Conditional large and moderate deviations for sums of discrete random variables. Combinatoric applications

Fabrice Gamboa, Thierry Klein, Clémentine Prieur

Conditional large and moderate deviations for sums of discrete random variables. Combinatoric applications.

Fabrice GAMBOA[1], Thierry KLEIN[2] and Clémentine PRIEUR[3]

Running title. Conditional large and moderate deviations.

Abstract. We prove large and moderate deviation principles for the distribution of an empirical mean conditioned by the value of the sum of discrete i.i.d. random variables. Some applications for combinatoric problems are discussed.

Key words: Large and moderate deviation principles; Conditional distribution; Combinatoric problems.

Mathematics Subject Classification: 60F10; 60F05; 62E20; 60C05; 60J65; 68W40.

1 Introduction

In many random combinatorial problems, the distribution of the interesting statistic is the law of an empirical mean built on an independent and identically distributed (i.i.d.) sample conditioned by some exogenous integer random variable (r.v.). In general, this exogenous r.v. is also itself a sample mean built on integer r.vs. Hence, a general frame for this kind of problem may be formalized as follows. Let (q_n) be a positive integer sequence. Further, let $\mathbf{X} = (X_j^{(n)})_{n\in\mathbb{N}^*,j=1,\dots,nq_n}$ and $\mathbf{Y} = (Y_j^{(n)})_{n\in\mathbb{N}^*,j=1,\dots,nq_n}$ be two triangular arrays of random variables. Both arrays are such that on their lines the r.vs are i.i.d.. Moreover, it is assumed that the elements of the array $\mathbf{X}$ are integer. The interesting distribution is then the law of $(nq_n)^{-1}T_n := (nq_n)^{-1}\sum_{j=1}^{nq_n} Y_j^{(n)}$ conditioned on a specific value of $S_n := \sum_{j=1}^{nq_n} X_j^{(n)}$. That is the conditional distribution

$$\mathcal{L}_n := \mathcal{L}((nq_n)^{-1}T_n|S_n = np_n),$$

where (p_n) is some given positive integer sequence. When the distribution of $(X_j^{(n)}, Y_j^{(n)})$ does not depends on n, the Gibbs conditioning principle ([22, 5, 6]) states that $\mathcal{L}_n$ converges weakly to the degenerated distribution concentrated on a point χ depending on the conditioning value (see Corollary 2.6). Around the Gibbs conditioning principle, general limit theorems yielding the asymptotic behavior of the conditioned sum are given in [21, 13, 17]. Asymptotic expansions for the distribution of the conditioned sum are proved in [11, 18]. In this paper our aim is

[1]IMT, Equipe de Statistique et Probabilités, UMR 5219, Université Paul Sabatier, 118 Route de Narbonne 31062 Toulouse cedex 4, France. gamboa@math.ups-tlse.fr

[2]IMT, Equipe de Statistique et Probabilités, UMR 5219, Université Paul Sabatier, 118 Route de Narbonne 31062 Toulouse cedex 4, France. tklein@math.ups-tlse.fr

[3]IMT, Equipe de Statistique et Probabilités, UMR 5219, INSA Toulouse, GMM, 135 avenue de Rangueil, 31077 Toulouse cedex 4, France. clementine.prieur@insa-toulouse.fr

to prove a large deviation principle for $\mathcal{L}_n$. Roughly speaking, this means that we will give an exponential equivalent for this conditional distribution. On a finer scale, we prove a large deviation principle for $\tilde{\mathcal{L}}_n := \mathcal{L}(\sqrt{\frac{a_n}{nq_n}}(T_n - b_n)|S_n = np_n)$, where b_n is a centering factor specified in Theorem 2.4 of Section 2.2.3 and a_n is a decreasing positive sequence of real numbers with $a_n \to 0$, $na_nq_n \to +\infty$. We then say that $\tilde{\mathcal{L}}_n$ satisfies a moderate deviation principle [6, Section 3.7]. Our work follows the nice ones of Janson [15, 14]. In these last papers, a central limit theorem with moment convergence is proved. The starting point in the proof is a simple representation of the conditional characteristic function as an inverse Fourier transform. This representation was first given by Bartlett [3, Equation (16)]. To establish large and moderate deviation principles we will make use of Gärtner-Ellis Theorem in which an asymptotic evaluation of the Laplace transform is needed. For this purpose, we first transcribe the Bartlett formula to get a simple integral representation for the conditional Laplace transform (see Lemma 3.1). The main result of [15] is quite general as it only requires assumptions on the three first moments of $(\mathbf{X}, \mathbf{Y})$. Here we need further assumptions. However, contrarily to [15, Section 2], we do not restrict to the central case (conditioning on $S_n = \mathbb{E}(S_n)$) nor on the "pseudo" central case (conditioning on $S_n = \mathbb{E}(S_n) + \mathcal{O}\left(\sqrt{nq_n}\sigma_{X_i^{(n)}}\right)$, with $\sigma_X^2 = \mathrm{Var}(X)$). In [18], the authors study general saddle point approximations for multidimensional discrete empirical means and obtain an approximation formula for conditional probabilities. We focus here on the exponential part of this formula, stating a full large deviation principle (see Theorem 2.3). Using some classical tools of convex analysis we give an explicit natural and elegant form for the rate function. Furthermore, we complement our study by stating a moderate deviation principle for the conditional law (see Theorem 2.4). As usual, the rate function is quadratic and the scaling factor is the asymptotic variance, which can be interpreted here as a residual variance in some linear regression model, generalizing the factor found in [15]. The paper is organized as follows. In the next section, to be self contained, we first recall some classical results on large deviation principles. Then we state our main results: a large deviation principle and a moderate deviation principle for conditioned sums. Section 3 is devoted to the proofs. In Section 4 we apply our main results to some combinatorial examples. We also discuss possible extensions to more general models.

2 Main results

2.1 Large and moderate deviations

2.1.1 Some generalities

Let us first recall what is a large deviation principle (L.D.P.) (see for example [6, 12]). In the whole paper, (a_n) is a decreasing positive sequence of real numbers with $\lim_{n\to\infty} a_n = 0$.

Definition 2.1. *We say that a sequence (R_n) of probability measures on a measurable Hausdorff space $(U, \mathcal{B}(U))$ satisfies a LDP with rate function I and speed (a_n) if:*

i) I is lower semi continuous (lsc), with values in $\mathbb{R}^+ \cup \{+\infty\}$.

ii) For any measurable set A of U:

$$-I(\operatorname{int} A) \le \liminf_{n\to\infty} a_n \log R_n(A) \le \limsup_{n\to\infty} a_n \log R_n(A) \le -I(\operatorname{clo} A),$$

where $I(A) = \inf_{\xi \in A} I(\xi)$ and $\operatorname{int} A$ (resp. $\operatorname{clo} A$) is the interior (resp. the closure) of A.

We say that the rate function I is good if its level set $\{x \in U : \ I(x) \leq a\}$ is compact for any $a \geq 0$. More generally, a sequence of U-valued random variables is said to satisfy a LDP if their distributions satisfy a LDP.

To be self-contained, we also recall some definitions and results which will be used in the sequel. (we refer to [6, 12] for more on large deviations).

Laplace and Fenchel-Legendre transforms To begin with, let Z be a non negative integer random variable and define the span of Z by $m_Z := \sup\{m \in \mathbb{N}, \ \exists b \in \mathbb{N}, \ \mathrm{Supp}(Z) \subset m\mathbb{N}+b\}$. Let φ_Z denote the characteristic function of Z. When Z is square integrable, σ_Z^2 denotes its variance. For τ lying in $\operatorname{dom} \psi_Z := \{\tau \in \mathbb{R} : \mathbb{E}[\exp(\tau Z)] < +\infty\}$, we define $\psi_Z(\tau) := \log \mathbb{E}[\exp(\tau Z)]$ as the cumulant generating function of Z. Obviously, $\operatorname{dom} \psi_Z$ contains at least $\mathbb{R}^-$ and ψ_Z is analytic in the interior of $\operatorname{dom} \psi_Z$. We denote by R_Z the interior of the range of ψ'_Z. It is well known that R_Z is a subset of the interior of the convex hull of the support of Z. These two subsets of $\mathbb{R}$ coincide whenever ψ_Z is essentially smooth (see definition below). Further, let ψ_Z^* denote the Fenchel-Legendre transform of ψ_Z [6, Definition 2.2.2 p. 26]. For any $\tau^* \in R_Z$, there exists a unique $\tau_{\tau^*} \in \operatorname{dom} \psi_Z$ such that $\psi'_Z(\tau_{\tau^*}) = \tau^*$ and we may define Z^{*,τ^*} as a r.v. on $\mathbb{N}$ having the following distribution

$$\mathbb{P}(Z^{*,\tau^*} = k) = \exp[k\tau_{\tau^*} - \psi_Z(\tau_{\tau^*})]\mathbb{P}(Z = k), \quad (k \in \mathbb{N}). \tag{1}$$

It is well known that $\mathbb{E}\left(Z^{*,\tau^*}\right) = \tau^*$. For more details on the relationships between $\psi_Z, \psi_Z^*, Z^{*,\tau^*}$ we refer to the book [2].
Let now (Z, W) be a random vector of $\mathbb{R}^2$. We naturally extend some of the previous notations to (Z, W). For example, $\psi_{Z,W}$ is the cumulant generating function built on (Z, W) defined on $\operatorname{dom} \psi_{Z,W} \subset \mathbb{R}^2$ and $\psi_{Z,W}^*$ denotes the Fenchel-Legendre transform of (Z, W).

Convex functions Let f be a proper convex function on $\mathbb{R}^k$. That is f is convex and valued in $\mathbb{R} \cup \{+\infty\}$. We say that f is essentially smooth whenever it is differentiable on the non empty interior of $\operatorname{dom} f$ and it is steep. That is, for any vector c lying on the boundary of $\operatorname{dom} f$

$$\lim_{x \to c, x \in \operatorname{int} \operatorname{dom} f} \|\nabla f(x)\| = +\infty,$$

where $\nabla f(x)$ denotes the gradient of f at point x.

Gärtner-Ellis Corollary.

Corollary 2.2. *[Gärtner-Ellis, [6, Theorem 2.3.6 c) p.44]] Let (Z_n) be a sequence of random variables valued in $\mathbb{R}$, $(a_n)_n$ a decreasing positive sequence of real numbers with $\lim_{n\to\infty} a_n = 0$. Define $\Lambda_n(\theta) = \ln \mathbb{E} e^{\theta Z_n}$. Assume that*

1. *for all $\theta \in \mathbb{R}$, $a_n \Lambda_n(\theta/a_n) \to \Lambda(\theta) \in]-\infty, +\infty]$,*
2. *0 lies in the interior of $dom(\Lambda(\theta))$ and $\Lambda(\theta)$ is essentially smooth and lower semi continuous.*

Then (Z_n) satisfies a LDP with good rate function Λ^* and speed a_n.

2.2 Main results

2.2.1 The model

For $n \in \mathbb{N}^*$, let $(X^{(n)}, Y^{(n)})$ be a random vector with $X^{(n)} \in \mathbb{N}$. We assume that $m_{X^{(n)}} = 1$ and that $(X^{(n)}, Y^{(n)})$ converges in law to (X, Y) where X is a non essentially constant non negative integer valued r.v. having also span 1. Note that it implies that ψ_X is strictly convex and that $X^{(n)}$ is not essentially constant and that $\psi_{X^{(n)}}$ is strictly convex for n large enough. Further let $\Big((X_i^{(n)}, Y_i^{(n)})\Big)_{1\leq i\leq n}$ be an i.i.d. sample having the same distribution as $(X^{(n)}, Y^{(n)})$.

Let, for $n \in \mathbb{N}^*$ and $q_n \in \mathbb{N}^*$, $S_n = X_1^{(n)} + \cdots + X_{nq_n}^{(n)}$ and $T_n = Y_1^{(n)} + \cdots + Y_{nq_n}^{(n)}$. In the whole paper p_n will be a sequence of positive integers such that $\mathbb{P}(S_n = np_n) > 0$.

2.2.2 Large deviations

Theorem 2.3. *Let $p, q, p_n, q_n \in \mathbb{N}^*$ such that $p_n/q_n \in R_{X^{(n)}} \to p/q \in R_X$. Assume that*

1. *the function $\psi_{X,Y}$ is essentially smooth, and let τ be the unique real such that $\psi_X'(\tau) = p/q$,*
2. *$dom\psi_Y = dom\psi_{Y^{(n)}} = \mathbb{R}$,*
3. *there exists $r > 0$ such that $I_\tau := [\tau - r, \tau + r] \subset (dom\, \psi_X) \cap (\cap_{n\geq 1} dom\, \psi_{X^{(n)}})$ and*

$$\forall\, u \in \mathbb{R}\,, \forall\, s \in I_\tau\,,\ \sup_{t\in\mathbb{R}} \left|\mathbb{E}\left[e^{(it+s)X^{(n)}+uY^{(n)}} - e^{(it+s)X+uY}\right]\right| \to 0\,. \tag{2}$$

Then the distribution of (T_n/nq_n) conditioned by the event $\{S_n = np_n\}$ satisfies a LDP with good rate function $\psi_{X,Y}^*(p/q, \cdot) - \psi_X^*(p/q)$ and speed $(nq_n)^{-1}$.

2.2.3 Moderate deviations

Let ξ lying in the interior of domψ_X and consider the random vector $(\check{X}_\xi, \check{Y}_\xi)$ whose distribution is given, for any $k \in \mathbb{N}$ and real Borel set A, by

$$\mathbb{P}(\check{X}_\xi = k, \check{Y}_\xi \in A) = \exp[-\psi_X(\xi) + k\xi]\mathbb{P}(X = k, Y \in A). \tag{3}$$

We define in the same way the random vector $(\check{X}_\xi^{(n)}, \check{Y}_\xi^{(n)})$. Obviously, $\check{X}_\xi$ has the same distribution as X^{*,ξ^*} with $\xi^* = \psi_X'(\xi)$. Further, let α_ξ^2 be the variance of the residual $\check{\varepsilon}_\xi$ for the linear regression of $\check{Y}_\xi$ on $\check{X}_\xi$:

$$\check{\varepsilon}_\xi := \Big(\check{Y}_\xi - \mathbb{E}\big(\check{Y}_\xi\big)\Big) - \frac{\mathrm{cov}(\check{X}_\xi, \check{Y}_\xi)}{\mathrm{var}(\check{X}_\xi)}\Big(\check{X}_\xi - \mathbb{E}\big(\check{X}_\xi\big)\Big). \tag{4}$$

Then we get the following result.

Theorem 2.4. *Let $p, q, p_n, q_n \in \mathbb{N}^*$ such that p/q (resp. p_n/q_n) lies in R_X (resp. $R_{X^{(n)}}$) and $p_n/q_n \to p/q$. Assume that*

1. *there exists $r_0 > 0$ such that*

$$B_0 :=]-r_0, r_0[\subset (dom\psi_Y) \cap (\cap_{n\geq 1} dom\psi_{Y^{(n)}}) \ ,$$

and let τ (resp. τ_n) be the unique real such that $\psi'_X(\tau) = p/q$ (resp. $\psi'_{X^{(n)}}(\tau_n) = p_n/q_n$),

2. *there exists $r > 0$ such that $I_\tau := [\tau - r, \tau + r] \subset (dom\ \psi_X) \cap (\cap_{n\geq 1} dom\ \psi_{X^{(n)}})$ and*

$$\forall\, s \in I_\tau\ ,\ \sup_{t\in\mathbb{R}} \left| \mathbb{E}\left[e^{(it+s)X^{(n)}} - e^{(it+s)X} \right] \right| \to 0\,, \tag{5}$$

$$\sup_n \sup_{(s,v)\in I_\tau\times B_0} \mathbb{E}\left(e^{sX^{(n)}+v(Y^{(n)}-\mathbb{E}(\check{Y}^{(n)}_{\tau_n}))} \right) < \infty\,, \tag{6}$$

3. *(a_n) satisfies $na_nq_n \to +\infty$.*

Then the distribution of $\left(\sqrt{\frac{a_n}{nq_n}}\left(T_n - nq_n\mathbb{E}(\check{Y}^{(n)}_{\tau_n})\right)\right)$ conditioned by the event $\{S_n = np_n\}$ satisfies a LDP with good rate function $J(\cdot) = \frac{(\cdot)^2}{2\alpha_\tau^2}$ and speed a_n.

Remark 2.5. As $(nq_n)^{-1} = o(a_n)$, we say that the distribution of $\left(\sqrt{\frac{a_n}{nq_n}}\left(T_n - nq_n\mathbb{E}(\check{Y}^{(n)}_{\tau_n})\right)\right)$ conditioned by the event $\{S_n = np_n\}$ satisfies a moderate deviation principle (MDP).

Corollary 2.6. *Under the assumptions of one of the last theorems, $\mathcal{L}_n$ converges in distribution toward's the degenerate distribution concentrated on $E(\check{Y}_{p/q})$.*

3 Proofs

For $p_n \in \mathbb{N}$, such that $\mathbb{P}(S_n = np_n) \neq 0$, let

$$f_n(u) := \frac{1}{nq_n} \log \mathbb{E}\left[\exp(uT_n)\,|S_n = np_n\right] \in \mathbb{R} \cup \{+\infty\}.$$

In order to apply Gärtner-Ellis Corollary, we have to prove that $f_n(u)$ converges when $n \to \infty$. The next two subsections yield a simple representation of $f_n(u)$ using the Fourier Transform.

3.1 A simple representation using Fourier Transform

Recall that we set $\varphi_Z(t) := \mathbb{E}\left(e^{itZ}\right)$. An obvious but useful lemma follows.

Lemma 3.1 (Bartlett's Formula, see Equation (16) in [3]). *Let Z be a non negative integer r.v. and W be an integrable r.v. Then, for any non negative integer k lying in the support of Z,*

$$\mathbb{E}\left[W|Z = k\right] = \frac{\int_{-\pi}^{\pi} \mathbb{E}[W\exp(itZ)]\exp(-ikt)dt}{\int_{-\pi}^{\pi} \varphi_Z(t)\exp(-ikt)dt}\,.$$

3.2 Laplace lemmas

We begin this section with a variation on a Lemma first due to Laplace, see [10] . To be self contained we give also the sketch of its proof.

Lemma 3.2. *Let $p_n, q_n, p, q \in \mathbb{N}^*$. Let Z be a non constant square integrable non negative integer r.v. with span $m_Z = 1$. Let $(Z^{(n)})$ be a sequence of non negative i.i.d. integer random variables also having span 1. Let $Z_1^{(n)}, \dots, Z_n^{(n)}$ be an i.i.d. sample distributed as $Z^{(n)}$. Assume that*

1. $\|\varphi_{Z^{(n)}} - \varphi_Z\|_\infty \xrightarrow[n\to+\infty]{} 0$,
2. *the means of Z and $Z^{(n)}$ are rational, equal respectively to p/q and p_n/q_n with $p_n/q_n \xrightarrow[n\to+\infty]{} p/q$,*
3. $\sigma^2_{Z^{(n)}} \xrightarrow[n\to+\infty]{} \sigma^2_Z$,
4. $\mathbb{E}\big(|Z^{(n)} - p_n/q_n|^3\big)$ *is uniformly bounded.*

Then, when n tends to infinity

$$\mathbb{P}\left(\sum_{j=1}^{nq_n} Z_j^{(n)} = np_n\right) = \frac{1}{\sqrt{2\pi n q_n}\sigma_Z}(1 + o(1)). \tag{7}$$

Proof: The inversion of the Fourier Transform yields

$$\mathbb{P}\left(\sum_{j=1}^{nq_n} Z_j^{(n)} = np_n\right) = \int_{-\pi}^{\pi} e^{-inp_n t}\varphi^{nq_n}_{Z^{(n)}}(t)\frac{dt}{2\pi} = \int_{-\pi}^{\pi} \left[e^{-i\frac{p_n}{q_n}t}\varphi_{Z^{(n)}}(t)\right]^{nq_n}\frac{dt}{2\pi}. \tag{8}$$

On one hand, using a Taylor expansion of order 2 for $e^{-i\frac{p_n}{q_n}t}\varphi_{Z^{(n)}}(t) = \varphi_{Z^{(n)}-\frac{p_n}{q_n}}(t)$, we get

$$\varphi_{Z^{(n)}-\frac{p_n}{q_n}}(t) = 1 - \frac{t^2}{2}\sigma_n^2 - \frac{t^3}{6}\mathbb{E}\big(i(Z^{(n)} - p_n/q_n)^3 e^{it^* Z^{(n)}}\big),$$

where t^* lies in $[0, t]$ and σ_n^2 states for $\sigma^2_{Z^{(n)}}$. Now as $\mathbb{E}|(Z^{(n)} - p_n/q_n)|^3$ is bounded and as $\sigma_n^2 \to \sigma_Z^2$ we can find a positive number (independent of n) $\delta < \pi$ such that for $|t| < \delta$ and for n large enough

$$\left|e^{-i\frac{p_n}{q_n}t}\varphi_{Z^{(n)}}(t)\right| = \left|\varphi_{Z^{(n)}-\frac{p_n}{q_n}}(t)\right| \le 1 - \frac{\sigma_Z^2 t^2}{4}. \tag{9}$$

On the other hand, as $m_{Z^{(n)}} = m_Z = 1$ one has both $\xi := \sup_{\delta\le t\le\pi}|\varphi_Z(t)| < 1$ and $\xi_n := \sup_{\delta\le t\le\pi}|\varphi_{Z^{(n)}}(t)| < 1$. Further, as $\|\varphi_{Z^{(n)}} - \varphi_Z\|_\infty \to 0$, we get $\xi_n \to \xi$. Let $\epsilon > 0$ be such that $\xi + \epsilon < 1$. For n large enough $\xi_n \le \xi + \epsilon$. Now to conclude, one splits the integral in (8) in two integrals I_1, I_2 integrating on $|t| < \delta$ and on $|t| \ge \delta$. $|I_2|$ is bounded by $\big(\xi + \epsilon\big)^{nq_n}/(2\pi)$, hence is exponentially small. To deal with I_1, one performs the variable change $u = \sqrt{nq_n}\sigma_Z t$, and use both (9) and inequality $\log(1-\theta) \le -\theta, (\theta \in [0, 1[)$ to conclude by using both central limit and Lebesgue Theorems. ■

Remark 3.3. Note that if $m_Z \neq 1$, the theorem above is in general not valid. Take for example $m_Z = 2$ then in the interval $[0, 2\pi[$, $\varphi_{Z^{(n)}}(t) = \varphi_Z(t) = 1$ if and only if $t = 0$ or $t = \pi$. Obviously,

$$\frac{1}{2\pi}\int_{-\pi}^{\pi} e^{-inp_n t}\varphi_{Z^{(n)}}^{nq_n}(t)dt = \frac{1}{2\pi}\int_{-\pi/2}^{3\pi/2} e^{-inp_n t}\varphi_{Z^{(n)}}^{nq_n}(t)dt.$$

Hence

$$\mathbb{P}\left(\sum_{j=1}^{nq_n} Z_j^{(n)} = np_n\right) = \frac{1}{2\pi}\int_{-\pi/2}^{3\pi/2} e^{-inp_n t}\varphi_{Z^{(n)}}^{nq_n}(t)dt. \tag{10}$$

Set $f_n(t) := e^{-inp_n t}\varphi_{Z^{(n)}}^{nq_n}(t)dt$. Now, we split the integral in the right hand side of equation (10) in five parts

$$\begin{aligned}\frac{1}{2\pi}\int_{-\pi/2}^{3\pi/2} f_n(t)dt =& \frac{1}{2\pi}\int_{-\pi/2}^{-\delta} f_n(t)dt + \frac{1}{2\pi}\int_{-\delta}^{\delta} f_n(t)dt + \frac{1}{2\pi}\int_{\delta}^{\pi-\delta} f_n(t)dt \\ &+ \frac{1}{2\pi}\int_{\pi-\delta}^{\pi+\delta} f_n(t)dt + \frac{1}{2\pi}\int_{\pi+\delta}^{3\pi/2} f_n(t)dt. \\ &= I_1 + I_2 + I_3 + I_4 + I_5.\end{aligned}$$

Using the same arguments as above, we can prove that I_1, I_3 and I_5 are exponentially small. We also get that $\frac{1}{2\pi}\int_{-\delta}^{\delta} f_n(t)dt = \frac{1}{\sqrt{2\pi nq_n}\sigma_Z}(1 + o(1))$. Let us deal now with $I_4 = \frac{1}{2\pi}\int_{\pi-\delta}^{\pi+\delta} f_n(t)dt$. There exist non negative integer valued random variables Y_n such that $Z^{(n)} = 2Y_n + b$, set $u = t - \pi$

$$I_4 = \frac{1}{2\pi}(-1)^{np_n + bnq_n}\int_{-\delta}^{\delta} e^{-inp_n u}\varphi_{Y_n}^{nq_n}(2u)du.$$

Hence $I_4 \neq \frac{1}{\sqrt{2\pi nq_n}\sigma_Z}(1 + o(1))$.

We now give an extension of the previous lemma involving not only the probability for the sum to be equal to the mean of $Z^{(n)}$ but to any *good* rational number.

Lemma 3.4. *Let Z be a non negative and non degenerated integer r.v. with span $m_Z = 1$. Let $(Z_j^{(n)})_j$ be a sequence of i.i.d. non negative integer random variables having also span 1. Let $p_n, q_n, p, q \in \mathbb{N}^*$ such that $p_n/q_n \in R_{Z^{(n)}} \to p/q \in R_Z$. Let τ (resp. τ_n) be the unique real such that $\psi'_Z(\tau) = p/q$ (resp. $\psi'_{Z^{(n)}}(\tau_n) = p_n/q_n$). We make the following assumptions.*

1. *There exists $r > 0$ such that $I_\tau := [\tau - r, \tau + r] \subset (\cap_{n\geq 1} dom\, \psi_{Z^{(n)}}) \cap (dom\, \psi_Z)$.*
2. $$\forall\, s \in I_\tau\,,\ \lim_n \sup_{t\in\mathbb{R}}\left|\mathbb{E}\left[\left(e^{(it+s)Z^{(n)}} - e^{(it+s)Z}\right)\right]\right| = 0\,. \tag{11}$$

Then, when n goes to infinity

$$\mathbb{P}\left(\sum_{j=1}^{nq_n} Z_j^{(n)} = np_n\right) = e^{-nq_n\psi^*_{Z^{(n)}}(p_n/q_n)}\frac{1}{\sqrt{2\pi nq_n}\sigma_{Z^{*,p/q}}}(1 + o(1)), \tag{12}$$

where $\sigma^2_{Z^{*,p/q}}$ is the variance of $Z^{*,p/q}$ defined in (1).

Proof: Using the multinomial formula, we may write

$$\mathbb{P}\left(\sum_{j=1}^{nq_n} Z_j^{(n),*,p_n/q_n} = np_n\right) = e^{nq_n(\frac{p_n}{q_n}\tau_n - \psi_{Z^{(n)}}(\tau_n))}\mathbb{P}\left(\sum_{j=1}^{nq_n} Z_j^{(n)} = np_n\right),$$

where τ_n is the unique real such that $\psi'_{Z(\tau_n)} = p_n/q_n$. Hence

$$\mathbb{P}\left(\sum_{j=1}^{nq_n} Z_j^{(n)} = np_n\right) = e^{-nq_n\psi^*_{Z^{(n)}}(p_n/q_n)}\mathbb{P}\left(\sum_{j=1}^{nq_n} Z_j^{(n),*,p_n/q_n} = np_n\right),$$

where $Z_1^{(n),*,p_n/q_n}, \ldots, Z_n^{(n),*,p/q}$ are i.i.d. r.vs. having the distribution defined by

$$\mathbb{P}\big(Z^{(n),*,p_n/q_n} = k\big) = \exp\big(k\tau_n - \psi_{Z_1^{(n)}}(\tau_n)\big)\mathbb{P}\big(Z_1^{(n)} = k\big).$$

Further, the expectation of this last r.v. is p_n/q_n. Let us now check the assumptions of Lemma 3.2.

- Assumption 2 of Lemma 3.2 is satisfied by construction of $Z^{(n),*,p_n/q_n}$.
- Let us prove that $\mathbb{E}\left|Z^{(n),*,p_n/q_n} - p_n/q_n\right|^3$ is bounded. Using Hölder inequality we get that

$$\begin{aligned}\mathbb{E}\left|Z^{(n),*,p_n/q_n} - p_n/q_n\right|^3 &= \sum_{k=0}^{\infty}\left|k - \frac{p_n}{q_n}\right|^3 e^{k\tau_n - \psi_{Z^{(n)}}(\tau_n)}\mathbb{P}\big(Z_1^{(n)} = k\big) \\ &= e^{\psi^*_{Z^{(n)}}(p_n/q_n)}\mathbb{E}\left(\left|Z^{(n)} - \frac{p_n}{q_n}\right|^3 e^{\tau_n(Z^{(n)} - p_n/q_n)}\right) \\ &\leq e^{\psi^*_{Z^{(n)}}(\frac{p_n}{q_n})}\left(\mathbb{E}\left((Z^{(n)} - \frac{p_n}{q_n})^4 e^{\tau_n(Z^{(n)} - \frac{p_n}{q_n})}\right)\right)^{3/4}\left(\mathbb{E}\left(e^{\tau_n(Z^{(n)} - \frac{p_n}{q_n})}\right)\right)^{1/4}.\end{aligned}$$

  Using classical arguments on convex functions [19], we get that $\tau_n \xrightarrow[n\to+\infty]{} \tau$. Hence, by Assumptions 1. and 2. of Lemma 3.4 we get that $\mathbb{E}\left|Z^{(n),*,p/q} - p/q\right|^3$ is bounded.
- Similar arguments yield that

$$\|\varphi_{Z_1^{(n),*,p_n/q_n}} - \varphi_{Z^{*,p/q}}\| \to 0 \tag{13}$$

  and that $\sigma^2_{Z_n^*} \to \sigma^2_{Z^*}$.

Hence all the assumptions of Lemma 3.2 are satisfied and we may conclude using Lemma 3.2. ■

3.3 Some changes of probability

One of the main tool to prove large deviation results is the use of changes of probability. In this section, we review the different changes of probability used in this paper.

(a) Let $p/q \in R_Z$. We define $\tau \in$ dom ψ_Z by $\psi_Z'(\tau) = p/q$. We then introduce $Z^{*,p/q}$ as a random variable valued on $\mathbb{N}$:

$$\mathbb{P}(Z^{*,p/q} = k) = \exp[k\tau - \psi_Z(\tau)]\mathbb{P}(Z = k), \quad (k \in \mathbb{N}). \tag{14}$$

We have $\mathbb{E}(Z^{*,p/q}) = \frac{p}{q}$. This change of probability is quite classical in large deviation theory. In order to prove Lemma 3.4, we also define $Z^{(n),*,p_n/q_n}$, replacing Z by $Z^{(n)}$, p/q by p_n/q_n and τ by τ_n. Then $\mathbb{E}\big(Z^{(n),*,p_n/q_n}\big) = p_n/q_n$, as needed to apply Lemma 3.2.

(b) For u in domψ_Y, define $\widehat{X}_u$ by

$$\mathbb{P}\left(\widehat{X_u} = k\right) = \exp\left[-\psi_Y(u)\right]\mathbb{E}\left[\exp\left(uY\right)\mathbb{1}_{\{X=k\}}\right]. \tag{15}$$

Similarly, replacing (X,Y) by $\big(X^{(n)}, Y^{(n)}\big)$, we define $\widehat{X}_u^{(n)}$. The r.v. $\widehat{X_u}$ and $\widehat{X}_u^{(n)}$ appear naturally when applying the inversion of Fourier transform in the proof of Theorem 2.3.

(c) For the moderate deviations, the asymptotic is different (see Theorem 2.4). Therefore the r.v. $Y^{(n)}$ have to be centered. The centering factor and the rate function are closely related to the following change of probability. Let ξ lying in the interior of domψ_X and consider the random vector $(\check{X}_\xi, \check{Y}_\xi)$ whose distribution is given, for any $k \in \mathbb{N}$ and real Borel set A, by

$$\mathbb{P}(\check{X}_\xi = k, \check{Y}_\xi \in A) = \exp[-\psi_X(\xi) + k\xi]\mathbb{P}(X = k, Y \in A). \tag{16}$$

We define in the same way the random vector $(\check{X}_\xi^{(n)}, \check{Y}_\xi^{(n)})$. Obviously, $\check{X}_\xi$ has the same distribution as X^{*,ξ^*} with $\xi^* = \psi_X'(\xi)$. Further, let α_ξ^2 be the variance of the residual $\check{\varepsilon}_\xi$ for the linear regression of $\check{Y}_\xi$ on $\check{X}_\xi$:

$$\check{\varepsilon}_\xi := \Big(\check{Y}_\xi - \mathbb{E}\big(\check{Y}_\xi\big)\Big) - \frac{\operatorname{cov}(\check{X}_\xi, \check{Y}_\xi)}{\operatorname{var}(\check{X}_\xi)}\Big(\check{X}_\xi - \mathbb{E}\big(\check{X}_\xi\big)\Big). \tag{17}$$

Note that $J(y) = \frac{y^2}{\alpha_\tau^2}$ is the rate function in Theorem 2.4. Moreover, the centering factor is $\mathbb{E}(\check{Y}_{\tau_n}^{(n)})$. Hence, the change of probability used in the proof of Theorem 2.3 (see change of probability (15) above) has to be modified, according to this centering factor. This leads to the change of probability (18) below.

(d) Let τ (resp. τ_n) be such that $\psi_X'(\tau) = p/q$ (resp. $\psi_{X^{(n)}}'(\tau_n) = p_n/q_n$). Define the random variable $\widetilde{\widehat{X_u^{(n)}}}$ distributed on $\mathbb{N}$ by:

$$\mathbb{P}\left(\widetilde{\widehat{X_u^{(n)}}} = k\right) = e^{-\psi_{\widetilde{Y}^{(n)}}(u_n)}\mathbb{E}\left[e^{\frac{u\widetilde{Y}^{(n)}}{\sqrt{n a_n q_n}}}\mathbb{1}_{\{X^{(n)}=k\}}\right], \tag{18}$$

where $\widetilde{Y}^{(n)} = Y^{(n)} - \mathbb{E}(\check{Y}_{\tau_n}^{(n)})$, $u_n = u/\sqrt{n a_n q_n}$.

3.4 Proof of Theorem 2.3

Let, for $t \in \mathbb{R}$ and $u \in \mathbb{R}$,

$$\Phi_{X^{(n)},Y^{(n)}}(t,u) := \mathbb{E}\left(\exp[itX^{(n)} + uY^{(n)}]\right).$$

On one hand, using Lemma 3.1, we may write, for $u \in \mathbb{R}$ and n large enough,

$$f_n(u) = \frac{1}{nq_n} \log \frac{\int_{-\pi}^{\pi} e^{-inp_n t} \Phi^{nq_n}_{X^{(n)},Y^{(n)}}(t,u)dt}{\int_{-\pi}^{\pi} e^{-inp_n t} \Phi^{nq_n}_{X^{(n)},Y^{(n)}}(t,0)dt}. \tag{19}$$

Using twice equation (8) we may rewrite (19) as

$$f_n(u) = \frac{1}{nq_n} \left[\log \mathbb{P}\left(\sum_{j=1}^{nq_n} \widehat{X}^{(n)}_{u,j} = np_n \right) - \log \mathbb{P}\left(S_n = np_n\right) \right] + \psi_{Y^{(n)}}(u), \tag{20}$$

where $\widehat{X}^{(n)}_{u,1}, \cdots, \widehat{X}^{(n)}_{u,n}$ are independent copies of $\widehat{X}^{(n)}_u$ defined in Subsection 3.3 by equation (15). In order to apply Lemma 3.4 to $\widehat{X}^{(n)}_{u,i}$, let us prove that

$$\forall\, s \in I_\tau \ \lim_n \sup_{t\in\mathbb{R}} \left| \mathbb{E}\left[\left(e^{(it+s)\widehat{X}^{(n)}_u} - e^{(it+s)\widehat{X}_u} \right) \right] \right| = 0. \tag{21}$$

We have, for all $s \in I_\tau$,

$$\left| \mathbb{E}\left[\left(e^{(it+s)\widehat{X}^{(n)}_u} - e^{(it+s)\widehat{X}_u} \right) \right] \right| \leq Ce^{-\psi_Y(u)} \left| \mathbb{E}\left[e^{uY^{(n)}} e^{(s+it)X^{(n)}} - e^{uY} e^{(s+it)X} \right] \right|.$$

The right hand side of this last inequality tends to 0 by assumption 3. of Theorem 2.3. It remains to prove that p_n/q_n (resp. p/q) belongs to $R_{\hat{X}^{(n)}_u}$ (resp. $R_{\hat{X}_u}$). Using the fact that $\psi_{X,Y}$ is essentially smooth and Assumption 3. of Theorem 2.3 it is easy to see that $R_{\hat{X}_u} = R_X$ (resp. $R_{\hat{X}^{(n)}_u} = R_{X^{(n)}}$ at least for n large enough).

Applying Lemma 3.4 we obtain, for $u \in \mathbb{R}$

$$f(u) := \lim_{n\to\infty} \frac{1}{nq_n} \log \mathbb{E}\left[\exp(uT_n)\,|S_n = np_n\right] = -[\psi^*_{\widehat{X}_u}(p/q) - \psi_Y(u) - \psi^*_X(p/q)].$$

The convex dual function f^* of f is given by

$$\begin{aligned} f^*(y) := \sup_{u\in\mathbb{R}}[uy - f(u)] &= \sup_{u\in\mathbb{R}} \left(uy + \left[\psi^*_{\widehat{X}_u}(p/q) - \psi_Y(u) \right] \right) - \psi^*_X(p/q) \\ &= \sup_{(u,\xi)\in \mathrm{dom}\psi_{X,Y}} \left[uy + \xi\frac{p}{q} - \psi_{X,Y}(\xi,u) \right] - \psi^*_X(p/q) \\ &= \psi^*_{X,Y}\left(\frac{p}{q}, y\right) - \psi^*_X(p/q). \end{aligned} \tag{22}$$

As $\psi_{X,Y}$ is essentially smooth, using Theorem 26.3 in [19], we deduce that $\psi^*_{X,Y}$ is essentially strictly convex. Hence, using once more Theorem 26.3 in [19], we may deduce that f is essentially smooth. Therefore we can apply Gärtner-Ellis Corollary 2.2 (see Theorem 2.3.6. (c) in [6]) and conclude. ■

3.5 Proof of Theorem 2.4

Let $\tilde{T}_n = T_n - nq_n\mathbb{E}(\check{Y}^{(n)}_{\tau_n})$ and

$$g_n(u) = a_n \log\Big(\mathbb{E}\big(e^{\tilde{T}_n u/\sqrt{na_nq_n}}|S_n = np_n\big)\Big).$$

Proceeding as in the proof of Theorem 2.3, we have

$$\begin{aligned}
g_n(u) &= a_n \log \frac{\int_{-\pi}^{\pi} e^{-inp_nt}\Phi^{nq_n}_{X^{(n)},Y^{(n)}-\mathbb{E}(\check{Y}^{(n)}_{\tau_n})}(t, u/\sqrt{na_nq_n})}{\int_{-\pi}^{\pi} e^{-inp_nt}\Phi^{nq_n}_{X^{(n)},Y^{(n)}-\mathbb{E}(\check{Y}^{(n)}_{\tau_n})}(t, 0)} \\
&= a_n\Big(\log\mathbb{P}\big(\sum_{j=1}^{nq_n}\widehat{\widehat{X^{(n)}_{u,j}}} = np_n\big) - \log\mathbb{P}\big(S_n = np_n\big)\Big) + a_nnq_n\psi_{Y^{(n)}-\mathbb{E}(\check{Y}^{(n)}_{\tau_n})}(u/\sqrt{na_nq_n}),
\end{aligned}$$

where $\widehat{\widehat{X^{(n)}_{u,j}}}$ are i.i.d. r.v. on $\mathbb{N}$ with distribution defined in Subsection 3.3 by Equation (18).

In order to use Lemma 3.4 we first have to prove that

$$\forall\, s\in I_\tau\,,\ \limsup_n \sup_{t\in\mathbb{R}}\left|\mathbb{E}\left[e^{(it+s)\widehat{\widehat{X^{(n)}_u}}} - e^{(it+s)\widehat{\widehat{X}}_u}\right]\right| = 0. \tag{23}$$

We have

$$\left|\mathbb{E}\left[e^{(it+s)\widehat{\widehat{X^{(n)}_u}}} - e^{(it+s)\widehat{\widehat{X}}_u}\right]\right| \le Ce^{-\psi_{Y-\mathbb{E}(\check{Y}_{\xi_0})}(0)}\left|\mathbb{E}\left[e^{(s+it)X^{(n)}} - e^{(s+it)X}\right]\right|,$$

which tends to zero by assumption (5). As in the proof of Theorem 2.3 it is easy to prove that $R_{\widehat{\widehat{X}}_u} = R_X$ and $R_{\widehat{\widehat{X^{(n)}_u}}} = R_{X^{(n)}}$.

Using Lemma 3.4 we obtain, for $u\in\mathbb{R}$

$$g_n(u) \overset{n\to+\infty}{\sim} -q_nna_n\left[\psi^*_{\widehat{\widehat{X^{(n)}_u}}}(p_n/q_n) - \psi_{Y^{(n)}-\mathbb{E}(\check{Y}^{(n)}_{\tau_n})}(u/\sqrt{na_nq_n}) - \psi^*_{X^{(n)}}(p_n/q_n)\right].$$

Define

$$H_n(h) = \sup_{\xi\in\mathbb{R}}\left[\xi\frac{p_n}{q_n} - \psi_{X^{(n)},Y^{(n)}-\mathbb{E}(\check{Y}^{(n)}_{\tau_n})}(\xi,h)\right].$$

As

$$\psi^*_{\widehat{\widehat{X^{(n)}_u}}}(p_n/q_n) = \sup_x\left(\frac{p_n}{q_n}x - \psi_{\widehat{X}^{(n)}_u}(x)\right),$$

and

$$\psi_{\widehat{X}^{(n)}_u}(x) = \psi_{X^{(n)},Y^{(n)}-\mathbb{E}(\check{Y}^{(n)}_{\tau_n})}(x, u/\sqrt{na_nq_n}) - \psi_{Y^{(n)}-\mathbb{E}(\check{Y}^{(n)}_{\tau_n})}(u/\sqrt{na_nq_n}),$$

we get

$$g_n(u) \overset{n\to+\infty}{\sim} -q_nna_n\big(H_n(u/\sqrt{na_nq_n}) - H_n(0)\big). \tag{24}$$

We claim that if $\lim_n h_n = 0$, then

$$\lim_n \frac{H_n(h_n) - H_n(0)}{h_n^2} = -\frac{\alpha_{\tau_n}^2}{2} + O(1). \tag{25}$$

Assuming that (25) is true, and as $\alpha_{\tau_n}^2 \to \alpha_\tau^2$, we get that

$$\lim_n g_n(u) = -u^2 \frac{\alpha_\tau^2}{2}.$$

We easily conclude, since

$$g^*(y) = \sup_u \left\{ uy + \lim_n g_n(u) \right\} = \frac{y^2}{2\alpha_\tau^2}.$$

It remains to prove that (25) is true. Recall that

$$H_n(h) = \sup_{\xi \in \mathbb{R}} \left[\xi \frac{p_n}{q_n} - \psi_{X^{(n)}, Y^{(n)} - \mathbb{E}(\check{Y}_{\tau_n}^{(n)})}(\xi, h) \right].$$

In the sequel ψ'_x (resp. ψ'_y) will denote the partial derivative of $\psi_{X^{(n)}, Y^{(n)} - \mathbb{E}(\check{Y}_{\tau_n}^{(n)})}(\xi, h)$ with respect to the first (resp. second) variable. On one hand, by assumption (6), we can define on $I_\tau \times B_0$ the function F_n by:

$$F_n(\xi, h) = \psi'_x(\xi, h) - p_n/q_n \,.$$

We then deduce from the implicit function Theorem that there exists a neighborhood of $(\tau_n, 0)$ on which:

$$H_n(h) = \xi_n(h) \frac{p_n}{q_n} - \psi_{X^{(n)}, Y^{(n)} - \mathbb{E}(\check{Y}_{\tau_n}^{(n)})}(\xi_n(h), h)\,,$$

with

$$\xi'_n(h) = -\frac{\psi''_{xy}(\xi_n(h), h)}{\psi''_{xx}(\xi_n(h), h)}\,.$$

We can then calculate the derivatives of H_n (in the sequel we omit the argument $(\xi_n(h), h)$ in the derivatives). We have (with obvious notations)

$$\begin{aligned}
H'_n(h) &= -\psi'_y, \\
H''_n(h) &= \frac{\left(\psi''_{x,y}\right)^2}{\psi''_{x,x}} - \psi''_{y,y}, \\
H_n^{(3)}(h) &= \left(\frac{\psi''_{x,y}}{\psi''_{x,x}}\right)^3 \psi^{(3)}_{x,x,x} - 3\left(\frac{\psi''_{x,y}}{\psi''_{x,x}}\right)^2 \psi^{(3)}_{x,x,y} + 3\frac{\psi''_{x,y}}{\psi''_{x,x}} \psi^{(3)}_{x,y,y} - \psi^{(3)}_{y,y,y}\,.
\end{aligned}$$

Replacing the partial derivative of ψ by its expression, we get

$$H'_n(0) = 0 \text{ and } H''_n(0) = -\alpha_{\tau_n}^2\,.$$

On the other hand, using a Taylor expansion, we get

$$H_n(h_n) - H_n(0) = h_n H'_n(0) + \frac{h_n^2}{2} H''_n(0) + \frac{h_n^3}{6} H_n^{(3)}(z_n), \qquad z_n \in [0, h_n]. \tag{26}$$

Hence (26) becomes

$$H_n(h_n) - H_n(0) = -\frac{h_n^2\alpha_{\tau_n}^2}{2} + \frac{h_n^3}{6}H_n^{(3)}(z_n),\ z_n \in [0, h_n]. \tag{27}$$

Now the expression of $H_n^{(3)}$ is a rational fraction of some partial derivatives of $\mathbb{E}\left(e^{\xi X^{(n)}+h(Y^{(n)}-\mathbb{E}(\check{Y}_{\tau_n}^{(n)}))}\right)$. The denominator of this rational fraction is bounded away from 0 as it converges to a variance and numerator is bounded by (6). Hence $H_n^{(3)}$ is bounded and the claim is proved. ■

4 Examples

In this section we give two examples of applications and one counter example. These examples are borrowed from [15].

4.0.1 Occupancy problem

In the classical occupancy problem (see [15] and the references therein for more details), m balls are distributed at random into N urns. The resulting numbers of balls $Z_1, \cdots, Z_N$ have a multinomial distribution, and it is well-known that this equals the distribution of $\left(X_1, \cdots, X_N\right)$ conditioned on $\sum_{i=1}^N X_i = m$, where $X_1, \cdots, X_N$ are i.i.d. with $X_i \sim \mathcal{P}(\lambda)$, for an arbitrary $\lambda > 0$. The classical occupancy problem studies the number W of empty urns; this is thus $\sum_{i=1}^N \mathbb{1}_{\{X_i=0\}}$ conditioned on $\sum_{i=1}^N X_i = m$.
Now suppose that $m = np_n \to \infty$ and $N = nq_n \to \infty$ with $\frac{p_n}{q_n} \to \frac{p}{q}$. Take $X_i^{(n)} \sim \mathcal{P}(\lambda_n)$. Note that we do not assume that $\lambda_n = p_n/q_n$ and $\lambda = p/q$ which is the case in Janson's work. It is easy to see that Assumption 3. of Theorem 2.3 is fulfilled and that $\psi_{X,Y}$ is essentially smooth. Moreover, for $(x, y) \in \mathbb{R}^2$ and $(p, q) \in (\mathbb{R}^*)^2$, we have

$$\begin{aligned}
\psi_X(x) &= -\lambda + \lambda e^x, \\
\psi_X^*(p/q) &= p/q \log(\frac{p}{q\lambda}) + \lambda - \frac{p}{q}, \\
\psi_{X,Y}(x, y) &= -\lambda + \log\left(e^{\lambda \exp(x)} - 1 + e^y\right).
\end{aligned}$$

Hence we can apply Theorem 2.3. Here the function $\psi_{X,Y}^*$ does not have any explicit form. We give in Appendix the graph of the rate function for some particular values of p/q and λ. Assumptions of Theorem 2.4 are obviously fulfilled. We have

$$\begin{aligned}
\mathbb{E}\big(\check{Y}_{\tau_n}^{(n)}\big) &= e^{-\lambda_n \exp(\tau_n)}, \\
\mathbb{P}(\check{X}_\tau = k) &= e^{-\lambda \exp(\tau)}(e^\tau\lambda)^k/k!.
\end{aligned}$$

Hence $\check{X}_\tau$ is Poisson with parameter λe^τ. An easy calculation gives

$$\begin{aligned}
\operatorname{cov}\big(\check{X}_\tau, \check{Y}_\tau\big) &= -\lambda e^\tau e^{-\lambda \exp(\tau)}, \\
\operatorname{Var}(\check{Y}_\tau) &= e^{-\lambda \exp(\tau)}(1 - e^{-\lambda \exp(\tau)}).
\end{aligned}$$

Hence $\alpha_\tau^2 = e^{-\lambda\exp(\tau)}\Big(1-e^{-\lambda\exp(\tau)}+\lambda e^\tau e^{-\lambda\exp(\tau)}\Big)$. Now, as $\tau = \log(\frac{p}{q\lambda})$, we get $J(.) = \frac{(.)^2 e^\lambda}{1-e^{-\lambda}+\lambda e^{-\lambda}}$ in the particular case where $\lambda = p/q$. Note that functional L.D.P. is given in [4].

Remark 4.1. Theorem 2.3 allows us to deal with other statistics than $\sum_{i=1}^n \mathbb{1}_{\{X_i=0\}}$. For example, statistics of the form $\sum_{i=1}^n f(X_i, Z_i)$ where $Z_1, \dots, Z_n$ are i.i.d. and independent from $X_1, \dots, X_n$. Let us describe the particular case of bootstrap see [7]. Let $Z_1, \dots, Z_n$ be i.i.d. real valued random variables, independent of $X_1, \dots, X_n$. We choose at random with replacement a sample $Z_1^*, \dots, Z_n^*$. Then $\sum_{i=1}^n f(Z_i^*)$ is distributed as $\sum_{i=1}^n X_i f(Z_i)$ conditioned on $\sum_{i=1}^n X_i = m$, where $X_i \sim \mathcal{P}(\lambda)$, $i = 1, \dots, n$ for any $\lambda > 0$. Hence we get the same kind of conditioning as for the occupancy problem.

4.0.2 Branching processes

Consider a Galton-Watson process, beginning with one individual, where the number of children of an individual is given by a random variable X having finite moments. Assume further that $\mathbb{E}(X) = 1$. We number the individuals as they appear. Let X_i be the number of children of the $i-$th individual. It is well known (see example 3.4 in [15] and the references therein) that the total progeny is $n \geq 1$ if and only if

$$S_k := \sum_{i=1}^k X_i \geq k \text{ for } 0 \leq k < n \text{ but } S_n = n-1\,. \tag{28}$$

This type of conditioning is different from the one studied in the present paper, but Janson proves [15, Example 3.4] that if we ignore the order of $X_1, \dots, X_n$, conditioning on (28) is equivalent to conditioning on $S_n = n-1$. Hence we can study variables of the kind $Y_i = f(X_i)$. Considering the case where $Y_i = \mathbb{1}_{\{X_i=3\}}$, the $\sum_{i=1}^n Y_i$ is the number of families with three children. Now choosing $X_i \sim \mathcal{P}(\lambda)$, we compute the rate function as in Example 4.0.1.

4.0.3 Hashing

The model

Hashing with linear probing can be regarded as throwing n balls sequentially into m urns at random; the urns are arranged in a circle and a ball that lands in an occupied urn is moved to the next empty urn, always moving in a fixed direction. The length of the move is called the displacement of the ball, and we are interested in the sum of all displacements which is a random variable noted $d_{m,n}$. We assume $n < m$.

After throwing all balls, there are $N = m - n$ empty urns. These divide the occupied urns into blocks of consecutive urns. For convenience, we consider the empty urn following a block as belonging to this block. Janson [14] proved that the length of the blocks (counting the empty urn) and the sum of displacements inside each block are distributed as $(X_1, Y_1), \dots, (X_N, Y_N)$ $(N = m-n)$ conditioned on $\sum_{i=1}^N X_i = m$, where (X_i, Y_i) are i.i.d. copies of a pair (X, Y) of random variables. X has the Borel distribution

$$\mathbb{P}\big(X = l\big) = \frac{1}{T(\lambda)}\frac{l^{l-1}}{l!}\lambda^l, \quad l \in \mathbb{N}^*,\ z \in [0, e^{-1}[\,, \tag{29}$$

where $T(\lambda) = \sum_{l=1}^\infty \frac{l^{l-1}}{l!}\lambda^l$ is the well-known tree function and λ is an arbitrary number with $0 < \lambda \leq e^{-1}$. The conditional distribution of Y given $X = l$ is the same as the distribution of

$d_{l,l-1}$.
Using Janson's results [9, 14, 15], unfortunately we can prove that the joint Laplace transform of (X_1, Y_1) is defined only on $(-\infty, a) \times (-\infty, 0)$ for some positive a. Hence our results can not be applied. Nevertheless, in a forthcoming work, we will study conditioned L.D.P for self-normalized sums in the spirit of [20]. In that case the Laplace will be defined.

4.0.4 Bose-Einstein statistics

This example is borrowed from [13]. Consider N urns. Put n indistinguishable balls in the urns in such a way that each distinguishable outcome has the same probability i.e.,

$$1/\binom{n+N-1}{n},$$

see for example [8]. Let Z_k be the number of balls in the kth urn. It is well known that $(Z_1, \ldots, Z_N)$ is distributed as $\left(X_1, \cdots, X_N\right)$ conditioned on $\sum_{i=1}^N X_i = n$, where $X_1, \cdots, X_N$ are i.i.d. with a geometric distribution. As for Example 4.0.1, we can get a L.D.P for variables of the form $\sum_i h(X_i)$ if $\text{dom}\psi_{h(X_i)} = \mathbb{R}$.

4.0.5 Possible extensions

Among possible extensions, let us mention the case where the variables Y_i are independent but do not have the same distribution. This case occurs in [13, Examples 2 and 3], where the quantity of interest is the law of $\sum_{i=1}^N h_i(X_i)$ conditioned on the event $\sum_{i=1}^N X_i = n$. Another way to extend our work is to deal with the case where the variables X_i are independent but not i.i.d.. This case occurs when counting from a random permutation the number of cycles of a fixed size see for example [1, Chapter 1]

In the present paper we assume that p_n, q_n, p and q are positive. The other cases will be considered in a forthcoming work.

References

[1] Arratia, R. , Barbour, A.D. , Tavare, S. (2003) *Logarithmic combinatorial structures: a probabilistic approach* , European Mathematical Society

[2] Barndorff-Nielsen, O. (1978) *Information and exponential families in statistical theory.* Wiley. Chichester.

[3] Bartlett, M.S. (1938) *The characteristic function of a conditional statistic.* J. London Math Soc. **13** 62-67.

[4] Boucheron, S., Leonard, C. and Gamboa, F. (2002) *Bins and Balls: Large Deviations of the Empirical Occupancy Process.* The Annals of Applied Probability. **12** 607-636.

[5] Csiszár, I. (1984) *Sanov Property, Generalized I-Projection and a Conditional Limit Theorem.* The Annals of Probability. **12** 768-793.

[6] Dembo, A. and Zeitouni, O.(1998). *Large deviations techniques and applications.* Applications of Mathematics (New York), second edition. ISBN 0-387-98406-2. Springer-Verlag. New York.

[7] Efron, B. and Tibshirani, R. J. (1993). *An introduction to the bootstrap.* Monographs on Statistics and Applied Probability, vol 57, Chapman and Hall, New York, 1993. ISBN 0-412-04231-2.

[8] Feller, W. (1968). *An Introduction to Probability Theory and its Applications 1 3rd ed.* Willey. New York.

[9] Flajolet, P., Poblete, P. and Viola, A. (1997). *On the Analysis of Linear Probing Hashing.* ISSN 0249-6399 Rapport de recherche N 3265 Septembre 1997 de l'INRIA

[10] Hald, A. (1998) *A history of mathematical statistics from 1750 to 1930.* Wiley Series in Probability and Statistics: Texts and References Section. A Wiley-Interscience Publication. John Wiley & Sons, Inc., New York, 1998. xx+795 pp. ISBN: 0-471-17912-4

[11] Hipp, C. (1984). *Asymptotic expansions for conditional distributions: The lattice case Prob. Math. Stat*, Vol. 4, 207-219.

[12] Hollander, F. D. (2000) *Large deviation* ISBN 0-8218-1989-5 *A.M.S. 2000*

[13] Holst, L. (1979). *Two conditional limit theorems with applications The annals of Statistics*, Vol. 7, No. 3, 551-557.

[14] Janson, S. (2001). Asymptotic distribution for the cost of linear probing hashing. *Random Struct. Appl.*, 19(2):no.3-4, 438–471.

[15] Janson, S. (2001). Moment convergence in conditional limit theorems. *J. Appl. Probab.*, 38 (2):421–437.

[16] Kim, G.W. and Truax, D.R. (1998). Conditional large deviations for density case. *Stat. Prob. Let.*, 38:137–144.

[17] Kudlaev, E. M. (1984) *On conditional limit distributions of sums of random vectors.* in Russian Teor. Veroyatn. Primen. 29, No.4, 743-752 (1984). ISSN0040-361X English trans: Theory Probab Appl **29** (1984), no. 4, 776-786.

[18] Robinson, J., Höglund, T., Holst, L. and Quine, P.(1990). On approximating probabilities for small and large deviations in $\mathbb{R}^d$. *Ann. Prob.*, 18(2):727–753.

[19] Rockafellar, R.T. (1972). *Convex analysis.* Princeton University Press, second edition. ISBN 0-491-08069-0.

[20] B. Y. Ying, Q. M. Shao and Q. Y. Wang (2003) *Self-normalized Cramer type large deviations for independent random variables.* Ann. Probab. 31 (2003), 2167-2215.

[21] Steck G.P. (1957) *Limit theorems for conditional distributions* Univ. California Publ. statist. **2** 237-284.

[22] van Campenhout, J. Cover, T. (1981) *Maximum entropy and conditional probability* , IEEE Trans. Inf. Th. **27** 483-489.

5 Appendix

Here we give the shape of the large deviation rate function in the example of the Occupancy problem when $\lambda = 1$ and $p/q = 1$, $p/q = 0.4$ or $p/q = 3$