mean



Estimation rather than identification

Fact:
With finite sample, it is impossible to recover change-point in noisy regions.



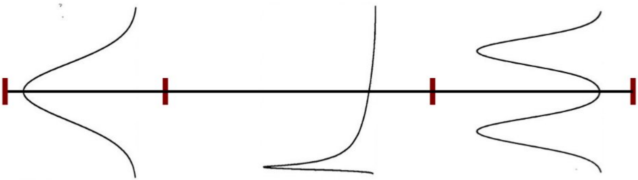
Purpose:

Estimate the regression function.

Idea:

$\longrightarrow$ Without too strong noise, recover true change-points.

Example 1: Changes in the distribution

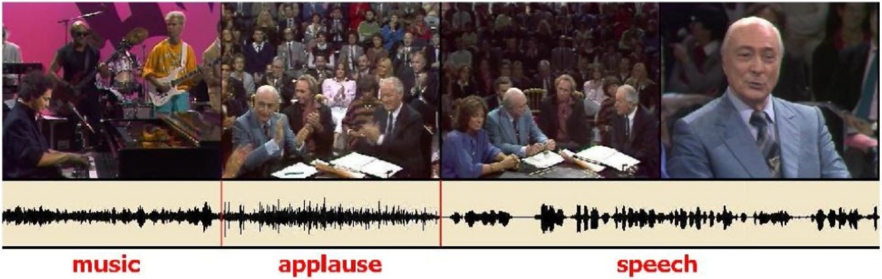


Description:

- Observations generated along the time.
- Observation distribution is piecewise constant along the time.
- Observations have the same mean and variance.

$\longrightarrow$ Detecting changes in the mean and variance is useless.

Example 2: Working with non-vectorial objects



Description:

- Video sequences from "Le grand échiquier", 70s-80s French talk show.
- At each time, one observes an image (high-dimensional).
- Each image is summarized by a histogram.

Example 2: Working with non-vectorial objects

- Preprocessing images (patches in yellow).
- Each histogram bin corresponds to a patch.

Non-vectorial object:
Histograms with D bins belong to

$$\left\{(p_1, \ldots, p_D) \in [0,1]^D,\ \sum_{i=1}^{D} p_i = 1\right\}.$$

$\longrightarrow$ Algorithms for vectorial data are useless.

Detect abrupt changes. . .

General purposes:

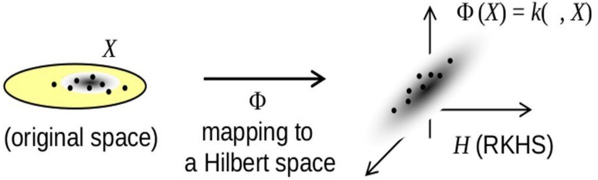
1. Detect changes in the whole distribution (not only in the mean)

2. High-dimensional data of different nature:
   - Vectorial: measures in $\mathbb{R}^d$, curves (sound recordings,. . .)
   - Non vectorial: phenotypic data, graphs, DNA sequence,. . .
   - Both vectorial and non vectorial data.

3. Efficient algorithm allowing to deal with large data sets

Kernel and Reproducing Kernel Hilbert Space (RKHS)

- $X_1, \ldots, X_n \in \mathcal{X}$: initial observations.
- $k(\cdot, \cdot) : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$: reproducing kernel ($\mathcal{H}$: RKHS).
- $\phi(\cdot) : \mathcal{X} \to \mathcal{H}$ s.t. $\phi(x) = k(x, \cdot)$: canonical feature map.
- $< \cdot, \cdot >_{\mathcal{H}}$: inner-product in $\mathcal{H}$.



Asset:

Enables to work with high-dimensional heterogeneous data.

Model

Mapping of the initial data

$$\forall 1 \leq i \leq n, \quad Y_i = \phi(X_i) \in \mathcal{H} \enspace .$$

$$\longrightarrow (t_1, Y_1), \ldots, (t_n, Y_n) \in [0,1] \times \mathcal{H} : \quad \text{independent} \enspace .$$

Model

Mapping of the initial data

$$\forall 1 \leq i \leq n, \quad Y_i = \phi(X_i) \in \mathcal{H} \ .$$

$$\longrightarrow (t_1, Y_1), \ldots, (t_n, Y_n) \in [0,1] \times \mathcal{H} : \quad \text{independent} \ .$$

Mean element

The mean element of P_{X_i} (distribution of X_i) is $\mu_i^\star$:

$$< \mu_i^\star, f >_{\mathcal{H}} = \mathbb{E}_{X_i}\left[< \phi(X_i), f >_{\mathcal{H}} \right], \quad \forall f \in \mathcal{H} \ .$$

Fact:

With characteristic kernels,

$$P_{X_i} \neq P_{X_j} \quad \Rightarrow \quad \mu_i^\star \neq \mu_j^\star \ .$$

Model

$$\forall 1 \le i \le n, \qquad Y_i = \mu_i^\star + \varepsilon_i \quad \in \mathcal{H} \ ,$$

where

- $\mu_i^\star \in \mathcal{H}$: mean element of P_{X_i} (distribution of X_i)
- $\forall i, \quad \varepsilon_i := Y_i - \mu_i^\star, \quad$ with $\quad \mathbb{E}\varepsilon_i = 0, \quad v_i := \mathbb{E}\left[\|\varepsilon_i\|_{\mathcal{H}}^2 \right]$.

Model

$$\forall 1 \leq i \leq n, \qquad Y_i = \mu_i^\star + \varepsilon_i \quad \in \mathcal{H} \ ,$$

where

- $\mu_i^\star \in \mathcal{H}$: mean element of P_{X_i} (distribution of X_i)
- $\forall i, \quad \varepsilon_i := Y_i - \mu_i^\star, \quad$ with $\quad \mathbb{E}\varepsilon_i = 0, \quad v_i := \mathbb{E}\left[\|\varepsilon_i\|_{\mathcal{H}}^2 \right]$.

Assumptions

1. $\max_i \|Y_i\|_{\mathcal{H}} \leq M \quad a.s. \qquad$ **(Db)** .
2. $\max_i v_i \leq v_{\max} \qquad$ **(Vmax)** .
3. $\mu^\star = (\mu_1^\star, \ldots, \mu_n^\star)' \in \mathcal{H}^n$: piecewise constant.

$$\|\mu^\star - \mu\|^2 := \sum_{i=1}^n \|\mu_i^\star - \mu_i\|_{\mathcal{H}}^2.$$

Goal: $\longrightarrow$ Estimate $\mu^\star$ to recover change-points.

Model selection

Models:

- $\mathcal{M}_n = \{m, \text{ segmentation of } \{1, \ldots, n\}\}, \quad D_m = \mathrm{Card}(m).$
- $S_m = \{\mu : (t_1, \ldots, t_n) \to \mathcal{H}, \text{ piecewise const. on } (I_\lambda)_{\lambda \in m}\},$
$(I_\lambda)_{\lambda \in m}$: $I_1 = [0, t_{\lambda_1}], \; I_2 =]t_{\lambda_1}, t_{\lambda_2}], \ldots, \; I_{D_m} =]t_{\lambda_{D_m - 1}}, 1].$

Strategy:

$$(S_m)_{m \in \mathcal{M}_n} \longrightarrow (\widehat{\mu}_m)_{m \in \mathcal{M}_n} \longrightarrow \widehat{\mu}_{\widehat{m}} \quad ???$$

Model selection

Models:

- $\mathcal{M}_n = \{m, \text{ segmentation of } \{1, \ldots, n\}\}, \quad D_m = \mathrm{Card}(m)$.
- $S_m = \{\mu : (t_1, \ldots, t_n) \to \mathcal{H}, \text{ piecewise const. on } (I_\lambda)_{\lambda \in m}\}$, $(I_\lambda)_{\lambda \in m}$: $I_1 = [0, t_{\lambda_1}]$, $I_2 =]t_{\lambda_1}, t_{\lambda_2}], \ldots, I_{D_m} =]t_{\lambda_{D_m - 1}}, 1]$.

Strategy:

$$(S_m)_{m \in \mathcal{M}_n} \longrightarrow (\widehat{\mu}_m)_{m \in \mathcal{M}_n} \longrightarrow \widehat{\mu}_{\widehat{m}} \quad ???$$

Goal:

Oracle inequality (in expectation, or with large probability):

$$\left\|\mu^\star - \widehat{\mu}_{\widehat{m}}\right\|^2 \leq C \inf_{m \in \mathcal{M}_n} \left\{ \left\|\mu^\star - \widehat{\mu}_m\right\|^2 \right\} + r_n \quad .$$

Least-squares estimator

- Empirical risk minimizer over S_m (= model):

$$\widehat{\mu}_m \in \arg\min_{u \in S_m} \frac{1}{n} \sum_{i=1}^{n} \| u(t_i) - Y_i \|_{\mathcal{H}}^2 \left(=: \arg\min_{u \in S_m} P_n \gamma(u) \right) .$$

- Regressogram:

$$\widehat{\mu}_m = \sum_{\lambda \in m} \widehat{\beta}_\lambda \mathbb{1}_{I_\lambda} \qquad \widehat{\beta}_\lambda = \frac{1}{\mathrm{Card}\left\{ t_i \in I_\lambda \right\}} \sum_{t_i \in I_\lambda} Y_i .$$

Choose $D - 1$ change-points...

Assumption: (Harchaoui, Cappé (2007))

The number $D - 1$ of change-points is known.

Strategy:

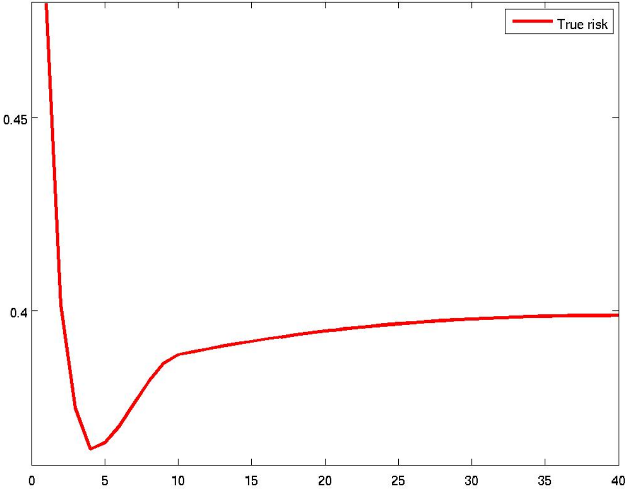
Choose $\widehat{m}(D)$ among $\{m \in \mathcal{M}_n,\ D_m = D\}$.

ERM algorithm:

$$\widehat{m}(D) = \widehat{m}_{\mathrm{ERM}}(D) = \mathrm{Argmin}_{m \mid D_m = D} \left\| Y - \widehat{\mu}_m \right\|^2 .$$

(dynamic programming).

Quality of the segmentations



Elementary calculations

Expectations $\left(v_\lambda = \frac{1}{\mathrm{Card}(\lambda)} \sum_{i \in \lambda} v_i \right)$

$$\mathbb{E}\left[\left\| \mu^\star - \widehat{\mu}_m \right\|^2 \right] = \left\| \mu^\star - \Pi_m \mu^\star \right\|^2 + \sum_{\lambda \in m} v_\lambda \enspace ,$$

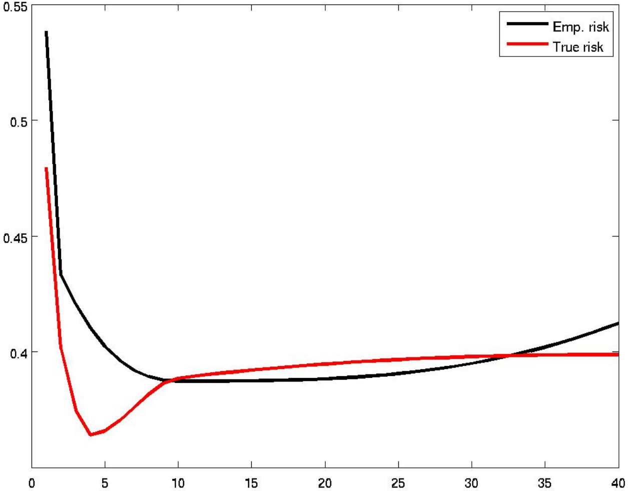
$$\mathbb{E}\left[\left\| Y - \widehat{\mu}_m \right\|^2 \right] = \left\| \mu^\star - \Pi_m \mu^\star \right\|^2 - \sum_{\lambda \in m} v_\lambda + Cste \enspace ,$$

(Π_m: orthog. proj. operator onto S_m).

Conclusion:

$\longrightarrow$ ERM prefers models with large $\sum_{\lambda \in m} v_\lambda$ (overfitting).

Overfitting illustration



Choose the number of change-points

From $\{\widehat{\mu}_{\widehat{m}_D}\}_D$, choose D amounts to choose the "best model".

Ideal penalty:

$$m^* \in \mathrm{Argmin}_{m\in\mathcal{M}} \|\mu^\star - \widehat{\mu}_m\|^2 \quad \text{(oracle segmentation)}$$
$$= \mathrm{Argmin}_{m\in\mathcal{M}} \left\{\|Y - \widehat{\mu}_m\|^2 + \mathrm{pen}_{\mathrm{id}}(m)\right\} \quad ,$$

with $\quad \mathrm{pen}_{\mathrm{id}}(m) := 2\|\Pi_m\varepsilon\|^2 - 2 < (I - \Pi_m)\mu^\star, \varepsilon >.$

Strategy

1. Concentration inequalities for linear and quadratic terms.
2. Derive a tight upper bound $\mathrm{pen} \geq \mathrm{pen}_{\mathrm{id}}$ with high probability.

Previous work:

Birgé, Massart (2001): Gaussian assump. + real valued functions.

$\longrightarrow$ cannot be extended to Hilbert framework.

Concentration of the linear term

Theorem (Linear term)

Let us consider a segmentation m and assume $(\mathbf{Db}) - (\mathbf{Vmax})$ *hold true. Then for every $x > 0$ with probability at least $1 - 2e^{-x}$,*

$$\left| < \Pi_m \mu^\star - \mu^\star, \varepsilon > \right| \le \theta \left\| \Pi_m \mu^\star - \mu^\star \right\|^2 + \left(\frac{v_{max}}{\theta} + \frac{4M^2}{3} \right) x \enspace ,$$

for every $\theta > 0$.

Concentration of the quadratic term

Theorem (Quadratic term)

Assuming (**Db**)-(**Vmax**), *and*

$$\exists \kappa \geq 1, \quad 0 < \frac{M^2}{\kappa} \leq \min_i v_i \qquad \textbf{(Vmin)}$$

for every $m \in \mathcal{M}_n$, $x > 0$, *and* $\theta \in (0,1]$,

$$\left| \left\| \Pi_m \varepsilon \right\|^2 - \mathbb{E}\left[\left\| \Pi_m \varepsilon \right\|^2 \right] \right| \leq \theta \mathbb{E}\left[\left\| \Pi_m \mu^\star - \widehat{\mu}_m \right\|^2 \right] + \theta^{-1} L(\kappa) v_{max} x \ ,$$

with probability at least $1 - 2e^{-x}$, *where* $L(\kappa)$ *is a constant.*

Idea of the proof:

- Pinelis-Sakhanenko's inequality ($\left\| \sum_{i \in \lambda} \varepsilon_i \right\|_{\mathcal{H}}$).
- Bernstein's inequality (upper bounding moments)

Oracle inequality

Theorem

Let us assume (**Db**)-(**Vmin**)-(**Vmax**) *and for every* $x > 0$, *define*

$$\widehat{m} \in \operatorname{Argmin}_m \left\{ \|Y - \widehat{\mu}_m\|^2 + \mathrm{pen}(m) \right\} \quad ,$$

where $\mathrm{pen}(m) = v_{max} D_m \left[C_1 \ln\left(\frac{n}{D_m}\right) + C_2 \right]$ *for constants* $C_1, C_2 > 0$. *Then, there exists an event of probability larger than* $1 - 2e^{-x}$ *on which*

$$\|\mu^\star - \widehat{\mu}_{\widehat{m}}\|^2 \leq \Delta_1 \inf_m \left\{ \|\mu^\star - \widehat{\mu}_m\|^2 + \mathrm{pen}(m) \right\} + \Delta_2 \quad ,$$

where $\Delta_1 \geq 1$ *and* $\Delta_2 > 0$ *is a remainder term.*

Rk:
In Birgé, Massart (2001), $\mathrm{pen}(m) = \sigma^2 D_m \left[c_1 \ln\left(\frac{n}{D_m}\right) + c_2 \right]$.

Model selection procedure

Penalty:

$$\mathrm{pen}(m) = v_{max} D_m \left[C_1 \ln\left(\frac{n}{D_m}\right) + C_2 \right] = \mathrm{pen}(D_m) \quad .$$

Algorithm

1. For every $1 \leq D \leq D_{\max}$,

$$\widehat{m}_D \in \mathrm{Argmin}_{m,\ D_m = D} \left\{ \|Y - \widehat{\mu}_m\|^2 \right\} \quad ,$$

2. Define

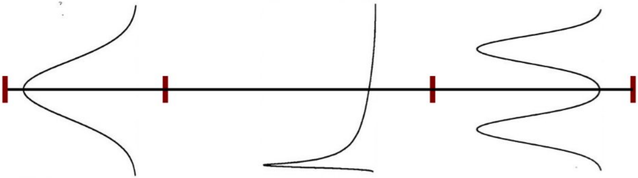
$$\widehat{D} = \mathrm{Argmin}_D \left\{ \left\|Y - \widehat{\mu}_{\widehat{m}_D}\right\|^2 + v_{max} D \left[C_1 \ln\left(\frac{n}{D}\right) + C_2 \right] \right\} \quad .$$

where C_1, C_2: computed by simulation experiments.

3. Final estimator:

$$\widehat{\mu}_{\widehat{m}} := \widehat{\mu}_{\widehat{m}_{\widehat{D}}} .$$

Changes in the distribution (synthetic data)

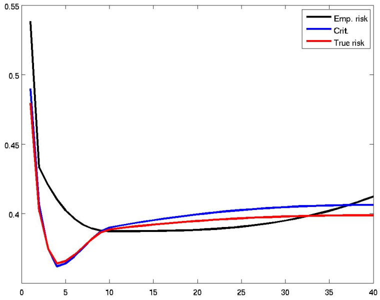


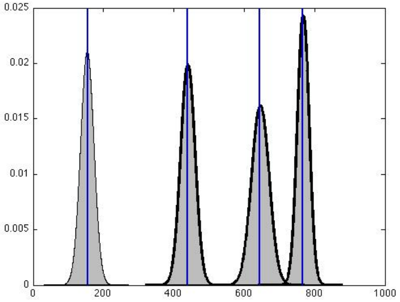
Description:

1. $n = 1\,000$, $D^* = 4$, $N_{rep} = 100$.
2. In each segment, observations generated according to one distribution within a pool of 10 distributions with same mean and variance.
3. Our kernel based approach enables to distinguish them (higher order moments)

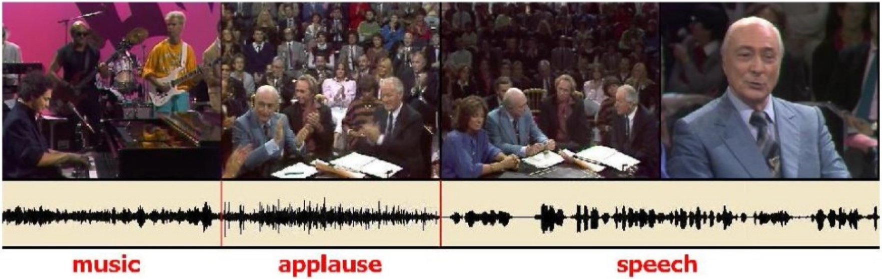
Changes in the distribution (synthetic data) (Cont'.)

Results





"Le grand échiquier", 70s-80s French talk show



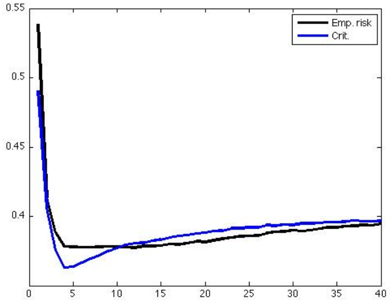
- Audio and video recordings.
- Audio: different situations can be distinguished from sound recordings (music, applause, speech,. . .).
- Video: different video scenes can be distinguished by their backgrounds or specific actions of people (clapping hands, discussing,. . .).

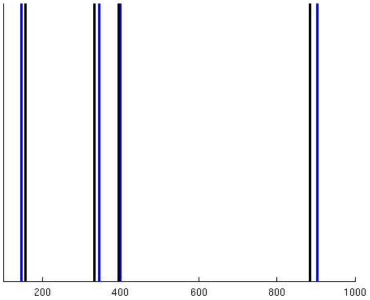
Video sequence

Description:

- $n = 10\,000$, $D^* = 4$.
- Each image summarized by a histogram with 1 024 bins.
- χ^2 kernel: $k_d(x, y) = \sum_{i=1}^{d} \frac{(x_i - y_i)^2}{x_i + y_i}$.

Results:





Concluding remarks

Open questions:

1. Relax the assumption on the variance.
2. Use resampling strategies (hetoroscedasticity).
3. Influence of the choice of kernel.

Preprint:

- ArXiv
- `http://www.math.univ-lille1.fr/~celisse/`

Concluding remarks

Open questions:

1. Relax the assumption on the variance.
2. Use resampling strategies (hetoroscedasticity).
3. Influence of the choice of kernel.

Preprint:

- ArXiv
- `http://www.math.univ-lille1.fr/~celisse/`

Thank you!

Sketch of proof

1. $\|\Pi_m \varepsilon\|^2 = \sum_{\lambda \in m} \frac{1}{n_\lambda} \left\| \sum_{i \in \lambda} \varepsilon_i \right\|_{\mathcal{H}}^2 = \sum_{\lambda \in m} T_\lambda$.
2. $\left\{ \left\| \sum_{i \in \lambda} \varepsilon_i \right\|_{\mathcal{H}}^2 \right\}_{\lambda \in m}$ are independent r.v. .
3. Bernstein's inequality to $\|\Pi_m \varepsilon\|^2 \quad (\star)$.
4. For every $q \geq 2$, upper bound of $\mathbb{E}\left[T_\lambda^q \right]$.
5. Pinelis-Sakhanenko's inequality on $\left\| \sum_{i \in \lambda} \varepsilon_i \right\|_{\mathcal{H}}$:

$$\forall x > 0, \quad \mathbb{P}\left[\left\| \sum_{i \in \lambda} \varepsilon_i \right\|_{\mathcal{H}} > x \right] \leq 2 \exp\left[-\frac{x^2}{2\left(\sigma_\lambda^2 + b_\lambda x\right)} \right] ,$$

with $b_\lambda = 2M/3$ and $\sigma_\lambda^2 = \sum_{i \in \lambda} v_i$.

Bernstein rather than Talagrand

Talagrand's inequality

$\|\Pi_m \varepsilon\| = \sup_{f \in B_n} < f, \Pi_m \varepsilon >= \sup_{f \in B_n} \sum_{i=1}^n < f_i, (\Pi_m \varepsilon)_i >_{\mathcal{H}}$

$$\mathbb{P}\left[\|\Pi_m \varepsilon\| \leq \mathbb{E}\left[\|\Pi_m \varepsilon\| \right] + \sqrt{2vx} + \frac{b}{3}x \right] \quad ,$$

with $v = \sum_{i=1}^n \sup_f \mathbb{E}\left(< f_i, (\Pi_m \varepsilon)_i >_{\mathcal{H}}^2\right) + 16b\mathbb{E}\left[\|\Pi_m \varepsilon\| \right]$.

Bernstein's inequality

$$\sigma^2 = \sup_f \sum_{i=1}^n \mathbb{E}\left(< f_i, (\Pi_m \varepsilon)_i >_{\mathcal{H}}^2\right) = \mathbb{E}\left[\|\Pi_m \varepsilon\|^2 \right] \quad .$$