Malliavin calculus for marked binomial processes Chen-Stein method

Hélène Halconruy

Introduction

Application 1 : The longest head run

Large number of independent tosses of a coin, $C_i \sim \mathscr{B}(p)$.

0001100101010111111100001010011011111111111000001101111111110010

Question : Law of R_n longest series of 1's starting in the first n tosses = *longest head run* ?

Introduction

Application 1 : The longest head run

Large number of independent tosses of a coin, $C_i \sim \mathscr{B}(p)$. m_n : test length

0001100101010$\underbrace{11111}_{m_n=5}0100001010011\underbrace{1111111111}_{m_n=5 \quad m_n=5}1000001101\underbrace{11111}_{m_n=5}$1010010

Question : Law of R_n longest series of 1's starting in the first n tosses = *longest head run* ?

$$\boxed{\{R_n < m_n\} = \{U_n = 0\}}$$

U_n : nb of starting positions in $\{1,\ldots,n\}$ of a head run

Runs of length m_n at positions $j = 14, 33, 38, 52.$

Introduction

Application 1 : The longest head run

Large number of independent tosses of a coin, $C_i \sim \mathscr{B}(p)$. m_n : test length

0001100101010**11111**0100001010011011111111111000001101**11111**010010

m_n=5 m_n=5 m_n=5 m_n=5

Question : Law of R_n longest series of 1's starting in the first n tosses = *longest head run* ?

$$\{R_n < m_n\} = \{U_n = 0\}$$

U_n : nb of starting positions in $\{1,\ldots,n\}$ of a head run

Runs of length m_n at positions $j = 14, 33, 34, 38, 52, 53.$

Introduction

Application 1 : The longest head run

Large number of independent tosses of a coin, $C_i \sim \mathcal{B}(p)$. m_n : test length

0001100101010**11111**0100001010011011**111111**1111000001101**11111**1010010

m_n=5 m_n=5 m_n=5

Question : Law of R_n longest series of 1's starting in the first n tosses = *longest head run* ?

$$\{R_n < m_n\} = \{U_n = 0\}$$

U_n : nb of starting positions in $\{1,\ldots,n\}$ of a head run

Runs of length m_n at positions $j =$ 14 ,33 , 34 , 35 , 36 ,37 , 38 ,39 ,52 , 53

Clumps $\mathbf{P}(\text{run at } j+1 \,|\, \text{run at } j) = p$

Introduction

Application 1 : The longest head run

Large number of independent tosses of a coin, $C_i \sim \mathcal{B}(p)$. m_n : test length

0001100101010**11111**0100001010011011**111111**1111000001101**11111**1010010

m_n=5 m_n=5 m_n=5

Question : Law of R_n longest series of 1's starting in the first n tosses = *longest head run* ?

$$\{R_n < m_n\} = \{U_n = 0\}$$

U_n : nb of starting positions in $\{1,\ldots,n\}$ of a head run

Runs of length m_n at positions $j =$ 14 ,33 , 34 , 35 , 36 ,37 , 38 ,39 ,52 , 53

Clumps $\mathbf{P}(\text{run at } j+2 \,|\, \text{run at } j) = p^2$

Introduction

Application 1 : The longest head run

Large number of independent tosses of a coin, $C_i \sim \mathcal{B}(p)$. m_n : test length

0001100101010**11111**0100001010011011**111111**1111000001101**11111**1010010

m_n=5 m_n=5 m_n=5

Question : Law of R_n longest series of 1's starting in the first n tosses = *longest head run* ?

$$\{R_n < m_n\} = \{U_n = 0\}$$

U_n : nb of starting positions in $\{1,\ldots,n\}$ of a head run

Runs of length m_n at positions $j =$ 14 ,33 , 34 , 35 , 36 ,37 , 38 ,39 ,52 , 53

Clumps $\mathbf{P}(\text{run at } j+k \mid \text{run at } j) = p^k$

Introduction

Application 1 : The longest head run

Large number of independent tosses of a coin, $C_i \sim \mathcal{B}(p)$. m_n : test length

0001100101010**11111**0100001010011011**1111111**11110000011011**11111**010010

m_n=5 m_n=5 m_n=5

Question : Law of R_n longest series of 1's starting in the first n tosses = *longest head run* ?

$$\{R_n < m_n\} = \{U_n = 0\}$$

U_n : nb of starting positions in $\{1,\ldots,n\}$ of a head run

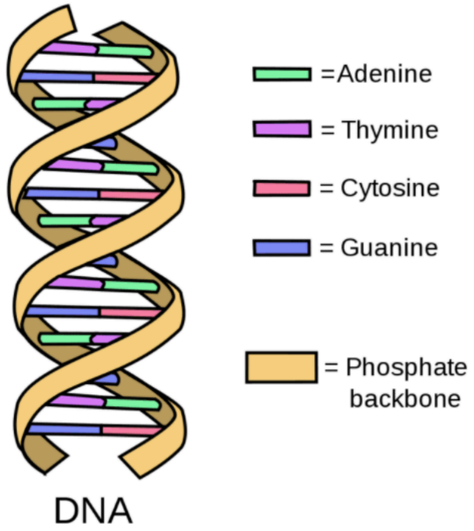
Runs of length m_n at positions $j =$ 14 ,33 , 34 , 35 , 36 ,37 , 38 ,39 ,52 , 53

Clumps Average size of clump : $1 + p + p^2 + \ldots$

Application 2 : a rare word in a DNA sequence

Series $X_1 X_2 \ldots X_n$ from $\mathscr{A} = \{\text{A,C,G,T}\}$

Rare word : $W_n = w_1 w_2 \ldots w_{h_n}$ of length $h_n \sim \log(n)$



Question : number of occurrences of W_n ?

$$\mathfrak{T}(W_n) = \sum_{j \in I} Z_j = \sum_{j \in I} \mathbf{1}_{\{X_j = w_1, \ldots, X_j + h_n - 1 = w_{h_n}\}},$$

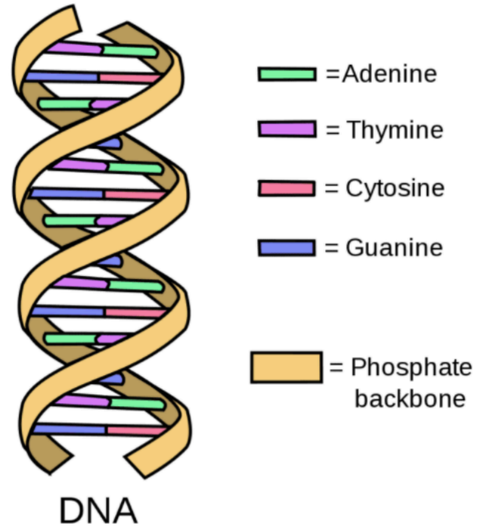
$$\mathbf{E}[Z_j] = \mu(W_n)$$

$$I = \{1, \ldots, n - h_n + 1\}$$

Example : $W_n = \text{ACTAA}$.

GACTAACTAAACTAATGAAACTAACG

Application 2 : a rare word in a DNA sequence



Series $X_1X_2\ldots X_n$ from $\mathscr{A} = \{\text{A,C,G,T}\}$

Rare word : $W_n = w_1w_2\ldots w_{h_n}$ of length $h_n \sim \log(n)$

Question : number of occurrences of W_n ?

$$\mathfrak{T}(W_n) = \sum_{j\in I} Z_j = \sum_{j\in I} \mathbf{1}_{\{X_j=w_1,\ldots,X_j+h_n-1=w_{h_n}\}},$$

$$\mathbf{E}[Z_j] = \mu(W_n)$$

$$I = \{1,\ldots,n-h_n+1\}$$

Example : $W_n = \text{ACTAA}$.

GACTAACTAAACTAATGAAACTAACG
ACTAA

GACTAACTAAACTAATGAAACTAACG

Clumps

3-clump at $j = 2$, 1-clump at $j = 20$

Motivation Poisson approximation ?

Head runs

Average size of clump : $1 + p + p^2 + \dots$

Declump $X_k^n = (1 - C_{k-1}) \prod_{i=k}^{k+m_n-1} C_i, C_i \sim \text{Ber}(p),$

$U_n = \sum_{k \in I} X_k^n$: nb of runs of lenght m_n

DNA sequence

Motivation Poisson approximation ?

Head runs $U_n \sim \mathscr{P}(\lambda_n)$?

Declump $X_k^n = (1 - C_{k-1}) \prod_{i=k}^{k+m_n-1} C_i, C_i \sim \mathrm{Ber}(p),$

Average size of clump : $1 + p + p^2 + \ldots$

$$U_n = \sum_{k \in I} X_k^n : \text{nb of runs of lenght } m_n$$

DNA sequence

Motivation Poisson approximation ?

Head runs $U_n \sim \mathscr{P}(\lambda_n)$?

Average size of clump : $1 + p + p^2 + \ldots$

Declump $X_k^n = (1 - C_{k-1}) \prod_{i=k}^{k+m_n-1} C_i, C_i \sim \mathrm{Ber}(p),$

$U_n = \sum_{k \in I} X_k^n$: nb of runs of lenght m_n

DNA sequence

$\mathbf{P}(\text{occ . word once at } j) = \lambda_n$

Motivation Poisson approximation ?

Average size of clump : $1 + p + p^2 + \ldots$

Head runs $U_n \sim \mathscr{P}(\lambda_n)$?

Declump $X_k^n = (1 - C_{k-1}) \prod_{i=k}^{k+m_n-1} C_i, C_i \sim \mathrm{Ber}(p),$

$U_n = \sum_{k \in I} X_k^n$: nb of runs of lenght m_n

DNA sequence $\mathfrak{T}(W_n) \sim \mathscr{PC}(\lambda_n, \mathbf{V})$?

$\mathbf{P}(\text{occ . word } k \text{ times at } j) = \lambda_n \mathbf{V}(\{k\})$

Motivation Poisson approximation ?

Average size of clump : $1 + p + p^2 + \dots$

Head runs $U_n \sim \mathscr{P}(\lambda_n)$?

Declump $X_k^n = (1 - C_{k-1}) \prod_{i=k}^{k+m_n-1} C_i, C_i \sim \mathrm{Ber}(p),$

$U_n = \sum_{k \in I} X_k^n$: nb of runs of lenght m_n

DNA sequence $\mathfrak{T}(W_n) \sim \mathscr{PC}(\lambda_n, \mathbf{V})$?

$\mathbf{P}(\text{occ . word } k \text{ times at } j) = \lambda_n \mathbf{V}(\{k\})$

Question : How to quantify the convergence rate for TV- distance ?

For $\mathbf{P}, \mathbf{Q}$ probability measures on $\mathbb{N} = \{0,1,\dots\}$ $d_{\mathrm{TV}}(\mathbf{P}, \mathbf{Q}) = \sup_{A \subseteq \mathbb{N}} |\mathbf{P}(A) - \mathbf{Q}(A)|$

Chen-Stein method

Equivalent problem $\mathbf{P}^\star = \mathscr{PC}(\lambda_0, \mathbf{V}), F$: $\mathbb{N}$**-valued**

L : Stein operator $L\varphi(F) = \lambda_0 \sum_{k\geqslant 1} k\varphi(F+k)\mathbf{V}(\{k\}) - F\varphi(F)$

$(\psi_A, A \subset \mathbb{N})$ Stein's class

$$d_{\mathrm{TV}}(\mathbf{P}_F, \mathbf{P}^\star) = \sup_{A\subset\mathbb{N}} \left| \mathbf{E}\left[\lambda_0 \sum_{k\geqslant 1} k\psi_A(F+k)\mathbf{V}(\{k\}) - F\psi_A(F)\right]\right|$$

Original approach

$$d_{\mathrm{TV}}(\mathbf{P}_F, \mathbf{P}^\star) = \sup_{A \subset \mathbb{N}} \left| \mathbf{E}\left[\lambda_0 \sum_{k \geqslant 1} k \psi_A(F+k) \mathbf{V}(\{k\}) - F\psi_A(F) \right] \right|$$

Frame $\quad F = \sum_{i \in I} \sum_{k \geqslant 1} X_i \mathbf{1}_{\{\text{occ. at } i \text{ of size } k\}}, \quad X_i = \mathbf{1}_{\{\text{occ. at } i\}} \sim \mathrm{Ber}(p), \quad I \subset \mathbb{N}.$

Neighborhoods of dependence $\quad (\mathcal{N}_i, i \in I) \quad X_i \quad \text{is independent of} \quad \{X_j, j \notin \mathcal{N}_i \cup \{i\}\}$

Approximation

$$d_{\mathrm{TV}}(\mathbf{P}_F, \mathbf{P}^\star) \leqslant b_1 + b_2 + b_3$$

$$b_1 = \sum_{i \in I} (\mathbf{E}[X_i])^2 + \sum_{i \in I} \sum_{j \in \mathcal{N}_i, j \neq i} (\mathbf{E}[X_i])(\mathbf{E}[X_j])$$

$$b_2 = \sum_{i \in I} \sum_{j \in \mathcal{N}_i} \mathbf{E}[X_i X_j]$$

$$b_3 = \sum_{i \in I} \mathbf{E}\left[\left| \mathbf{E}[X_i - p_i \,|\, \sigma(X_j, j \notin \mathcal{N}_i)] \right| \right]$$

New approach

$$d_{\mathrm{TV}}(\mathbf{P}_F, \mathbf{P}^\star) = \sup_{A \subseteq \mathbb{N}} \left| \mathbf{E}\left[\lambda_0 \sum_{k \geqslant 1} k \psi_A(F+k) \mathbf{V}(\{k\}) - F \psi_A(F) \right] \right|$$

Idea 1.

$$F = \sum_{i \in I} \sum_{k \geqslant 1} X_i \mathbf{1}_{\{\text{occ. at } i \text{ of size } k\}}$$

$$= f\Big((\Delta N_1, V_1), \ldots, (\Delta N_n, V_n), \ldots \Big)$$

$$(\Delta N_j)_j \perp\!\!\!\perp (V_j)_j \,.$$

$$\Delta N_j \sim \mathscr{B}(\lambda), V_j \sim \mathbf{V}$$

New approach

$$d_{\mathrm{TV}}(\mathbf{P}_F, \mathbf{P}^\star) = \sup_{A \subseteq \mathbb{N}} \left| \mathbf{E}\left[\lambda_0 \sum_{k \geqslant 1} k \psi_A(F+k) \mathbf{V}(\{k\}) - F \psi_A(F) \right] \right|$$

Idea 1.

$$F = \sum_{i \in I} \sum_{k \geqslant 1} X_i \mathbf{1}_{\{\text{occ. at } i \text{ of size } k\}}$$

$$= f\Big((\Delta N_1, V_1), \ldots, (\Delta N_n, V_n), \ldots \Big)$$

$$(\Delta N_j)_j \perp\!\!\!\perp (V_j)_j \,.$$

$$\Delta N_j \sim \mathscr{B}(\lambda), V_j \sim \mathbf{V}$$

Idea 2.

Nourdin Peccati 2009

New approach

$$d_{\mathrm{TV}}(\mathbf{P}_F, \mathbf{P}^\star) = \sup_{A \subseteq \mathbb{N}} \left| \mathbf{E}\left[\lambda_0 \sum_{k \geqslant 1} k \psi_A(F+k) \mathbf{V}(\{k\}) - F\psi_A(F) \right] \right|$$

Idea 1.

$$F = \sum_{i \in I} \sum_{k \geqslant 1} X_i \mathbf{1}_{\{\text{occ. at } i \text{ of size } k\}}$$

$$= f\big((\Delta N_1, V_1), \ldots, (\Delta N_n, V_n), \ldots\big)$$

$$(\Delta N_j)_j \perp\!\!\!\perp (V_j)_j \,.$$

$$\Delta N_j \sim \mathscr{B}(\lambda), V_j \sim \mathbf{V}$$

Idea 2.

$\mathbf{E}[F\psi_A(F)]$

Nourdin Peccati 2009

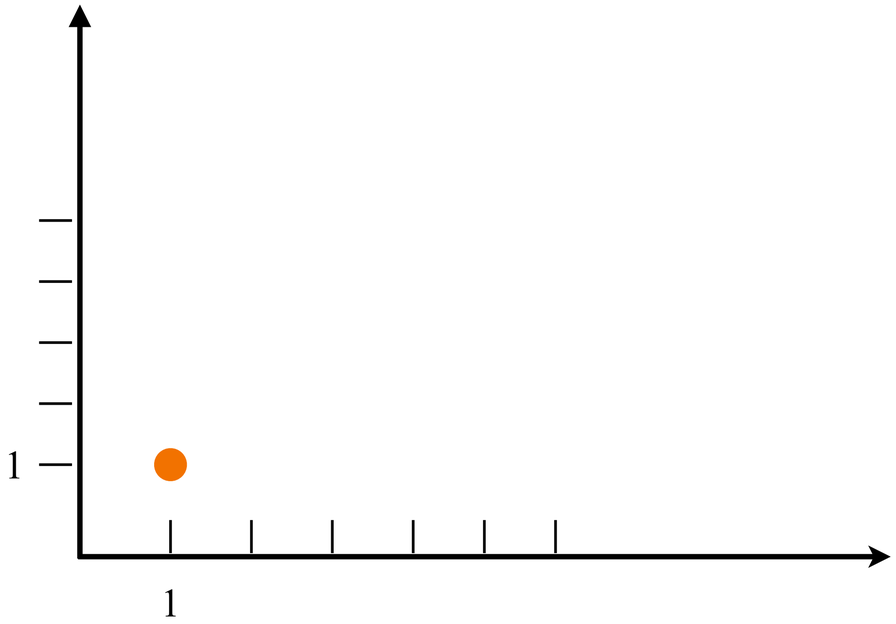
Integration by parts ?

Malliavin calculus for marked binomial processes (MBP)

Framework

Marked binomial processes $(\Omega, \mathscr{F}, \mathbf{P})$

E countable = *mark space*.



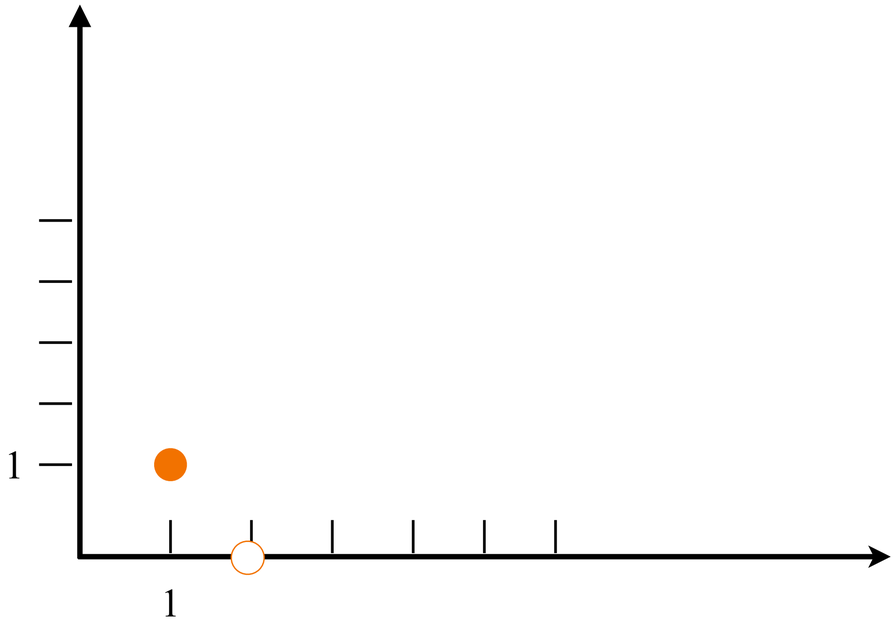
$\underline{n = 1}$ $\quad \Delta N_1 = 1,\ V_1 = 1$

$\eta(1,1) = 1$

Framework

Marked binomial processes $(\Omega, \mathscr{F}, \mathbf{P})$

E countable = *mark space*.



$\underline{n = 1}$ $\quad \Delta N_1 = 1,\ V_1 = 1$

$\eta(1,1) = 1$

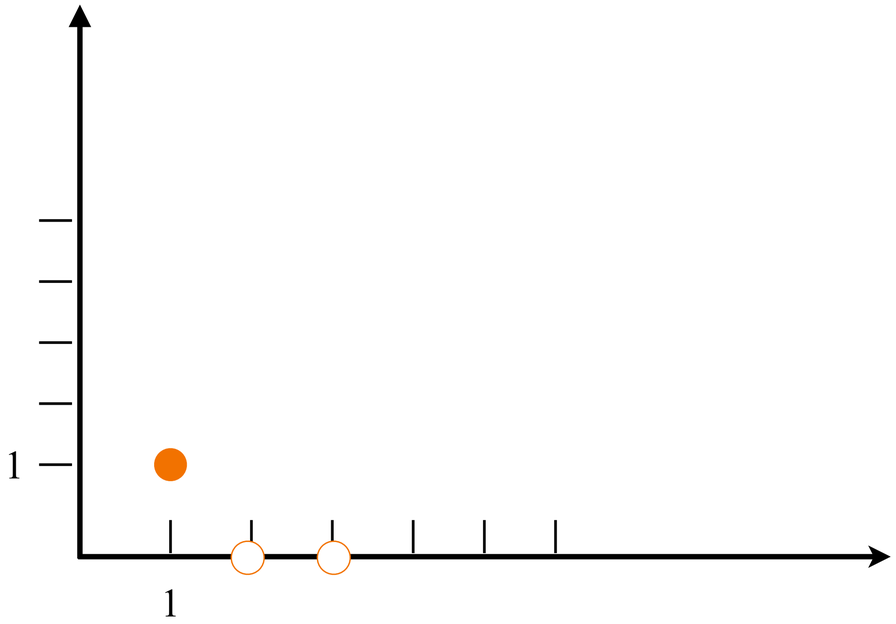
$\underline{n = 2}$ $\quad \Delta N_2 = 0$

$\eta(2,\cdot) = \sum_{k \in \mathrm{E}} \eta(2,k) = 0$

Framework

Marked binomial processes $(\Omega, \mathscr{F}, \mathbf{P})$

E countable = *mark space*.



$\underline{n = 1}$ $\quad \Delta N_1 = 1,\ V_1 = 1$

$$\eta(1,1) = 1$$

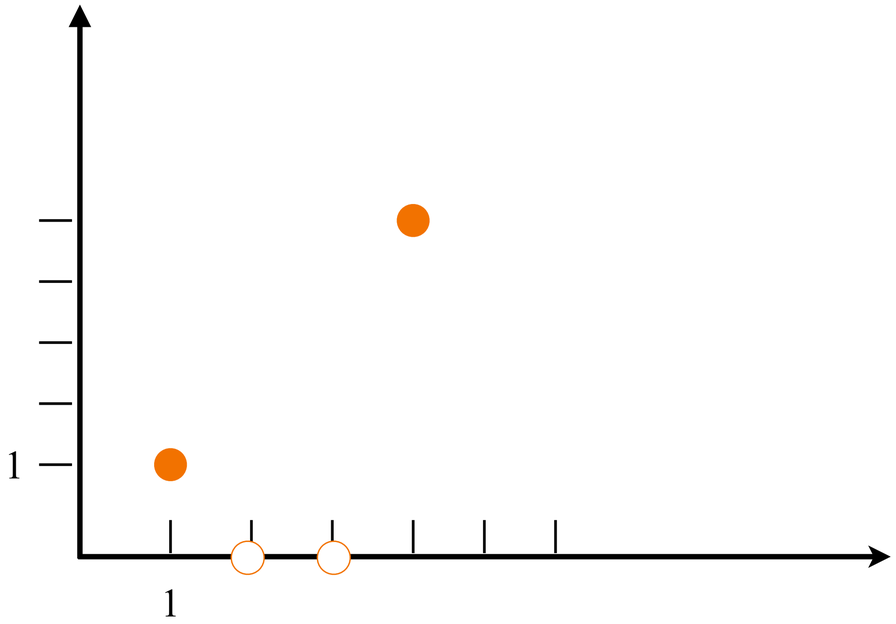
$\underline{n = 2,3}$ $\quad \Delta N_2, \Delta N_3 = 0$

$$\eta(2,\cdot) = \sum_{k \in \mathrm{E}} \eta(2,k) = 0$$

Framework

Marked binomial processes $(\Omega, \mathscr{F}, \mathbf{P})$

E countable = *mark space*.



$\underline{n = 1}$ $\quad \Delta N_1 = 1,\ V_1 = 1$

$\eta(1,1) = 1$

$\underline{n = 2,3}$ $\quad \Delta N_2, \Delta N_3 = 0$

$$\eta(2,\cdot) = \sum_{k \in \mathrm{E}} \eta(2,k) = 0$$

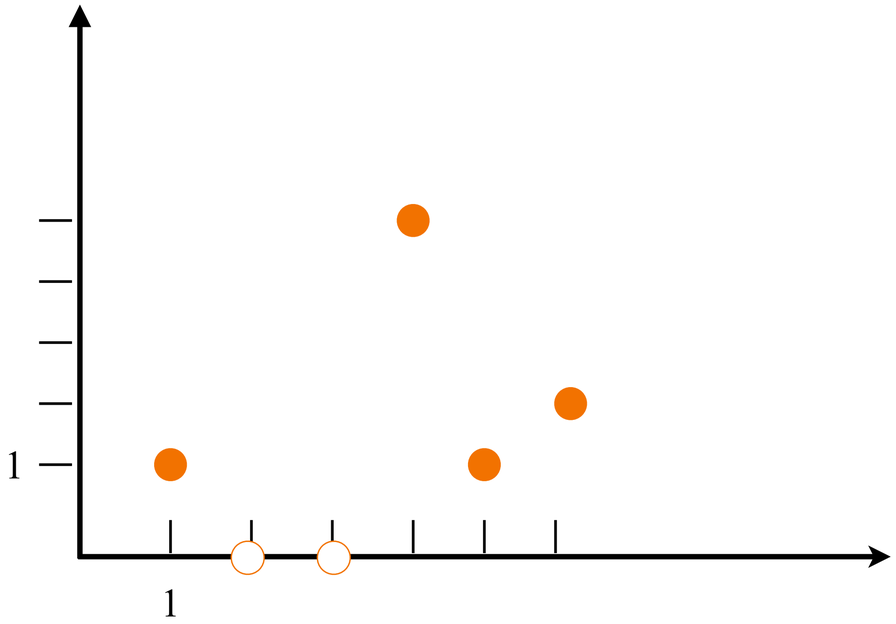
$\underline{n = 4}$ $\quad \Delta N_4 = 1,\ V_2 = 5$

$\eta(4,5) = 1$

Framework

Marked binomial processes $(\Omega, \mathscr{F}, \mathbf{P})$

E countable = *mark space*.



$\underline{n = 1}$ $\Delta N_1 = 1,\ V_1 = 1$

$\eta(1,1) = 1$

$\underline{n = 2,3}$ $\Delta N_2, \Delta N_3 = 0$

$\eta(2,\cdot) = \sum_{k \in \mathrm{E}} \eta(2,k) = 0$

$\underline{n = 4}$ $\Delta N_4 = 1,\ V_2 = 5$

$\eta(4,5) = 1$

$\underline{n = 5,6}$ Etc.

Framework

Marked binomial processes $(\Omega, \mathcal{F}, \mathbf{P})$

MBP $(\lambda, \mathbf{V})$ $\mathbb{X} = \mathbb{N}^* \times \mathrm{E}$

$$\eta = \sum_{t=1}^{\infty} \mathbf{1}_{\{T_t < \infty\}} \, \delta_{(T_t, V_t)}$$

T_t t-th jump, $T_0 = 0$ and $T_t - T_{t-1} \sim \mathcal{G}(\lambda)$, i.i.d. $\quad N_0 = 0,\ \Delta N_t = N_t - N_{t-1} = \mathbf{1}_{\{\text{jump at } t\}}$

V_t mark of the t-th jump i.i.d. $V_t \sim \mathbf{V}, (V_t)_t \perp\!\!\!\perp (T_t)_t$

Intensity of MBP $(\lambda, \mathbf{V})$ $\quad d\nu(t,k) = \lambda \mathbf{V}(\{k\}) \delta_{(t,k)} \quad \Delta N_t \sim \mathrm{Ber}(\lambda)$

Functionals and processes $(\Omega, \mathscr{A}, \mathbf{P})$

Marked binomial functionals

$F = \mathfrak{f}(\eta)$ $\mathfrak{f}$ representative of F

$L^p(\mathbf{P}) = \{F \ : \ \mathbf{E}[\,|F|^p\,] < \infty\}$ $p \geqslant 1$

Marked binomial processes

$u = \displaystyle\sum_{(t,k)\in\mathbb{X}} \mathfrak{u}(\eta, (t,k))\mathbf{1}_{(t,k)}$ $\mathfrak{u}$ representative of u

$L^p(\mathbf{P}\otimes\nu) = \left\{ u \ : \ \mathbf{E}\left[\displaystyle\sum_{(t,k)} |\, u(\,\cdot\,, (t,k))\,|^p \right] < \infty \right\}$

Filtration $(\mathscr{F}_t)_{t\geqslant 1}$ $\mathscr{F}_t = \sigma\{\eta(s,k), s \leqslant t, k \in \mathrm{E}\}\,.$

L^1-theory

Mecke formula

$\mathfrak{u}$ representative of u

Lemma (H)

For $\mathfrak{u}$ non-negative,

$$\mathbf{E}\left[\sum_{(t,k)\in\eta}\mathfrak{u}(\eta,(t,k))\right]=\mathbf{E}\left[\int_{\mathbb{X}}\mathfrak{u}\left(\pi_t(\eta)+\delta_{(t,k)},(t,k)\right)d\nu(t,k)\right]$$

L^1-theory

Mecke formula

Lemma (H)

For $\mathfrak{u}$ non-negative,

$$\mathbf{E}\left[\sum_{(t,k)\in\eta}\mathfrak{u}(\eta,(t,k))\right]=\mathbf{E}\left[\int_{\mathbb{X}}\mathfrak{u}\left(\pi_t(\eta)+\delta_{(t,k)},(t,k)\right)d\nu(t,k)\right]$$

where

$$\pi_t(\eta)=\sum_{s\neq t}\sum_{k\in\mathrm{E}}\eta(s,k)$$

L^1-Operators and integration by parts

Add-one cost operator

$$D^+_{(t,k)}F = \mathfrak{f}(\pi_t(\eta) + \delta_{(t,k)})) - \mathfrak{f}(\pi_t(\eta))$$

« Skorokhod integral »

$$\widetilde{\delta}(u) = \sum_{(t,k)\in\mathbb{X}} u_{(t,k)}\Delta Z_{(t,k)}$$

Proposition. L^1 **integration by parts (H)**

For u predictable, F s.t. $\widetilde{D}Fu \in L^1(\mathbf{P}\otimes\nu)$,

$$\mathbf{E}\left[\int_{\mathbb{X}} \widetilde{D}_{(t,k)}F\, u_{(t,k)}\, d\nu(t,k)\right] = \mathbf{E}[F\widetilde{\delta}(u)]$$

where $\widetilde{D}_{(t,k)}F = \mathfrak{f}(\pi_t(\eta) + \delta_{(t,k)}) - \mathfrak{f}(\eta)$.

$$\Delta Z_{(t,k)} := \mathbf{1}_{\{\eta(s,k)=1\}} - \lambda\mathbf{V}(\{k\})$$

- $\mathbf{E}[\Delta Z_{(t,k)}] = 0$
- $\Delta Z_{(t,\cdot)} \perp\!\!\!\perp \Delta Z_{(s,\cdot)}$ for $t \neq s$
- $\mathbf{E}[\Delta Z_{(t,k)}\Delta Z_{(t,\ell)}] \neq 0.$

L^1-Operators and integration by parts

Add-one cost operator

$$D^+_{(t,k)}F = \mathfrak{f}(\pi_t(\eta) + \delta_{(t,k)})) - \mathfrak{f}(\pi_t(\eta))$$

« Skorokhod integral »

$$\widetilde{\delta}(u) = \sum_{(t,k)\in\mathbb{X}} u_{(t,k)}\Delta Z_{(t,k)}$$

Proposition. L^1 **integration by parts (H)**

For u predictable, F s.t. $\widetilde{D}Fu \in L^1(\mathbf{P}\otimes\nu)$,

$$\mathbf{E}\left[\int_{\mathbb{X}} \widetilde{D}_{(t,k)}F\, u_{(t,k)}\, d\nu(t,k)\right] = \mathbf{E}[F\widetilde{\delta}(u)]$$

where $\widetilde{D}_{(t,k)}F = \mathfrak{f}(\pi_t(\eta) + \delta_{(t,k)}) - \mathfrak{f}(\eta)$.

$$\Delta Z_{(t,k)} := \mathbf{1}_{\{\eta(s,k)=1\}} - \lambda\mathbf{V}(\{k\})$$

- $\mathbf{E}[\Delta Z_{(t,k)}] = 0$
- $\Delta Z_{(t,\cdot)} \perp\!\!\!\perp \Delta Z_{(s,\cdot)}$ for $t \neq s$
- $\mathbf{E}[\Delta Z_{(t,k)}\Delta Z_{(t,\ell)}] \neq 0.$

$$\Delta Z_{(t,k)} = \mathbf{1}_{\{(t,k)\in\eta\}} - \lambda \mathbf{V}(\{k\})$$

L^2-theory

The family $\mathscr{R}$

$$\Delta R_0 = 1, \quad \Delta R_{(t,1)} = \Delta Z_{(t,1)} \quad \text{and} \quad \Delta R_{(t,k)} = \Delta Z_{(t,k)} - \sum_{j=1}^{k-1} \frac{\mathbf{E}[\Delta Z_{(t,k)} \Delta R_{(t,j)}]}{\mathbf{E}[(\Delta R_{(t,j)})^2]} \Delta R_{(t,j)}$$

The $\Delta R_{(\cdot,\cdot)}$ satisfy

- $\mathbf{E}[\Delta R_{(t,k)}] = 0$
- $\Delta R_{(t,\cdot)} \perp\!\!\!\perp \Delta R_{(s,\cdot)}$ for $t \neq s$
- $\mathbf{E}[\Delta R_{(t,k)} \Delta R_{(t,\ell)}] = 0$ for $k \neq \ell$ and $\mathbf{E}[(\Delta R_{(t,k)})^2] =: \kappa_k$
- $\Delta R_{(t,k)} = \Delta Z_{(t,k)} - \sum_{j=1}^{k-1} \rho_j \ \Delta Z_{(t,j)}$

Chaotic decomposition $(\mathbf{t}_n, \mathbf{k}_n) = ((t_1, k_1), \ldots, (t_n, k_n))$

$$\Delta Z_{(t,k)} = \mathbf{1}_{(t,k)} - \lambda \mathbf{V}(\{k\})$$
$$\Delta R_{(t,k)} = \Delta Z_{(t,k)} - \sum_{j=1}^{k-1} \rho_j \Delta Z_{(t,j)}$$
$$\mathbf{E}[(\Delta R_{(t,k)})^2] =: \kappa_k$$

Hilbert space of symmetric functions

$$\langle f_n, g_n \rangle_{L^2(\widetilde{\nu})^{\circ n}} = n! \int_{\mathbb{X}^{n,<}} f_n(\mathbf{t}_n, \mathbf{k}_n) g_n(\mathbf{t}_n, \mathbf{k}_n) d\widetilde{\nu}^{\otimes n}(\mathbf{t}_n, \mathbf{k}_n) \text{ with } \widetilde{\nu}(t,k) = \kappa_k \, \nu(t,k)$$

Proposition. $\mathscr{R}$-integration (H)

The $\mathscr{R}$-integral of $f_n \in \mathscr{C}^n(\mathbb{X}^n, \mathbb{R})$ is

$$J_0(f_0) = f_0 \quad \text{and} \quad J_n(f_n) = n! \sum_{(\mathbf{t}_n, \mathbf{k}_n)} f_n(\mathbf{t}_n, \mathbf{k}_n) \prod_{i=1}^{n} \Delta R_{(t_i,k_i)} \quad (n \geqslant 1)$$

Isometry property : for $f_n \in L^2(\widetilde{\nu})^{\circ n}, g_m \in L^2(\widetilde{\nu})^{\circ m}$,

$$\mathbf{E}[J_n(f_n) J_m(g_m)] = \mathbf{1}_n(m) n! \, \langle f_n, g_n \rangle_{L^2(\widetilde{\nu})^{\circ n}}$$

Chaotic decomposition

$$J_n(f_n) = n! \sum_{(\mathbf{t}_n,\mathbf{k}_n)} f_n(\mathbf{t}_n,\mathbf{k}_n) \prod_{i=1}^{n} \Delta R_{(t_i,k_i)}$$

Marked binomial functionals

$$F = f_0 \mathbf{1}_{\{\eta(\mathbb{X})=0\}} + \sum_{n\geqslant 1} \sum_{(\mathbf{t}_n,\mathbf{k}_n)\in\mathbb{X}^n} f_n(\mathbf{t}_n,\mathbf{k}_n) \mathbf{1}_{\{\eta(\mathbb{X})=n\}} \prod_{i=1}^{n} \mathbf{1}_{\{\eta(t_i,k_i)=1\}} \; ; \quad f_n \in L^2(\widetilde{\nu})^{\circ n}$$

Marked binomial chaoses

$$\mathscr{H}_0 = \mathbf{R} \quad \text{and} \quad \mathscr{H}_n = \{J_n(f_n)\,;\, f_n \in L^2(\widetilde{\nu})^{\circ n}\}$$

Theorem. Chaotic decomposition (H)

$$L^2(\mathbf{P}) = \bigoplus_{n\in\mathbf{N}} \mathscr{H}_n$$

In other words, $F \in L^2(\mathbf{P})$ can be uniquely expanded

$$F = \mathbf{E}[F] + \sum_{n\in\mathbf{N}^*} J_n(f_n)$$

L^2-Integration by parts - Malliavin calculus

Annihilation operator

$$D_{(t,k)}J_n(f_n) = nJ_{n-1}\big(f_n(\star,(t,k))\mathbf{1}_{\{1,\dots,t-1\}^n}\big)$$

Itô-Wiener integral

$$\delta(u) = J_1(u) = \sum_{(t,k)} u_{(t,k)}\Delta R_{(t,k)}$$

Proposition. (Extended) L^2-Integration by parts (H)

For any $(F,u) \in \mathrm{dom}D \times \mathrm{dom}\delta$,

$$\mathbf{E}[F\delta u] = \mathbf{E}\big[\langle DF, u\rangle_{L^2(\mathbb{X},\widetilde{\nu})}\big]$$

$L^1 - L^2$ correspondence?

$$\Delta Z_t = \mathbf{1}_{\{\eta(t)=1\}} - \lambda$$

A particular case : $\mathrm{E} = \{1\}$

	L^1 -theory	L^2 -theory
$F \in L^2(\mathbf{P})$ s.t. $D^+F \in L^2(\mathbf{P} \otimes \nu)$	$D^+F = DF$, $\mathbf{P}$-a.s.	
$u \in \mathrm{dom}\delta$	$\widetilde{\delta}(u) = \sum_t u(\eta, t)\Delta Z_t = \delta(u)$	
$F \in \mathrm{dom}L$ $D^+F \in L^2(\mathbf{P} \otimes \nu)$	$\widetilde{L}F = -\widetilde{\delta}(D^+F) = -\delta(DF) = -\sum_{n \geqslant 1} nJ_n(f_n) = LF$	

$L^1 - L^2$ correspondence?

The general case : E ***countable***

	L^1 -theory	L^2 -theory
$F \in L^2(\mathbf{P})$ s.t. $D^+F \in L^2(\mathbf{P} \otimes \nu)$	$D^+F = DF$, $\mathbf{P}$-a.s.	
u simple	$\widetilde{\delta}(u) = \sum_{(t,k)} u(\eta,(t,k))\, \Delta Z_{(t,k)}$	$\neq \quad \delta(u) = \sum_{(t,k)} u(\eta,(t,k)) \Delta R_{(t,k)}$
$F \in \mathrm{dom} L$ $D^+F \in L^2(\mathbf{P} \otimes \nu)$	$\widetilde{L}F = -\widetilde{\delta}(D^+F)$	$\neq \quad LF = -\sum_{n \geqslant 1} nJ_n(f_n) = -\delta(DF)$
$F \in L^2(\mathbf{P})$	The operator $\widetilde{L}$ is not the generator of $(P_\tau)_{\tau \geqslant 0}$	$P_\tau F = \sum_{n \geqslant 0} e^{-n\tau} J_n(f_n)$

No chance to get a L^1/L^2 unified theory

Mehler's formula and corollaries

Thinning

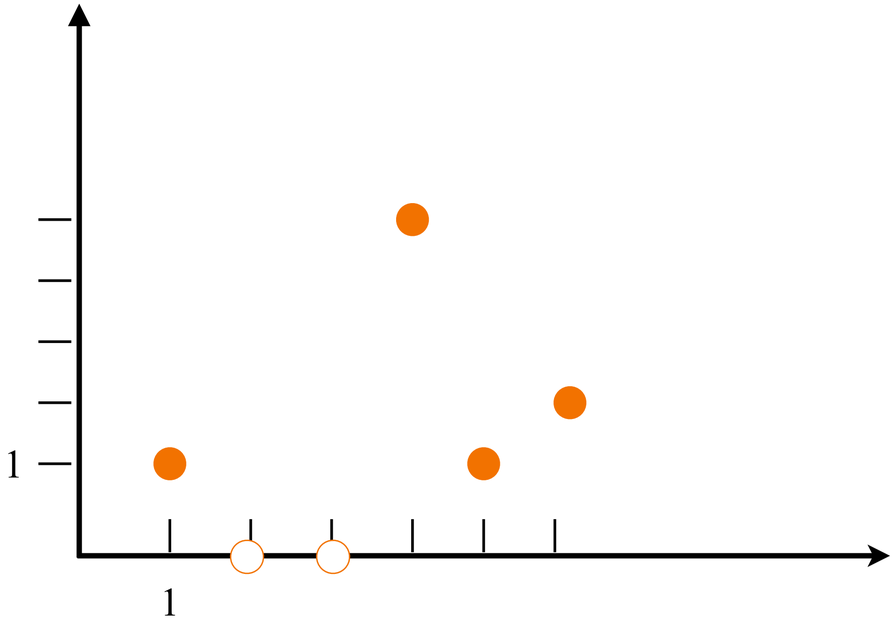
Let $\varepsilon_t^\tau := \mathbf{1}_{\{\theta_t \leqslant \tau\}}$ where $\theta_t \sim \mathscr{E}(1)$ i.i.d.

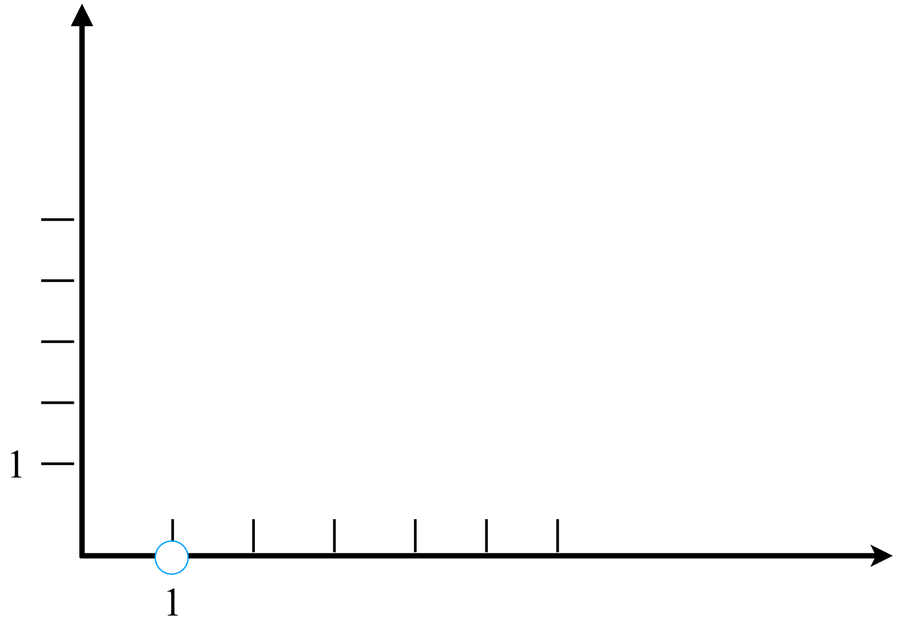
$\eta^{\tau,0} := \eta \mathbf{1}_{\{\varepsilon_t^\tau = 0\}}$ is a MBP of intensity $e^{-\tau}\nu$.

Original process

$\varepsilon_1^\tau = 1$

Thinned process





Mehler's formula and corollaries

Thinning

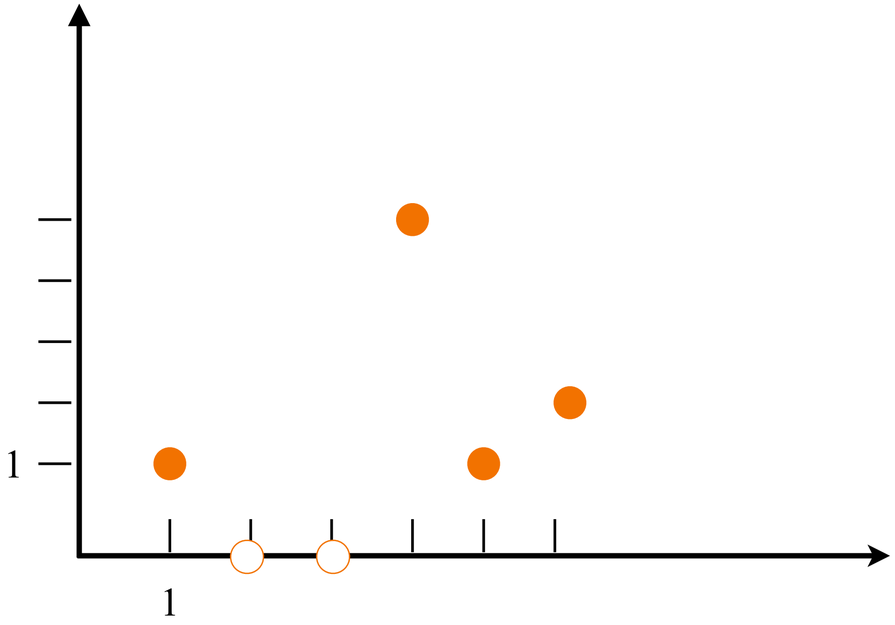
Let $\varepsilon_t^\tau := \mathbf{1}_{\{\theta_t \leqslant \tau\}}$ where $\theta_t \sim \mathscr{E}(1)$ i.i.d.

$\eta^{\tau,0} := \eta \mathbf{1}_{\{\varepsilon_t^\tau = 0\}}$ is a MBP of intensity $e^{-\tau}\nu$.

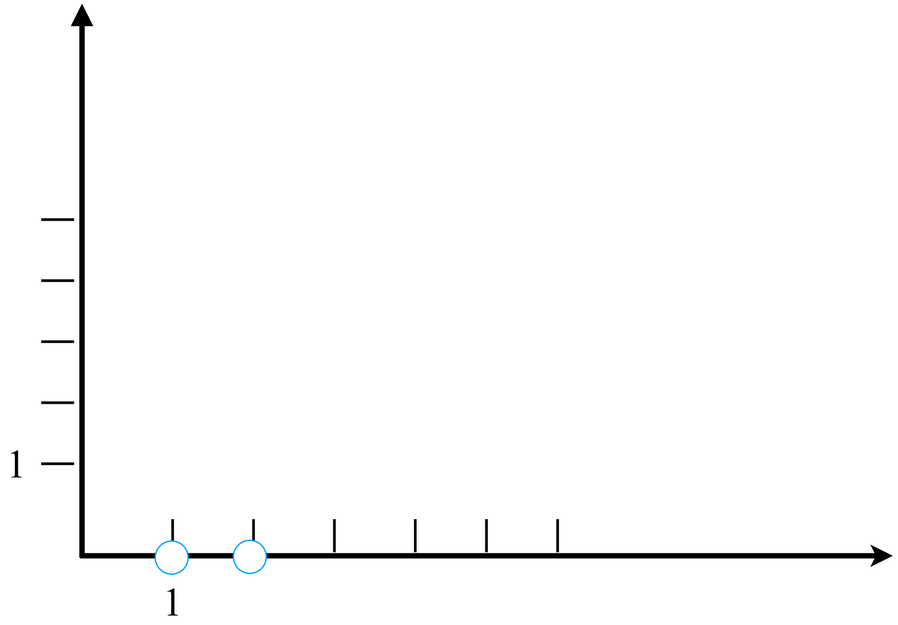
Original process

$\varepsilon_1^\tau = 1$

$\varepsilon_2^\tau = 0$



Thinned process



Mehler's formula and corollaries

Thinning

Let $\varepsilon_t^\tau := \mathbf{1}_{\{\theta_t \leqslant \tau\}}$ where $\theta_t \sim \mathscr{E}(1)$ i.i.d.

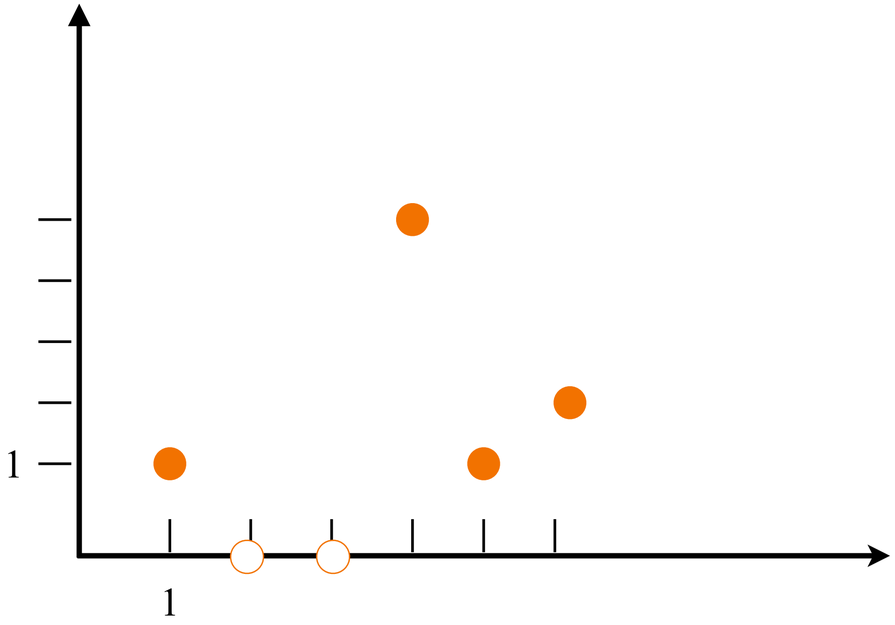
$\eta^{\tau,0} := \eta \mathbf{1}_{\{\varepsilon_t^\tau = 0\}}$ is a MBP of intensity $e^{-\tau}\nu$.

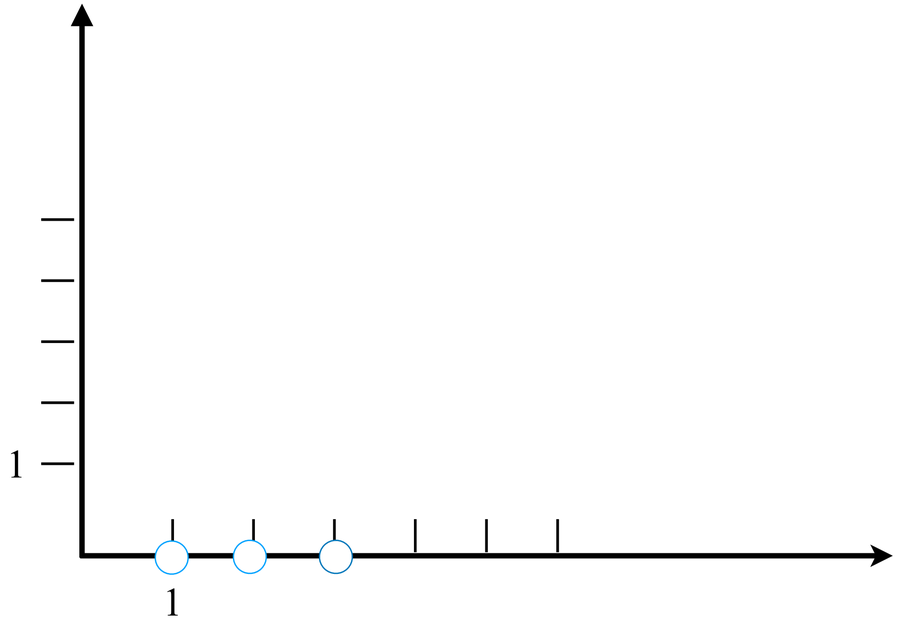
Original process

$\varepsilon_1^\tau = \varepsilon_3^\tau = 1$

$\varepsilon_2^\tau = 0$

Thinned process





Mehler's formula and corollaries

Thinning

Let $\varepsilon_t^\tau := \mathbf{1}_{\{\theta_t \leqslant \tau\}}$ where $\theta_t \sim \mathscr{E}(1)$ i.i.d.

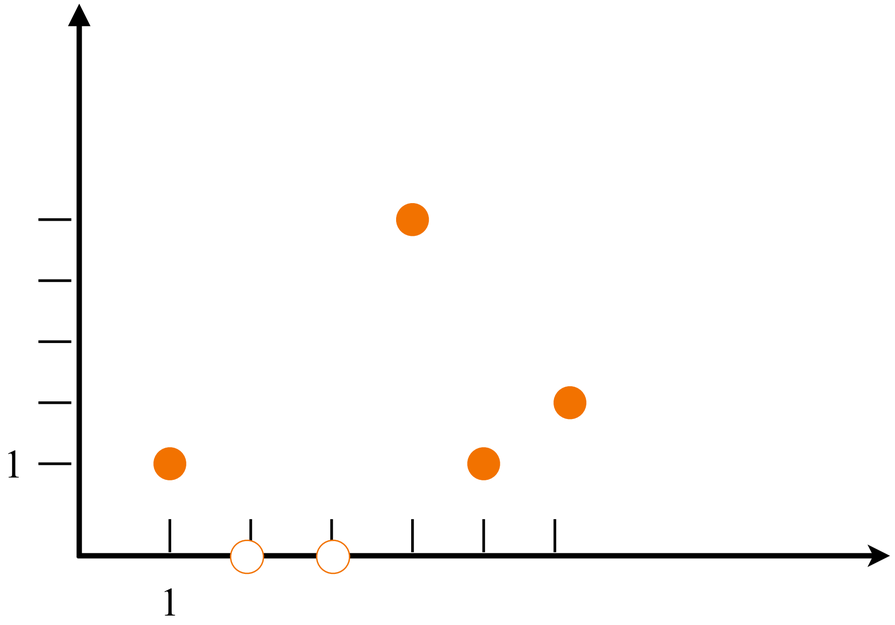
$\eta^{\tau,0} := \eta \mathbf{1}_{\{\varepsilon_t^\tau = 0\}}$ is a MBP of intensity $e^{-\tau}\nu$.

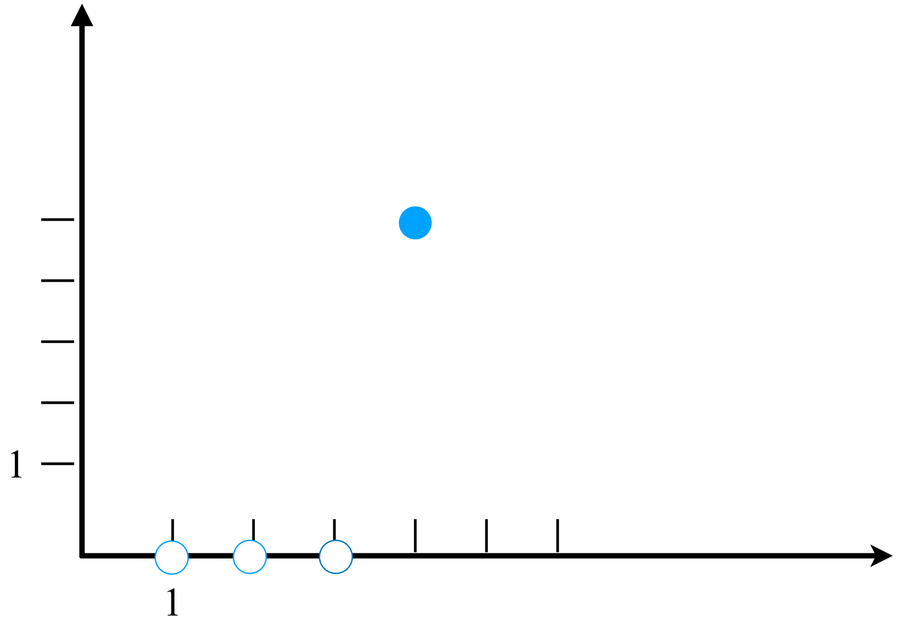
Original process

$$\varepsilon_1^\tau = \varepsilon_3^\tau = 1$$

$$\varepsilon_2^\tau = \varepsilon_4^\tau = 0$$

Thinned process





Mehler's formula and corollaries

Thinning

Let $\varepsilon_t^\tau := \mathbf{1}_{\{\theta_t \leqslant \tau\}}$ where $\theta_t \sim \mathscr{E}(1)$ i.i.d.

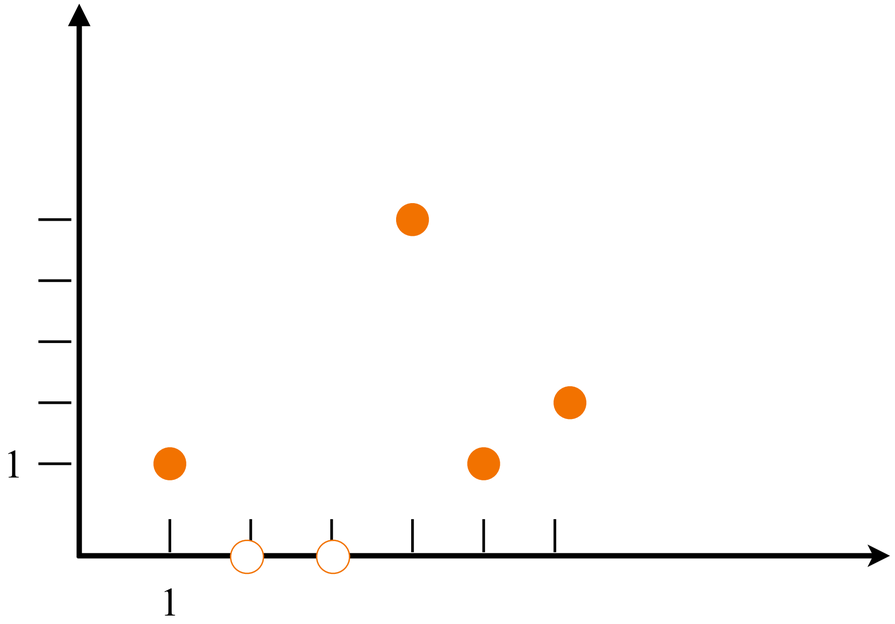
$\eta^{\tau,0} := \eta \mathbf{1}_{\{\varepsilon_t^\tau = 0\}}$ is a MBP of intensity $e^{-\tau}\nu$.

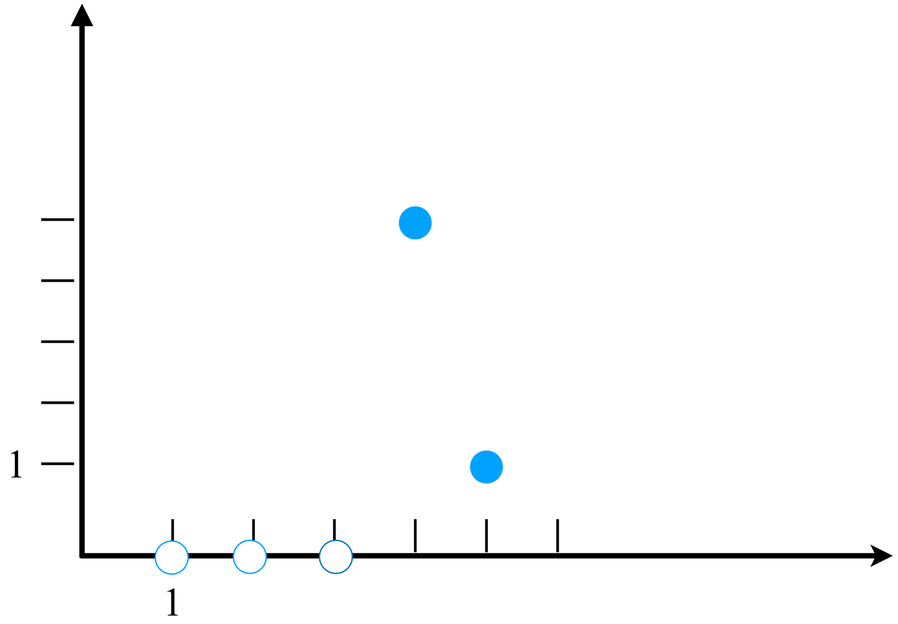
Original process

$$\varepsilon_1^\tau = \varepsilon_3^\tau = 1$$

$$\varepsilon_2^\tau = \varepsilon_4^\tau = \varepsilon_5^\tau = 0$$

Thinned process





Mehler's formula and corollaries

Thinning

Let $\varepsilon_t^\tau := \mathbf{1}_{\{\theta_t \leqslant \tau\}}$ where $\theta_t \sim \mathscr{E}(1)$ i.i.d.

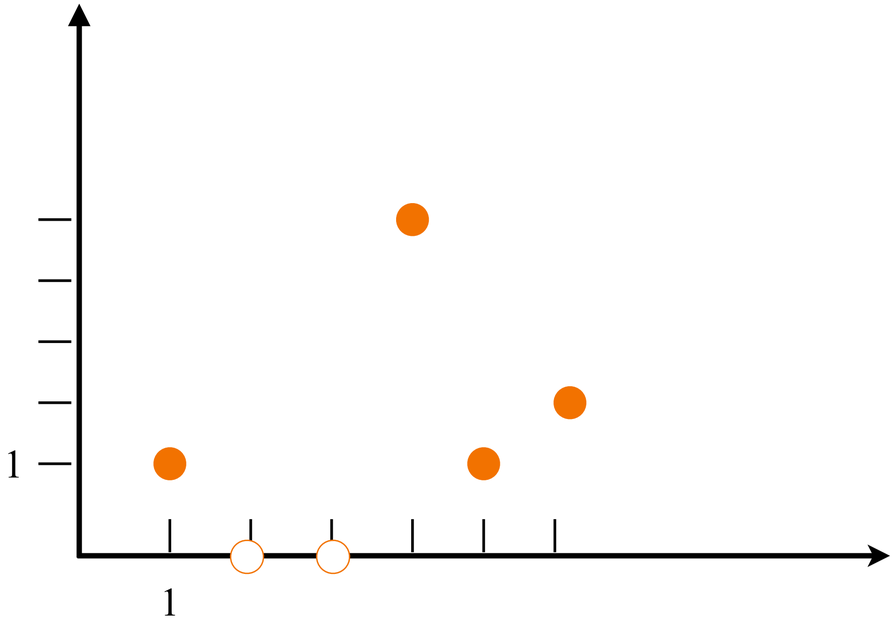
$\eta^{\tau,0} := \eta \mathbf{1}_{\{\varepsilon_t^\tau = 0\}}$ is a MBP of intensity $e^{-\tau}\nu$.

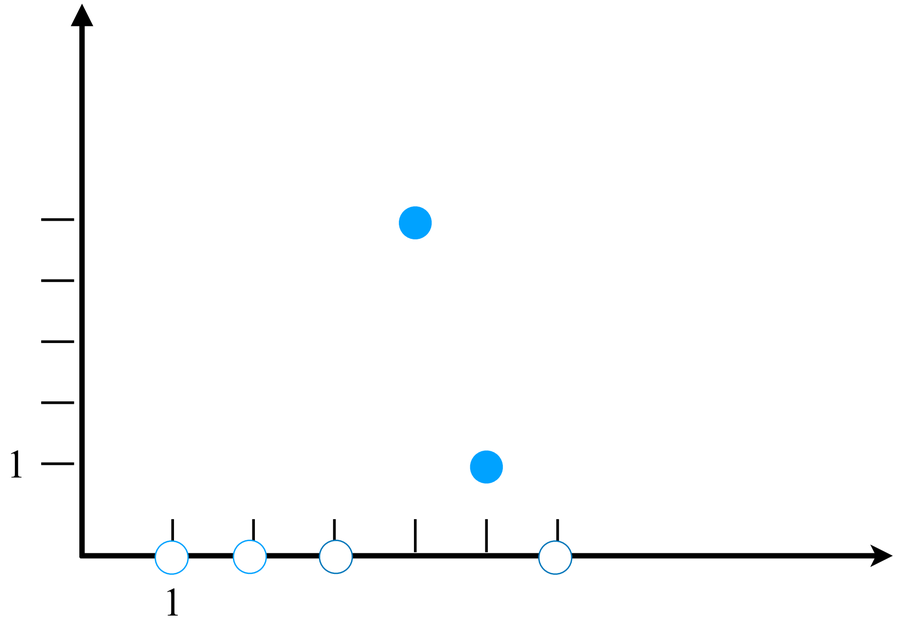
Original process

$$\varepsilon_1^\tau = \varepsilon_3^\tau = \varepsilon_6^\tau = 1$$

$$\varepsilon_2^\tau = \varepsilon_4^\tau = \varepsilon_5^\tau = 0$$

Thinned process





Integral representation of $(P_\tau)_{\tau\geqslant 0}$

Proposition (H)

Let η **finite** and $F = \mathfrak{f}(\eta) \in L^1(\mathbf{P})$. For $\tau \in \mathbf{R}_+$,

$$P_\tau F = \int \mathbf{E}\left[\mathfrak{f}(\eta^{\tau,0} + \varepsilon\widetilde{\eta}) \,|\, \eta\right] \Pi_\nu(d\widetilde{\eta}) \qquad \mathbf{P}\text{-a.s.}$$

where Π_ν is the distribution of a MBP(ν) and the distribution of $\widetilde{\eta} \,|\, \eta$ is s.t.

$$\mathbf{P}\Big((t,k) \in \widetilde{\eta} \,|\, (t,k) \notin \eta\Big) = \lambda \mathbf{V}(\{k\}) \quad \text{and} \quad \mathbf{P}\Big((t,k) \notin \widetilde{\eta} \,|\, (t,k) \in \eta\Big) = 1 - \lambda \mathbf{V}(\{k\})$$

Integral representation of $(P_\tau)_{\tau\geqslant 0}$

Proposition (H)

Let η **finite** and $F = \mathfrak{f}(\eta) \in L^1(\mathbf{P})$. For $\tau \in \mathbf{R}_+$,

$$P_\tau F = \int \mathbf{E}\left[\mathfrak{f}(\eta^{\tau,0} + \varepsilon\widetilde{\eta}) \,|\, \eta\right] \Pi_\nu(d\widetilde{\eta}) \qquad \mathbf{P}\text{-a.s.}$$

where Π_ν is the distribution of a MBP(ν) and the distribution of $\widetilde{\eta} \,|\, \eta$ is s.t.

$$\mathbf{P}\Big((t,k) \in \widetilde{\eta} \,|\, (t,k) \notin \eta\Big) = \lambda \mathbf{V}(\{k\}) \quad \text{and} \quad \mathbf{P}\Big((t,k) \notin \widetilde{\eta} \,|\, (t,k) \in \eta\Big) = 1 - \lambda \mathbf{V}(\{k\})$$

$$\mathbf{P}\Big((t,k) \in \widetilde{\eta} \,|\, (t,k) \in \eta\Big) = \mathbf{P}\Big((t,k) \notin \widetilde{\eta} \,|\, (t,k) \notin \eta\Big) = 0$$

Corollaries

Corollary 1 **Contraction property**

For $F \in L^2(\mathbf{P}), \tau \in \mathbf{R}_+$,

$$DP_\tau F = e^{-\tau} P_\tau DF$$

Corollary 2 **Melher's formula**

For $F \in L^2(\mathbf{P})$ s.t. $\mathbf{E}[F] = 0$,

$$L^{-1}F = -\int_0^{+\infty} P_\tau F \, d\tau, \quad \mathbf{P} \otimes \nu\text{-a.s.} \quad \text{and} \quad DL^{-1}F = -\int_0^{+\infty} e^{-\tau} P_\tau DF \, d\tau,$$

Application : Poisson approximations

Poisson approximation

General bound

$$\widetilde{D}_t F = \mathfrak{f}(\pi_t(\eta) + \delta_t) - \mathfrak{f}(\eta)$$
$$D_t F = D_t^+ F = \mathfrak{f}(\pi_t(\eta) + \delta_t) - \mathfrak{f}(\pi_t(\eta))$$

Theorem (H)

Let a $\mathbb{N}$-valued RV $F \in L^2(\mathbf{P})$ s.t. $\mathbf{E}[F] = \lambda_0 > 0$. Then

$$d_{\mathrm{TV}}(\mathbf{P}_F, \mathscr{P}(\lambda_0)) \leqslant \frac{1 - e^{-\lambda_0}}{\lambda_0} \mathbf{E}\Big[\big| \lambda_0 - \big\langle \widetilde{D} F, -DL^{-1}(F - \mathbf{E}[F]) \big\rangle_{L^2(\nu)} \big| \Big]$$
$$+ \frac{1 - e^{-\lambda_0}}{\lambda_0} \mathbf{E}\Big[\int_{\mathbb{N}^*} \big| \widetilde{D}_t F (\widetilde{D}_t F - 1) D_t L^{-1}(F - \mathbf{E}[F]) \big| \, d\nu(t) \Big] .$$

$$\nabla \varphi_A(F) = \varphi_A(F+1) - \varphi_A(F)$$

Sketch of Proof

1. Prove that (Mehler's formula + contraction property)

$$D^+\varphi_A(F)\, D^+L^{-1}(F - \mathbf{E}[F]) \in L^1(\mathbf{P} \otimes \nu)$$

2. IBP

$$\begin{aligned}
\mathbf{E}\Big[F\varphi_A(F) - \lambda_0\varphi_A(F+1)\Big] &= \mathbf{E}\Big[\big(\widetilde{L}\widetilde{L}^{-1}(F - \mathbf{E}[F])\big)\varphi_A(F)\Big] - \lambda_0\mathbf{E}[\nabla\varphi_A(F)] \\
&= -\mathbf{E}\Big[\int_{\mathbb{N}} \widetilde{D}_t(\varphi_A(F))D_tL^{-1}(F - \mathbf{E}[F])d\nu(t)\Big] - \lambda_0\mathbf{E}[\nabla\varphi_A(F)] \\
&= -\mathbf{E}\Big[\nabla\varphi_A(F)\int_{\mathbb{N}} (\widetilde{D}_tF)D_tL^{-1}(F - \mathbf{E}[F])d\nu(t) + \mathrm{rem_A}\Big] \\
&\qquad - \lambda_0\mathbf{E}[\nabla\varphi_A(F)]
\end{aligned}$$

3. Bound $\mathbf{E}[\,|\mathrm{rem_A}|\,]$ + take $\sup_{A \in \subset \mathbb{N}}$

Application : the longest head run

$$X_k^n = (1 - C_{k-1}) \prod_{i=k}^{k+m_n-1} C_i, \;\; C_i \sim \mathrm{Ber}(p), \text{ i.i.d.}$$

$$U_n = \sum_{k \in I} X_k^n$$

$$U_n = \prod_{i=1}^{m_n} \Delta N_i + \sum_{i=1}^{n} (1 - \Delta N_i) \Delta N_{i+1} \Delta N_{i+2} \dots \Delta N_{i+m_n}$$

Test length $m_n = \log_{1/p}((n-1)(1-p) + 1) + \mathrm{cst}$

Theorem (H)

Let $\lambda_n = p^{m_n}((n-1)(1-p) + 1)$. Then,

$$d_{\mathrm{TV}}(\mathbf{P}_{U_n}, \mathscr{P}(\lambda_n)) \leqslant \frac{p^{m_n}}{\lambda_n} \Big[p^{m_n}[2(m_n - 1)q^2 + 2m_n q + 1] + c m_n^2 (1-p)^2 p^{m_n - 1} + 2(n - m_n + 1)(1-p)p^{m_n} + \mathrm{o}(p^{m_n}) \Big] .$$

Remark The CDF of the longest head run can be approximated by

$$\mathbf{P}(R_n < m_n) = \mathbf{P}(U_n = 0) = e^{-\lambda_n} .$$

Application : the longest head run

$$X_k^n = (1 - C_{k-1}) \prod_{i=k}^{k+m_n-1} C_i, \quad C_i \sim \mathrm{Ber}(p), \text{ i.i.d.}$$

$$U_n = \sum_{k \in I} X_k^n$$

$$U_n = \prod_{i=1}^{m_n} \Delta N_i + \sum_{i=1}^{n} (1 - \Delta N_i) \Delta N_{i+1} \Delta N_{i+2} \dots \Delta N_{i+m_n}$$

Test length $m_n = \log_{1/p}((n-1)(1-p) + 1) + \mathrm{cst}$

Theorem (H)

Let $\lambda_n = p^{m_n}((n-1)(1-p) + 1)$. Then,

$$d_{\mathrm{TV}}(\mathbf{P}_{U_n}, \mathscr{P}(\lambda_n)) \leqslant \frac{p^{m_n}}{\lambda_n} \Big[p^{m_n}[2(m_n - 1)q^2 + 2m_n q + 1] + c m_n^2 (1-p)^2 p^{m_n - 1} + 2(n - m_n + 1)(1-p)p^{m_n} + \mathrm{o}(p^{m_n}) \Big] .$$

$\frac{p^{m_n}}{\lambda_n} \sim 1/n$

Remark The CDF of the longest head run can be approximated by

$$\mathbf{P}(R_n < m_n) = \mathbf{P}(U_n = 0) = e^{-\lambda_n} .$$

Compound Poisson approximation

General bound

$$\widetilde{D}_t F = \mathfrak{f}(\pi_t(\eta) + \delta_{t,k}) - \mathfrak{f}(\eta)$$
$$\widetilde{L} F = -\widetilde{\delta} D^+ F$$
$$D^+_{t,k} F = \mathfrak{f}(\pi_t(\eta) + \delta_{t,k}) - \mathfrak{f}(\pi_t(\eta))$$

Theorem (H)

Let $\lambda_0 > 0$, $\mathbf{V}$ probability distribution on $\mathbb{N}^*$.

Let a $\mathbb{N}$-valued RV $F \in L^2(\mathbf{P})$ s.t. $\mathbf{E}[F] = \lambda_0 \mathbf{E}[V_1] > 0$. Then,

$$d_{\mathrm{TV}}(\mathbf{P}_F, \mathscr{PC}(\lambda_0, \mathbf{V})) \leqslant \sup_{A \subset \mathbb{Z}_+} \left| \int_{\mathbb{X}} \left[D^+ \widetilde{L}^{-1} (\mathfrak{f}(\eta) - \mathbf{E}[\mathfrak{f}(\eta)]) \psi_A (\mathfrak{f}(\pi_t(\eta) + \delta_{(t,k)})) - k \psi_A (\mathfrak{f}(\eta) + k) \right] d\nu(t,k) \right|$$

$$+ c_{\mathscr{PC}} \left| \int_{\mathbb{X}} \left[D^+ \widetilde{L}^{-1} (\mathfrak{f}(\eta) - \mathbf{E}[\mathfrak{f}(\eta)]) - k \right] d\nu(t,k) \right|$$

Application : count of a rare word in a DNA sequence

$$\mathfrak{T}(W_n) = \sum_{j\in I} Z_j := \sum_{j\in I} \mathbf{1}_{\{X_j=w_1,\ldots,X_j+h_n-1=w_{h_n}\}}$$

Let $F_n = \sum_{j\in I} V_j \Delta N_j$ where $V_j \sim \mathbf{V} = \mathscr{G}(1-\alpha)$ i.i.d.

$\Delta N_j \sim \mathrm{Ber}((1-\alpha)\mu(W_n))$ i.i.d.

Proposition (H)

Let $\lambda_n = (n-h_n+1)(1-\alpha)\mu(W_n)$ and $\mathbf{V} = \mathscr{G}(1-\alpha)$. Then,

$$d_{\mathrm{TV}}(\mathbf{P}_{\mathfrak{T}(W_n)}, \mathscr{PC}(\lambda_n, \mathbf{V})) \leqslant d_{\mathrm{TV}}(\mathbf{P}_{\mathfrak{T}(W_n)}, \mathbf{P}_{F_n}) + d_{\mathrm{TV}}(\mathbf{P}_{F_n}, \mathscr{PC}(\lambda_n, \mathbf{V}))$$

$$\leqslant 2h_n\mu(W_n) + (n-h_n+1)c_{\mathscr{PC}}\mu(W_n)^2$$

~~Conclusion~~ Open questions

What about inserting dependence in the process ?

Possible applications in stochastic geometry ?

References

- R. Arratia, L. Goldstein, L. Gordon, *Poisson approximation and the Chen-Stein method*, Statistical Science, 1990.
- A. D. Barbour, L. H. Y. Chen, W. L. Loh, *Compound Poisson approximation for nonnegative random variables via Stein's method,* Annals of Probability, 1992.
- G. Last, G. Peccati, M. Schulte, *Normal approximation on Poisson spaces: Mehler's formula, second order Poincaré inequalities and stabilization*, Probability Theory and Related Fields, 2016.
- G. Last, M. Penrose, *Lecture on the Poisson Process*, Cambridge University Press, 2017.
- I. Nourdin, G. Peccati, *Normal Approximations with Malliavin calculus: from Stein's method to universality*, Cambridge University Press, 2012.
- S. Schbath, *Compound approximation of word counts in DNA sequences*, ESAIM: probability and statistics, 2016.
- H. Halconruy, *Malliavin calculus for marked binomial processes and applications*, Electronic Journal of Probability, 2022.

Merci pour votre attention !