



**HAL**  
open science

# On strict sub-Gaussianity, optimal proxy variance and symmetry for bounded random variables

Julyan Arbel, Olivier Marchal, Hien T Nguyen

► **To cite this version:**

Julyan Arbel, Olivier Marchal, Hien T Nguyen. On strict sub-Gaussianity, optimal proxy variance and symmetry for bounded random variables. ESAIM: Probability and Statistics, 2020, 24, pp.39-55. 10.1051/ps/2019018 . hal-01998252

**HAL Id: hal-01998252**

**<https://hal.science/hal-01998252>**

Submitted on 29 Jan 2019

**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.

# On strict sub-Gaussianity, optimal proxy variance and symmetry for bounded random variables

Julyan Arbel<sup>1\*</sup>, Olivier Marchal<sup>2</sup>, Hien D. Nguyen<sup>3</sup>

\*Corresponding author, email: [julyan.arbel@inria.fr](mailto:julyan.arbel@inria.fr).

<sup>1</sup>Université Grenoble Alpes, Inria, CNRS, Grenoble INP, LJK, 38000 Grenoble, France.

<sup>2</sup>Université de Lyon, CNRS UMR 5208, Université Jean Monnet, Institut Camille Jordan, 69000 Lyon, France.

<sup>3</sup>Department of Mathematics and Statistics, La Trobe University, Bundoora Melbourne 3086, Victoria Australia.

## Abstract

We investigate the sub-Gaussian property for almost surely bounded random variables. If sub-Gaussianity *per se* is *de facto* ensured by the bounded support of said random variables, then exciting research avenues remain open. Among these questions is how to characterize the optimal sub-Gaussian proxy variance? Another question is how to characterize *strict* sub-Gaussianity, defined by a proxy variance equal to the (standard) variance? We address the questions in proposing conditions based on the study of functions variations. A particular focus is given to the relationship between strict sub-Gaussianity and symmetry of the distribution. In particular, we demonstrate that symmetry is neither sufficient nor necessary for strict sub-Gaussianity. In contrast, simple necessary conditions on the one hand, and simple sufficient conditions on the other hand, for strict sub-Gaussianity are provided. These results are illustrated via various applications to a number of bounded random variables, including Bernoulli, beta, binomial, uniform, Kumaraswamy, and triangular distributions.

# 1 Introduction

Sub-Gaussian distributions are probability distributions that have tail probabilities that are upper bounded by Gaussian tails. More specifically, a random variable  $X$  with finite mean  $\mu = \mathbb{E}[X]$  is sub-Gaussian if there exists  $\sigma^2 > 0$  such that:

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

The constant  $\sigma^2$  is called a *proxy variance* and  $X$  is termed  $\sigma^2$ -sub-Gaussian. For a sub-Gaussian random variable  $X$ , the smallest proxy variance is called the **optimal proxy variance** and is denoted  $\sigma_{\text{opt}}^2(X)$ , or simply  $\sigma_{\text{opt}}^2$ . The variance always provides a lower bound on the optimal proxy variance:  $\mathbb{V}[X] \leq \sigma_{\text{opt}}^2(X)$ . When  $\sigma_{\text{opt}}^2(X) = \mathbb{V}[X]$ ,  $X$  is said to be *strictly* sub-Gaussian.

The sub-Gaussian property is increasingly studied and used in various fields of probability and statistics, primarily due to its intricate link with concentration inequalities (Boucheron et al., 2013, Raginsky and Sason, 2013), transportation inequalities (Bobkov and Götze, 1999, van Handel, 2014) and PAC-Bayes inequalities (Catoni, 2007). Applications include the missing mass problem (McAllester and Ortiz, 2003, Berend and Kontorovich, 2013, Ben-Hamou et al., 2017), multi-armed bandit problems (Bubeck and Cesa-Bianchi, 2012) and singular values of random matrices (Rudelson and Vershynin, 2010).

This paper focuses on the study of almost surely bounded random variables, where Bernoulli, beta, binomial, Kumaraswamy (Jones, 2009) or triangular (Kotz and Van Dorp, 2004) distributions are taken as standard and common examples. If sub-Gaussianity *per se* is *de facto* ensured because the support of said random variables is bounded, then exciting research avenues remain open in the area. Among these questions are (a) how to obtain the optimal sub-Gaussian proxy variance, and (b) how to characterize *strict* sub-Gaussianity?

Regarding question (a), we propose general conditions characterizing the optimal sub-Gaussian proxy variance, thus generalizing previous work (Marchal and Arbel, 2017) that was tailored to the beta and Dirichlet distributions. Several techniques based on studying variations of functions are proposed. In illustrating our results with the Bernoulli distribution, we prove as a by-product of Proposition 4.1 the uniqueness of a global maximum of a function that was observed by Berend and Kontorovich (2013) “as an intriguing open problem”.

As for question (b), it turns out that the *symmetry* of the distribution plays a crucial role. By symmetry, we mean symmetry with respect to the mean  $\mu = \mathbb{E}[X]$ . That is, we say that  $X$  is symmetrically distributed if  $X$  and  $2\mu - X$  have the same distribution. Thus, if

$X$  has a density, this means that the density is symmetric with respect to  $\mu$ . A simple, and remarkable, equivalence holds for most of the standard bounded random variables.

**Proposition 1.1.** *Let  $X$  be a Bernoulli, beta, binomial, Kumaraswamy or triangular random variable. Then,*

$$X \text{ is symmetric} \iff X \text{ is strictly sub-Gaussian.}$$

The result is known for the beta distribution (Marchal and Arbel, 2017). In this article, we provide proofs for the Bernoulli, binomial, Kumaraswamy and triangular distributions.

From Proposition 1.1, it may be tempting to conjecture that the equivalence holds true for *any* random variable having a bounded support. However, we establish that this is not the case. This was actually one of the starting points for the present work. More precisely, we shall provide a proof of the following result.

**Proposition 1.2.** *Symmetry of  $X$  is neither*

*(i) a sufficient condition, nor*

*(ii) a necessary condition,*

*for the strict sub-Gaussian property.*

The proof of this result is presented in Section 3.2, where we demonstrate that (i) there exists simple symmetric mixtures of distributions (e.g., a two-components mixture of beta distribution and a three-components mixture distribution of Dirac masses) which are not strictly sub-Gaussian, and that (ii) there exists an asymmetric three-components mixture of Dirac masses which is strictly sub-Gaussian.

Before delving into detailing the strict sub-Gaussianity property in Section 3, we first investigate some conditions that characterize the optimal proxy variance  $\sigma_{\text{opt}}^2$ , in Section 2. The results of Sections 2 and 3 are then illustrated on a number of standard random variables on bounded supports, in Section 4. Technical results are presented in Appendix A.

## 2 Characterizations of the optimal proxy variance $\sigma_{\text{opt}}^2$

Let  $X$  be an almost surely bounded random variable with mean  $\mu = \mathbb{E}[X]$ . Then,  $X$  is sub-Gaussian and satisfies Definition 1 for some  $\sigma^2 > 0$ .

An equivalent definition is that

$$\forall \lambda \in \mathbb{R} : \sigma^2 \geq \frac{2}{\lambda^2} \mathcal{K}(\lambda),$$

where the function  $\mathcal{K}$ , defined on  $\mathbb{R}$  by:  $\mathcal{K}(\lambda) = \ln \mathbb{E}[\exp(\lambda[X - \mu])]$ , corresponds to the cumulants generating function of  $X - \mu$ . Thus the optimal proxy variance  $\sigma_{\text{opt}}^2$  can be defined as the supremum

$$\sigma_{\text{opt}}^2 = \sup_{\lambda \in \mathbb{R}} \frac{2}{\lambda^2} \mathcal{K}(\lambda). \quad (2)$$

If  $X$  is almost surely bounded, then this supremum is attained, see Lemma A.1 for details. Note that the function  $h$ , defined on  $\mathbb{R}$  by

$$h(\lambda) = \frac{2}{\lambda^2} \mathcal{K}(\lambda), \quad (3)$$

is continuous at  $\lambda = 0$ , since a standard series expansion demonstrates that:

$$h(\lambda) \stackrel{\lambda \rightarrow 0}{\cong} \mathbb{V}[X] + o(1). \quad (4)$$

Moreover,  $h$  may never vanish. In fact, since the logarithm function is strictly concave, Jensen's inequality implies that for any  $\lambda \in \mathbb{R}$ ,

$$h(\lambda) = \frac{2}{\lambda^2} \ln \mathbb{E}[e^{\lambda(X-\mu)}] > \frac{2}{\lambda^2} \mathbb{E}[\ln e^{\lambda(X-\mu)}] = 0. \quad (5)$$

Equation (4) also explains directly why  $\sigma_{\text{opt}}^2 \geq \mathbb{V}[X]$ , since the variance is the value of the right-hand side (r.h.s.) function at  $\lambda = 0$  and thus the maximum is always greater or equal to it. We therefore have the following result.

**Proposition 2.1** (Characterization of  $\sigma_{\text{opt}}^2$  by  $h$ ). *The optimal proxy variance is given by:*

$$\sigma_{\text{opt}}^2 = \max_{\lambda \in \mathbb{R}} h(\lambda) = \max_{\lambda \in \mathbb{R}} \frac{2}{\lambda^2} \mathcal{K}(\lambda). \quad (6)$$

We may now present a necessary (but not always sufficient) system of equations for  $\sigma_{\text{opt}}^2$ . Indeed, since the maximum is achieved at a finite point, then this point must necessarily be a zero of the derivative of  $h$ , if  $h$  is differentiable (we will denote by  $\mathcal{D}^k$  the space of

functions that are  $k$  times differentiable on  $\mathbb{R}$  and by  $\mathcal{C}^k$  the space of functions that are  $k$  times differentiable on  $\mathbb{R}$  and for which the  $k^{\text{th}}$  derivative is continuous on  $\mathbb{R}$ ).

Thus, we obtain the following corollary.

**Corollary 2.2** (Necessary condition for  $\sigma_{\text{opt}}^2$ , with respect to  $h$ ). *Let  $\sigma_{\text{opt}}^2$  be the optimal proxy variance, and assume that  $h$  and  $\mathcal{K}$  are  $\mathcal{D}^1$ . Then there exists a finite  $\lambda_0$ , such that*

$$\sigma_{\text{opt}}^2 = h(\lambda_0) \text{ and } h'(\lambda_0) = 0, \quad (7)$$

which is equivalent to

$$\sigma_{\text{opt}}^2 = \frac{2}{\lambda_0^2} \mathcal{K}(\lambda_0) \text{ and } \lambda_0 \mathcal{K}'(\lambda_0) = 2\mathcal{K}(\lambda_0), \quad (8)$$

using only the centered cumulants generating function  $\mathcal{K}$ .

In practice, the previous set of equations has to be used with caution, since there may be more than one solution to the second equation involving the derivative of  $h$  (or that of  $\mathcal{K}$ ), and a global maximizer is required to be picked among the stationary points, instead of a minimizer or a local maximizer. On a case-by-case basis, the following approach based on ordinary differential equations (ODEs), satisfied by  $h$ , can be used to demonstrate that it has a unique global maximum.

**Proposition 2.3.** *If the function  $h$  is  $\mathcal{C}^2$ , then it is the unique solution of the ordinary differential equations:*

$$h'(\lambda) + \frac{2}{\lambda} h(\lambda) = \frac{2}{\lambda^2} \mathcal{K}'(\lambda) \text{ with } h(0) = \mathbb{V}[X], \quad (9)$$

or

$$h''(\lambda) + \frac{3}{\lambda} h'(\lambda) = \frac{2}{\lambda} \left( \frac{\mathcal{K}'(\lambda)}{\lambda} \right)' \text{ with } h(0) = \mathbb{V}[X] \text{ and } h'(0) = \frac{1}{3} \mathbb{E}[(X - \mu)^3]. \quad (10)$$

*Proof.* The result is directly obtained by differentiating  $h$  and via standard analysis theorems.  $\square$

*Remark.* For cases such as the Bernoulli and uniform distributions, we may prove that the r.h.s. of (10) is strictly negative on  $\mathbb{R}^* := \mathbb{R} \setminus \{0\}$ . This implies that if  $\lambda_0$  is extremal (i.e.,  $h'(\lambda_0) = 0$ ), then it satisfies  $h''(\lambda_0) < 0$  so that it is a local maximum. This implies that  $h$  has no local minimum and thus may only have one critical point which is necessarily the unique global maximum.

We conclude this section with another possible methodology for deriving a necessary and sufficient condition for  $\sigma_{\text{opt}}^2$ . To this end, the problem needs to be addressed from a different point of view, by studying the difference of the terms of Definition 1:

$$\Delta : (\sigma^2, \lambda) \in \mathbb{R}_+^* \times \mathbb{R} \mapsto \exp\left(\frac{\lambda^2 \sigma^2}{2}\right) - \mathbb{E}[\exp(\lambda[X - \mu])]. \quad (11)$$

**Proposition 2.4** (Characterization of  $\sigma_{\text{opt}}^2$ , with respect to  $\Delta$ ). *If  $\Delta$  is  $\mathcal{C}^1$ , then the optimal proxy variance is characterized by:*

$$\lambda \mapsto \Delta(\sigma_{\text{opt}}^2, \lambda) \geq 0 \text{ and } \exists \lambda_0 \in \mathbb{R}, \text{ such that } \Delta(\sigma_{\text{opt}}^2, \lambda_0) = 0 \text{ and } \partial_\lambda \Delta(\sigma_{\text{opt}}^2, \lambda_0) = 0. \quad (12)$$

*Proof.* See Section A.2, in Appendix. □

This proof technique was used by Marchal and Arbel (2017) for obtaining the optimal proxy variance of the beta and Dirichlet distributions. However we find more convenient to use the conditions stated in Proposition 2.3 using the function  $h$  to address the issues presented in this article, except for the triangular distribution in Section 4.2 where this method is employed for a numerical evaluation of  $\sigma_{\text{opt}}^2$ .

*Remark.* In general, we would like to remove the condition:  $\lambda \mapsto \Delta(\sigma^2, \lambda) \geq 0$  on the r.h.s. of Proposition 2.4, in order to have a simpler (and local) characterization of the optimal proxy variance, as a solution of (12). However, this is not possible, since we may not exclude that there exists a value  $\sigma^2 < \sigma_{\text{opt}}^2$  for which  $\Delta(\sigma^2, \lambda)$  presents a double zero  $\lambda_0$  where locally it remains non-negative but at the same time a whole interval far from  $\lambda_0$  where it would be strictly negative.

## 3 On strict sub-Gaussianity

### 3.1 Conditions based on the cumulants

Strict sub-Gaussianity is fulfilled when the optimal proxy variance equals the variance. In view of Equation (4), Proposition 2.1 can be rewritten as the following corollary in order to characterize the strict sub-Gaussianity property.

**Corollary 3.1** (Corollary of Proposition 2.1). *A distribution is strictly sub-Gaussian if and only if the maximum of function  $h$ , defined in (3), is attained in zero (and is automatically equal to  $\mathbb{V}[X]$ ). That is:*

$$\max_{\lambda \in \mathbb{R}} h(\lambda) = h(0) = \mathbb{V}[X]. \quad (13)$$

This characterization provides necessary conditions, based on cumulants, that are required for strict sub-Gaussianity to hold.

**Proposition 3.2** (Necessary conditions based on cumulants). *If  $X$  is strictly sub-Gaussian, then the 3<sup>rd</sup> and 4<sup>th</sup> cumulants of  $X$  must satisfy*

$$\kappa_3 = \mathbb{E}[(X - \mathbb{E}[X])^3] = 0, \text{ and} \quad (14)$$

$$\kappa_4 = \mathbb{E}[(X - \mathbb{E}[X])^4] - 3\mathbb{V}[X]^2 \leq 0. \quad (15)$$

*Proof.* By definition of the cumulant generating function  $\mathcal{K}(\lambda)$  of  $X - \mu$ ,

$$\mathcal{K}(\lambda) = \sum_{i=1}^{\infty} \kappa_i \frac{\lambda^i}{i!}, \quad (16)$$

where  $\kappa_i$  are the cumulants of  $X - \mu$ . Since  $\kappa_1 = \mu - \mu = 0$  and  $\kappa_2 = \mathbb{V}[X]$ , and using values for the third and fourth cumulants given in (14) and (15), we may write (locally around  $\lambda \rightarrow 0$ ):

$$h(\lambda) = \mathbb{V}[X] + \mathbb{E}[(X - \mu)^3] \frac{\lambda}{3} + (\mathbb{E}[(X - \mu)^4] - 3\mathbb{V}[X]^2) \frac{\lambda^2}{12} + O(\lambda^3). \quad (17)$$

Therefore if  $\mathbb{E}[(X - \mu)^3] \neq 0$ , the maximum of  $h(\lambda)$  cannot be  $h(0)$  and thus strict sub-Gaussianity cannot be achieved. We conclude the proof by noting that if  $\mathbb{E}[(X - \mu)^3] = 0$ , we have the fact that  $\lambda = 0$  can be a local maximum, only if  $\mathbb{E}[(X - \mu)^4] \leq 3\mathbb{V}[X]^2$ .  $\square$

Condition (14) requires that the third centered moment is zero and Condition (15) imposes a relation between the second and fourth centered moments. Note that the latter condition can be compactly formulated via an alternative condition on the kurtosis of  $X$ :

$$\text{Kurt}[X] = \frac{\mathbb{E}[(X - \mathbb{E}[X])^4]}{\mathbb{E}[(X - \mathbb{E}[X])^2]^2} \leq 3.$$

More specifically, sub-Gaussianity requires that the random variable has kurtosis less than or equal to three, which is the kurtosis of a standard Gaussian random variable. Such distributions are referred to as *platycurtic*. The fourth cumulant defined in (15) is also termed *excess kurtosis*. Thus, strict sub-Gaussianity requires negative excess kurtosis.

When the above necessary conditions (14) and (15) hold, we are not able to obtain simple additional necessary conditions on the next cumulants. In particular, note that strict sub-Gaussianity *does not* imply symmetry (i.e.,  $\mathbb{E}[(X - \mathbb{E}[X])^{2j+1}] = 0$ , for any  $j \geq 0$ ), as will be discussed in the next section.

In contrast, more can be said when the distribution is symmetric. In fact, in the symmetric case, the moments of odd order are zero, and a simple sufficient condition can be readily obtained by comparing the Taylor expansions at  $\lambda = 0$  of both terms of inequality (1), as stated in the following proposition.

**Proposition 3.3** (Sufficient condition based on moments). *If  $X$  is symmetric with respect to its mean  $\mu = \mathbb{E}[X]$ , then a sufficient condition for  $X$  to be strictly sub-Gaussian can be stated in terms of all its even moments. That is, for  $X$  to be strictly sub-Gaussian, it is sufficient that*

$$\forall j \geq 2, \quad \frac{\mathbb{E}[(X - \mu)^{2j}]}{(2j)!} \leq \frac{(\mathbb{V}[X])^j}{2^j j!} \quad (18)$$

holds.

*Proof.* The proof is based on series expansions at  $\lambda = 0$  of both terms of inequality (1), when the proxy variance  $\sigma^2$  is set to the variance  $\mathbb{V}[X]$ . Namely:

$$\mathbb{E}[\exp(\lambda X)] = \sum_{j=0}^{\infty} \mathbb{E}[X^{2j}] \frac{\lambda^{2j}}{(2j)!}, \quad \text{and} \quad \exp\left(\frac{\lambda^2 \mathbb{V}[X]}{2}\right) = \sum_{j=0}^{\infty} \frac{(\mathbb{V}[X])^j \lambda^{2j}}{2^j j!}, \quad (19)$$

when compared term-by-term, leads to inequality (1), under assumption (18). Note that inequality (18) needs be checked only for  $j \geq 2$ , as it trivially holds for  $j = 0, 1$ .  $\square$

This technique was used by [Marchal and Arbel \(2017\)](#) (Section 2.2) for showing that a (symmetric) Beta( $\alpha, \alpha$ ) random variable is strictly sub-Gaussian. We also use it to address the cases of Bernoulli and binomial, and triangular distributions in Section 4.

## 3.2 Link with symmetry

The relationship between strict sub-Gaussianity and symmetry was discussed in the Introduction. Here, we provide a proof of Proposition 1.2, while the proof of Proposition 1.1 is deferred to Section 4.

### 3.2.1 Symmetry is neither a sufficient condition...

Simple symmetric distributions which break the necessary condition of negative excess kurtosis can easily be constructed by hand. One such construction is by means of mixture of

Dirac masses. First, consider the discrete random variable

$$X \sim \frac{\eta}{2}(\delta_{-1} + \delta_1) + (1 - \eta)\delta_0, \quad (20)$$

which is a three-component mixture of Dirac masses at locations  $-1$ ,  $0$  and  $1$ , with  $\eta \in [0, 1]$ . It is symmetric, by construction, and its excess kurtosis equals

$$\kappa_4 = \mathbb{E}[X^4] - 3\mathbb{V}[X]^2 = \eta - 3\eta^2 = \eta(1 - 3\eta), \quad (21)$$

which is *strictly positive* for all values  $\eta \in (0, \frac{1}{3})$ , hence  $X$  is not strictly sub-Gaussian for these values by virtue of Proposition 3.2. On the other hand when  $\eta \rightarrow 1$ , the distribution of  $X$  degenerates to that of the so-called Rademacher random variable, which leads to the least possible excess kurtosis of  $-2$ .

Similar counter-examples to the sufficientness of symmetry can be built in the form of mixtures of two symmetric beta variables:

$$X \sim \eta \text{Beta}(\alpha, \alpha) + (1 - \eta) \text{Beta}(\beta, \beta),$$

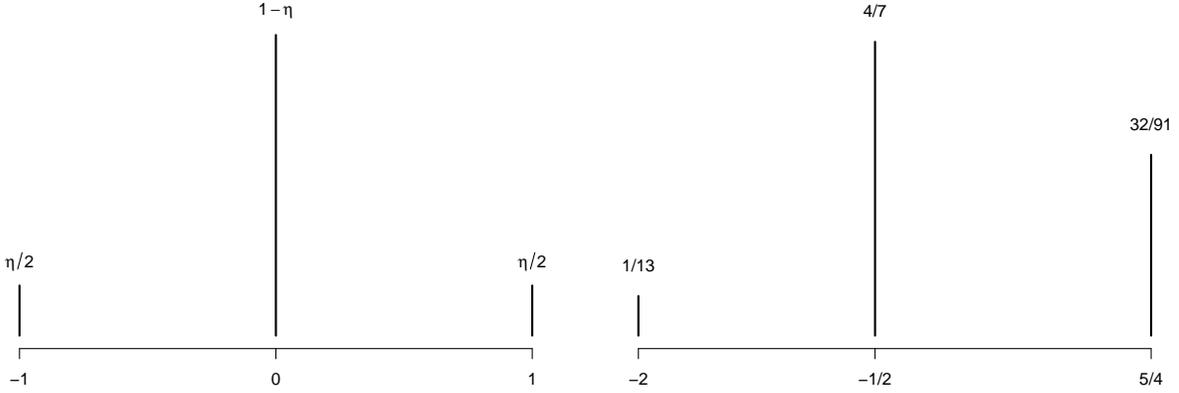
for  $\eta \in (0, 1)$  and  $\alpha, \beta > 0$ . For any value of  $\eta \in (0, 1)$ , values for  $\alpha, \beta$  leading to positive excess kurtosis can be obtained. For instance, we may set  $(\eta, \alpha, \beta) = (0.1, 1.5, 9)$ , to obtain the excess kurtosis  $\kappa_4 \approx 1.1 \times 10^{-4}$ .

### 3.2.2 ... nor a necessary condition for strict sub-Gaussianity

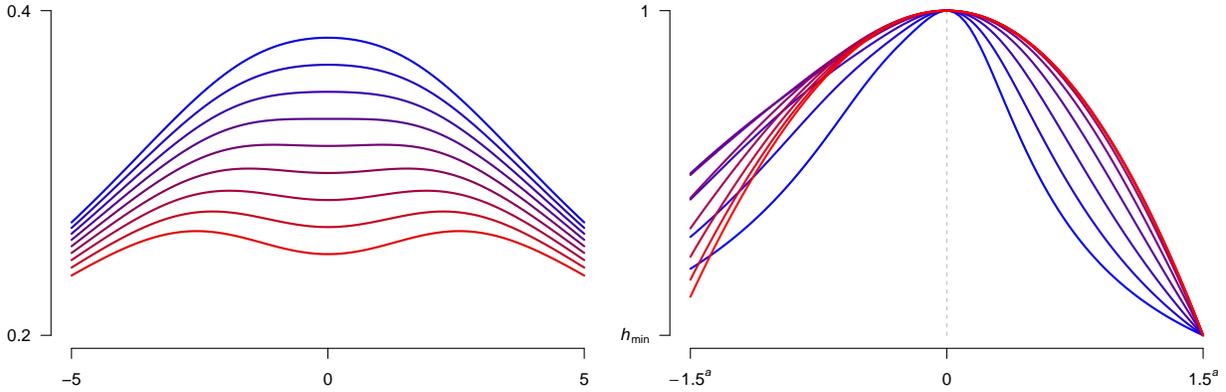
Although most typical bounded random variables that are strictly sub-Gaussian are symmetric (see, e.g., Proposition 1.1), the symmetry of the distributions of such variables is not a necessary condition for strict sub-Gaussianity. Examples of such distributions include mixtures of Dirac masses. For example,

$$X \sim \sum_{i=1}^3 p_i \delta_{x_i} \text{ with } \sum_{i=1}^3 p_i = 1 \quad (22)$$

with  $(x_1, x_2, x_3) = (-2, -\frac{1}{2}, \frac{5}{4})$  and  $(p_1, p_2, p_3) = (\frac{1}{13}, \frac{4}{7}, \frac{32}{91})$ . The function  $h$  for the random variable characterized by (22) is plotted in Figure 1b. Note that it attains its maximum in  $\lambda = 0$ .



(a) Distribution (20): symmetric but not strictly sub-Gaussian when  $\eta \in (0, \frac{1}{3})$  (b) Distribution (22): asymmetric but strictly sub-Gaussian



(c) Function  $h$  for (20) with  $\eta \in [0.25, 0.4]$  (d) Function  $h$  for (22), zoomed-in for red curves

Figure 1: Illustration of the mixtures of Dirac masses, proving Proposition 1.2 described in Sections 3.2.1 (a,c) and 3.2.2 (b,d). In (c),  $\eta$  varies from 0.25 (red curve, maximum not at zero) to 0.4 (blue curve, maximum at zero). In (d), we illustrate the function  $h$  with different zooming scales (around  $\lambda = 0$ ): from  $[-1.5^5, 1.5^5]$  (blue curve, maximum zoom-out) to  $[-1.5^{-5}, 1.5^{-5}]$  (red curve, maximum zoom-in), with an adapted  $y$ -scale, showing that the maximum is attained in zero.

## 4 Results and applications to standard distributions

### 4.1 Bernoulli and binomial distributions

Consider a Bernoulli random variable,  $X \sim \text{Ber}(\mu)$  with  $\mu \in (0, 1)$  and a binomial random variable,  $Y \sim \text{Bin}(n, \mu)$  which can be obtained as the sum of  $n$  independent  $\text{Ber}(\mu)$  random variables,  $n$  a positive integer.

*Proof of Proposition 1.1 for the Bernoulli and binomial distributions.*

Starting with the Bernoulli: the third cumulant is equal to

$$\mathbb{E}[(X - \mathbb{E}[X])^3] = 2\mu(1 - \mu) \left( \frac{1}{2} - \mu \right),$$

thus, by virtue of Proposition 3.2, a non-degenerate Bernoulli random variable may only be strictly sub-Gaussian when  $\mu = \frac{1}{2}$ . That is, when it is symmetric.

Conversely, verifying the sufficient condition for the symmetric Bernoulli distribution  $\text{Ber}(1/2)$  is equivalent to assessing the condition for the Rademacher distribution instead. That is, the distribution of random variable  $X$ , where the events  $X = -1$  and  $X = 1$  have equal probability

$$\mathbb{P}[X = -1] = \mathbb{P}[X = 1] = \frac{1}{2}.$$

Since  $X^2 = 1$ , the variance of  $X$  and all of its even moments are  $\mathbb{V}[X] = \mathbb{E}[X^{2j}] = 1$ . Therefore, to verify the sufficient condition of Proposition 3.3, we are required to demonstrate that

$$(2j)! \geq 2^j j!,$$

for each  $j \geq 2$ , which follows from the expansion

$$(2j)! = \underbrace{2j \times \cdots \times (j+1)}_j \times j! \geq 2^j j!.$$

Thus, we have verification of the sufficient condition for the Rademacher distribution and hence the symmetric Bernoulli distribution, as a consequence.

Turning to the binomial distribution, we observe that the optimal proxy variance of a sum of i.i.d. (independent and identically distributed) variables is the sum of the optimal proxy variances. Thus, we immediately obtain the result that

$$\sigma_{\text{opt}}^2[\text{Bin}(n, \mu)] = n\sigma_{\text{opt}}^2[\text{Ber}(\mu)]. \quad (23)$$

In particular,  $X \sim \text{Bin}(n, \mu)$  is strictly sub-Gaussian if and only if  $\mu = \frac{1}{2}$ . □

We now turn to the optimal proxy variance of a Bernoulli, which has the form

$$\sigma_{\text{opt}}^2 = \frac{\frac{1}{2} - \mu}{\ln\left(\frac{1}{\mu} - 1\right)}. \quad (24)$$

This fact is known via Theorem 2.1 and Theorem 3.1 of [Buldygin and Moskvichova \(2013\)](#); see also the discussion in the introduction of [Marchal and Arbel \(2017\)](#). Here, we focus on

a rather different approach, based on function  $h$  and Corollary 2.2, where

$$h(\lambda) = \frac{2}{\lambda^2} (\ln[\mu e^\lambda + (1 - \mu)] - \mu\lambda). \quad (25)$$

Note that the study of the variations of  $h$  is observed by Berend and Kontorovich (2013) (cf. their function  $g$ ; Equation (2.1)). However, a formal proof that  $h$  has a single global maximum is left “as an intriguing open problem” by Berend and Kontorovich. This is stated in the next proposition, and formally proved, below. An illustration of this result is presented in Figure 2.

**Proposition 4.1.** *If  $X \sim \text{Ber}(\mu)$ , then the function*

$$h : \lambda \mapsto \frac{2}{\lambda^2} (\ln(\mu e^\lambda + 1 - \mu) - \mu\lambda)$$

*admits a unique critical point which is a global maximum. The global maximizer is obtained at  $\lambda_0 = 2 \ln \frac{1-\mu}{\mu}$ , which leads to the optimal proxy variance of form*

$$\sigma_{\text{opt}}^2 = h(\lambda_0) = \frac{\frac{1}{2} - \mu}{\ln\left(\frac{1}{\mu} - 1\right)}.$$

*Proof.* Let us first prove that  $h$  admits a unique critical point, which is a global maximum, by using Proposition 2.3 and the remark that follows. ODEs (9) and (10) are respectively

$$h'(\lambda) + \frac{2}{\lambda} h(\lambda) := r(\lambda) = \frac{2\mu(1-\mu)(e^\lambda - 1)}{\lambda^2(\mu e^\lambda + 1 - \mu)}, h(0) = \mu(1 - \mu) \quad (26)$$

and

$$h''(\lambda) + \frac{3}{\lambda} h'(\lambda) = r'(\lambda) + \frac{r(\lambda)}{\lambda} = \frac{2\mu(1-\mu)}{\lambda^3(\mu e^\lambda + 1 - \mu)} ((\lambda + 2\mu - 1)e^\lambda - \mu e^{2\lambda} + 1 - \mu), \quad (27)$$

with  $h'(0) = \frac{\mu(1-\mu)(2\mu-1)}{3}$ . Let us denote  $g(\lambda) = (\lambda + 2\mu - 1)e^\lambda - \mu e^{2\lambda} + 1 - \mu$ ,  $u = e^\lambda > 0$  and  $G(u) = u \ln(u) + (2\mu - 1)u - \mu u^2 + 1 - \mu$ . We have  $G''(u) = \frac{1}{u} - 2\mu$  so that  $G''$  is positive on  $\left[0, \frac{1}{2\mu}\right]$  and negative on  $\left[\frac{1}{2\mu}, +\infty\right)$ . Since  $G'\left(\frac{1}{2\mu}\right) = -\ln(2\mu) + 2\mu(1 - 2\mu) < 0$ , for  $\mu > \frac{1}{2}$ , we have the fact that  $G'$  is always strictly negative on  $\mathbb{R}_+$ . Thus,  $G$  is a strictly decreasing function of  $u$  and hence  $g$  is also a strictly decreasing function of  $\lambda$ . Note that  $g(0) = 0$ , so that  $g$  is positive on  $\mathbb{R}_-$  and negative on  $\mathbb{R}_+$ . Thus,  $r'(\lambda) + \frac{r(\lambda)}{\lambda}$  is always strictly negative (there is a factor  $\lambda^3$  in the denominator that changes sign at  $\lambda = 0$ ). We conclude that a point  $\lambda_2 \neq 0$ , where  $h'(\lambda_2) = 0$ , always satisfies  $h''(\lambda_2) < 0$  and therefore it is always a local maximum of  $h$ .

By differentiating  $h$ , we observe that the global maximizer of  $h$  is obtained as the unique solution (in  $\lambda_0$ ) of the equation

$$2(\mu e^{\lambda_0} + 1 - \mu) \ln(\mu e^{\lambda_0} + 1 - \mu) - \mu \lambda_0 ((1 + \mu)e^{\lambda_0} + 1 - \mu) = 0.$$

It is easy to verify that  $\lambda_0 = 2 \ln \left( \frac{1-\mu}{\mu} \right)$ , and that this leads to the optimal proxy variance as stated.  $\square$

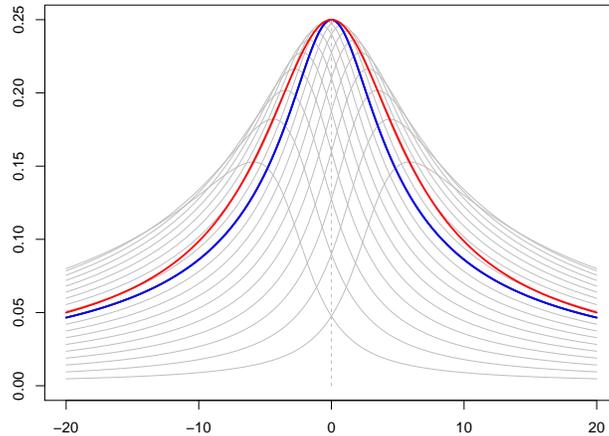


Figure 2: Illustration of the function  $h$  for  $\text{Ber}(\mu)$  random variables with  $\mu$  varying from 0.05 to 0.95 (gray curves), and of the particular case of  $\text{Ber}(\frac{1}{2})$  (blue curve). The maximum of  $h$  is realized at zero, only in the  $\text{Ber}(\frac{1}{2})$  case, which corresponds to a symmetric distribution. Maxima of  $h$ , corresponding to the optimal proxy variance, are depicted by the red curve. Varying  $\lambda$  on the  $x$ -axis.

## 4.2 Triangular distribution

We say that  $X \sim \text{Tri}(a, b)$  is a triangular random variable on  $(-a, b)$ , for any  $a, b > 0$ , if it is characterized by a density equal to

$$f(x; a, b) = \begin{cases} \frac{2}{a(a+b)}(x+a) & \text{if } -a < x < 0, \\ \frac{2}{b(a+b)}(b-x) & \text{if } 0 < x < b. \end{cases} \quad (28)$$

See [Kotz and Van Dorp \(2004\)](#) for details and properties of such distributions. A recent review of research developments regarding the triangular distribution appears in [Nguyen and McLachlan \(2017\)](#).

*Proof of Proposition 1.1 for the triangular distribution.*

The third cumulant is equal to

$$\mathbb{E}[(X - \mathbb{E}[X])^3] = \frac{(b-a)(2a+b)(2b+a)}{270},$$

so by virtue of [Proposition 3.2](#), a triangular random variable may only be strictly sub-Gaussian when  $a = b$ . That is when it is symmetric.

Conversely, when the distribution is symmetric with  $a = b$ , we can easily express the moments of even order in the form

$$\mathbb{E}[X^{2j}] = \frac{2a^{2j}}{(2j+1)(2j+2)},$$

so that the sufficient moment condition of [Proposition 3.3](#) is equivalent to

$$\frac{\mathbb{E}[X^{2j}]}{(2j)!} \leq \frac{(\mathbb{V}[X])^j}{2^j j!} \Leftrightarrow \frac{2}{(2j+2)!} \leq \frac{1}{12^j j!}.$$

In other words, the only remaining inequality is to show that

$$12^j j! \leq \frac{(2j+2)!}{2}, \quad \forall j \geq 1.$$

The result is true for  $j \in \{1, 2, 3, 4\}$  by direct computation. We then make the decomposition:

$$\frac{(2j+2)!}{2} = (j+1)2^j \prod_{i=1}^j (2i+1) \geq (j+1)2^j 3 \times 5 \times \prod_{i=3}^j 7 = \frac{15}{49}(j+1)2^j 7^j,$$

which yields

$$12^j j! \leq \frac{(2j+2)!}{2} \Leftrightarrow j+1 \geq \frac{15}{49} \Leftrightarrow j \geq 3,$$

thus verifying the sufficient condition in the symmetric case.  $\square$

For the general case, we first observe that

$$\mathbb{E}[e^{\lambda(X-\mathbb{E}[X])}] = \frac{2}{ab(a+b)\lambda^2} \left[ ae^{\frac{\lambda(a+2b)}{3}} + be^{\frac{-\lambda(b+2a)}{3}} - (a+b)e^{\frac{\lambda(a-b)}{3}} \right].$$

We further observe, numerically, that the difference  $\Delta$  introduced in (11) admits a unique minimum, and also that the  $h$  function (3) admits a unique (global) maximum; see Figure 3. The optimal proxy variance can be obtained numerically, by minimizing (11), as detailed in Proposition 2.4.

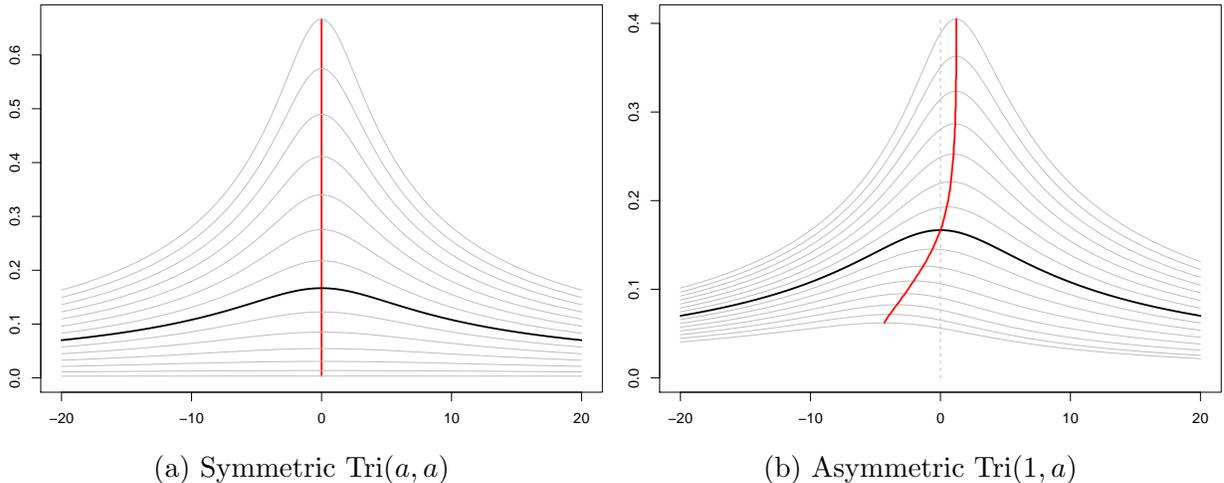


Figure 3: Illustrations of function  $h$  for  $\text{Tri}(a, a)$  (left), and  $\text{Tri}(1, a)$  (right) distributions, with  $a$  varying from 0 to 2. The black curve represents the  $h$  function of the  $\text{Tri}(1, 1)$  distribution. Maxima of  $h$  corresponding to the optimal proxy variances are depicted by the red curve. Varying  $\lambda$  on the  $x$ -axis.

### 4.3 Uniform distribution

In this section, we prove that the uniform distribution is strictly sub-Gaussian using a similar proof as for obtaining the optimal proxy variance in the Bernoulli case (i.e., Proposition 4.1). First, we observe that after translation/dilatation, we may always reduce the problem to the case of  $X \sim \text{Unif}([0, 1])$ . In this case, the moment generating function is straightforward to

compute and we may write

$$h(\lambda) = \frac{2}{\lambda^2} \ln \left[ \frac{2 \operatorname{sh}(\frac{\lambda}{2})}{\lambda} \right], \quad (29)$$

which is a symmetric and is a  $\mathcal{C}^\infty$  function. It remains to prove that it attains its global maximum at  $\lambda = 0$ .

To this end, we use Proposition 2.3 and the remark following it. ODEs (9) and (10) are respectively

$$h'(\lambda) + \frac{2}{\lambda} h(\lambda) = r(\lambda) = \frac{\operatorname{ch}(\frac{\lambda}{2})}{\lambda^2 \operatorname{sh}(\frac{\lambda}{2})} - \frac{2}{\lambda^3}, \quad \text{with } h(0) = \mathbb{V}[X] = \frac{1}{12}, \quad (30)$$

and

$$h''(\lambda) + \frac{3}{\lambda} h'(\lambda) = r'(\lambda) + \frac{r(\lambda)}{\lambda} = -\frac{2s(\lambda)}{\lambda^4 \operatorname{sh}^2(\frac{\lambda}{2})}, \quad (31)$$

with  $h(0) = \frac{1}{12}$ ,  $h'(\lambda) = 0$ , and  $s(\lambda) := \lambda^2 + \lambda \operatorname{sh}(\lambda) - 4\operatorname{ch}(\lambda) + 4$ . We now apply the same method as for the Bernoulli case. That is, we need to show that the r.h.s. of (31) is negative, which amounts to showing that the function  $s$  is positive. Let us first observe that the result holds locally around  $\lambda = 0$ . Indeed, we have  $h(0) = \frac{1}{12}$ ,  $h'(0) = 0$  and  $h''(0) = -\frac{1}{720} < 0$ . Then, observe that

$$\begin{aligned} s(\lambda) &= \lambda^2 + \lambda \operatorname{sh}(\lambda) - 4 \operatorname{ch}(\lambda) + 4, \\ s'(\lambda) &= 2\lambda - 3 \operatorname{sh}(\lambda) + \lambda \operatorname{ch}(\lambda), \\ s''(\lambda) &= 2 - 2 \operatorname{ch}(\lambda) + \lambda \operatorname{sh}(\lambda), \\ s^{(3)}(\lambda) &= \lambda \operatorname{ch}(\lambda) - \operatorname{sh}(\lambda), \text{ and} \\ s^{(4)}(\lambda) &= \lambda \operatorname{sh}(\lambda), \end{aligned} \quad (32)$$

from which we immediately conclude that

$$s(0) = 0, \quad s'(0) = 0, \quad s''(0) = 0, \quad s^{(3)}(0) = 0, \quad \text{and } s^{(4)}(0) = 0. \quad (33)$$

Obviously  $s^{(4)}$  is strictly positive on  $\mathbb{R}^*$ , thus  $s^{(3)}$  is a strictly increasing function on  $\mathbb{R}$ . Since  $s^{(3)}(0) = 0$ , we conclude that  $s^{(3)}$  is strictly negative on  $\mathbb{R}_-^*$  and strictly positive on  $\mathbb{R}_+^*$ . Thus,  $s''$  is strictly decreasing on  $\mathbb{R}_-^*$  and strictly increasing on  $\mathbb{R}_+^*$ . Finally, since  $s''(0) = 0$ , we conclude that  $s''$  is strictly positive on  $\mathbb{R}^*$ , therefore  $s'$  is a strictly increasing function on  $\mathbb{R}$ . Since  $s'(0) = 0$  then  $s'$  is strictly negative on  $\mathbb{R}_-^*$  and strictly positive on  $\mathbb{R}_+^*$  so that  $s$  is strictly decreasing on  $\mathbb{R}_-^*$  and strictly increasing on  $\mathbb{R}_+^*$ . Since  $s(0) = 0$ , we conclude that  $s$  is strictly positive on  $\mathbb{R}^*$ . This proves that  $h$  has only one unique critical point, which is therefore the global maximizer.

In conclusion for  $X \sim \text{Unif}([a, b])$  with  $a < b$ , we have the celebrated result that

$$\sigma_{\text{opt}}^2 = \mathbb{V}[X] = \frac{1}{12}(b - a)^2.$$

### 4.3.1 Sum of independent uniform variables

We may now consider the sum of independent (but not necessarily identically distributed) uniform random variables. Let  $(X_1, \dots, X_n)$  be independent variables with  $X_i \sim \text{Unif}([a_i, b_i])$  for  $i \in \{1, \dots, n\}$ , with  $a_i < b_i$  and denote  $S_n = \sum_{i=1}^n X_i$ . Since the family of uniform distributions is invariant under translation and multiplication by a constant, we have the standard result that

$$Z_i = \frac{X_i - a_i}{b_i - a_i} \sim \text{Unif}([0, 1]).$$

Thus, since  $X_i = (b_i - a_i)Z_i + a_i$ , we have

$$h_{X_i}(\lambda) = (b_i - a_i)^2 h_{\text{Unif}([0,1])}((b_i - a_i)\lambda),$$

and then, since the  $h$  function of a sum of independent random variables is the sum of the  $h$  functions of the variables and by independence of the variables  $(X_i)_{i \leq n}$ , we obtain

$$h_{S_n}(\lambda) = \sum_{i=1}^n (b_i - a_i)^2 h_{\text{Unif}([0,1])}((b_i - a_i)\lambda).$$

The sum of the r.h.s. of the equation above is composed of functions that are all strictly increasing on  $\mathbb{R}_-$  and all strictly decreasing on  $\mathbb{R}_+$ . Thus, it too is strictly increasing on  $\mathbb{R}_-$  and strictly decreasing on  $\mathbb{R}_+$ . In particular, the global maximum is unique and obtained at  $\lambda = 0$  for which we find:

$$\sigma_{\text{opt}}^2[S_n] = \sum_{i=1}^n \mathbb{V}[X_i] = \frac{1}{12} \sum_{i=1}^n (b_i - a_i)^2$$

*Remark.* Note in particular that the sum of two independent uniform variables is generically a trapezoid distribution, with symmetric triangular parts, or a symmetric (up to translation) triangular distribution. However the general asymmetric triangular case, considered in Section 4.2, cannot be expressed as a sum of independent uniform distributions.

## 4.4 Kumaraswamy distribution

Kumaraswamy distribution is characterized by the density on  $(0, 1)$ :

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

for  $\alpha, \beta > 0$ , which yields the simple distribution function of form

$$F(x; \alpha, \beta) = 1 - (1 - x^\alpha)^\beta.$$

The distribution was first studied in [Kumaraswamy \(1980\)](#) and was considered in details by [Jones \(2009\)](#).

*Proof of Proposition 1.1 for the Kumaraswamy distribution.*

The Kumaraswamy distribution is symmetric if and only if  $\alpha = \beta = 1$  ([Jones, 2009](#)). In this case, it reduces to the uniform distribution, which is strictly sub-Gaussian, as was proved in [Section 4.3](#).

Conversely, let us now consider any potentially strictly sub-Gaussian Kumaraswamy distribution. It must then satisfy the necessary conditions of [Proposition 3.2](#). The third cumulant  $\kappa_3$  vanishes if and only if the parameters satisfy the relation

$$\alpha = \frac{1}{\beta - (\beta - 1)2^{\frac{1}{\beta}}}. \quad (34)$$

In such a case, a numerical evaluation of the 4<sup>th</sup> cumulant  $\kappa_4 = \mathbb{E}[(X - \mathbb{E}[X])^4] - 3\mathbb{V}[X]^2$  demonstrates that it is negative, thus both necessary conditions of [Proposition 3.2](#) hold. However, a numerical evaluation also shows that the maximizer of the  $h$  function is never located at zero (i.e., the condition of [Corollary 3.1](#) is not satisfied), except for  $\alpha = \beta = 1$  (i.e., the uniform distribution). This is illustrated in [Figure 4](#), where the function  $h$  is plotted for  $(\alpha, \beta)$ , satisfying relation (34), with  $\beta$  varying in the interval  $[10^{-3}, 5]$ . The maximum of  $h$  is illustrated with the red curve, showing that the global maximizer always deviates from zero, except for the case of the uniform distribution (the black curve) and the degenerate symmetric Bernoulli distribution (the blue curve). This proves the necessity of symmetry and concludes the proof.  $\square$

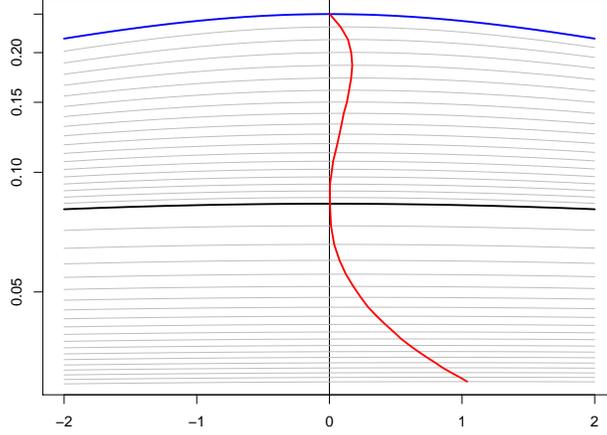


Figure 4: Function  $h$  for Kumaraswamy distributions for  $(\alpha, \beta)$  satisfying relation (34), with  $\beta$  varying from  $10^{-3}$  (top gray curve) to 5 (bottom gray curve). The black curve represents the  $h$  function of the uniform distribution. The blue curve represents the  $h$  function of the symmetric  $\text{Ber}(\frac{1}{2})$ . Maxima of  $h$ , corresponding to the optimal proxy variances, are depicted by the red curve. In particular, the maxima are located at zero only for the symmetric distributions (i.e., the uniform distribution and  $\text{Ber}(\frac{1}{2})$ ). Varying  $\lambda$  on the  $x$ -axis. Log-scale on the  $y$ -axis.

## 4.5 Beta distribution

The optimal proxy variance for the  $\text{Beta}(\alpha, \beta)$  distribution was derived in [Marchal and Arbel \(2017\)](#), Theorem 2.1. In particular, this theorem states that the optimal proxy variance is equal to the variance if and only  $\alpha = \beta$ . That is, if and only if the beta distribution is symmetric. This proves Proposition 1.1 for the beta distribution. The  $h$  function and the optimal proxy variance is illustrated on Figure 5.

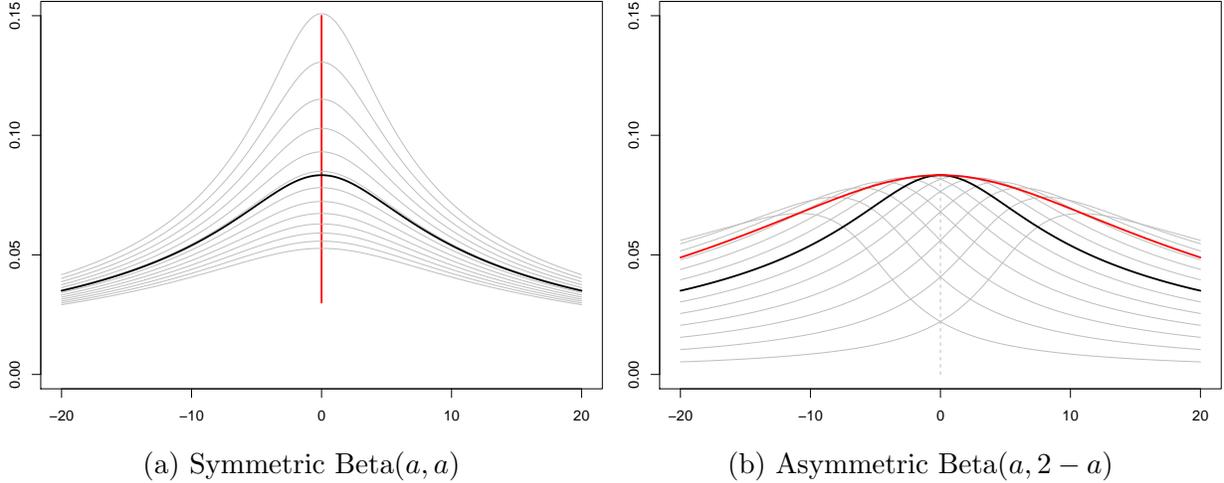


Figure 5: Illustrations of function  $h$  for Beta( $a, a$ ) (left), and Beta( $a, 2 - a$ ) (right) distributions, with  $a$  varying from 0.2 to 1.8. The black curve represents the  $h$  function of the uniform distribution. Maxima of  $h$  corresponding to the optimal proxy variance are depicted by the red curve. Varying  $\lambda$  on the  $x$ -axis.

## A Technical results

### A.1 A lemma regarding the supremum of $h$

**Lemma A.1.** *For variables with bounded support, the supremum in (2) is a maximum.*

*Proof.* We have

$$\exp(\lambda[X - \mu]) \leq \exp(|\lambda|[M + |\mu|]) \leq \exp(2|\lambda|M),$$

where  $M$  is the maximum value of  $|X|$  (which is finite since  $X$  is almost surely bounded).

Thus we obtain

$$\frac{2}{\lambda^2} \mathcal{K}(\lambda) \leq \frac{4M|\lambda|}{\lambda^2} = \frac{4M}{|\lambda|} \xrightarrow{\lambda \rightarrow \infty} 0. \quad (35)$$

Therefore, the supremum is not at infinity and since the function  $\lambda \in \mathbb{R} \mapsto \frac{2\mathcal{K}(\lambda)}{\lambda^2}$  is continuous and positive, it must achieve its maximal value at finite values of  $\lambda$ .  $\square$

### A.2 Proof of Proposition 2.4

The proof of Proposition 2.4 is based on the study of the variations of the  $\Delta$  function, defined in Equation (11), which is the object of the next lemma.

We first observe that for any  $\lambda \in \mathbb{R}^*$ , the function  $\sigma^2 \mapsto \Delta(\sigma^2, \lambda)$  is strictly increasing.

Moreover, at  $\lambda = 0$  we have:

$$\Delta(\sigma^2, \lambda) = (\sigma^2 - \mathbb{V}[X]) \frac{\lambda^2}{2} - \mathbb{E}[(X - \mu)^3] \frac{\lambda^3}{3!} + (3\sigma^4 - \mathbb{E}[(X - \mu)^4]) \frac{\lambda^4}{4!} + O(\lambda^5). \quad (36)$$

Therefore, for  $\sigma^2 > \mathbb{V}[X]$ , the function  $\lambda \mapsto \Delta(\sigma^2, \lambda)$  is strictly positive in a neighborhood of  $\lambda = 0$ , while for  $\sigma^2 < \mathbb{V}[X]$ , the function is strictly negative in a neighborhood of  $\lambda = 0$ . Thus we obtain the following lemma.

**Lemma A.2.** *The variations of  $\lambda \mapsto \Delta(\sigma^2, \lambda)$  depend on the value of  $\sigma^2$ , with respect to  $\sigma_{\text{opt}}^2$ , as follows:*

1. for  $\sigma^2 > \sigma_{\text{opt}}^2$ , the function  $\lambda \mapsto \Delta(\sigma^2, \lambda)$  is strictly positive on  $\mathbb{R}$ ,
2. for  $\sigma^2 = \sigma_{\text{opt}}^2$ , the function  $\lambda \mapsto \Delta(\sigma_{\text{opt}}^2, \lambda)$  is non-negative and there exists at least one point  $\lambda_0 \in \mathbb{R}$  for which  $\Delta(\sigma_{\text{opt}}^2, \lambda_0) = 0$ . In particular, since the function remains non-negative, it implies that  $\partial_\lambda \Delta(\sigma_{\text{opt}}^2, \lambda_0) = 0$  and  $\partial_\lambda^2 \Delta(\sigma_{\text{opt}}^2, \lambda_0) \geq 0$ , and
3. for  $\sigma^2 < \sigma_{\text{opt}}^2$ , there exists an interval not reduced to a point on which  $\lambda \mapsto \Delta(\sigma^2, \lambda) < 0$ ,

where the first and second derivatives of  $\lambda \mapsto \Delta(\sigma^2, \lambda)$  are denoted by  $\partial_\lambda \Delta$  and  $\partial_\lambda^2 \Delta$ , respectively.

*Proof.* The proof is based on fact that  $\sigma^2 \mapsto \Delta(\sigma^2, \lambda)$  is strictly increasing, and the fact that  $(\sigma^2, \lambda) \mapsto \Delta(\sigma^2, \lambda)$  is continuous.

1. Assume by contradiction that  $\Delta$  is not strictly positive. Then, since it is non-negative, there must exist at least one point  $\lambda_1$  for which  $\Delta(\sigma^2, \lambda_1) = 0$  with  $\lambda_1 \neq 0$  (because for  $\sigma^2 > \sigma_{\text{opt}}^2 \geq \mathbb{V}[X]$ , we know that  $\Delta$  is strictly positive around  $\lambda = 0$ ). Thus, for any  $\tilde{\sigma}^2 < \sigma^2$  we have  $\Delta(\tilde{\sigma}^2, \lambda_1) < \Delta(\sigma^2, \lambda_1) = 0$ , so that  $X$  is not  $\tilde{\sigma}^2$ -sub-Gaussian, hence  $\tilde{\sigma}^2 < \sigma_{\text{opt}}^2$  and by taking the limit  $\tilde{\sigma}^2 \rightarrow \sigma^2$  from below, we get  $\sigma^2 \leq \sigma_{\text{opt}}^2$ , which is a contradiction.

2. At  $\sigma^2 = \sigma_{\text{opt}}^2$ , the function  $\Delta$  must vanish at least at one point  $\lambda_0 \in \mathbb{R}$ , while remaining non-negative on  $\mathbb{R}$ . Indeed, the function  $\Delta$  is non-negative by the sub-Gaussian definition, but if  $\Delta$  was strictly positive on  $\mathbb{R}$ , then the continuity of  $\Delta$ , relatively to  $(\sigma^2, \lambda)$ , would imply that we may lower  $\sigma^2$  without  $\Delta$  vanishing. This would be in contradiction with the minimality of  $\sigma_{\text{opt}}^2$ .

3. For  $\sigma^2 < \sigma_{\text{opt}}^2$ , there exists at least a point  $\lambda_1 \in \mathbb{R}$  for which  $\Delta(\sigma^2, \lambda_1) < 0$ . By continuity of the function  $\lambda \mapsto \Delta(\sigma^2, \lambda)$ , this implies that there exists a neighborhood of  $\lambda_1$  in which  $\Delta$  is strictly negative, which concludes the proof.  $\square$

We are now ready to prove Proposition 2.4.

*Proof of Proposition 2.4.* Proposition A.2 indicates that if  $\sigma^2 = \sigma_{\text{opt}}^2$ , then  $\Delta \geq 0$ , and there exists  $\lambda_0 \in \mathbb{R}$ , such that  $\Delta(\sigma^2, \lambda_0) = 0$  and  $\partial_\lambda \Delta(\sigma^2, \lambda_0) = 0$ . Conversely, let us assume that  $\Delta \geq 0$  and  $\exists \lambda_0 \in \mathbb{R}$ , such that  $\Delta(\sigma^2, \lambda_0)$  and  $\partial_\lambda \Delta(\sigma^2, \lambda_0) = 0$ . Then, since  $\Delta \geq 0$ , we have  $\sigma^2 \geq \sigma_{\text{opt}}^2$ , and since for  $\tilde{\sigma}^2 < \sigma^2$ ,  $\Delta(\tilde{\sigma}^2, \lambda_0) < 0$ , we also have  $\sigma^2 \leq \sigma_{\text{opt}}^2$ . Thus  $\sigma^2 = \sigma_{\text{opt}}^2$ , which concludes the proof.  $\square$

## Acknowledgements

O.M. would like to thank Université Lyon 1, Université Jean Monnet and Institut Camille Jordan for material support. H.D.N. is funded by Australian Research Council grants: DE170101134 and DP180101192. 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).

## References

- 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*, 23:249–287.
- Berend, D. and Kontorovich, A. (2013). On the concentration of the missing mass. *Electronic Communications in Probability*, 18:1–7.
- Bobkov, S. G. and Götze, F. (1999). Exponential integrability and transportation cost related to logarithmic Sobolev inequalities. *Journal of Functional Analysis*, 163:1–28.
- Boucheron, S., Lugosi, G., and Massart, P. (2013). *Concentration inequalities: A nonasymptotic theory of independence*. Oxford University Press, Oxford.
- Bubeck, S. and Cesa-Bianchi, N. (2012). Regret analysis of stochastic and nonstochastic multi-armed bandit problems. *Foundations and Trends in Machine Learning*, 5:1–122.
- 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.
- Catoni, O. (2007). *PAC-Bayesian supervised classification: the thermodynamics of statistical learning*, volume 56 of *Monograph Series*. Institute of Mathematical Statistics Lecture Notes.

- Jones, M. C. (2009). Kumaraswamy’s distribution: A beta-type distribution with some tractability advantages. *Statistical Methodology*, 6:70–81.
- Kotz, S. and Van Dorp, J. R. (2004). *Beyond Beta: Other Continuous Families of Distributions with Bounded Support and Applications*. World Scientific, Singapore.
- Kumaraswamy, P. (1980). A generalized probability density function for double-bounded random processes. *Journal of Hydrology*, 46:79–88.
- Marchal, O. and Arbel, J. (2017). On the sub-Gaussianity of the beta and Dirichlet distributions. *Electronic Communications in Probability*, 22.
- 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.
- Nguyen, H. D. and McLachlan, G. J. (2017). Progress on a conjecture regarding the triangular distribution. *Communications in Statistics - Theory and Methods*, 46:11261–11271.
- Raginsky, M. and Sason, I. (2013). Concentration of measure inequalities in information theory, communications, and coding. *Foundations and Trends in Communications and Information Theory*, 10:1–246.
- Rudelson, M. and Vershynin, R. (2010). Non-asymptotic theory of random matrices: extreme singular values. In *Proceedings of the International Congress of Mathematicians 2010 (ICM 2010) (In 4 Volumes) Vol. I: Plenary Lectures and Ceremonies Vols. II–IV: Invited Lectures*, pages 1576–1602, Singapore. World Scientific.
- van Handel, R. (2014). Probability in high dimension. Technical report, Princeton University.