# Inégalités de déviation pour des U-statistiques à valeurs dans un espace de Hilbert

**Séminaire "Probabilités et Statistique", Lille**
Davide Giraudo (Université de Strasbourg)

18 mars 2025

# Plan

# Introduction aux $U$-statistiques

Afin d'estimer par une moyenne empirique des paramètres qui s'expriment sous la forme $\mathbb{E}\left[h\left(\xi_1, \ldots, \xi_m\right)\right]$, où $\left(\xi_i\right)_{i \geqslant 1}$ est i.i.d. et $h \colon S^m \to \mathbb{R}$, $(S, \mathcal{S})$ est un espace mesurable, la $U$-statistique de noyau $h$, définie par

$$U_{m,n}(h) := \sum_{1 \leqslant i_1 < \cdots < i_m \leqslant n} h\left(\xi_{i_1}, \ldots, \xi_{i_m}\right), \quad n \geqslant m,$$

où $\left(\xi_j\right)_{j \geqslant 1}$ est i.i.d., fut introduite par **Hoeffding** (1948).

On peut ainsi prendre comme estimateur $U_{m,n}(h) / C_n^m$, qui est non biaisé.

**Objectif :** comprendre le comportement asymptotique de $\left(U_{m,n}(h)\right)_{n \geqslant m}$.

Les termes que l'on somme sont identiquement distribués. Par exemple, pour $m = 2$, pour tous $i < j$, $h\left(\xi_i, \xi_j\right)$ a la même loi que $h\left(\xi_1, \xi_2\right)$. La $U$-statistique $U_{2,n}(h)$ est alors la somme partielle des variables aléatoires dépendantes $D_j := \sum_{i=1}^{j-1} h\left(\xi_i, \xi_j\right)$.

# Exemples de noyaux

1. Estimateur de variance : $h(x,y) = (x-y)^2/2$.
2. Moyenne de Gini : $h(x,y) = |x-y|$.
3. Estimateur de Grassberger-Procaccia : pour $t > 0$ fixé, $h(x,y) = \mathbf{1}\{|x-y| \leqslant t\}$.
4. $h(x,y) = \operatorname{sgn}(x-y)$, avec $\operatorname{sgn}(t) = \mathbf{1}_{t>0} - \mathbf{1}_{t<0}$.
5. Mesure de dissymétrie :

   $$h(x_1,x_2,x_3) = \operatorname{sgn}(2x_1 - x_2 - x_3) + \operatorname{sgn}(2x_2 - x_1 - x_3) + \operatorname{sgn}(2x_3 - x_1 - x_2).$$

6. Distance de covariance : étant donné un espace métrique $(S,d)$, soit $f : S^4 \to \mathbb{R}$ défini par

   $$f(z_1,z_2,z_3,z_4) = d(z_1,z_2) - d(z_1,z_3) - d(z_2,z_4) + d(z_3,z_4),$$

   le noyau $g : \left(S^2\right)^6 \to \mathbb{R}$ par

   $$g((x_1,y_1),\ldots,(x_6,y_6)) = f(x_1,x_2,x_3,x_4)\, f(y_1,y_2,y_5,y_6)$$

   ainsi que sa version symétrisée

   $$h((x_1,y_1),\ldots,(x_6,y_6)) = \frac{1}{6!}\sum_{\sigma\in\mathcal{S}_6} g\left(\left(x_{\sigma(1)},y_{\sigma(1)}\right),\ldots,\left(x_{\sigma(6)},y_{\sigma(6)}\right)\right).$$

# Décomposition d'Hoeffding (1)

Soit $h\colon S^2 \to \mathbb{R}$ une fonction mesurable et soit $(\xi_i)_{i\geqslant 1}$ une suite i.i.d. Soit

$$U_{2,n}(h) = \sum_{1\leqslant i<j\leqslant n} h(\xi_i, \xi_j).$$

On définit $\theta := \mathbb{E}[h(\xi_1, \xi_2)]$,

$$h_1(x) = \mathbb{E}[h(x, \xi_2)] - \theta, \quad h_2(y) = \mathbb{E}[h(\xi_1, y)] - \theta,$$

$$h_3(x, y) = h(x, y) - h_1(x) - h_2(y) - \theta.$$

Alors

$$U_{2,n}(h) = C_n^2\theta + \sum_{i=1}^{n-1}(n-i)\, h_1(\xi_i) + \sum_{j=2}^{n}(j-1)\, h_2(\xi_j) + \sum_{1\leqslant i<j\leqslant n} h_3(\xi_i, \xi_j)$$

et

$$\mathbb{E}[h_3(\xi_i, \xi_j) \mid \xi_1, \ldots, \xi_{j-1}] = \mathbb{E}[h_3(\xi_i, \xi_j) \mid \xi_{i+1}, \ldots, \xi_n] = 0.$$

# Décomposition d'Hoeffding (2)

**Définition**

*On dit que le noyau $h\colon S^m \to \mathbb{R}$ est dégénéré par rapport à la suite i.i.d. $(\xi_i)_{i\geqslant 1}$ si pour tout $\ell \in [\![1, m]\!]$,*

$$\mathbb{E}\left[h\left(\xi_1, \ldots, \xi_m\right) \mid \sigma\left(\xi_k, k \in [\![1, m]\!] \setminus \ell\right)\right] = 0.$$

La décomposition d'Hoeffding est valide pour les $U$-statistiques de tout ordre.

Si $h$ est symétrique, c'est-à-dire, $h\left(x_{\sigma(1)}, \ldots, x_{\sigma(m)}\right) = h\left(x_1, \ldots, x_m\right)$ pour toute bijection $\sigma\colon [\![1, m]\!] \to [\![1, m]\!]$, alors

$$U_{m,n}\left(h\right) = C_n^m \sum_{c=0}^{m} \frac{C_m^c}{C_n^c} U_{c,n}\left(h_c\right),$$

où $h_c\colon S^c \to \mathbb{R}$ est dégénéré par rapport à $(\xi_i)_{i\geqslant 1}$.

**Définition**

*On dit que $h$ est dégénéré d'ordre $d$ si la décomposition d'Hoeffding s'écrit*

$$U_{m,n}\left(h\right) = C_n^m \sum_{c=d}^{m} \frac{C_m^c}{C_n^c} U_{c,n}\left(h_c\right)$$

# Théorème limite central

**Hoeffding** (1962) a montré que si $h(\xi_1, \ldots, \xi_m) \in \mathbb{L}^2$, alors

$$\sqrt{n}\left(\frac{U_{m,n}(h) - \mathbb{E}[U_{m,n}(h)]}{C_n^m}\right) \to \mathcal{N}\left(0, m^2 \operatorname{Var}(h_1(\xi_1))\right) \text{ en loi.}$$

Notons qu'il est possible que la variable aléatoire $h_1(\xi_1)$ soit constante. Dans ce cas, une autre normalisation doit être prise pour avoir une convergence vers une variable aléatoire non-dégénérée.

# Théorème limite central

**Hoeffding** (1962) a montré que si $h(\xi_1, \ldots, \xi_m) \in \mathbb{L}^2$, alors

$$\sqrt{n}\left(\frac{U_{m,n}(h) - \mathbb{E}[U_{m,n}(h)]}{C_n^m}\right) \to \mathcal{N}\left(0, m^2 \operatorname{Var}(h_1(\xi_1))\right) \text{ en loi.}$$

Notons qu'il est possible que la variable aléatoire $h_1(\xi_1)$ soit constante. Dans ce cas, une autre normalisation doit être prise pour avoir une convergence vers une variable aléatoire non-dégénérée.

Pour $m = 2$ et $h(\xi_i, \xi_j) = \xi_i \xi_j$ avec $\xi_i$ centrée et de variance $1$,

$$U_{2,n}(h) = \frac{1}{2}\left(\sum_{i=1}^{n} \xi_i\right)^2 - \frac{1}{2}\sum_{i=1}^{n} \xi_i^2$$

donc

$$\frac{2}{n} U_{2,n}(h) \to \mathcal{N}^2 - 1,$$

où $\mathcal{N}$ est de loi normale centrée réduite.

# Théorème limite central fonctionnel

Soit le processus sommes partielles

$$\mathcal{U}_{2,n,h}^{\mathrm{pl}}(t) = \begin{cases} \sum_{1 \leqslant i < j \leqslant k} h(\xi_i, \xi_j) & \text{si } t = \frac{k}{n} \text{ pour un certain } k \in [\![0, n]\!], \\ \text{interpolation linéaire} & \text{sur } \left]\frac{k}{n}, \frac{k+1}{n}\right[, k \in [\![0, n-1]\!]. \end{cases}$$

**Mandelbaum et Taqqu** (1984) :

- Si $\mathbb{E}\left[\mathbb{E}\left[h(\xi_1, \xi_2) \mid \xi_1\right]^2\right] = \sigma^2 > 0$, alors

$$n^{-3/2}\left(\mathcal{U}_{2,n,h}^{\mathrm{pl}}(t) - \mathbb{E}\left[\mathcal{U}_{2,n,h}^{\mathrm{pl}}(t)\right]\right) \to \sigma W(t) \text{ en loi dans } C[0,1],$$

  où $W$ est un mouvement brownien standard.

- Si $\mathbb{E}\left[\mathbb{E}\left[h(\xi_1, \xi_2) \mid \xi_1\right]^2\right] = 0$, alors il existe une suite de réels de carré sommable $(a_i)_{i \geqslant 1}$ et une suite de mouvements browniens standard indépendants $\left(B^{(i)}\right)_{i \geqslant 1}$ tels que

$$n^{-1}\mathcal{U}_{2,n,h}^{\mathrm{pl}}(t) \to \sum_{i=1}^{\infty} a_i \left(\left(B_t^{(i)}\right)^2 - t\right) \text{ en loi dans } C[0,1].$$

# $U$-statistiques dont le noyau dépend de l'indice

On considérera des $U$-statistiques de la forme

$$U_{m,n}\left((h_{\boldsymbol{i}})\right) = \sum_{\boldsymbol{i} \in \mathrm{Inc}_n^m} h_{\boldsymbol{i}}\left(\xi_{\boldsymbol{i}}\right), \quad n \geqslant m,$$

avec $h_{\boldsymbol{i}} \colon S^m \to \mathbb{H}$,

$$\mathrm{Inc}_n^m = \left\{ \boldsymbol{i} = \left(i_\ell\right)_{\ell \in [\![1,m]\!]}, 1 \leqslant i_1 < i_2 < \cdots < i_m \leqslant n \right\},$$

où $\mathbb{H}$ est un espace de Hilbert, $[\![1, m]\!] = \{1, \ldots, m\}$ et $\xi_{\boldsymbol{i}} = (\xi_{i_1}, \ldots, \xi_{i_m})$.

# Pourquoi travailler dans un espace de Hilbert/mesuré ?

Nous allons considérer par la suite des $U$-statistiques à valeurs dans un espace de Hilbert séparable $\mathbb{H}$.

- Certains tests robustes (**Chakraborty et Chaudhuri** (2015,2017) ; **Wegner et Wendler** (2023), **Jiang, Wang et Shao** (2023)) se basent sur une généralisation du noyau à valeurs réelles $h(x,y) = \operatorname{sgn}(x-y)$, donnée par

$$h\colon \mathbb{H}\times\mathbb{H}\to\mathbb{H}, \quad h(x,y) = \begin{cases} \frac{x-y}{\|x-y\|_{\mathbb{H}}} & \text{si } x \neq y, \\ 0 & \text{si } x = y. \end{cases}$$

- Le fait de considérer des variables aléatoires $\xi_i$ à valeurs dans un espace mesuré permet, entre autre, de considérer des données fonctionnelles.

# Plan

# Inégalités de déviation : motivations

On cherche à majorer, à $N \geqslant m$ et $t > 0$ fixés, la quantité

$$\mathbb{P}\left( \max_{m \leqslant n \leqslant N} \left\| \sum_{i \in \mathrm{Inc}_n^m} h_i(\xi_i) \right\|_{\mathbb{H}} > t \right)$$

par une autre dépendant de $t$ au travers de la queue de certaines variables aléatoire s'exprimant comme une somme de variables aléatoires positives.

On peut alors :

- formuler une vitesse de convergence dans la loi des grands nombres à l'aide de condition sur l'intégrabilité des variables aléatoires positives intervenant dans la majoration ;
- vérifier des critères de tension pour des processus dans certains espaces de Hölder.

## Pour les martingales

Soit $(D_i)_{i\geqslant 1}$ une suite d'accroissement d'une martingale à valeurs réelles par rapport à la filtration $(\mathcal{F}_i)_{i\geqslant 0}$. L'inégalité suivante due à **Fan, Grama et Liu** (2015)

$$\mathbb{P}\left(\max_{1\leqslant n\leqslant N}\left|\sum_{j=1}^{n} D_j\right| > x\right) \leqslant 2\exp\left(-\frac{x^2}{y^2}\right) + \mathbb{P}\left(\sum_{j=1}^{N}\left(D_j^2 + \mathbb{E}\left[D_j^2 \mid \mathcal{F}_{j-1}\right]\right) > \frac{y^2}{2}\right)$$

## Pour les martingales

Soit $(D_i)_{i\geqslant 1}$ une suite d'accroissement d'une martingale à valeurs réelles par rapport à la filtration $(\mathcal{F}_i)_{i\geqslant 0}$. L'inégalité suivante due à **Fan, Grama et Liu** (2015)

$$\mathbb{P}\left(\max_{1\leqslant n\leqslant N}\left|\sum_{j=1}^{n} D_j\right| > x\right) \leqslant 2\exp\left(-\frac{x^2}{y^2}\right) + \mathbb{P}\left(\sum_{j=1}^{N}\left(D_j^2 + \mathbb{E}\left[D_j^2 \mid \mathcal{F}_{j-1}\right]\right) > \frac{y^2}{2}\right)$$

et l'inégalité de **Nagaev** (2003)

$$\mathbb{P}\left(\max_{1\leqslant n\leqslant N}\left|\sum_{j=1}^{n} D_j\right| > t\right) \leqslant C_q \int_0^1 u^{q-1}\mathbb{P}\left(\max_{1\leqslant j\leqslant N}|D_j| > tu\right) du + C_q \int_0^1 u^{q-1}\mathbb{P}\left(\left(\sum_{j=1}^{N}\mathbb{E}\left[D_j^2 \mid \mathcal{F}_{j-1}\right]\right)^{1/2} > tu\right) du$$

ont lieu pour tous $t, x, y > 0$.

# Découplage

Afin d'établir des inégalités sur $\|U_{m,n}(h)\|_{\mathbb{H}}$, il est possible d'utiliser le découplage. **De la Peña, Montgomery-Smith** (1995) ont montré que pour tout $m \geqslant 2$ il existe une constante $C_m$ telle que pour tous noyaux $h_i$ et toute suite $(\xi_i)_{i\geqslant 1}$ i.i.d.,

$$\mathbb{P}\left(\left\|\sum_{\boldsymbol{i}\in \mathrm{Inc}_n^m} h_{\boldsymbol{i}}(\xi_{\boldsymbol{i}})\right\|_{\mathbb{H}} > t\right) \leqslant C_m \mathbb{P}\left(C_m \left\|\sum_{\boldsymbol{i}\in \mathrm{Inc}_n^m} h_{\boldsymbol{i}}\left(\xi_{\boldsymbol{i}}^{\mathrm{dec}}\right)\right\|_{\mathbb{H}} > t\right),$$

où $\xi_{\boldsymbol{i}}^{\mathrm{dec}} = \left(\xi_{i_1}^{(1)}, \ldots, \xi_{i_m}^{(m)}\right)$ et $\left(\xi_i^{(\ell)}\right)_{i\in\mathbb{Z}}$, $\ell \in [\![1, m]\!]$ sont des copies indépendantes de $(\xi_i)_{i\in\mathbb{Z}}$.

# Inégalité exponentielle

**Théorème (G. (2025))**

*Soit $m \geqslant 1$ un entier, $(S, \mathcal{S})$ un espace mesurable, $h_i : S^m \to \mathbb{H}$ des fonctions mesurables, où $\mathbb{H}$ est un espace de Hilbert séparable et $(\xi_i)_{i \geqslant 1}$ est une suite i.i.d. à valeurs dans $S$. Supposons que pour chaque $\boldsymbol{i}$ et chaque $\ell_0 \in [\![1, m]\!]$,*

$$\mathbb{E}\left[h_i\left(\xi_{[\![1,m]\!]}\right) \mid \sigma\left(\xi_k, k \in [\![1, m]\!] \setminus \ell_0\right)\right] = 0 \quad \text{où } \xi_{[\![1,m]\!]} := (\xi_1, \ldots, \xi_m).$$

*L'inégalité suivante a lieu pour tous $N \geqslant 1, m, x, y > 0$ :*

$$\mathbb{P}\left(\max_{m \leqslant n \leqslant N} \left\| \sum_{\boldsymbol{i} \in \mathrm{Inc}_n^m} h_i\left(\xi_i\right) \right\|_{\mathbb{H}} > x\right) \leqslant A_m \exp\left(-\left(\frac{x}{y}\right)^{\frac{2}{m}}\right)$$

$$+ B_m \int_1^\infty u\left(1 + \log u\right)^{\frac{m(m+1)}{2} - 1} \mathbb{P}\left(\sqrt{\sum_{\boldsymbol{i} \in \mathrm{Inc}_N^m} \left\| h_i\left(\xi_{\boldsymbol{i}}^{\mathrm{dec}}\right) \right\|_{\mathbb{H}}^2} > C_m y u\right) du,$$

*où $A_m$, $B_m$ et $C_m$ ne dépendent que de $m$.*

# Remarques sur l'inégalité exponentielle

- L'exposant $2/m$ dans le terme $\exp\left(-\left(\frac{x}{y}\right)^{\frac{2}{m}}\right)$ est optimal, même dans le cas $h_i = h$ et $h$ borné, voir **Arcones** (1995).
- En posant

$$Y := \sqrt{\sum_{i \in \mathrm{Inc}_N^m} \left\| h_i\left(\xi_i^{\mathrm{dec}}\right)\right\|_{\mathbb{H}}^2},$$

l'inégalité peut se reformuler ainsi : pour tous $x, y > 0$,

$$\mathbb{P}\left(\max_{m \leqslant n \leqslant N} \left\| \sum_{i \in \mathrm{Inc}_n^m} h_i\left(\xi_i\right)\right\|_{\mathbb{H}} > x\right) \leqslant A_m \exp\left(-\left(\frac{x}{y}\right)^{\frac{2}{m}}\right) + B'_m \mathbb{E}\left[\varphi_{2,m(m+1)/2}\left(\frac{Y}{y}\right)\right] du,$$

avec $\varphi_{2,q}(u) = u^2 (\log(1+u))^q$.

# Autres inégalités, mais qui supposent les noyaux bornés

- Cas i.i.d. : **Houdré et Reynaud-Bouret** (2002) pour $m = 2$, avec des noyaux pouvant dépendre de $(i, j)$ et des constantes explicites.

- Suite $\alpha$-mélangeantes :
  - **Shen, Han et Witten** (2020) sous une condition sur la transformée de Fourier du noyau
  - **F. Han** (2018)

- Fonctions de chaînes de Markov : **Duchemin, de Castro et Lacour** (2021) pour $m = 2$, avec des noyaux pouvant dépendre de $(i, j)$.

# Inégalité de déviation, décroissance polynomiale

**Théorème (G. (2025))**

*Si* $(\xi_i)_{i\in\mathbb{Z}}$ *est i.i.d. à valeurs dans* $S$, $h_{i,j}\colon S^2 \to \mathbb{H}$ *sont telles que pour tous* $i<j$, $\mathbb{E}\left[h_{i,j}(\xi_1,\xi_2) \mid \xi_1\right] = \mathbb{E}\left[h_{i,j}(\xi_1,\xi_2) \mid \xi_2\right] = 0$, *alors pour tous* $1 < p \leqslant 2$, $N \geqslant 2$, $q, t > 0$,

$$\mathbb{P}\left(\max_{2\leqslant n\leqslant N}\left\|\sum_{1\leqslant i<j\leqslant n} h_{i,j}(\xi_i,\xi_j)\right\|_{\mathbb{H}} > t\right) \leqslant \frac{K_{p,q}}{t^q}\left(\sum_{1\leqslant i<j\leqslant N}\mathbb{E}\left[\|h_{i,j}(\xi_1,\xi_2)\|_{\mathbb{H}}^p\right]\right)^{\frac{q}{p}}$$
$$+K_{p,q}\sum_{i=1}^{N-1}\int_0^1 u^{q-1}\mathbb{P}\left(\left(\sum_{j=i+1}^{N}\mathbb{E}\left[\|h_{i,j}(\xi_1,\xi_2)\|_{\mathbb{H}}^p \mid \xi_1\right]\right)^{\frac{1}{p}} > tu\right)du$$
$$+K_{p,q}\sum_{j=2}^{N}\int_0^1 u^{q-1}\mathbb{P}\left(\left(\sum_{i=1}^{j-1}\mathbb{E}\left[\|h_{i,j}(\xi_1,\xi_2)\|_{\mathbb{H}}^p \mid \xi_2\right]\right)^{\frac{1}{p}} > tu\right)du$$
$$+K_{p,q}\sum_{1\leqslant i<j\leqslant N}\int_0^1 u^{q-1}\mathbb{P}\left(\|h_{i,j}(\xi_1,\xi_2)\|_{\mathbb{H}} > tu\right)du.$$

# Remarques sur l'inégalité de déviation précédente

- L'inégalité précédente s'étend aux $U$-statistiques d'ordre $m$, où l'hypothèse $\mathbb{E}\left[h_{i,j}\left(\xi_1,\xi_2\right) \mid \xi_1\right] = \mathbb{E}\left[h_{i,j}\left(\xi_1,\xi_2\right) \mid \xi_2\right] = 0$ est remplacée par

$$\mathbb{E}\left[h_i\left(\xi_{[\![1,m]\!]}\right) \mid \sigma\left(\xi_\ell, \ell \in [\![1,m]\!] \setminus \{\ell_0\}\right)\right] = 0.$$

  Le majorant suit un schéma similaire : une somme indexée par les sous-ensembles $I$ de $[\![1,m]\!]$, avec une somme de probabilités sur les indices de $I$ et les queues des sommes des puissances d'ordre $p$ sur les coordonnées de $[\![1,m]\!] \setminus I$.

- Des inégalités de moments similaires à celles des articles d'**Adamczak** (2006) et **Giné, Latała et Zinn** (2000) peuvent être déduites de notre inégalité, mais avec des constantes non explicites. Cependant, il est possible de formuler des inégalités de moments pour les moments faibles, donnés par

$$\|Y\|_{q,\infty} := \left(\sup_{t>0} t^q \mathbb{P}\left(|Y| > t\right)\right)^{1/q}.$$

# Élements communs et différences entre les inégalités présentées

Les résultats présentés sont établis par récurrence sur l'ordre de la $U$-statistique. On passe de l'ordre $m$ à l'ordre $m+1$ grâce à des inégalités sur les martingales.

- Dans le cas de l'inégalité exponentielle, il s'agit d'une extension aux espaces de Hilbert d'une inégalité de **Fan, Grama, Liu** (2015) : si $(D_j)_{j\geqslant 1}$ est une suite d'accroissements d'une martingale par rapport à la filtration $(\mathcal{F}_j)_{j\geqslant 0}$, alors pour tous $x, y > 0$,

$$\mathbb{P}\left(\max_{1\leqslant k\leqslant n}\left\|\sum_{j=1}^{k} D_j\right\|_{\mathbb{H}} > x\right) \leqslant 4\exp\left(-\frac{x^2}{y^2}\right) + 2\mathbb{P}\left(\sum_{j=1}^{n}\left(\|D_j\|_{\mathbb{H}}^2 + \mathbb{E}\left[\|D_j\|_{\mathbb{H}}^2 \mid \mathcal{F}_{j-1}\right]\right) > \frac{y^2}{8}\right).$$

# Élements communs et différences entre les inégalités présentées

Les résultats présentés sont établis par récurrence sur l'ordre de la $U$-statistique. On passe de l'ordre $m$ à l'ordre $m+1$ grâce à des inégalités sur les martingales.

- Dans le cas de l'inégalité exponentielle, il s'agit d'une extension aux espaces de Hilbert d'une inégalité de **Fan, Grama, Liu** (2015) : si $(D_j)_{j\geqslant 1}$ est une suite d'accroissements d'une martingale par rapport à la filtration $(\mathcal{F}_j)_{j\geqslant 0}$, alors pour tous $x, y > 0$,

$$\mathbb{P}\left(\max_{1\leqslant k\leqslant n}\left\|\sum_{j=1}^{k} D_j\right\|_{\mathbb{H}} > x\right) \leqslant 4\exp\left(-\frac{x^2}{y^2}\right) + 2\mathbb{P}\left(\sum_{j=1}^{n}\left(\|D_j\|_{\mathbb{H}}^2 + \mathbb{E}\left[\|D_j\|_{\mathbb{H}}^2 \mid \mathcal{F}_{j-1}\right]\right) > \frac{y^2}{8}\right).$$

- Dans le cas de la décroissance polynomiale des queues, on établit une inégalité sur les moments conditionnels d'une martingale.

Les inégalités présentées sont complémentaires : l'une sera plus pertinente lorsque la queue de $h_i\left(\xi_{[\![1,m]\!]}\right)$ est à décroissance polynomiale, l'autre quand la queue est à décroissance exponentielle.

## Et lorsque $h_i = h$ ?

On suppose que $h_i = h$ avec $h$ symétrique et dégénérée d'ordre $d$. Alors pour tous $x, y > 0$,

$$\mathbb{P}\left(\max_{m\leqslant n\leqslant N}\left\|\sum_{i\in \mathrm{Inc}_n^m} h(\xi_i)\right\|_{\mathbb{H}} > xN^{m-d/2}\right) \leqslant A_m \exp\left(-\left(\frac{x}{y}\right)^{\frac{2}{d}}\right)$$
$$+B_m \sum_{k=d}^{m}\int_1^\infty \mathbb{P}\left(H > C_m u N^{\frac{k-d}{2}} x^{1-\frac{k}{d}} y^{\frac{k}{d}}\right) u\,(1+\log u)^{\frac{m(m+1)}{2}}\,du,$$

## Et lorsque $h_i = h$ ?

On suppose que $h_i = h$ avec $h$ symétrique et dégénérée d'ordre $d$. Alors pour tous $x, y > 0$,

$$\mathbb{P}\left(\max_{m\leqslant n\leqslant N}\left\|\sum_{i\in \mathrm{Inc}_n^m} h\left(\xi_i\right)\right\|_{\mathbb{H}} > xN^{m-d/2}\right) \leqslant A_m \exp\left(-\left(\frac{x}{y}\right)^{\frac{2}{d}}\right)$$
$$+B_m \sum_{k=d}^{m}\int_1^{\infty} \mathbb{P}\left(H > C_m u N^{\frac{k-d}{2}} x^{1-\frac{k}{d}} y^{\frac{k}{d}}\right) u\left(1+\log u\right)^{\frac{m(m+1)}{2}} du,$$

et pour tout $t > 0$,

$$\mathbb{P}\left(\max_{m\leqslant n\leqslant N}\left\|\sum_{i\in \mathrm{Inc}_n^m} h\left(\xi_i\right)\right\|_{\mathbb{H}} > tN^{m-d/2}\right)$$
$$\leqslant K_{m,q}\left(\int_0^1 u^{q-1}\left(\sum_{j=0}^{d-1} N^j \mathbb{P}\left(H > tN^{\frac{j}{2}} u\right) + \sum_{j=d}^{m} N^j \mathbb{P}\left(H > tN^{\frac{2j-d}{2}} u\right)\right) du\right),$$

où

$$H = \max_{k\in[\![0,m]\!]} \sqrt{\mathbb{E}\left[\left\|h\left(\xi_{[\![1,m]\!]}\right)\right\|_{\mathbb{H}}^2 \mid \xi_{[\![1,k]\!]}\right]} \approx \left\|h\left(\xi_{[\![1,m]\!]}\right)\right\|_{\mathbb{H}}.$$

# Plan

# Application au principe d'invariance dans des espaces de Hölder

On considère le processus sommes partielles

$$\mathcal{U}_{m,n,h}^{\mathrm{pl}}(t) = \begin{cases} \sum_{1\leqslant i_1<\cdots<i_m\leqslant k} h(\xi_{i_1},\ldots,\xi_{i_m}) & \text{si } t=\frac{k}{n} \text{ pour un certain } k\in[\![0,n]\!], \\ \text{interpolation linéaire} & \text{sur } \left]\frac{k}{n},\frac{k+1}{n}\right[, k\in[\![0,n-1]\!]. \end{cases}$$

On peut étudier sa convergence dans des espaces de Hölder de la forme

$$\mathcal{H}_{\alpha,\beta}^{o} = \left\{ x\in[0,1], \sup_{0\leqslant s,t\leqslant 1, t-s\leqslant\delta} \frac{|x(t)-x(s)|}{\rho_{\alpha,\beta}(t-s)} \underset{\delta\to 0}{\to} 0 \right\}, \quad \rho_{\alpha,\beta}(t) = t^{\alpha}\left(\log\left(\frac{c}{t}\right)\right)^{\beta}.$$

- Si $\alpha < 1/2$ et

$$\lim_{t\to\infty} t^{\frac{1}{1/2-\alpha}} \mathbb{P}\left(|h(\xi_1,\ldots,\xi_m)|>t\right) = 0,$$

alors $n^{-m+1/2}\left(\mathcal{U}_{m,n,h}^{\mathrm{pl}}(\cdot) - \mathbb{E}\left[\mathcal{U}_{m,n,h}^{\mathrm{pl}}(\cdot)\right]\right) \to \sigma W(\cdot)$ en loi $\mathcal{H}_{\alpha,0}^{o}$.

# Application au principe d'invariance dans des espaces de Hölder

On considère le processus sommes partielles

$$\mathcal{U}^{\mathrm{pl}}_{m,n,h}(t) = \begin{cases} \sum_{1\leqslant i_1<\cdots<i_m\leqslant k} h(\xi_{i_1},\ldots,\xi_{i_m}) & \text{si } t=\frac{k}{n} \text{ pour un certain } k\in[\![0,n]\!], \\ \text{interpolation linéaire} & \text{sur } \left]\frac{k}{n},\frac{k+1}{n}\right[, k\in[\![0,n-1]\!]. \end{cases}$$

On peut étudier sa convergence dans des espaces de Hölder de la forme

$$\mathcal{H}^{o}_{\alpha,\beta} = \left\{ x\in[0,1], \sup_{0\leqslant s,t\leqslant 1, t-s\leqslant\delta} \frac{|x(t)-x(s)|}{\rho_{\alpha,\beta}(t-s)} \underset{\delta\to 0}{\to} 0 \right\}, \quad \rho_{\alpha,\beta}(t) = t^{\alpha}\left(\log\left(\frac{c}{t}\right)\right)^{\beta}.$$

- Si $\alpha < 1/2$ et

$$\lim_{t\to\infty} t^{\frac{1}{1/2-\alpha}} \mathbb{P}\left(|h(\xi_1,\ldots,\xi_m)|>t\right) = 0,$$

alors $n^{-m+1/2}\left(\mathcal{U}^{\mathrm{pl}}_{m,n,h}(\cdot) - \mathbb{E}\left[\mathcal{U}^{\mathrm{pl}}_{m,n,h}(\cdot)\right]\right) \to \sigma W(\cdot)$ en loi $\mathcal{H}^{o}_{\alpha,0}$.

- Si $\alpha = 1/2$, $\beta > m/2$ et

$$\forall A>0, \quad \mathbb{E}\left[\exp\left(A\,|h(\xi_1,\ldots,\xi_m)|^{\frac{1}{\beta-m/2}}\right)\right] < \infty,$$

alors $n^{-m+1/2}\left(\mathcal{U}^{\mathrm{pl}}_{m,n,h}(\cdot) - \mathbb{E}\left[\mathcal{U}^{\mathrm{pl}}_{m,n,h}(\cdot)\right]\right) \to \sigma W(\cdot)$ en loi dans $\mathcal{H}^{o}_{1/2,\beta}$.

# Comment les inégalités précédents mènent à la convergence de processus ?

Les conditions trouvées sont analogues à celles de **Račkauskas et Suquet** (2003,2004) pour des processus sommes partielles basés sur une suite i.i.d.

L'intérêt de la norme hölderienne réside dans le fait d'être sensible à un changement de paramètre sur un très court intervalle de temps.

La convergence des lois fini-dimensionnelles est bien connue. L'enjeu est d'établir l'équi-tension. Pour des processus de la forme $W_n(t) := \frac{1}{a_n}\left(\sum_{i=1}^{[nt]} X_i + (nt - [nt])\, X_{[nt]+1}\right)$ un critère a été donné dans **G.** (2021), qui se base uniquement sur des termes de la forme $\mathbb{P}(|S_b - S_a| > t)$, où $S_b = \sum_{i=1}^{b} X_i$.

Dans le contexte des processus sommes partielles basés sur des $U$-statistiques, ces quantités s'expriment comme la queue d'une $U$-statistique pondérée.

# $U$-statistiques incomplètes, introduction

Le calcul d'une $U$-statistique d'ordre $m$ basée sur un échantillon de $m$ éléments met en jeu $C_n^m$ termes, ce qui peut être contraignant dans la pratique. C'est pour cette raison que **Blom** (1976) a introduit le concept de $U$-statistique incomplète. L'idée est de mettre devant $h(\xi_{\boldsymbol{i}})$ un poids aléatoire de la manière suivante :

- échantillonnage sans remplacement : on choisit sans remplacement $N$ $m$-uplets de la forme $\boldsymbol{i} = (i_\ell)_{\ell \in [\![1,m]\!]}$ où $1 \leqslant i_1 < \cdots < i_m \leqslant n$.
- échantillonnage avec remplacement : on choisit avec remplacement $N$ $m$-uplets de la forme $\boldsymbol{i} = (i_\ell)_{\ell \in [\![1,m]\!]}$ où $1 \leqslant i_1 < \cdots < i_m \leqslant n$ et on met devant $h(\xi_{\boldsymbol{i}})$ le nombre de fois où $\boldsymbol{i}$ a été choisi.
- échantillonnage de Bernoulli : pour chaque $\boldsymbol{i} = (i_\ell)_{\ell \in [\![1,m]\!]} \in \mathrm{Inc}_n^m$, on prend une variable aléatoire $Z_{n;\boldsymbol{i}}$ prenant la valeur $1$ avec probabilité $p_n$ et $0$ avec probabilité $1 - p_n$. De plus, on suppose que $(Z_{n,\boldsymbol{i}})_{\boldsymbol{i} \in \mathrm{Inc}_n^m}$ est indépendante et également indépendante de la suite $(\xi_i)_{i \in \mathbb{Z}}$.

# Inégalité exponentielle, échantillonage avec et sans remise

## Corollaire (Échantillonage avec et sans remplacement)

*Soit $m \geqslant 1$. Il existe des constantes $a_m$, $b_m$ et $c_m$ telles que si $(\mathbb{H}, \langle \cdot, \cdot \rangle)$ est un espace de Hilbert séparable, $(\xi_i)_{i \geqslant 1}$ est une suite i.i.d. à valeurs dans un espace mesurable $(S, \mathcal{S})$, $h \colon S^m \to \mathbb{H}$ est dégénéré d'ordre $d$, $N \geqslant 1$ et $n \geqslant m$ sont des entiers et $(Z_{n,\boldsymbol{i}})_{\boldsymbol{i} \in \mathrm{Inc}_n^m}$ est une collection de variables aléatoires à valeurs dans $\{0,1\}$ qui est indépendante de $(\xi_i)_{i \geqslant 1}$ et telle que $\sum_{\boldsymbol{i} \in \mathrm{Inc}_n^m} Z_{n,\boldsymbol{i}} = N$ et $x, y > 0$, alors*

$$\mathbb{P}\left( \left\| \sum_{\boldsymbol{i} \in \mathrm{Inc}_n^m} Z_{n,\boldsymbol{i}} h\left(\xi_{\boldsymbol{i}}\right) \right\|_{\mathbb{H}} > x\sqrt{N}\sqrt{\min\left\{N, n^{m-d}\right\}} \right) \leqslant a_m \exp\left( - \left( \frac{x}{y} \right)^{\frac{2}{m}} \right)$$
$$+ b_m \int_1^{\infty} u \left(1 + \log u\right)^{\frac{m(m+1)}{2}} \mathbb{P}\left( \left\| h\left( \xi_{[\![1,m]\!]} \right) \right\|_{\mathbb{H}} > c_m y u \right) du.$$

# Inégalité exponentielle, échantillonage de Bernoulli

## Corollaire (Échantillonage de Bernoulli)

*Soit $m \geqslant 1$. Il existe des constantes $a_m$, $b_m$ et $c_m$ telles que si $(\mathbb{H}, \langle \cdot, \cdot \rangle)$ est un espace de Hilbert séparable, $(\xi_i)_{i \geqslant 1}$ est une suite i.i.d. à valeurs dans un espace mesurable $(S, \mathcal{S})$, $h \colon S^m \to \mathbb{H}$ est dégénéré d'ordre $d$, $N \geqslant 1$ et $n \geqslant m$ sont des entiers et $(Z_{n,i})_{i \in \mathrm{Inc}_n^m}$ est une collection de variables aléatoires de Bernoulli de paramètre $p_n$ qui est indépendante de $(\xi_i)_{i \geqslant 1}$ et $x, y > 0$, alors*

$$\mathbb{P}\left( \left\| \sum_{i \in \mathrm{Inc}_n^m} Z_{n,i} h\left(\xi_i\right) \right\|_{\mathbb{H}} > x n^m \sqrt{p_n} \sqrt{\min\left\{p_n, n^{-d}\right\}} \right)$$

$$\leqslant a_m \exp\left( -\frac{n^m p_n^2}{2} \right) + a_m \exp\left( -\left( \frac{x}{y} \right)^{\frac{2}{m}} \right)$$

$$+ b_m \int_1^\infty u \left(1 + \log u\right)^{\frac{m(m+1)}{2}} \mathbb{P}\left( \left\| h\left(\xi_{[\![1,m]\!]}\right) \right\|_{\mathbb{H}} > c_m y u \right) du.$$

# Perspectives

- Traiter le cas de $U$-statistiques dont l'ordre est autorisé à dépendre de $n$.
- Fournir des inégalités de déviation pour des $U$-statistiques basées sur des variables aléatoire dépendantes (sous des conditions de mélange ou de $\tau$-dépendance par exemple).
- Utiliser la convergence dans les espaces de Hölder dans le but de mettre en place des tests statistiques pour détecter la présence d'un changement de paramètre dans un échantillon.