Concentration of measure for U-statistics

with applications

Radosław Adamczak[1]

Institute of Mathematics
Polish Academy of Sciences

International Workshop on Applied Probability 2006

The setup

- $(\Sigma, \mathcal{F})$ – a Polish space with the Borel σ-field
- $X_1, X_2, \ldots, X_n$ - i.i.d. Σ-valued random variables
- $h \colon \Sigma^d \to \mathbb{R}$ – a Borel measurable function
- $I_n^d = \{\mathrm{i} = (i_1, \ldots, i_d) \colon i_j \in \mathbb{N}, 1 \leq i_j \leq n, i_j \neq i_k \text{ if } j \neq k\}$
- a U-statistic:

$$Z = \sum_{\mathrm{i} \in I_n^d} h(X_{i_1}, \ldots, X_{i_d})$$

Additional assumptions

We assume

- **Symetry** – h invariant under permutation of coordinates
- **Complete degeneracy**

$$\mathbb{E}h(X_1, x_2, \ldots, x_d) = 0.$$

A natural assumption in view of the Hoeffding decomposition.

Aim

Our aim is to find good (exponential) estimates on

$$\mathbb{P}(|Z| \geq t)$$

How to do it?

- Estimate moments $\|Z\|_p$
- Use Chebyshev's Inequality and optimize in p

Previously known results

Theorem (Bernstein's inequality, $d = 1$)

$$Z = \sum_{i=1}^{n} h(X_i)$$

$$\mathbb{P}(|Z| \geq t) \leq K \exp\Big(-\frac{1}{K}\min\Big[\frac{t^2}{n\mathbb{E}h^2}, \frac{t}{\|h\|_\infty}\Big]\Big)$$

Previously known results

Theorem (Giné, Latała, Zinn (Houdré,Reynaud-Bouret with a different proof), $d = 2$)

$$Z = \sum_{i \neq j} h(X_i, X_j)$$

$$\mathbb{P}(|Z| \geq t) \leq$$

$$K \exp\Big(-\frac{1}{K}\min\Big[\frac{t^2}{n^2 \mathbb{E} h^2}, \frac{t}{n\|h\|_{L^2 \to L^2}}, \frac{t^{2/3}}{(n\|\mathbb{E}_{X_1} h^2\|_\infty)^{1/3}}, \frac{t^{1/2}}{\|h\|_\infty^{1/2}}\Big]\Big)$$

Previously known results

Actually Giné, Latała, Zinn prove

Theorem

$$\begin{aligned}\mathbb{E}|Z|^p \leq & K^p \Big[p^{p/2}(n^2\mathbb{E}h^2)^{p/2} + p^p(n\|h\|_{L^2\to L^2})^p \\ & + p^{3p/2}\mathbb{E}_X \max_{i\leq n}(n\mathbb{E}_Y h^2(X_i,Y))^{p/2} \\ & + p^{2p}\mathbb{E}\max_{i,j\leq n}|h(X_i,Y_j)|^p\Big],\end{aligned}$$

where (Y_i) - independent copy of (X_i).

Notation and definitions

- I - a finite, nonempty set,
- $\mathcal{P}_I$ - set of partitions of I into disjoint, nonempty sets
- $\mathcal{J} = \{J_1, \ldots, J_k\} \in \mathcal{P}_I$
- For $I = \emptyset$, $\mathcal{P}_I = \{\emptyset\}$
- $\deg \mathcal{J} = \#\mathcal{J}$.

Notation and definitions

With each partition we associate a norm $\|h\|_{\mathcal{J}}$.
It is best to define it by examples

$$\begin{aligned}
\|h(X_1,X_2,X_3)\|_{\{1,2,3\}} &= \sup\{\mathbb{E}h(X_1,X_2,X_3)f(X_1,X_2,X_3): \\
&\qquad \mathbb{E}f(X_1,X_2,X_3)^2 \leq 1\} \\
\|h(X_1,X_2,X_3)\|_{\{1,2\}\{3\}} &= \sup\{\mathbb{E}h(X_1,X_2,X_3)f(X_1,X_2)g(X_3): \\
&\qquad \mathbb{E}f(X_1,X_2)^2, \mathbb{E}g(X_3)^2 \leq 1\} \\
\|h(X_1,X_2,X_3)\|_{\{1\}\{2\}\{3\}} &= \sup\{\mathbb{E}h(X_1,X_2,X_3)f(X_1)g(X_2)k(X_3): \\
&\qquad \mathbb{E}f(X_1)^2, \mathbb{E}g(X_2)^2, \mathbb{E}k(X_3)^2 \leq 1\}.
\end{aligned}$$

Notation and definitions

We can see that

- $\|h\|_{\{1,2,3\}} = \|h\|_2$
- $\|h\|_{\{1\},\{2\},\{3\}}$ is the norm of h viewed as a 3-linear functional on $L^2(X_1) \times L^2(X_2) \times L^2(X_3)$.
- $\|h\|_{\{1,2\},\{3\}}$ is the norm of h as a 2-linear functional on $L^2(X_1, X_2) \times L^2(X_3)$.

Notation and definitions

Similarly we define e.g.

$$\|h(X_1, X_2, X_3)\|_{\{1\}\{2\}} = \sup\{\mathbb{E}_{X_1,X_2} h(X_1, X_2, X_3) f(X_1) g(X_2) \colon \mathbb{E}f^2, \mathbb{E}g^2 \leq 1\}$$

$$\|h(X_1, X_2, X_3)\|_{\{3\}} = \sup\{\mathbb{E}_{X_3} h(X_1, X_2, X_3) f(X_3) \colon \mathbb{E}f^2 \leq 1\},$$

but now they are **random variables.**
Finally (to simplify the statements of theorems) we define

$$\|h(X_1, X_2, X_3)\|_{\emptyset} = |h(X_1, X_2, X_3)|.$$

Moments estimates

$$Z = \sum_{i \neq j \neq k} h(X_i, X_j, X_k).$$

Theorem

For all $p \geq 2$ we have

$$\mathbb{E}|Z|^p$$
$$\leq K^p \sum_{I \subseteq \{1,2,3\}} \sum_{\mathcal{J} \in \mathcal{P}_I} n^{\#I p/2} p^{p(\deg(\mathcal{J})/2 + \#I^c)} \mathbb{E}_{I^c} \max_{\mathbf{i}_{I^c}} \|h(X_i, Y_j, Z_k)\|_{\mathcal{J}}^p,$$

where $(Y_j), (Z_k)$ – independent copies of (X_i), $\mathbf{i} = (i, j, k)$.

A closer look at the right-hand side

- $I = \{1,2,3\}$
  - $p^{p/2} n^{3p/2} \|h\|^p_{\{1,2,3\}} \sim p^{p/2} (\mathbb{E}|Z|^2)^{p/2}$
  - $p^p n^{3p/2} \|h\|^p_{\{1,2\},\{3\}}$,
  - $p^{3p/2} n^{3p/2} \|h\|^p_{\{1\}\{2\}\{3\}}$,
- $I = \{1,2\}$
  - $p^{3p/2} n^p \mathbb{E}_Y \max_{k \leq n} \|h(X_1, X_2, Y_k)\|^p_{\{1,2\}} =$ $p^{3p/2} n^p \mathbb{E}_Y \max_{k \leq n} (\mathbb{E}_{X_1,X_2} h(X_1, X_2, Y_k)^2)^{p/2}$
  - $p^{2p} n^p \mathbb{E}_Y \max_{k \leq n} \|h(X_1, X_2, Y_k)\|^p_{\{1\}\{2\}}$

A closer look at the right-hand side

- $I = \{1\}$
  - $p^{5p/2} n^{p/2} \mathbb{E}_{Y,Z} \max_{j,k\leq n} \|h(X_1, Y_j, Z_k)\|_{\{1\}}^p = p^{5p/2} n^{p/2} \mathbb{E}_{Y,Z} \max_{j,k\leq n} (\mathbb{E}_{X_1} h^2(X_1, Y_j, Z_k)^2)^{p/2}$
- $I = \emptyset$,
  - $p^{3p} \mathbb{E} \max_{i,j,k\leq n} |h(X_i, Y_j, Z_k)|^p$.

Tail estimates

Theorem

$$\mathbb{P}\left(\left|\sum_{\mathbf{i}} h_{\mathbf{i}}\right| \geq t\right)$$
$$\leq K \exp\left[-\frac{1}{K} \min_{I \subseteq I_d, \mathcal{J} \in \mathcal{P}_I} \left(\frac{t}{n^{\# I/2} \left\| \|h\|_{\mathcal{J}} \right\|_\infty}\right)^{2/(\deg(\mathcal{J})+2\# I^c)}\right].$$

Gaussian chaoses

Let (a_{ijk}) be a 3-indexed symmetric matrix with zeros on the diagonal, and $g_1, g_2, \ldots$ – independent $\mathcal{N}(0,1)$ Gaussian variables. Consider

$$Z = \sum_{ijk} a_{ijk} g_i g_j g_k.$$

Define for $\mathcal{J} = \{J_1, \ldots, J_m\} \in \mathcal{P}_{\{1,2,3\}}$

$$\|(a_{ijk})\|_{\mathcal{J}} = \sup\{\sum_{ijk} a_{ijk} \prod_{l=1}^{m} x^{(l)}_{\mathrm{i}_{J_l}} \colon \sum_{i_{J_l}} (x^{(l)}_{\mathrm{i}_{J_l}})^2 \leq 1\}$$

e.g.

$$\|(a_{ijk})\|_{\{1,2\}\{3\}} = \sup\{\sum_{ijk} a_{ijk} x_{ij} y_k \colon \sum x_{ij}^2 \leq 1, \sum y_k^2 \leq 1\}$$

Gaussian chaoses

Theorem (Latała)

For $p \geq 2$

$$\|Z\|_p \sim \sum_{\mathcal{J}\in\mathcal{P}_{\{1,2,3\}}} p^{\deg \mathcal{J}/2} \|(a_{ijk})\|_{\mathcal{J}}.$$

In consequence for $t \geq 0$

$$\mathbb{P}(|Z| \geq t) \leq K \exp\left[-\frac{1}{K}\min_{\mathcal{J}\in\mathcal{P}_{\{1,2,3\}}}\left(\frac{t}{\|(a_{ijk})\|_{\mathcal{J}}}\right)^{2/\deg \mathcal{J}}\right].$$

Tools

- Decoupling Inequalities (de la Peña, Montgomery-Smith)
- Talagrand's Inequality for suprema of empirical processes (in the moments version, Giné, Latała, Zinn & Boucheron, Bousquet, Lugosi, Massart)
- Estimates between weak and strong variance for empirical processes
- Estimates on Gaussian averages in operator spaces (Latała)

Crucial Lemma

Lemma

Let Z_k be independent random variables and $A_k(Z_k) = (a_{ijk}(Z_k))_{ij}$ - independent centered random matrices. Then for $p \geq 2$

$$\mathbb{E}\|\sum_k A_k(Z_k)\| \leq K\Big[\frac{1}{\sqrt{p}}\|(a_{ijk}(Z_k))\|_{\{1,2,3\}} + \|(a_{ijk}(Z_k))\|_{\{1,3\}\{2\}} + \|(a_{ijk}(Z_k))\|_{\{1\}\{2,3\}} + \sqrt{p}\|(a_{ijk}(Z_k))\|_{\{1\}\{2\}\{3\}} + p\sqrt{\mathbb{E}\max_k \|(A_k(Z_k))\|^2}\Big]$$

Crucial Lemma

$$\|(a_{ijk}(Z_k))\|_{\{1,2,3\}} = \sqrt{\mathbb{E}\sum_{ijk} a_{ijk}(Z_k)^2}$$

$$\|(a_{ijk}(Z_k))\|_{\{1,3\},\{2\}} = \sqrt{\sup_{\|x\|_2\leq 1} \sum_{i,k} \mathbb{E}(\sum_j a_{ijk}(Z_k)x_j)^2}$$

$$\|(a_{ijk}(Z_k))\|_{\{1\}\{2\}\{3\}} = \sqrt{\sup_{\|x\|_2,\|y\|_2\leq 1} \sum_k \mathbb{E}(\sum_{i,j} a_{ijk}(Z_k)x_iy_j)^2}.$$

- Tail estimates for multiple stochastic integrals with respect to processes with independent increments and uniformly bounded jumps (in the spirit of inequalities by Houdré, Reynaud-Bouret for $d = 2$)
- Law of the iterated logarithm for kernel density estimators (Giné, Mason), $d = 2$.
- Law of the iterated logarithm for canonical U-statistics

Definition

For $u \geq 0$ let us define

$$\|h(X_1,X_2,X_3)\|_{\{1,2,3\},u} = \sup\{\mathbb{E}h(X_1,X_2,X_3)f(X_1,X_2,X_3)\colon \|f\|_2 \leq 1, \|f\|_\infty \leq u\}$$

$$\|h(X_1,X_2,X_3)\|_{\{1,2\}\{3\},u} = \sup\{\mathbb{E}h(X_1,X_2,X_3)f(X_1,X_2)g(X_3)\colon \|f\|_2, \|g\|_2 \leq 1, \|f\|_\infty, \|g\|_\infty \leq u\}$$

Theorem (Latała, R.A. (Giné, Kwapień, Latała, Zinn for $d = 2$))

The $h\colon \Sigma^d \to \mathbb{R}$ be arbitrary kernel. Then the LIL

$$\limsup_{n\to\infty} \frac{1}{n^{d/2}\mathrm{loglog}^{d/2} n}\Big|\sum_{\mathrm{i}\in I_n^d} h(X_{i_1},\ldots,X_{i_d})\Big| < \infty$$

holds if and only if h is completely degenerated and for all $\mathcal{J} \in \mathcal{P}_{\{1,\ldots,d\}}$ we have

$$\limsup_{u\to\infty} \frac{\|h\|_{\mathcal{J},u}}{\mathrm{loglog}^{(d-\deg\mathcal{J})/2} u} < \infty$$

A few questions

- Prove estimates for suprema of U-statistics (U-processess) at least over VC classes of kernels or for U-statistics in Banach spaces of type 2 (known for Hilbert spaces).
- Identify the limit in the LIL for $d \geq 2$
- Prove tail estimates for chaoses generated by other variables (e.g. stable –> consequences in stochastic processes, Bernoulli –> consequences in random graphs theory)