Kolmogorov Berry-Esseen bounds for binomial functionals

Raphaël Lachièze-Rey,
Univ. South California, Univ. Paris 5 René Descartes,

Joint work with Giovanni Peccati, University of Luxembourg

New direction on Stein’s method, May 18-29, Singapore
Framework

- $n \geq 1$
- $X = (X_1, \ldots, X_n)$ IID variables on some space $(E, \mathcal{E})$
- $(x_1, \ldots, x_n) \mapsto f(x_1, \ldots, x_n)$ a real functional
- $W = f(X)$ has finite variance $\sigma^2$.
- $N$: Standard Gaussian variable. We investigate here

$$d_W(W, N) := \sup_{\|h\| \leq 1} |Eh(W) - h(N)|$$

$$d_K(W, N) := \sup_{t \in \mathbb{R}} |P(W \leq t) - P(N \leq t)|$$

- Note also

$$\tilde{W} = \frac{W - EW}{\sqrt{\text{Var}(W)}}$$
1 A general bound

2 Hoeffding decomposition and variance lower bound

3 Geometric applications

4 Analogy with Poisson results
Stein’s bound

It is known that there is $C > 0$ such that for any suitable $W$

$$d_N(W, N) \leq \sup_{g: \|g\| \leq C, \|g'\| \leq C, \|g''\| \leq C} |E_W g(W) - g'(W)|.$$

Recent results by Schulte 2012 and Eischelsbacher & Thaele 2014 yield

$$d_K(W, N) \leq \sup_{g: \|g\| \leq C', \|g'\| \leq C', (\ast)_t, t \in \mathbb{R}} |E_W g(W) - g'(W)|$$

where $(\ast)_t$ replaces Taylor expansion at the 2d order

$$|g(x + u) - g(x) - ug'(x)| \leq \frac{u^2}{2} (|x| + C) + |u| (1_{x \leq t < x + u} + 1_{x + u \leq t < x})$$

$$= \frac{u^2}{2} (|x| + C) + u (1_{x \leq t < x + u} - 1_{x + u \leq t < x})$$
Some difference operators and a decomposition

Introduce an additional vector $X' \overset{(d)}{=} X$ independent of $X$ and $X^A = (X^A_i)_{1 \leq i \leq n}, A \subseteq [n],

\[X^A_i = \begin{cases} X_i & \text{if } i \notin A \\ X'_i & \text{if } i \in A, \end{cases}\]

and the difference operator

\[\Delta_A f(X, X') = f(X) - f(X^A)\]
\[\Delta_j f(X, X') = \Delta_{\{j\}} f(X, X')\]

Then, given two $L^2$ functionals $\varphi, \psi$, Chatterjee 2008 derived the decomposition

\[\text{cov}(\varphi(X), \psi(X)) = \frac{1}{2} \sum_{A \subseteq [n]} \frac{1}{|A| (n - |A|)} \sum_{j \notin A} \mathbb{E}[\Delta_j \varphi(X, X') \Delta_j \psi(X^A, X')]\]

convex combination in the sense that $\sum_{A \subseteq [n]} \sum_{j \notin A} \kappa_{n, A} = 1$. 
General bounds if $E W = 0$

This decomposition is the starting point for proving the bound

$$d_W(W, N) \leq \sigma^{-2} \sqrt{\text{Var}(E(T|W))} + \sigma^{-3} \frac{\sqrt{2\pi}}{16} \sum_{j=1}^{n} E|\Delta_j f(X, X')|^3$$

$$d_K(W, N) \leq B_W + \sigma^{-2} \sqrt{\text{Var}(E(T'|W))}$$

$$+ \frac{\sigma^{-4}}{4} E \sum_{j,A,j \notin A} \kappa_{n,A} |f(X)\Delta_j f(X, X')^2 \Delta_j f(X^A, X')|$$

where

$$T = \frac{1}{2} \sum_{j,A:j \notin A} \kappa_{n,A} \Delta_j f(X, X') \Delta_j f(X^A, X'), \quad E T = \sigma^2$$

$$T' = \frac{1}{2} \sum_{j,A:j \notin A} \kappa_{n,A} \Delta_j f(X, X') |\Delta_j f(X^A, X')|, \quad E T' = 0,$$
In all the applications, $T'$ is bounded exactly like $T$

In many applications, the other new term is bounded by

$$\frac{\sigma^{-3}}{4} \sum_{j=1}^{n} \sqrt{\mathbb{E}|\Delta_j f(X, X')|^6}$$

(not optimal for Voronoi approximation)

If not optimal and if $f$ is symmetric, there is instead the bound

$$n \frac{\sigma^{-4}}{4} \sup_{A \subseteq [n]} \sqrt{\mathbb{E}|f(X) \Delta_1 f(X^A, X)|^3},$$

As often, it requires lower bounds on $\sigma$. 
A general bound

Hoeffding decomposition and variance lower bound

Geometric applications

Analogy with Poisson results
Hoeffding’s decomposition

It is known that any $L^2$ functional $W = f(X)$ admits a (unique) Hoeffding decomposition

$$f(X) = \sum_{k=0}^{n} f_k(X)$$

where $f_k(X)$ is a degenerate $U$-statistic of order $k$, that is

$$f_k(X) = \sum_{1 \leq i_1 < \ldots < i_k \leq n} h_{i_1, \ldots, i_k}(X_{i_1}, \ldots, X_{i_k})$$

where the integral of any of the arguments of $h_{i_1, \ldots, i_k}$ vanish, that is

$$Eh(x_1, \ldots, x_{i-1}, X_i, x_{i+1}, \ldots, x_k) = 0$$

for $\mu$-a.a. $x_j, j \neq i$. This implies the orthogonality

$$Ef_k(X)f_j(X) = 0, k \neq j.$$
The orthogonality yields

$$\text{Var}(f(X)) = \sum_{k=1}^{n} \text{Var}(f_k(X))$$

The kernel $h_{i_1,\ldots,i_k}$ has a representation in terms of difference operators:

$$h_{i_1,\ldots,i_k}(x_1,\ldots,x_k) = (-1)^k \mathbb{E} \Delta_{i_1} (\Delta_{i_2}(\ldots \Delta_{i_k} f(X, (x_1,\ldots,x_k))))$$

The previous decomposition implies the variance lower bound

$$\text{Var}(f(X)) \geq \text{Var}(f_1(X)) = \sum_{j=1}^{n} \mathbb{E} \left[ (\mathbb{E} \Delta_j f(X'), X)^2 | X \right]$$

The form of the kernels recalls the orthogonal Wiener-Ito chaotic decomposition of Poisson functionals.
1. A general bound

2. Hoeffding decomposition and variance lower bound

3. Geometric applications

4. Analogy with Poisson results
Geometric assumptions

In many geometric applications, $f$ can be seen as a functional taking in argument an unordered finite set of points. This has two formal consequences:

- $f$ is symmetric in its arguments
- $f$ can have any finite number of arguments.

Introduce $X^i = (X_j; j \neq i)$, and the derivatives

$$D_i f(X) = f(X) - f(X^i)$$

$$D_{i,j} f(X) = D_i(D_j f(X)) = f(X) - f(X^i) - f(X^j) + f(X^{ij}) = D_{j,i} f(X).$$

A bound that is useful on $D_{i,j} f(X)$ is

$$|D_{i,j} f(X)| \leq \left( |D_i f(X)| + |D_i f(X^j)| \right) 1_{\{D_{i,j} f(X) \neq 0\}}$$

where $D_{i,j} f(X) \neq 0$ means somehow that $i$ and $j$ interact, i.e. the contribution of $X_i$ is affected by the presence of $j$, i.e.

$$D_i f(X) \neq D_i f(X^j).$$
Example: The boolean model

Let $X_1, \ldots, X_n$ IID random compact sets in some space $E \subseteq \mathbb{R}^d$, 

$$f_{\text{vol}}(X) = \text{vol} \left( E \cap \bigcup_{i=1}^{n} X_i \right).$$

Then

$$|D_i f_{\text{vol}}(X)| \leq \text{vol}(X_i),$$

$$D_{i,j} f_{\text{vol}}(X) = 0 \text{ if } X_i \cap X_j = \emptyset.$$

$$f_{\text{iso}}(X) = \#\{j : X_j \cap X_i \cap E = \emptyset, j \neq i\},$$

then

$$|D_i f_{\text{iso}}(X)| \leq \#\{j : X_j \cap X_i \neq \emptyset\},$$

$$D_{i,j} f_{\text{iso}}(X) = 0 \text{ if no } X_k \text{ touches } X_i \text{ and } X_j.$$
Bound on $\text{Var}(E[T|W])$, $\text{Var}(E[T'|W])$

Back to the general case

- Introduce $\tilde{X}$ an independent copy of $X$ (also independent of $X'$). Call recombination of $\{X, X', \tilde{X}\}$ a vector $Y = (Y_1, \ldots, Y_n)$ where for each $j$, $Y_j \in \{X_j, X'_j, \tilde{X}_j\}$.
- If $f$ is symmetric, expanding $\text{Var}(E(T|W))$ gives, up to some universal $C > 0$

$$\text{Var}(E[T'|W]) \leq C \sum_{j,k=1}^{n} \sup_{(Y,Y',Z)} E \left[ 1_{\{D_{1,j}f(Y) \neq 0, D_{1,k}f(Y') \neq 0\}} D_j f(Z)^4 \right]$$

where the supremum is over recombinations $Y, Y', Z$ of $\{X, X', \tilde{X}\}$

- By symmetry, it suffices to treat the cases $j = k = 1$, $j \neq k = 1$, $j \neq k \neq 1$
Assume $E = n^{1/d}[0,1]^d$ and the $X_i$ are built in the following way:

- There are iid compact sets $K_1, \ldots, K_n$
- and IID points uniform $Y_1, \ldots, Y_n$ in $C_n := n^{1/d}[0,1]^d$ such that $X_i = (Y_i + K_i) \cap E_n$. Then, if $E \text{diam}(K_i)^{5d} < \infty$,

$$d_K(\tilde{f}_\text{vol}(X), N) \leq Cn^{-1/2}$$

and if $E \text{diam}(K_i)^{8d} < \infty$,

$$d_K(\tilde{f}_\text{iso}(X), N) \leq Cn^{-1/2}.$$ 

- Goldstein and Penrose 2010 obtained similar results with the size-bias method.
Set estimation

- $K$ is an unknown set in $[0, 1]^d$
- $X$ is a random sample of points
- We have the information $\{1_{x \in K}, x \in X\}$
- How to get a good idea of $K$? Measure the quality of the approximation?
Voronoi approximation

Reconstruct $K$ with

$$K^X = \{ x \in \mathbb{R}^d : x \text{ is closer from a point of } X \text{ inside } K \text{ than outside } K \}.$$
Regularity hypotheses

- Many applications (cardiology, oncology, chemistry) are concerned with the detection of irregular boundaries.
- We developed a general method allowing the set to be irregular in some sense. Assume that
  - The tubular neighborhood of $K$’s boundary satisfies for some $\alpha > 0$, $0 < C^- \leq C^+ < \infty$
    $$C^- r^\alpha \leq \text{Vol}(\partial K \oplus B(0, r)) \leq C^+ r^\alpha$$
  - Around most points of the boundary, 
    $$\frac{\text{Vol}(K^c \cap B(x, r))}{r^d}$$
    is not too small.
- These assumptions are satisfied by some fractal sets, such as for instance the Von Koch flake and anti flake, and also by some smooth sets which boundaries are piecewise $C^1$. 

R. Lachieze-Rey

Kolmogorov Berry-Esseen bounds for binomial functionals
New direction on Stein’s method, May 18-29, Singapore
When one removes the point $X_1$ from $X$, the approximation $\text{Vol}(K^X)$ can only vary by the volume of the Voronoi cell $\text{Vor}(X_1; X)$ of $X_1$ of $X$, and it does so only if $X_1$ is “close” to the boundary.

$$|D_1 f(X)| \leq \text{Vol}(\text{Vor}(X_1; X))1\{X \text{ is at “Vorono-distance” less than 2 from } \partial K\}$$

Also, two points $X_i, X_j$ interact if they are at “Vorono-distance” less than 2 and one of them is at distance less than 2 from $\partial K$, which gives a sharp bound on $1\{D_{i,j}f(X) \neq 0\}$. 
Second order theory

Given \( K \) satisfying the previous regularity hypotheses, the variance and Berry-Esseen bounds presented above apply and we have for some \( C, C', C'' > 0 \),

\[
C n^{-1 - \alpha / d} \leq \text{Var}(\text{Vol}(K^X)) \leq C' n^{-1 - \alpha / d}
\]

and

\[
d_K(\widetilde{\text{Vol}}(K^X), N) \leq C'' n^{-1/2 + \alpha / 2d} \log(n)^{3 + \alpha / d + \varepsilon}.
\]

Consistent with Yukich results with Poisson input and sets with a \( d - 1 \)-dimensional boundary.
1. A general bound

2. Hoeffding decomposition and variance lower bound

3. Geometric applications

4. Analogy with Poisson results
Second order Poincaré inequality

The bound involving second order derivatives bears some similarity with the “second order Poincare inequality” derived by Last, Peccati, and Schulte (2014), where for some Poisson process $X$ and $L^2$ functional $f(X)$,

$$d_K(\tilde{f}(X), N) \leq \sum_{i=1}^{6} \gamma_i$$

where the $\gamma_i$ are estimates depending on the first and second add-one cost derivatives, for instance

$$\gamma_1 = 4 \left[ \int \left[ \mathbb{E} (D_{x_1}F)^2 (D_{x_2}F)^2 \right]^{1/2} \left[ \mathbb{E} (D_{x_1,x_3}^2F)^2 (D_{x_2,x_3}^2F)^2 \right]^{1/2} dx_1 dx_2 dx_3 \right]$$

The first “Second-order Poincaré inequality” was derived by Chatterjee 2009 in the Gaussian framework, and generalized by Nourdin and Peccati (2009)
Variance bounds

**Poisson input** $X$ with intensity $\mu$: \( \text{Var}(f(X)) \) is between

\[
\int (\mathbb{E} [f(X \cup x) - f(X)])^2 \mu(dx) \quad \text{and} \quad \int \mathbb{E} (f(X \cup x) - f(X))^2 \mu(dx)
\]

First term of the orthogonal chaos decomposition

Poincaré inequality

**Binomial input** $X$: Define $X^i_x$ the vector $X$ where the $i$-th component has been replaced by $x$. Then \( \text{Var}(f(X)) \) is between

\[
\sum_{i=1}^{n} \int (\mathbb{E} [f(X^i_x) - f(X)])^2 \mu(dx) \quad \text{and} \quad \frac{1}{2} \sum_{i=1}^{n} \int \mathbb{E} (f(X^i_x) - f(X))^2 \mu(dx)
\]

1st term of Hoeffding decomposition

Stein-Efron inequality