HAL
open science

On the sub-Gaussianity of the Beta and Dirichlet distributions

Olivier Marchal, Julyan Arbel

To cite this version:

Olivier Marchal, Julyan Arbel. On the sub-Gaussianity of the Beta and Dirichlet distributions. Electronic Communications in Probability, 2017, 22 (paper no. 54), pp.1-14. 10.1214/17-ECP92 . hal-01521300

HAL Id: hal-01521300

https://hal.science/hal-01521300

Submitted on 26 Jan 2018

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

arXiv:1705.00048v2 [math.ST] 25 Sep 2017

On the sub-Gaussianity of the Beta and Dirichlet distributions

Olivier Marchal[1] and Julyan Arbel[2]

[1] Université de Lyon, CNRS UMR 5208, Université Jean Monnet,
Institut Camille Jordan, France

[2] Inria Grenoble Rhône-Alpes,
Laboratoire Jean Kuntzmann, Université Grenoble Alpes, France

September 26, 2017

Abstract

We obtain the optimal proxy variance for the sub-Gaussianity of Beta distribution, thus proving upper bounds recently conjectured by Elder (2016). We provide different proof techniques for the symmetrical (around its mean) case and the non-symmetrical case. The technique in the latter case relies on studying the ordinary differential equation satisfied by the Beta moment-generating function known as the confluent hypergeometric function. As a consequence, we derive the optimal proxy variance for the Dirichlet distribution, which is apparently a novel result. We also provide a new proof of the optimal proxy variance for the Bernoulli distribution, and discuss in this context the proxy variance relation to log-Sobolev inequalities and transport inequalities.

1 Introduction

The sub-Gaussian property (Buldygin and Kozachenko, 1980, 2000; Pisier, 2016) and related concentration inequalities (Boucheron et al., 2013; Raginsky and Sason, 2013) have attracted a lot of attention in the last couple of decades due to their applications in various areas such as pure mathematics, physics, information theory and computer sciences. Recent interest focused on deriving the optimal proxy variance for discrete random variables like the Bernoulli distribution (Buldygin and Moskvichova, 2013; Kearns and Saul, 1998; Berend and Kontorovich, 2013) and the missing mass (McAllester and Schapire, 2000; McAllester and Ortiz, 2003; Berend and Kontorovich, 2013; Ben-Hamou et al., 2017). Our focus is instead on two continuous random variables, the Beta and Dirichlet distributions, for which the optimal proxy variance was not known to the best of our knowledge. Some upper bounds were recently conjectured by Elder (2016) that we prove in the present article by providing the optimal proxy variance for both Beta and Dirichlet distributions. Similar concentration properties of the Beta distribution have been recently used in many contexts including Bayesian adaptive data analysis (Elder, 2016), Bayesian nonparametrics (Castillo, 2016) and spectral properties of random matrices (Perry et al., 2016).

We start by reminding the definition of sub-Gaussian property for random variables:

Definition 1 (Sub-Gaussian variables)**.** *A random variable X with finite mean $\mu = \mathbb{E}[X]$ is sub-Gaussian if there is a positive number σ such that:*

$$\mathbb{E}[\exp(\lambda(X-\mu))] \leq \exp\left(\frac{\lambda^2\sigma^2}{2}\right) \text{ for all } \lambda \in \mathbb{R}. \tag{1}$$

Such a constant σ^2 is called a proxy variance (or sub-Gaussian norm), and we say that X is σ^2-sub-Gaussian. If X is sub-Gaussian, one is usually interested in the optimal proxy variance:

$$\sigma^2_{opt}(X) = \min\{\sigma^2 \geq 0 \text{ such that } X \text{ is } \sigma^2\text{-sub-Gaussian}\}.$$

Note that the variance always gives a lower bound on the optimal proxy variance: $Var[X] \leq \sigma^2_{opt}(X)$. In particular, when $\sigma^2_{opt}(X) = Var[X]$, X is said to be strictly sub-Gaussian.

Every compactly supported distribution, as is the Beta(α, β) distribution, is sub-Gaussian. This can be seen by Hoeffding's classic inequality: any random variable X supported on $[0, 1]$ with mean μ satisfies

$$\forall \lambda \in \mathbb{R}, \quad \mathbb{E}\left[\mathrm{e}^{\lambda(X-\mu)}\right] \leq \mathrm{e}^{\frac{\lambda^2}{8}},$$

thus exhibiting $\frac{1}{4}$ as an upper bound to the proxy variance. This bound can be improved by taking into account the location of the mean μ within the interval $[0, 1]$. An early step in this direction is the second inequality in Hoeffding (1963) paper, indexed (2.2). It states that if $\mu < 1/2$, then for any positive ϵ, $\mathbb{P}(X - \mu > \epsilon) \leq \mathrm{e}^{-\epsilon^2 g(\mu)}$, where

$$g(\mu) = \frac{1}{1-2\mu} \ln \frac{1-\mu}{\mu} \tag{2}$$

thus indicating that X has a right tail lighter than a Gaussian tail of variance $\frac{1}{2g(\mu)}$. Hoeffding's result was strengthened by Kearns and Saul (1998) to comply with Definition 1 of sub-Gaussianity[1] as follows

$$\mathbb{E}[\exp(\lambda(X-\mu))] \leq \exp\left(\frac{\lambda^2}{4g(\mu)}\right) \text{ for all } \lambda \in \mathbb{R}, \tag{3}$$

thus indicating that $\frac{1}{2g(\mu)}$ is a distribution-sensitive proxy variance for any $[0, 1]$-supported random variable with mean μ (see also Berend and Kontorovich, 2013, for a detailed proof of this result). If this is the optimal proxy variance for the Bernoulli distribution (see Theorem 2.1 and Theorem 3.1 of Buldygin and Moskvichova, 2013), it is clear from our result that it does not hold true for the Beta distribution. However, fixing $\frac{\alpha}{\alpha+\beta} = \mu$ and letting $\alpha \to 0$, $\beta \to 0$, the Beta(α, β) distribution concentrates to the Bern(μ) distribution, and we show that we recover the optimal proxy variance for the Bernoulli distribution (Theorem 2).

An interesting common feature between optimal proxy variances for the Bernoulli distribution: $\frac{1}{2g(\mu)}$, and that of the Beta distribution derived later on, is that they deteriorate in a similar fashion as the mean μ goes to 0 or 1, see for instance the left panel of Figure 1. We briefly present here classical proof techniques for sub-Gaussianity hinging on certain tools from functional analysis. We show how they apply in the Bernoulli setting, and let as an interesting open problem how our proof in the Beta distribution setting could be supplemented by these same functional analysis tools.

Essentially two (related) functional inequalities allow one to derive a sub-Gaussian property: log-Sobolev inequalities, which date back to Gross (1975), and transport inequalities. The relation with the former inequalities is called Herbst's argument. It states that if a probability measure satisfies a log-Sobolev inequality with some constant, then it is sub-Gaussian with the same constant as a proxy variance[2] (see for instance Ledoux, 1999, Section 2.3 and Proposition 2.3). The optimal constant in the log-Sobolev inequality satisfied by the Bernoulli distribution also produces its optimal proxy variance (Ledoux, 1999, Corollary 5.9).

The relation with transport inequalities is usually referred to as Marton's argument (see for instance Raginsky and Sason, 2013, Section 3.4). Define the Wasserstein distance between two probability measures P and Q on a space $\mathcal{X}$ by

$$W(P, Q) = \inf_{\pi \in \Pi(P,Q)} \int_{\mathcal{X}\times\mathcal{X}} d(x, y)\pi(\mathrm{d}x, \mathrm{d}y),$$

where $\Pi(P, Q)$ is the set of probability measures on $\mathcal{X} \times \mathcal{X}$ with fixed marginal distributions respectively P and Q. The Wasserstein distance depends on some choice of a distance d on $\mathcal{X}$. A probability measure P is said to satisfy a transport inequality with constant c, if for any probability measure Q dominated by P,

$$W(P, Q) \leq \sqrt{2cD(Q||P)}, \tag{4}$$

where $D(Q||P)$ is the entropy, or Kullback–Leibler divergence, between P and Q. The transport inequality (4) is denoted by $\mathsf{T}(c)$.

Bobkov and Götze (1999) proved that $\mathsf{T}(c)$ implies c-sub-Gaussianity. See also Proposition 3.6 and Theorem 3.4.4 of Raginsky and Sason (2013) for general results. Further developments in the

[1] Note indeed that Equation (1), together with Markov inequality, imply $\mathbb{P}(X - \mu > \epsilon) \leq \mathrm{e}^{-\frac{\epsilon^2}{2\sigma^2}}$.

[2] The implied predicate is actually stronger than sub-Gaussianity, but it is not useful for our purposes.

discrete $\mathcal{X}$ setting are interesting for our purposes. Equip a discrete space $\mathcal{X}$ with the Hamming metric, $d(x,y) = \mathbb{1}_{\{x \neq y\}}$. The induced Wasserstein distance then reduces to the total variation distance, $W(P,Q) = \|P - Q\|_{\mathrm{TV}}$. In that setting, Ordentlich and Weinberger (2005) proved the distribution-sensitive transport inequality:

$$\|P - Q\|_{\mathrm{TV}} \leq \sqrt{\frac{1}{g(\mu_P)} D(Q||P)}, \tag{5}$$

where the function g is defined in Equation (2) and the coefficient μ_P is called the balance coefficient of P, and is defined by $\mu_P = \max_{A \subset \mathcal{X}} \min\{P(A), 1 - P(A)\}$. In particular, the Bernoulli balance coefficient is easily shown to coincide with its mean. Hence, applying the result of Bobkov and Götze (1999) to the $\mathsf{T}\left(\frac{1}{2g(\mu_P)}\right)$ transport inequality (5) yields a distribution-sensitive proxy variance of $\frac{1}{2g(\mu)}$ for the Bernoulli with mean μ. It is optimal, see for instance Theorem 3.4.6 of Raginsky and Sason (2013). This viewpoint highlights the key role played by the balance coefficient in the non-uniformity of the optimal proxy variance for discrete distributions such as the Bernoulli. However, it is not clear how this argument would carry over to non discrete distributions such as the Beta distribution for explaining similar sensitivity to the mean. However, to quote Raginsky and Sason (2013), the general approach may not produce optimal concentration estimates, that often require case-by-case treatments. This is the route followed in this note for the Beta distribution.

The outline of the note is as follows. We introduce the Beta distribution and state the main result (Theorem 1) in Section 2.1. We then prove our result depending on whether $\alpha = \beta$ (Section 2.2) or $\alpha \neq \beta$ (Section 2.3). In the first case, the proof is elementary and based on comparing the coefficients of the entire series representations of the functions of both sides of inequality (1). However, it does not directly carry over to the second case, whose proof requires some finer analysis tool: the study of the ordinary differential equation (ODE) satisfied by the confluent hypergeometric function ${}_1F_1$. Although the second proof also covers the case $\alpha = \beta$ upon slight modifications, the independent proof for the symmetric case is kept owing to its simplicity. As a by-product, we derive the optimal proxy variance for the Bernoulli and the Dirichlet distributions in Section 3. The R code for the plots presented in this note and for a function deriving the optimal proxy variance in terms of α and β is available at http://www.julyanarbel.com/software.

2 Optimal proxy variance for the Beta distribution

2.1 Notations and main result

The Beta(α, β) distribution, with $\alpha, \beta > 0$, is characterized by a density on the segment $[0,1]$ given by:

$$f(x) = \frac{1}{B(\alpha,\beta)} x^{\alpha-1}(1-x)^{\beta-1},$$

where $B(\alpha,\beta) = \int_0^\infty x^{\alpha-1}(1-x)^{\beta-1}\mathrm{d}x = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}$ is the Beta function. The moment-generating function of a Beta(α,β) distribution is given by a confluent hypergeometric function (also known as Kummer's function):

$$\mathbb{E}[\exp(\lambda X)] = {}_1F_1(\alpha; \alpha+\beta; \lambda) = \sum_{j=0}^{\infty} \frac{\Gamma(\alpha+j)\Gamma(\alpha+\beta)}{(j!)\Gamma(\alpha)\Gamma(\alpha+\beta+j)} \lambda^j. \tag{6}$$

This is equivalent to say that the jth raw moment of a Beta(α,β) random variable X is given by:

$$\mathbb{E}[X^j] = \frac{(\alpha)_j}{(\alpha+\beta)_j}, \tag{7}$$

where $(x)_j = x(x+1)\cdots(x+j-1) = \frac{\Gamma(x+j)}{\Gamma(x)}$ is the *Pochhammer symbol*, also known in the literature as a *rising factorial*. In particular, the mean and variance are given by:

$$\mathbb{E}[X] = \frac{\alpha}{\alpha+\beta}, \quad \mathrm{Var}[X] = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}.$$

The Beta distribution is ubiquitous in statistics. It plays a central role in the binomial model in Bayesian statistics where it is a conjugate prior distribution (the associated posterior distribution is also Beta): if $X \sim \text{Binomial}(\theta, N)$ and $\theta \sim \text{Beta}(\alpha, \beta)$, then $\theta | X \sim \text{Beta}(\alpha + X, \beta + N - X)$. It is also key to Bayesian nonparametrics where it embodies, among others, the distribution of the breaks in the stick-breaking representation of the Dirichlet process and the Pitman–Yor process; marginal distributions of Polya trees (Castillo, 2016); the posterior distribution of discovery probabilities under a Bayesian nonparametrics model (Arbel et al., 2017). Our main result opens new research avenues for instance about asymptotic (frequentist) assessments of these procedures.

Our main result regarding the Beta distribution is the following:

Theorem 1 (Optimal proxy variance for the Beta distribution)**.** *For any $\alpha, \beta > 0$, the Beta distribution* $\text{Beta}(\alpha, \beta)$ *is $\sigma^2_{opt}(\alpha, \beta)$-sub-Gaussian with optimal proxy variance $\sigma^2_{opt}(\alpha, \beta)$ given by:*

$$\begin{cases} \sigma^2_{opt}(\alpha, \beta) = \frac{\alpha}{(\alpha+\beta)x_0} \left(\frac{{}_1F_1(\alpha+1;\alpha+\beta+1;x_0)}{{}_1F_1(\alpha;\alpha+\beta;x_0)} - 1 \right) \\ \qquad \textit{where } x_0 \textit{ is the unique solution of the equation} \\ \ln({}_1F_1(\alpha; \alpha+\beta; x_0)) = \frac{\alpha x_0}{2(\alpha+\beta)} \left(1 + \frac{{}_1F_1(\alpha+1;\alpha+\beta+1;x_0)}{{}_1F_1(\alpha;\alpha+\beta;x_0)} \right). \end{cases} \tag{8}$$

A simple and explicit upper bound to $\sigma^2_{opt}(\alpha, \beta)$ is given by $\sigma^2_0(\alpha, \beta) = \frac{1}{4(\alpha+\beta+1)}$:
- for $\alpha \neq \beta$ we have $\text{Var}[\text{Beta}(\alpha, \beta)] < \sigma^2_{opt}(\alpha, \beta) < \frac{1}{4(\alpha+\beta+1)}$
- for $\alpha = \beta$ we have $\text{Var}[\text{Beta}(\alpha, \alpha)] = \sigma^2_{opt}(\alpha, \alpha) = \frac{1}{4(2\alpha+1)}$.

Equation (8) defining x_0 is a transcendental equation, the solution of which is not available in closed form. However, it is simple to evaluate numerically. The values of the variance, optimal proxy variance and its simple upper bound are illustrated on Figure 1. Note that for a fixed value of the sum of the parameters, $\alpha + \beta = S$, the optimal proxy variance deteriorates when α, or equivalently β, gets close to 0 or to S. This is reminiscent of the Bernoulli optimal proxy variance behavior which deteriorates when the success probability moves away from $\frac{1}{2}$ (Buldygin and Moskvichova, 2013). The intuition of the proof can be seen from Figure 2 (Section 2.3.3) where we represent the difference $\lambda \mapsto \exp\left(\mathbb{E}[X]\lambda + \frac{\sigma^2}{2}\lambda^2\right) - \mathbb{E}[\exp(\lambda X)]$ for various values of σ^2. The main argument is that the optimal proxy variance is obtained for the curve (in magenta) whose positive local minimum equals zero, thus leading to the system of equations of Theorem 1.

Corollary 1. *The Beta distribution* $\text{Beta}(\alpha, \beta)$ *is strictly sub-Gaussian if and only if $\alpha = \beta$.*

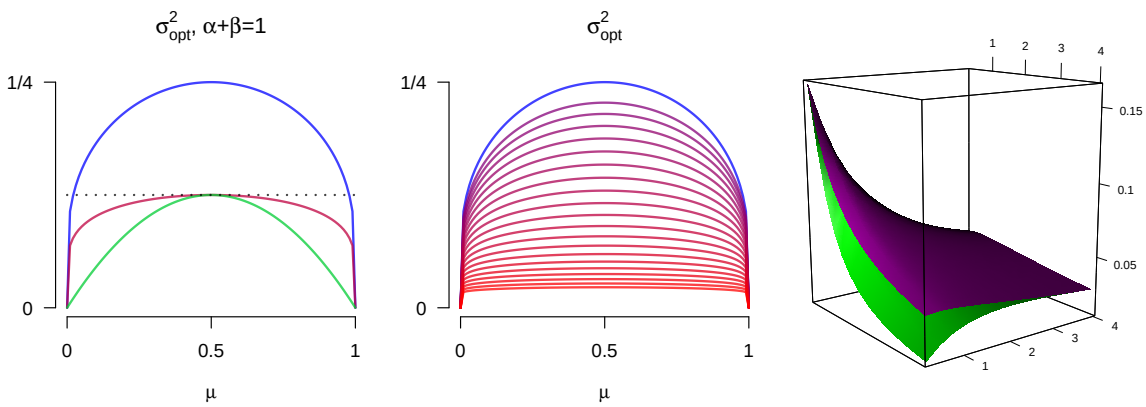

Figure 1: *Left*: curves of $\text{Var}[\text{Beta}(\alpha, \beta)]$ (green), $\sigma^2_{\text{opt}}(\alpha, \beta)$ (purple) and $\frac{1}{4(\alpha+\beta+1)}$ (dotted black) for the $\text{Beta}(\alpha, \beta)$ distribution with $\alpha + \beta$ set to 1, $\sigma^2_{\text{opt}}(\mu)$ for the $\text{Bern}(\mu)$ distribution (blue); varying mean μ on the x-axis. *Center*: curves of $\sigma^2_{\text{opt}}(\mu)$ for the $\text{Bern}(\mu)$ distribution (blue), and of $\sigma^2_{\text{opt}}(\alpha, \beta)$ for the $\text{Beta}(\alpha, \beta)$ distribution with $\alpha + \beta$ varying on a log scale from 0.1 (purple) to 10 (red); varying mean μ on the x-axis. *Right*: surfaces of $\text{Var}[\text{Beta}(\alpha, \beta)]$ (green) and $\sigma^2_{\text{opt}}(\alpha, \beta)$ (purple), for values of α and β varying in $[0.2, 4]$.

As a direct consequence, we obtain the strict sub-Gaussianity of the uniform, the arc-sine and the Wigner semicircle distributions, as special cases up to a trivial rescaling of the Beta(α,α) distribution respectively with α equal to 1, $\frac{1}{2}$ and $\frac{3}{2}$.

2.2 The Beta(α,α) distribution is strictly sub-Gaussian

Let $\sigma_0^2(\alpha) = \mathrm{Var}[\mathrm{Beta}(\alpha,\alpha)] = \frac{1}{4(2\alpha+1)}$. Since a random variable $X \sim \mathrm{Beta}(\alpha,\alpha)$ is symmetric around $\frac{1}{2}$, only its even centered moments are non-zero. The reason why $\mathbb{E}[\exp(\lambda(X-\mathbb{E}[X]))] \leq \exp\left(\frac{\lambda^2\sigma_0^2(\alpha)}{2}\right)$ is because the coefficients of the series expansions at $\lambda = 0$ of each side:

$$\mathbb{E}[\exp(\lambda(X-\mathbb{E}[X]))] = \sum_{j=0}^{\infty} \mathbb{E}\left[(X-1/2)^{2j}\right] \frac{\lambda^{2j}}{(2j)!}, \tag{9}$$

$$\exp\left(\frac{\lambda^2\sigma_0^2(\alpha)}{2}\right) = \sum_{j=0}^{\infty} \frac{\sigma^{2j}}{2^j}\frac{\lambda^{2j}}{j!} = \sum_{j=0}^{\infty} \frac{\lambda^{2j}}{2^{2j}2^j(2\alpha+1)^j(j!)}, \tag{10}$$

satisfy the inequalities:

$$\mathbb{E}\left[\left(X-\frac{1}{2}\right)^{2j}\right]\frac{1}{(2j)!} \leq \frac{\sigma_0^{2j}(\alpha)}{2^j}\frac{1}{j!}. \tag{11}$$

Indeed, algebra yields:

$$\mathbb{E}\left[\left(X-\frac{1}{2}\right)^{2j}\right] = \frac{(2j)!}{2^{2j}j!}\frac{\Gamma(2\alpha)\Gamma(\alpha+j)}{\Gamma(\alpha)\Gamma(2(\alpha+j))} = \frac{(2j)!}{2^{2j}j!}\frac{(\alpha)_j}{(2\alpha)_{2j}}. \tag{12}$$

Combining the expression of the raw moments (7) with the following inequality:

$$\frac{(\alpha)_j}{(2\alpha)_{2j}} = \frac{1}{2^j\prod_{l=1}^{j}(2\alpha+2l-1)} \leq \frac{1}{(2(2\alpha+1))^j}, \tag{13}$$

in (9) concludes the proof.

Remark 1. *The non-symmetrical distribution with $\alpha \neq \beta$ has even centered moments whose expressions are not as simple as (12). Moreover, it has obviously non-zero odd centered moments. For this last reason, the present proof does not carry over to the case $\alpha \neq \beta$.*

2.3 Optimal proxy variance for the Beta(α,β) distribution

2.3.1 Connection with ordinary differential equations

In this section, we assume that $X \sim \mathrm{Beta}(\alpha,\beta)$ with $\beta \neq \alpha$. We denote $\sigma_0^2 = \frac{1}{4(\alpha+\beta+1)}$ (we omit the dependence on α and β for compactness) and define for all $t \in \mathbb{R}$:

$$\sigma_t^2 = \frac{(1-t)}{4(1+\alpha+\beta)} + t\frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)} = \frac{1}{4(1+\alpha+\beta)} + \frac{(\beta-\alpha)^2}{4(\alpha+\beta)^2(1+\alpha+\beta)}t.$$

In other words, the decreasing function $t \mapsto \sigma_t^2$ maps the interval $[0,1]$ to the interval $[\sigma_1^2,\sigma_0^2]$ with $\sigma_1^2 = \mathrm{Var}[X]$. Then, we introduce the function u_t defined by:

$$u_t(x) \stackrel{\text{def}}{=} \exp\left(\frac{\alpha}{\alpha+\beta}x + \frac{\sigma_t^2}{2}x^2\right) - \mathbb{E}[\exp(xX)]\ ,\ \forall x \in \mathbb{R},$$

where σ_t^2-sub-Gaussianity amounts to non negativity of u_t on $\mathbb{R}$. Since the confluent hypergeometric function $y : x \mapsto y(x) = {}_1F_1(\alpha,\alpha+\beta;x)$ satisfies the linear second order ordinary differential equation $xy''(x) + (\alpha+\beta-x)y'(x) - \alpha y(x) = 0$, we obtain together with equation (6) that u_t is the unique solution of the Cauchy problem:

$$\begin{cases} xu_t''(x) + (\alpha+\beta-x)u_t'(x) - \alpha u_t(x) \\ \qquad\qquad = \frac{x}{16(\alpha+\beta)^4(1+\alpha+\beta)^2}\exp\left(\frac{\alpha}{\alpha+\beta}x + \frac{\sigma_t^2}{2}x^2\right)P_2(x;t), \\ u_t(0) = 0 \text{ and } u_t'(0) = 0, \end{cases} \tag{14}$$

where P_2 is a polynomial of degree 2 in x:

$$P_2(x;t) = 4(1-t)(\alpha^2-\beta^2)^2(1+\alpha+\beta)^2 \\ - 4(\beta^2-\alpha^2)(1+\alpha+\beta)\left((\alpha+\beta)^2 - t(\beta-\alpha)^2\right)x + \left((\alpha+\beta)^2 - t(\beta-\alpha)^2\right)^2 x^2.$$

For normalization purposes, we also define:

$$\begin{aligned} v_t(x) &= 16(\alpha+\beta)^4(1+\alpha+\beta)^2 u_t(x)\exp\left(-\frac{\alpha}{\alpha+\beta}x - \frac{\sigma_t^2}{2}x^2\right), \\ &= 16(\alpha+\beta)^4(1+\alpha+\beta)^2\left(1 - \mathbb{E}[\exp(xX)]\exp\left(-\frac{\alpha}{\alpha+\beta}x - \frac{\sigma_t^2}{2}x^2\right)\right). \end{aligned} \tag{15}$$

The function v_t is the unique solution of the Cauchy problem:

$$\begin{cases} xv_t''(x) + Q_1(x;t)v_t'(x) + Q_0(x;t)v_t(x) = xP_2(x;t), \\ v_t(0) = 0 \text{ and } v_t'(0) = 0, \end{cases} \tag{16}$$

with:

$$\begin{aligned} Q_1(x;t) &= \alpha+\beta - \frac{\beta-\alpha}{\alpha+\beta}x + \frac{(\alpha+\beta)^2 - t(\beta-\alpha)^2}{2(\alpha+\beta)^2(1+\alpha+\beta)}x^2, \\ Q_0(x;t) &= \frac{1}{16(\alpha+\beta)^4(1+\alpha+\beta)^2}\,xP_2(x;t). \end{aligned}$$

Note that u_t and v_t have the same sign hence proving that u_t is positive (resp. negative) is equivalent to proving that v_t is positive (resp. negative). From standard theory on ODEs (Birkhoff and Rota, 1989; Robinson, 2004), we get that the functions u_t and v_t are $\mathcal{C}^\infty(\mathbb{R})$. Indeed, the only possible singularity is at $x = 0$ but the initial conditions imply that the function is regular at this point. In particular, a Taylor expansion at $x = 0$ shows that:

$$v_t''(0) = \frac{P_2(0;t)}{1+Q_1(0;t)} = 4(\beta^2-\alpha^2)^2(1+\alpha+\beta)(1-t). \tag{17}$$

We also observe that the discriminant of the polynomial $x \mapsto P_2(x;t)$ is given by:

$$\Delta_t = 16t(1+\alpha+\beta)^2(\beta^2-\alpha^2)^2\left((\alpha+\beta)^2 - t(\beta-\alpha)^2\right)^2.$$

Hence we conclude that for $t > 0$, P_2 admits two distinct real zeros that are positive, while for $t < 0$ it remains strictly positive on $\mathbb{R}$. For $t = 0$, P_2 admits a double zero and thus remains positive on $\mathbb{R}$ appart from its zero.

By definition (15), we want to study the sign of v_t on $\mathbb{R}$. Indeed, showing that v_t is positive on $\mathbb{R}$ then X is equivalent to showing σ_t^2-sub-Gaussianity. We first observe that we may restrict the sign study on $\mathbb{R}_+$. Indeed, if we prove that:

$$\forall\, \lambda \geq 0\ ,\ \mathbb{E}[\exp(\lambda X)] = {}_1F_1(\alpha;\alpha+\beta;\lambda) \leq \exp\left(\frac{\lambda\alpha}{\alpha+\beta} + \frac{\lambda^2\sigma_t^2}{2}\right). \tag{18}$$

then, the case $\lambda < 0$ is automatically obtained by noting that $1 - X \sim \text{Beta}(\beta,\alpha)$, whose mean is $\frac{\beta}{\alpha+\beta} = 1 - \frac{\alpha}{\alpha+\beta}$. Therefore, applying (18) to $1 - X$ gives that for all $\lambda < 0$:

$$\begin{aligned} \mathbb{E}[\exp(\lambda X)] &= \exp(\lambda)\mathbb{E}[\exp(-\lambda(1-X))] \\ &\leq \exp(\lambda)\exp\left(-\lambda(1-\frac{\alpha}{\alpha+\beta}) + \frac{\sigma_t^2\lambda^2}{2}\right) = \exp\left(\lambda\frac{\alpha}{\alpha+\beta} + \frac{\sigma_t^2\lambda^2}{2}\right). \end{aligned}$$

Eventually, in agreement with the general theory, we observe that for $t > 1$ (i.e. $\sigma_t^2 < \text{Var}[X]$), X is not σ_t^2-sub-Gaussian. Indeed, the series expansion at $x = 0$ (17), shows that for $t > 1$, v_t is strictly negative in a neighborhood of 0. On the contrary, for $t < 1$, the function v_t is strictly positive in a neighborhood of 0 so that we may not directly conclude. Note also that for any value of t, we always have $\lim_{x\to\infty} v_t(x) = +\infty$.

2.3.2 Proof that the Beta(α, β) distribution is σ_0^2-sub-Gaussian

In this section, we take $t = 0$. As explained above, this corresponds to a case where P_2 is positive on $\mathbb{R}$ (apart from its double zero). We prove that $u_0(x) > 0$ for $x > 0$ by proceeding by contradiction. Let us assume that there exists $x_1 > 0$ such that $u_0(x_1) = 0$. Since the non-empty set $\{x > 0 \,/\, u_0(x) = 0\}$ is compact (because $u_0 \in \mathcal{C}^\infty(\mathbb{R})$) and excludes a neighborhood of 0, we may define $x_0 = \min\{x > 0 \text{ such that } u_0(x) = 0\} > 0$. Let us now define the set:

$$M = \{0 < x < x_0 \text{ such that } u_0'(x) = 0 \text{ and } u_0' \text{ changes sign at } x\}.$$

Since $u_0(0) = u_0(x_0) = 0$ and the facts that u_0 is strictly positive in a neighborhood of 0 and u_0' is continuous on $\mathbb{R}$, Rolle's theorem shows that M is not empty and that:

$$m = \min\{x \in M\} \text{ exists and } 0 < m < x_0.$$

Evaluating the ODE (14) at $x = m$ and using the fact that the polynomial P_2 is positive on $\mathbb{R}$ (appart from its double zero) leads to:

$$m u_0''(m) + 0 - \alpha u_0(m) \geq 0 \Rightarrow u_0''(m) \geq \frac{\alpha}{m} u_0(m) > 0. \tag{19}$$

However, combined with $u_0'(m) = 0$, this contradicts the fact that u_0' changes sign at $x = m$.

Thus, we conclude that there cannot exist $x_1 > 0$ such that $u_0(x_1) = 0$. Since u_0 is strictly positive in a neighborhood of 0 and continuous on $\mathbb{R}$, we conclude that it must remain strictly positive on $\mathbb{R}_+^*$.

Remark 2. *In this proof, the case $\beta = \alpha$ requires an adaptation since $u_0''(0) = 0$. Thus, we must determine $u_0^{(4)}(0) > 0$ ($u_0^{(3)}(0) = 0$ by symmetry) to ensure that the function u_0 is locally convex and remains strictly positive in a neighborhood of 0. Apart from this minor verification, the rest of the proof applies also to this case.*

2.3.3 Proof of the optimal proxy variance for the Beta(α, β) distribution

In this section we assume that $\beta \neq \alpha$. From general theorems regarding ODEs, we have that the application:

$$g : \left\{ \begin{array}{ccc} [0, +\infty) \times [0, +\infty) & \to & \mathbb{R} \\ (x, t) & \mapsto & g(x, t) = v_t(x), \end{array} \right. \tag{20}$$

is smooth ($g \in \mathcal{C}^\infty([0, +\infty) \times [0, +\infty))$). Indeed, the t-dependence of the coefficients of the ODE (16) is polynomial and thus smooth. The x-dependence of the coefficients of the ODE (16) is polynomial and as explained above, the only possible singularity in x is at $x = 0$ but initial conditions ensure that the solutions $x \mapsto v_t(x)$ are always regular there. Since for all $t \geq 0$ we have $\lim_{x \to \infty} v_t(x) = +\infty$, we also have that the function:

$$h : \left\{ \begin{array}{ccc} [0, +\infty) & \to & \mathbb{R} \\ t & \mapsto & \min\{v_t(x),\, x \in \mathbb{R}_+^*\}, \end{array} \right.$$

is continuous.

We now observe that for any $0 \leq t < 1$, the functions v_t are strictly positive in a neighborhood of 0. More precisely, if we choose a segment $[0, t_0]$ with $t_0 < 1$, then for all $0 \leq t \leq t_0$ we have that $v_t''(0) \geq v_{t_0}''(0) = 4(\beta^2 - \alpha^2)^2(1 + \alpha + \beta)(1 - t_0) > 0$. Hence, we may choose $\eta > 0$ such that for all $0 \leq t \leq t_0$, we have v_t is strictly positive on $]0, \eta]$. Moreover, since $\lim_{x \to \infty} v_0(x) = +\infty$, v_0 is bounded from below on $[\eta, +\infty[$ by a constant $A > 0$. Thus, since g is continuous, there exists a neighborhood of $t = 0$ in which all solutions v_t remain greater than $\frac{A}{2}$ on $[\eta, +\infty[$ and thus strictly positive on $\mathbb{R}_+^*$. This shows that for $t > 0$, σ_0^2 is not optimal.

Let us now introduce the set:

$$T_+ = \{t \geq 0 \text{ such that } v_t \text{ is positive on } \mathbb{R}_+^*\}.$$

Then, from the results presented above, we know that T_+ is non-empty, that it contains a neighborhood of 0 and that it is bounded from above by 1. Moreover, by connection with the initial problem (15), T_+ is an interval and thus is of the form $[0, t_{\text{opt}}]$ with $0 < t_{\text{opt}} < 1$. Indeed $t \in T_+$ implies

by construction that for all $s \leq t$, $s \in T_+$. Note also that $t < 1$ since v_1 is strictly negative in a neighborhood of 0 and thus $\min\{v_1(x), x \in \mathbb{R}_+^*\} < 0$. Hence, the continuity of h shows that there exists a neighborhood of $t = 1$ in which the solutions v_t are non-positive on $\mathbb{R}_+^*$. For $t = t_{\text{opt}}$ the function $v_{t_{\text{opt}}}$ must have a zero on $\mathbb{R}_+^*$ otherwise by continuity of h we may find a neighborhood of t_{opt} for which $\min\{v_t(x),\, x \in \mathbb{R}_+^*\}$ remains strictly positive thus contradicting the maximality of t_{opt}. Since $v_{t_{\text{opt}}}$ must remain positive, the zero is at least a double zero and therefore we find that there exists $x_0 > 0$ such that $v_{t_{\text{opt}}}(x_0) = 0$, $v'_{t_{\text{opt}}}(x_0) = 0$ and $v''_{t_{\text{opt}}}(x_0) \geq 0$. From (6) and (15), the conditions $v_{t_{\text{opt}}}(x_0) = 0$, $v'_{t_{\text{opt}}}(x_0) = 0$ are equivalent to the following system of equations (we use here the contiguous relations for the confluent hypergeometric function: ${}_1F_1'(a;b;x) = \frac{a}{b}\, {}_1F_1(a;b;x)$):

$$\begin{cases} {}_1F_1(\alpha;\alpha+\beta;x_0) = \exp\left(\frac{\alpha}{\alpha+\beta}x_0 + \frac{\sigma^2_{t_{\text{opt}}}}{2}x_0^2\right), \\ \frac{\alpha}{\alpha+\beta}\, {}_1F_1(\alpha+1;\alpha+\beta+1;x_0) = \left(\frac{\alpha}{\alpha+\beta} + \sigma^2_{t_{\text{opt}}}x_0\right)\, {}_1F_1(\alpha;\alpha+\beta;x_0). \end{cases} \tag{21}$$

This is equivalent to say that $x_0 \equiv x_0(\alpha,\beta)$ is the solution of the transcendental equation:

$${}_1F_1(\alpha;\alpha+\beta;x_0) = \exp\left(\frac{\alpha x_0}{2(\alpha+\beta)}\left(1 + \frac{{}_1F_1(\alpha+1;\alpha+\beta+1;x_0)}{{}_1F_1(\alpha;\alpha+\beta;x_0)}\right)\right),$$

and that $\sigma^2_{t_{\text{opt}}}$ is given by:

$$\sigma^2_{t_{\text{opt}}} = \frac{\alpha}{(\alpha+\beta)x_0}\left(\frac{{}_1F_1(\alpha+1;\alpha+\beta+1;x_0)}{{}_1F_1(\alpha;\alpha+\beta;x_0)} - 1\right).$$

Note that by symmetry, we have $x_0(\beta,\alpha) = -x_0(\alpha,\beta)$ hence, $\sigma^2_{t_{\text{opt}}}(\beta,\alpha) = \sigma^2_{t_{\text{opt}}}(\alpha,\beta)$. Moreover, if $\beta > \alpha$ then $x_0(\alpha,\beta) > 0$ while $\alpha > \beta$ implies $x_0(\alpha,\beta) < 0$. We may illustrate the situation with Figure 2 which displays the difference function $x \mapsto u_t(x)$.

Remark 3. *The system of equations (21) admits only one solution on $\mathbb{R}_+^*$. Indeed, let us transpose the problem from $v_{t_{opt}}$ to $u_{t_{opt}}$ using (15) and assume that there exist two points $0 < x_0 < x_1$ such that $u_{t_{opt}}(x_0) = 0$, $u'_{t_{opt}}(x_0) = 0$ and $u_{t_{opt}}(x_1) = 0$, $u'_{t_{opt}}(x_1) = 0$ with $u_{t_{opt}}$ strictly positive on (x_0, x_1) (hence $u''_{t_{opt}}(x_0) \geq 0$ and $u''_{t_{opt}}(x_1) \geq 0$). Using (14), this implies that $P_2(x_0;t_{opt}) \geq 0$ and $P_2(x_1;t_{opt}) \geq 0$. If we denote $x_- < x_+$ the potential distinct positive zeros of $x \mapsto P_2(x;t_{opt})$ we may exclude that $x_0 \leq x_-$. Indeed, if $x_0 \leq x_-$ then we may apply the same argument to $u_{t_{opt}}$ on the interval $[0, x_0]$ as the one developed for u_0 in Section 2.3.2 and obtain a contradiction. Thus, the only remaining case is to assume $x_1 > x_0 > x_+$. In that case, since $u_{t_{opt}}(x_0) = 0$, $u'_{t_{opt}}(x_0) = 0$, $u_{t_{opt}}(x_1) = 0$ and $x \mapsto P_2(x;t_{opt})$ is positive on $[x_0, x_1]$, we may apply the same argument to $u_{t_{opt}}$ on the interval $[x_0, x_1]$ as the one developed for u_0 in Section 2.3.2 and obtain a contradiction.*

3 Relations to other distributions

3.1 Optimal proxy variance for the Bernoulli distribution

We show that our proof technique can be used to recover the optimal proxy variance for the Bernoulli distribution, known since Kearns and Saul (1998). This is illustrated by the center panel of Figure 1.

Theorem 2 (Optimal proxy variance for the Bernoulli distribution)**.** *For any $\mu \in (0,1)$, the Bernoulli distribution with mean μ is sub-Gaussian with optimal proxy variance $\sigma^2_{opt}(\mu)$ given by:*

$$\sigma^2_{opt}(\mu) = \frac{(1-2\mu)}{2\ln\frac{1-\mu}{\mu}}. \tag{22}$$

Proof. In the limit $\alpha \to 0$ with $\frac{\alpha}{\alpha+\beta}$ fixed equal to μ, the differential equation (14) simplifies into:

$$u''_{t,\mu}(x) - u_{t,\mu}(x) = \exp\left(\mu x + \frac{x^2}{8} - \frac{x^2(2\mu-1)^2 t}{8}\right) \left(\frac{(2\mu-1)^2(1-t)}{4} + \frac{(2\mu-1)(1-t+4t\mu(1-\mu))}{4}x - \frac{(1-t+4t\mu(1-\mu))^2}{16}x^2\right)$$

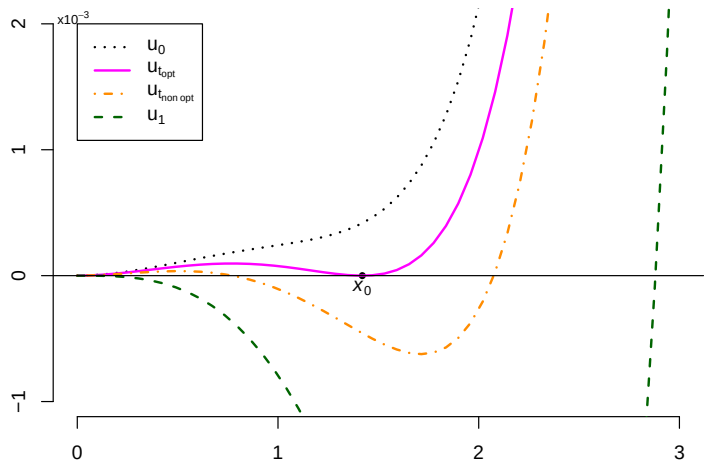

Figure 2: Difference function $x \mapsto u_t(x)$. For $t = 0$ (simple upper bound σ_0^2), the curve [dotted black] remains strictly positive. For $t = t_{\text{opt}}$ (optimal proxy variance σ^2_{opt}), the curve [magenta] has zero minimum (at x_0). For $t = 1$ (leading to the variance), the curve [dashed green] has negative second derivative at $x = 0$, hence is directly negative around 0. The intermediate case with $t_{\text{non opt}}$ in the interval $(t_{\text{opt}}, 1)$ produces a curve [orange, dash and dots] which is first positive, then negative, and positive again.

with the Cauchy initial conditions $u_{t,\mu}(0) = 0$ and $u'_{t,\mu}(0) = 0$. The solution of this Cauchy problem is explicit and given by:

$$u_{t,\mu}(x) = \exp\left(\mu x + \frac{x^2}{8} - \frac{x^2(2\mu-1)^2 t}{8}\right) - \mu e^x + \mu - 1 \tag{23}$$

Therefore the optimal proxy variance is given by $\sigma^2_{\text{opt}}(\mu) = \frac{1}{4} - \frac{1}{4}x(2\mu-1)^2 t_0$ where t_0 is determined by the system of equations: $u_{t_0,\mu}(x_0) = 0$ and $u'_{t_0,\mu}(x_0)$, thus defining implicitly t_0 and x_0 as functions of μ. In order to solve explicitly the last system of equations, we perform the change of variables: $(\mu, t) = \left(\frac{s}{s+1}, \frac{2\tilde{t}}{s+1} + 1\right)$ so that the solution (23) is now given by:

$$u_{\tilde{t},s}(x) = \exp\left(\frac{s}{s+1}x - \frac{\tilde{t}}{4(s+1)}x^2\right) - \frac{s}{s+1}e^x - \frac{1}{s+1}$$

Consequently, we have to solve the system:

$$\begin{cases} u_{\tilde{t}_0,s}(x_0) = 0 \\ u'_{\tilde{t}_0,x}(x_0) = 0 \end{cases} \Leftrightarrow \begin{cases} (s+1)\exp\left(\frac{s}{s+1}x_0 - \frac{\tilde{t}_0}{4(s+1)}x_0^2\right) = se^{x_0} + 1 \\ \left(s - \frac{x_0\tilde{t}_0}{2}\right)\exp\left(\frac{s}{s+1}x_0 - \frac{\tilde{t}_0}{4(s+1)}x_0^2\right) = se^{x_0} \end{cases}$$

Introducing another change of variable $(x_0, y_0) = (x_0, x_0\tilde{t}_0)$, the last system is equivalent to:

$$\begin{cases} s - y_0 = s(1+y_0)e^{x_0} \\ 1 + y_0 = \exp\left(-\frac{x_0}{s+1}\left(s - \frac{y_0}{2}\right)\right) \end{cases} \Leftrightarrow \begin{cases} s - y_0 = s\,\exp\left(\frac{y_0+2}{y_0-2s}\ln(1+y_0)\right) \\ x_0 = \frac{s+1}{\frac{y_0}{2}-s}\ln(1+y_0) \end{cases}$$

We now observe that $y_0 = s - 1$ and $x_0 = -2\ln s$ is a solution of the former system. Performing back the various changes of variables, this is equivalent to say that $\tilde{t}_0 = \frac{y_0}{x_0} = \frac{2(1-s)}{\ln s}$ so that $t_0 = \frac{4(1-s)}{(s+1)\ln s} + 1$ or equivalently $t_0 = \frac{1}{(2\mu-1)^2} + \frac{2}{(2\mu-1)\ln\frac{1-\mu}{\mu}}$. Consequently, the optimal proxy variance is given by:

$$\sigma^2_{\text{opt}}(\mu) = \frac{1}{4} - \frac{1}{4}(2\mu-1)^2 t_0 = \frac{(1-2\mu)}{2\ln\frac{1-\mu}{\mu}}$$

which is precisely the optimal proxy variance of a Bernoulli random variable with mean μ. □

3.2 Optimal proxy variance for the Dirichlet distribution

We start by reminding the definition of sub-Gaussian property for random vectors:

Definition 2 (Sub-Gaussian vectors)**.** *A random d-dimensional vector $\boldsymbol{X}$ with finite mean $\boldsymbol{\mu} = \mathbb{E}[\boldsymbol{X}]$ is σ^2-sub-Gaussian if the random variable $\boldsymbol{u}^\top \boldsymbol{X}$ is σ^2-sub-Gaussian for any unit vector $\boldsymbol{u}$ in the simplex $\mathcal{S}^{d-1} = \{\boldsymbol{u} \in [0,1]^d \,/\, \sum_{i=1}^d u_i = 1\}$. This is equivalent to say that:*

$$\mathbb{E}[\exp(\boldsymbol{\lambda}^\top(\boldsymbol{X} - \boldsymbol{\mu}))] \leq \exp\left(\frac{\|\boldsymbol{\lambda}\|^2\sigma^2}{2}\right) \text{ for all } \boldsymbol{\lambda} \in \mathbb{R}^d.$$

where $\|\boldsymbol{\lambda}\|^2 = \sum_{i=1}^d \lambda_i^2$. Eventually, a random vector $\boldsymbol{X}$ is said to be strictly sub-Gaussian, if the random variables $\boldsymbol{u}^\top \boldsymbol{X}$ are strictly sub-Gaussian for any unit vector $\boldsymbol{u} \in \mathcal{S}^{d-1}$.

Let $d \geq 2$. The Dirichlet distribution $\mathrm{Dir}(\boldsymbol{\alpha})$, with positive parameters $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_d)^\top$, is characterized by a density on the simplex $\mathcal{S}^{d-1}$ given by:

$$f(x_1, \ldots, x_d; \alpha_1, \ldots, \alpha_d) = \frac{1}{B(\boldsymbol{\alpha})} \prod_{i=1}^d x_i^{\alpha_i - 1},$$

where $B(\boldsymbol{\alpha}) = \frac{1}{\Gamma(\bar{\boldsymbol{\alpha}})} \prod_{i=1}^d \Gamma(\alpha_i)$ and $\bar{\boldsymbol{\alpha}} = \sum_{i=1}^d \alpha_i$. It generalizes the Beta distribution in the sense that the components are Beta distributed. More precisely, for any non-empty and strict subset $\mathcal{I}$ of $\{1, \ldots, d\}$:

$$\boldsymbol{X} = (X_1, \ldots, X_d)^\top \sim \mathrm{Dir}(\boldsymbol{\alpha}) \implies \sum_{i \in \mathcal{I}} X_i \sim \mathrm{Beta}\left(\sum_{i \in \mathcal{I}} \alpha_i, \sum_{j \notin \mathcal{I}} \alpha_j\right).$$

However, we remind the reader that the components $(X_i)_{1 \leq i \leq d}$ are not independent and the variance/covariance matrix is given by:

$$\forall\, i \neq j \,:\, \mathrm{Cov}[X_i, X_j] = -\frac{\alpha_i \alpha_j}{\bar{\boldsymbol{\alpha}}^2(1 + \bar{\boldsymbol{\alpha}})} \text{ and } \forall\, 1 \leq i \leq d \,,\ \mathrm{Var}[X_i] = \frac{\alpha_i(\bar{\boldsymbol{\alpha}} - \alpha_i)}{\bar{\boldsymbol{\alpha}}^2(1 + \bar{\boldsymbol{\alpha}})}.$$

Eventually, if we define $\boldsymbol{n} = (n_1, \ldots, n_d)^\top \in \mathbb{N}^d$, then the moments of the Dirichlet distribution are given by:

$$\mathbb{E}\left[\prod_{i=1}^d X_i^{n_i}\right] = \frac{B(\boldsymbol{\alpha} + \boldsymbol{n})}{B(\boldsymbol{\alpha})}.$$

This is equivalent to say that the moment-generating function of the Dirichlet distribution is:

$$\begin{aligned}
\mathbb{E}[\exp(\boldsymbol{\lambda}^\top \boldsymbol{X})] &= \sum_{m=0}^{\infty} \sum_{n_1 + \cdots + n_d = m} \frac{B(\boldsymbol{\alpha} + \boldsymbol{n})}{B(\boldsymbol{\alpha})} \frac{\lambda_1^{n_1} \ldots \lambda_d^{n_d}}{(n_1)! \ldots (n_d)!}, \\
&= \sum_{m=0}^{\infty} \sum_{n_1 + \cdots + n_d = m} \frac{\lambda_1^{n_1} \ldots \lambda_d^{n_d}}{(n_1)! \ldots (n_d)!} \frac{\Gamma(\bar{\boldsymbol{\alpha}})}{\Gamma(\bar{\boldsymbol{\alpha}} + \bar{\boldsymbol{n}})} \prod_{i=1}^d \frac{\Gamma(\alpha_i + n_i)}{\Gamma(\alpha_i)}, \\
&= \sum_{m=0}^{\infty} \sum_{n_1 + \cdots + n_d = m} \frac{\lambda_1^{n_1} \ldots \lambda_d^{n_d}}{(n_1)! \ldots (n_d)!} \frac{1}{(\bar{\boldsymbol{\alpha}})_{\bar{\boldsymbol{n}}}} \prod_{i=1}^d (\alpha_i)_{n_i},
\end{aligned}$$

where we have defined $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_d)^\top \in \mathbb{R}^d$ and $\bar{\boldsymbol{n}} = \sum_{i=1}^d n_i$.

Let us define $\boldsymbol{e}_i$ the ith canonical vector of $\mathbb{R}^d$ and $\boldsymbol{X} = (X_1, \ldots, X_d)^\top \sim \mathrm{Dir}(\boldsymbol{\alpha})$. From Definition 1 and the results regarding the $\mathrm{Beta}(\alpha, \beta)$ distribution obtained in Section 2.3, we immediately get that $\boldsymbol{e}_i^\top \boldsymbol{X} = X_i$ is σ_i^2-sub-Gaussian with $\sigma_i^2 \stackrel{\text{def}}{=} \sigma_{\text{opt}}^2(\alpha_i, \bar{\boldsymbol{\alpha}} - \alpha_i)$ defined from Theorem 1. Moreover, in direction $\boldsymbol{e}_i$, σ_i^2 is the optimal proxy variance. Therefore, the remaining issue is to generalize these results for arbitrary unit vectors on $\mathcal{S}^{d-1}$. We obtain the following result:

Theorem 3 (Optimal proxy variance for the Dirichlet distribution)**.** *For any parameter $\boldsymbol{\alpha}$, the Dirichlet distribution* $\mathrm{Dir}(\boldsymbol{\alpha})$ *is sub-Gaussian with optimal proxy variance $\sigma^2_{opt}(\boldsymbol{\alpha})$ given from Theorem 1 and:*

$$\sigma^2_{opt}(\boldsymbol{\alpha}) = \sigma^2_{opt}(\alpha_{max}, \bar{\boldsymbol{\alpha}} - \alpha_{max}) \quad \textit{where} \quad \alpha_{max} = \max_{1\leq i\leq d}\{\alpha_i\}.$$

Proof. We first observe that the computations of $\sigma^2_{\mathrm{opt}}(\alpha_i, \beta_i = \bar{\boldsymbol{\alpha}} - \alpha_i)$ correspond to cases where the sum $\alpha_i + \beta_i$ is fixed to $\bar{\boldsymbol{\alpha}}$ and thus independent of i. Therefore, $\sigma^2_{\mathrm{opt}}(\alpha_i, \bar{\boldsymbol{\alpha}} - \alpha_i)$ is maximal when $|(\bar{\boldsymbol{\alpha}} - \alpha_i) - \alpha_i|$ is minimal, i.e. when the distance from α_i to $\frac{1}{2}\bar{\boldsymbol{\alpha}}$ is minimal. It is easy to see that this corresponds to choosing $\alpha_i = \alpha_{\max} = \max\{\alpha_i\ ,\ 1 \leq i \leq d\}$ (by looking at the two possible cases $\alpha_{\max} \leq \frac{1}{2}\bar{\boldsymbol{\alpha}}$ and $\alpha_{\max} > \frac{1}{2}\bar{\boldsymbol{\alpha}}$).

We then observe that $\sigma^2_{\max}$ cannot be improved. Indeed, let us denote i_0 one of the components for which the maximum is obtained. Then, if we take $\boldsymbol{u} = \boldsymbol{e}_{i_0}$, the discussion presented above shows that $\sigma^2_{i_0} = \sigma^2_{\max}$ is the optimal proxy variance in this direction. Hence the optimal proxy variance cannot be lower than $\sigma^2_{\max}$.
Let us now prove that $\boldsymbol{X}$ is $\sigma^2_{\max}$-sub-Gaussian. Let $\boldsymbol{u} = (u_1, \dots, u_d)^\top$ be a unit vector on $\mathcal{S}^{d-1}$ and $\lambda \in \mathbb{R}$. We define for clarity $\boldsymbol{\lambda} = \lambda\boldsymbol{u}$. We have:

$$\begin{aligned}\mathbb{E}\left[\exp\left(\lambda\boldsymbol{u}^\top\boldsymbol{X}\right)\right] = \mathbb{E}\left[\exp\left(\boldsymbol{\lambda}^\top\boldsymbol{X}\right)\right] &= \mathbb{E}\left[\exp\left(\sum_{i=1}^d \lambda_i X_i\right)\right], \\ &= \sum_{m=0}^{\infty}\sum_{n_1+\dots+n_d=m}\frac{\lambda_1^{n_1}\dots\lambda_d^{n_d}}{(n_1)!\dots(n_d)!}\frac{1}{(\bar{\boldsymbol{\alpha}})_{\bar{\boldsymbol{n}}}}\prod_{i=1}^d(\alpha_i)_{n_i}. \quad (24)\end{aligned}$$

Note that we also have:

$$\begin{aligned}\prod_{i=1}^d \mathbb{E}\left[\exp(\lambda_i X_i)\right] &= \prod_{i=1}^d\left(\sum_{j=0}^{\infty}\frac{(\alpha_i)_j}{(j!)(\bar{\boldsymbol{\alpha}})_j}\lambda_i^j\right) \\ &= \sum_{m=0}^{\infty}\sum_{n_1+\dots+n_d=m}\frac{\lambda_1^{n_1}\dots\lambda_d^{n_d}}{(n_1)!\dots(n_d)!}\prod_{i=1}^d\frac{(\alpha_i)_{n_i}}{(\bar{\boldsymbol{\alpha}})_{n_i}}. \quad (25)\end{aligned}$$

Moreover we have the inequality:

$$\prod_{i=1}^d(\bar{\boldsymbol{\alpha}})_{n_i} \leq (\bar{\boldsymbol{\alpha}})_{\bar{\boldsymbol{n}}},$$

because both sides have the same number of terms in the product (i.e. $\bar{\boldsymbol{n}}$) but those of the right hand side are always greater or equal to those of the left hand side. Hence, from (24) and (25), we find:

$$\mathbb{E}\left[\exp\left(\boldsymbol{\lambda}^\top(\boldsymbol{X} - \boldsymbol{\mu})\right)\right] \leq \prod_{i=1}^d \mathbb{E}\left[\exp(\lambda_i(X_i - \mu_i))\right].$$

Using the optimal proxy variance of the Beta distribution proven in Theorem 1, we find:

$$\begin{aligned}\mathbb{E}\left[\exp\left(\boldsymbol{\lambda}^\top(\boldsymbol{X} - \boldsymbol{\mu})\right)\right] \leq \prod_{i=1}^d \mathbb{E}\left[\exp(\lambda_i(X_i - \mu_i))\right] &\leq \prod_{i=1}^d \exp\left(\frac{\lambda_i^2\sigma_i^2}{2}\right) \\ &\leq \prod_{i=1}^d \exp\left(\frac{\lambda_i^2\sigma^2_{\max}}{2}\right) = \exp\left(\frac{\sigma^2_{\max}\|\boldsymbol{\lambda}\|^2}{2}\right),\end{aligned}$$

thus showing that $\boldsymbol{X}$ is $\sigma^2_{\max}$-sub-Gaussian and concluding the proof. □

Note that using Theorem 3, we obtain the following corollary:

Corollary 2. *For any integer $d \geq 2$, the Dirichlet distribution* $\mathrm{Dir}(\alpha_1, \dots, \alpha_d)$ *is strictly sub-Gaussian if and only if $d = 2$ and $\alpha_1 = \alpha_2$.*

Indeed, we first need to require $\alpha_1 = \dots = \alpha_d \stackrel{\text{def}}{=} \alpha$ so that all directions have the same optimal proxy variance. Then, each component satisfies $X_i = \boldsymbol{e}_i^\top\boldsymbol{X} \sim \mathrm{Beta}(\alpha, (d-1)\alpha)$ and Theorem 1 shows that σ_i^2 is the optimal proxy variance for X_i if and only if $\alpha = (d-1)\alpha$, i.e. if and only if $d = 2$.

Acknowledgments

We wish to thank Stéphane Boucheron for an enlightening discussion of our results and an anonymous referee for insightful suggestions. O.M. would like to thank Université de Lyon, Université Jean Monnet and Institut Camille Jordan for financial support. This work was supported by the LABEX MILYON (ANR-10-LABX-0070) of Université de Lyon, within the program "Investissements d'Avenir" (ANR-11-IDEX-0007) operated by the French National Research Agency (ANR). J.A. would like to thank Inria Grenoble Rhône-Alpes and Laboratoire Jean Kuntzmann, Université Grenoble Alpes for financial support. J.A. is also member of Laboratoire de Statistique, CREST, Paris. This work was partially conducted during a scholar visit of J.A. at the Department of Statistics & Data Science of the University of Texas at Austin.

References

Arbel, J., Favaro, S., Nipoti, B., and Teh, Y. W. (2017). Bayesian nonparametric inference for discovery probabilities: credible intervals and large sample asymptotics. *Statistica Sinica*, 27:839–858.

Ben-Hamou, A., Boucheron, S., and Ohannessian, M. I. (2017). Concentration inequalities in the infinite urn scheme for occupancy counts and the missing mass, with applications. *Bernoulli*, 23(1):249–287.

Berend, D. and Kontorovich, A. (2013). On the concentration of the missing mass. *Electronic Communications in Probability*, 18(3):1–7.

Birkhoff, G. and Rota, G.-C. (1989). *Ordinary Differential Equations*. John Wiley and Sons Editions.

Bobkov, S. G. and Götze, F. (1999). Exponential integrability and transportation cost related to logarithmic sobolev inequalities. *Journal of Functional Analysis*, 163(1):1–28.

Boucheron, S., Lugosi, G., and Massart, P. (2013). *Concentration inequalities: A nonasymptotic theory of independence*. Oxford University Press.

Buldygin, V. V. and Kozachenko, Y. V. (1980). Sub-Gaussian random variables. *Ukrainian Mathematical Journal*, 32(6):483–489.

Buldygin, V. V. and Kozachenko, Y. V. (2000). *Metric characterization of random variables and random processes*, volume 188. American Mathematical Society, Providence, Rhode Island.

Buldygin, V. V. and Moskvichova, K. (2013). The sub-Gaussian norm of a binary random variable. *Theory of probability and mathematical statistics*, 86:33–49.

Castillo, I. (2016). Pólya tree posterior distributions on densities. *Annales de l'Institut Henri Poincaré, to appear*.

Elder, S. (2016). Bayesian adaptive data analysis guarantees from subgaussianity. *arXiv preprint: arXiv:1611.00065*.

Gross, L. (1975). Logarithmic sobolev inequalities. *American Journal of Mathematics*, 97(4):1061–1083.

Hoeffding, W. (1963). Probability inequalities for sums of bounded random variables. *Journal of the American statistical association*, 58(301):13–30.

Kearns, M. and Saul, L. (1998). Large deviation methods for approximate probabilistic inference. In *Proceedings of the Fourteenth conference on Uncertainty in artificial intelligence*, pages 311–319.

Ledoux, M. (1999). Concentration of measure and logarithmic sobolev inequalities. *Lecture notes in mathematics - Springer Verlag-*, pages 120–216.

McAllester, D. A. and Ortiz, L. (2003). Concentration inequalities for the missing mass and for histogram rule error. *Journal of Machine Learning Research*, 4:895–911.

McAllester, D. A. and Schapire, R. E. (2000). On the convergence rate of Good-Turing estimators. In *COLT*, pages 1–6.

Ordentlich, E. and Weinberger, M. J. (2005). A distribution dependent refinement of pinsker's inequality. *IEEE Transactions on Information Theory*, 51(5):1836–1840.

Perry, A., Wein, A. S., and Bandeira, A. S. (2016). Statistical limits of spiked tensor models. *arXiv preprint: arXiv:1612.07728*.

Pisier, G. (2016). Subgaussian sequences in probability and Fourier analysis. *arXiv preprint: arXiv:1607.01053*.

Raginsky, M. and Sason, I. (2013). Concentration of measure inequalities in information theory, communications, and coding. *Foundations and Trends in Communications and Information Theory*, 10(1-2):1–246.

Robinson, J. C. (2004). *An introduction to Ordinary Differential Equations*. Cambridge University Press.