Optimal sub-Gaussian variance proxy for 3-mass distributions

Soufiane Atouani, Olivier Marchal, Julyan Arbel

Optimal sub-Gaussian variance proxy for 3-mass distributions

Soufiane Atouani
Université Grenoble Alpes, Inria, CNRS, Grenoble INP, LJK, 38000 Grenoble, France
soufiane.atouani@inria.fr

Olivier Marchal
Université Jean Monnet Saint-Étienne, CNRS, Institut Camille Jordan UMR 5208,
Institut Universitaire de France, Les Forges 2, 20 Rue du Dr Annino, 42000 Saint-Étienne, France
olivier.marchal@univ-st-etienne.fr

Julyan Arbel
Université Grenoble Alpes, Inria, CNRS, Grenoble INP, LJK, 38000 Grenoble, France
julyan.arbel@inria.fr

Abstract

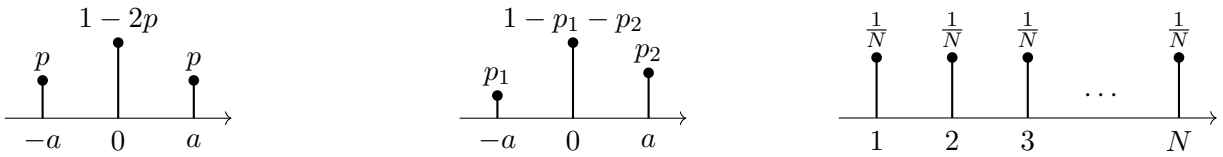
We investigate the problem of characterizing the optimal variance proxy for sub-Gaussian random variables,whose moment-generating function exhibits bounded growth at infinity. We apply a general characterization method to discrete random variables with equally spaced atoms. We thoroughly study 3-mass distributions, thereby generalizing the well-studied Bernoulli case. We also prove that the discrete uniform distribution over N points is strictly sub-Gaussian. Finally, we provide an open-source Python package that combines analytical and numerical approaches to compute optimal sub-Gaussian variance proxies across a wide range of distributions.

1 Introduction

The sub-Gaussian property, first characterized by Kahane (1960) and Buldygin and Kozachenko (1980), has become a critical tool for understanding the tail behavior of random variables. Since these pioneering works, this property has emerged as a fundamental concept in probability theory due to its profound implications in various mathematical disciplines, such as concentration inequalities (Hoeffding, 1963; Boucheron et al., 2013; Raginsky and Sason, 2013) and Bayesian statistics (Catoni, 2007). In machine learning, sub-Gaussian tails play a crucial role in bandit algorithms (Bubeck et al., 2012), in the study of the singular values of random matrices (Rudelson and Vershynin, 2010), and in Bayesian neural networks (Vladimirova et al., 2019, 2020).

Definition 1.1 (Sub-Gaussian variables)**.** *A random variable X with finite mean $\mu = \mathbb{E}[X]$ is sub-Gaussian if there exists a constant $\sigma^2 > 0$ such that $\mathbb{E}[\exp(\lambda(X-\mu))] \leq \exp(\lambda^2\sigma^2/2)$ for all $\lambda \in \mathbb{R}$. Such a constant σ^2 is called a variance proxy, and we say that X is σ^2-sub-Gaussian. The optimal variance proxy is $\sigma^2_{\text{opt}}(X) = \inf\left\{\sigma^2 > 0 \text{ such that } X \text{ is } \sigma^2\text{-sub-Gaussian}\right\}$. A variance proxy is always lower bounded by the variance, as shown by a Taylor expansion of the moment-generating function. When $\sigma^2_{\text{opt}}(X) = \text{Var}[X]$, X is called strictly sub-Gaussian.*

Extensive research on optimal variance proxy has focused on continuous distributions such as Beta and Dirichlet distributions (Marchal and Arbel, 2017), other bounded support distributions such as Kumaraswamy and triangular distributions (Arbel et al., 2020), as well as truncated Gaussian and exponential distributions (Barreto et al., 2025). Despite the prevalence of discrete distributions in modeling count data, binary outcomes, and combinatorial stochastic processes, their sub-Gaussianity remains largely underexplored. The first known



(a) Symmetric 3-mass (Section 3.1)

(b) Asymmetric 3-mass (Section 3.2)

(c) Uniform discrete (Section 4)

Figure 1: Probability mass function of the discrete distributions covered in the paper.

result on discrete distributions covers the Bernoulli distribution, the simplest discrete distribution, supported on two points, or atoms, 1 and 0, with masses p and $1-p$. Kearns and Saul (1998) derived the following "exquisitely delicate inequality", quoting Berend and Kontorovich (2013), for the Bernoulli moment-generating function:

$$(1-p)e^{-\lambda p} + pe^{\lambda(1-p)} \leq \exp\left(\frac{1-2p}{4\ln((1-p)/p)}\lambda^2\right), \quad p \in [0,1],\ \lambda \in \mathbb{R}, \tag{1}$$

which is tight and thus implies an optimal variance proxy of $\frac{1-2p}{2\ln((1-p)/p)}$. This result also provides the optimal variance proxy for the binomial distribution which is written as an i.i.d. sum of Bernoulli random variables. A natural generalization of the Bernoulli to more than two atoms is the categorical distribution with N atoms x_i, $i \in \{1, \ldots, N\}$, and weights $P(X = x_i) = p_i > 0$, with $p_1 + \cdots + p_N = 1$. While a general treatment of N-mass categorical distributions seems intractable, we address the case of the 3-mass distribution when atoms are equally spaced.

Contributions and outline. We first provide in Section 2 a characterization of the optimal sub-Gaussian variance proxy for random variables with bounded moment-generating functions, following a general methodology based on function variation analysis. This characterization yields a practical computational procedure via critical points identification and equation solving, enabling explicit computation of the optimal variance proxy (Theorem 2.2). These results are made even more precise when the number of critical points is at most two (Proposition 2.5 and Proposition 2.6). We then apply this approach with a focus on discrete distributions. We start in Section 3 with 3-point distributions, both symmetric and asymmetric, extending the classical Bernoulli case. In the symmetric setting, Theorem 3.1 uncovers a phase transition: for probabilities $p \geq \frac{1}{6}$, strict sub-Gaussianity holds, whereas for $p < \frac{1}{6}$, it does not, and we derive an explicit characterization of the optimal proxy through a pair of solvable equations. In the asymmetric case, we delineate two regimes depending on the relationship between the central mass and the edge probabilities, and in one of these, we provide a closed-form expression for the optimal proxy (Theorem 3.2). We establish in Section 4 that discrete uniform distribution over N equally spaced points is strictly sub-Gaussian for all $N \geq 2$, using a moment-based analysis of the exponential family induced by the log-partition function (Theorem 4.1). Finally, we describe in Section 5 the computational framework we developed to support reproducibility, providing an open-source Python package[1] that combines analytical and numerical approaches to compute optimal sub-Gaussian variance proxies across a wide range of distributions.

2 Characterization of optimal sub-Gaussian variance proxy

Let Y be a real random variable with finite moments and $\mu = \mathbb{E}[Y]$. Denote by $M_Y(\lambda) := \ln\left(\mathbb{E}[\exp(\lambda(Y-\mu))]\right)$ the cumulant-generating function of the centered random variable $Y - \mu$. In this pa-

[1]The package is available at https://github.com/jarbel/sub-Gaussian-implementation.git with comprehensive documentation, installation instructions, and usage examples.

per, we shall always assume that the random variable Y is such that M_Y is a smooth function. Define

$$g_Y(\lambda;\sigma^2) := \frac{1}{2}\lambda^2\sigma^2 - M_Y(\lambda) = \frac{1}{2}\lambda^2\sigma^2 - \ln\left(\mathbb{E}[\exp(\lambda(Y-\mu))]\right), \tag{2}$$

which is a smooth function of (λ,σ^2). We have $g'_Y(\lambda;\sigma^2) := \lambda\sigma^2 - M'_Y(\lambda)$ and $g''_Y(\lambda,\sigma^2) = \sigma^2 - M''_Y(\lambda)$. Observe that $g_Y(0,\sigma^2) = g'_Y(0,\sigma^2) = 0$. Moreover, for $\lambda \neq 0$, the equation $g'_Y(\lambda,\sigma^2) = 0$ is equivalent to $\sigma^2 = M'_Y(\lambda)/\lambda$. Thus, the system of equations $g_Y(\lambda;\sigma^2) = 0$ and $g'_Y(\lambda,\sigma^2) = 0$ with $\lambda \neq 0$ is equivalent to

$$\sigma^2 = \frac{1}{\lambda}M'_Y(\lambda) \text{ and } \lambda \neq 0 \text{ solution of } \lambda M'_Y(\lambda) - 2M_Y(\lambda) = 0. \tag{3}$$

Define the following sets:

$$\mathcal{L}_c^* := \left\{\lambda_c \in \mathbb{R}^* \;\middle|\; \lambda_c M'_Y(\lambda_c) - 2M_Y(\lambda_c) = 0 \text{ and } \lambda_c \text{ local minimum of } g_Y\left(.;\sigma^2 = \frac{M'_Y(\lambda_c)}{\lambda_c}\right)\right\},$$
$$\mathcal{S}_c^* := \left\{\frac{M'_Y(\lambda_c)}{\lambda_c} \;\middle|\; \lambda_c \in \mathcal{L}_c^*\right\}.$$

We complement the two previous sets by defining $\mathcal{L}_c := \mathcal{L}_c^* \cup \{0\}$, $\mathcal{S}_c := \mathcal{S}_c^* \cup \{\mathrm{Var}[Y]\}$. Let us first make the following observation.

Proposition 2.1 (Asymptotics at infinity and sufficient condition for sub-Gaussianity.)**.** *If the cumulant-generating function M_Y is a smooth function and satisfies $M_Y(\lambda) \overset{\lambda\to\pm\infty}{=} o(\lambda^2)$, then Y is* sub-Gaussian *and* $\lim_{\lambda\to\pm\infty} g_Y(\lambda;\sigma^2) = +\infty$, $\lim_{\lambda\to\pm\infty} g'_Y(\lambda;\sigma^2) = \pm\infty$, $\lim_{\lambda\to\pm\infty} g''_Y(\lambda;\sigma^2) = \sigma^2$.

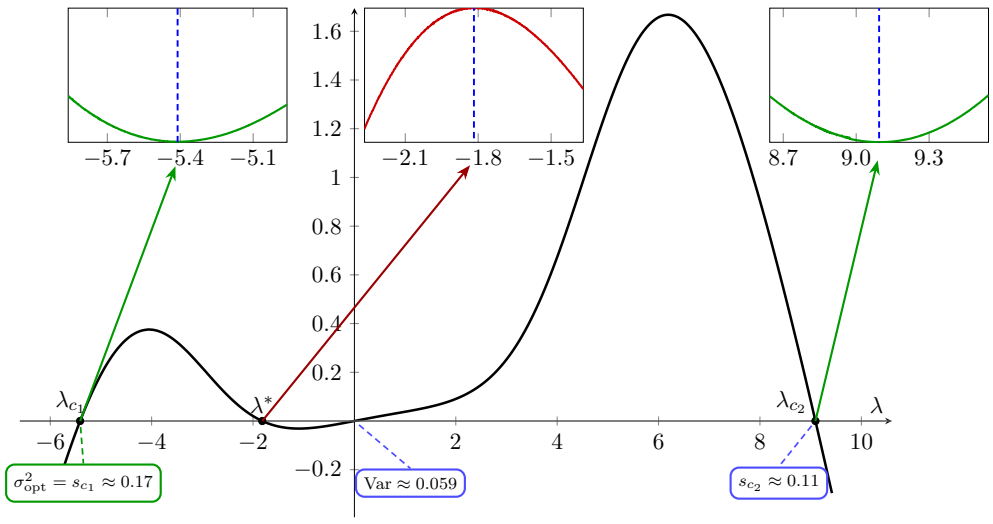


Figure 2: Illustration of Theorem 2.2 in the case of an asymmetric 3-mass distribution Y with parameters $p_1 = 0.05$ and $p_2 = 0.01$ (see Section 3.2). The black curve represents function $\lambda \mapsto \lambda M'_Y(\lambda) - 2M_Y(\lambda)$ of Equation (3). The green/red box plots represent the local behavior of $\lambda \mapsto g_Y\left(\lambda;\sigma^2 = \frac{M'_Y(\lambda^*)}{\lambda^*}\right)$ at each zero λ^* of the black curve to decide if λ^* is a local minimum of g_Y (in green) or not (in red). In this example, $\mathcal{L}_c^* = \{\lambda_{c_1} \approx -5.41, \lambda_{c_2} \approx 9.09\}$ yielding $\mathcal{S}_c^* = \{s_{c_1} \approx 0.17 \text{ and } s_{c_2} \approx 0.11\}$ while $\mathrm{Var}[Y] \approx 0.059$. Optimal variance proxy is thus $\sigma^2_{\mathrm{opt}} = s_{c_1} \approx 0.17$.

Our first main theoretical result is the following theorem that uses the sets $\mathcal{S}_c^*$ to characterize the optimal variance proxy, as illustrated in Figure 2.

Theorem 2.2 (Characterization of the optimal variance proxy.)**.** *Assume that M_Y is a smooth function and that $M_Y(\lambda) \overset{\lambda\to\pm\infty}{=} o(\lambda^2)$. Then, the optimal variance proxy is characterized by*

$$\sigma^2_{\text{opt}} = \max\{\text{Var}[Y],\ \sup \mathcal{S}^*_c\}.$$

Remark 2.3. *The main advantage of the characterization of the optimal variance proxy by Theorem 2.2 is that it provides a numerical way to obtain the optimal variance proxy in practice. Indeed, in order to determine it, one can numerically solve for the equation $\lambda M'_Y(\lambda) - 2M_Y(\lambda) = 0$. When numerical solutions are found, one should check numerically if λ is a local minimum of $g_Y\left(.;\sigma^2 = M'_Y(\lambda)/\lambda\right)$ (a sufficient condition being that $g''_Y\left(\lambda;\sigma^2 = M'_Y(\lambda)/\lambda\right) > 0$). Collecting all solutions, one then selects the optimal variance proxy by looking at the maximal values of $M'_Y(\lambda)/\lambda$ that can easily be computed numerically. In practice, such an algorithm is particularly efficient if the number of solutions of $\lambda M'_Y(\lambda) - 2M_Y(\lambda) = 0$ is low, and if one can bound the intervals on which to look for numerical solutions.*

Remark 2.4. *In practice, for a given family of distributions, one should study the elements of $\mathcal{L}^*_c$. If $\mathcal{L}^*_c$ is non-empty, one should compare its elements with the corresponding values of s_c to select the optimal variance proxy. This strategy is particularly efficient when $\mathcal{L}^*_c$ contains very few elements or when these elements can be explicitly expressed in closed-forms.*

The main numerical and theoretical difficulty to study elements of $\mathcal{L}^*_c$ is the fact that they must be local minima of $g_Y(.;\sigma^2)$. This property is tricky to verify because the second derivative may also vanish at these points making the analysis complicated. However, when $g''_Y(.;\sigma^2)$ has very few zeros, it is possible to study its sign and thus remove this complication.

Proposition 2.5 (Case when g''_Y has at most one zero on a half-line.)**.** *Assume that M_Y is a smooth function and that $M_Y(\lambda) \overset{\lambda\to\pm\infty}{=} o(\lambda^2)$. Assume that $g''_Y(.;\sigma^2)$ has no or one zero on $\mathbb{R}^*_+$ for some $\sigma^2 \geq \text{Var}[Y]$, then $g_Y(.;\sigma^2)$ is positive on $\mathbb{R}^*_+$. A similar result holds for $\mathbb{R}^*_-$.*

Unfortunately the situation becomes more involved when $g''_Y(.;\sigma^2)$ has more than one zero on a half-line. However, we can still get information when $g''_Y(.;\sigma^2)$ has exactly two zeros on a half-line. Two cases are studied in Theorem A.1 and Theorem A.2 provided in Appendix A, which can be put together to obtain the following proposition.

Proposition 2.6 (Case when g''_Y has exactly two zeros on a half-line.)**.** *Assume that M_Y is a smooth function and that $M_Y(\lambda) \overset{\lambda\to\pm\infty}{=} o(\lambda^2)$. Also assume that for any $\sigma^2 > 0$, $g''_Y(.;\sigma^2)$ has exactly two zeros $(\lambda_1(\sigma), \lambda_2(\sigma))$ such that $0 < \lambda_1(\sigma) < \lambda_2(\sigma)$. Then, with a similar result on $\mathbb{R}_-$:*

- *The equations $g_Y(\lambda,\sigma^2) = 0 = g'_Y(\lambda,\sigma^2)$ have at most one solution on $\mathbb{R}^*_+ \times \mathbb{R}^*_+$. When it exists, this unique solution (λ_0,σ^2_c) always satisfies $\sigma^2_c > \text{Var}[Y]$.*
- *$g_Y(.;\sigma^2)$ may only become negative on $\mathbb{R}_+$ if and only if the previous set of equations has exactly one solution (λ_0,σ^2_c) and $\sigma^2 < \sigma^2_c$.*

Theorem 2.5 and Theorem 2.6 are interesting to obtain a characterization of the optimal variance proxy when $g''_Y(.;\sigma^2)$ has at most two zeros on each half-lines $\mathbb{R}^*_\pm$. Indeed, in this case the optimal proxy variance is obtained as the maximum of $\text{Var}[Y]$, $\sigma^2_{c,+}$ and $\sigma^2_{c,-}$ where $\sigma^2_{c,+}$ (resp. $\sigma^2_{c,-}$) is the unique solution, when it exists, of the equations $g_Y(\lambda,\sigma^2) = 0 = g'_Y(\lambda,\sigma^2)$ on $\mathbb{R}^*_+$ (resp. $\mathbb{R}^*_-$). As we shall see, this situation happens for the 3-mass distributions. It also includes many other standard distributions.

3 Application to 3-mass distributions

In this section, we undertake a detailed analysis of the sub-Gaussian properties of three-mass discrete distributions supported on $\{-a, 0, a\}$, see Figure 1. Our objective is to characterize the optimal variance proxy σ^2_{opt} and to delineate the regimes in which strict sub-Gaussianity holds.

In the *symmetric case* (Section 3.1), where the outer masses are equally weighted, we establish two distinct behaviors. When the parameter satisfies $p \geq \frac{1}{6}$, the distribution is strictly sub-Gaussian and the optimal variance proxy coincides with the variance, $\sigma^2_{\text{opt}} = \text{Var}[X] = 2p$. In contrast, for $p < \frac{1}{6}$, strict sub-Gaussianity fails, and the optimal variance proxy is determined by a critical parameter $\lambda_c > 0$ solving a coupled system of equations (Theorem 3.1).

In the *asymmetric case* (Section 3.2), where the mass probabilities at $-a$ and a are not equal, the situation is more intricate. If the central mass satisfies $p_3 \leq 4\sqrt{p_1 p_2}$, then an explicit closed-form expression is available (Theorem 3.2 and Theorem 3.3), $\sigma^2_{\text{opt}} = 2(p_2 - p_1)/\ln(p_2/p_1)$. When $p_3 > 4\sqrt{p_1 p_2}$, the characterization of σ^2_{opt} requires the analysis of a nonlinear equation whose solution yields the critical value determining the transition between variance proxies (Theorem 3.5).

3.1 Symmetric 3-mass distribution

Let X be a discrete random variable on the set $\{-a, 0, a\}$, $a > 0$ and $\mathbb{P}(X = -a) = p\,, \mathbb{P}(X = 0) = 1 - 2p\,, \mathbb{P}(X = a) = p$, where $p \in \left(0, \frac{1}{2}\right)$, see Appendix 1. We may define $Y = \frac{1}{a}X$ and use the fact that $\sigma_{\text{opt}}[Y] = \frac{1}{a}\,\sigma_{\text{opt}}[X]$. Thus, we may restrict to $a = 1$. We have $\mu := \mathbb{E}[Y] = 0\ ,\ \sigma^2 := \text{Var}[Y] = 2p$. In the special case where $p = \frac{1}{2}$ the random variable X reduces to the symmetric Rademacher distribution, which is known to be strictly sub-Gaussian. For any $\sigma > 0$, we have that σ^2 is a variance proxy of Y if and only if

$$\mathbb{E}\left[e^{\lambda(Y-\mu)}\right] = pe^{\lambda} + pe^{-\lambda} + 1 - 2p \leq e^{\frac{\lambda^2\sigma^2}{2}}\ ,\ \forall\, \lambda \in \mathbb{R}.$$

This inequality is equivalent to

$$g_{\sigma,p}(\lambda) := \frac{\lambda^2\sigma^2}{2} - \ln(2p\cosh\lambda + 1 - 2p) \geq 0\ ,\ \forall\, \lambda \in \mathbb{R}.$$

Since $g_{\sigma,p}$ is an even function of λ, we only need to prove the former inequality for $\lambda \geq 0$. The general theory implies that $\text{Var}[Y] = 2p$ is a lower bound for the optimal variance proxy. Consequently, we shall only consider $\sigma^2 \geq 2p$.

The function $g_{\sigma,p}$ is obviously a smooth function of (λ, σ, p) and we have

$$\begin{aligned} g'_{\sigma,p}(\lambda) &= \lambda\sigma^2 - \frac{2p\sinh\lambda}{2p\cosh\lambda + 1 - 2p}\ ,\ g^{(2)}_{\sigma,p}(\lambda) = \sigma^2 + \frac{2p\,(2p\cosh\lambda - \cosh\lambda - 2p))}{(2p\cosh\lambda + 1 - 2p)^2} \\ g^{(3)}_{\sigma,p}(\lambda) &= \frac{2p\sinh\lambda\left(4p^2 + 4p - 1 + 2p(1-2p)\cosh\lambda\right)}{(2p\cosh\lambda + 1 - 2p)^3}. \end{aligned}$$

Let us now observe that $\forall\, \lambda \in \mathbb{R}$, $N_p(\lambda) := 4p^2 + 4p - 1 + 2p(1-2p)\cosh\lambda \geq 1$, and $g^{(3)}_{\sigma,p}(0) = 0$. Thus, $g^{(3)}_{\sigma,p}$ and N_p have the same sign on $\mathbb{R}_+$. Since $N'_p(\lambda) = 2p(1-2p)\sinh\lambda$, we get that N_p is a strictly increasing function on $\mathbb{R}_+$. Since $N_p(0) = 6p - 1$, we get two distinct cases.

First, $p \geq \frac{1}{6}$: In this case, N_p is a positive function on $\mathbb{R}_+$ and so is $g^{(3)}_{\sigma,p}$. It follows that $g^{(2)}_{\sigma,p}$ is a strictly increasing function on $\mathbb{R}_+$. Since $g^{(2)}_{\sigma,p}(0) = \sigma^2 - 2p \geq 0$, we conclude that $g^{(2)}_{\sigma,p}$ is positive on $\mathbb{R}_+$. Thus $g'_{\sigma,p}$ is a strictly increasing function on $\mathbb{R}_+$. Furthermore, since $g'_{\sigma,p}(0) = 0$ we end up with the fact that $g'_{\sigma,p}$ is a positive function on $\mathbb{R}_+$ and therefore $g_{\sigma,p}$ is an increasing function on $\mathbb{R}_+$. Finally, since $g_{\sigma,p}(0) = 0$ we conclude that $g_{\sigma,p}$ is positive on $\mathbb{R}_+$ so that σ is a variance proxy. This argument is valid for any $\sigma^2 \geq \text{Var}[Y] = 2p$ so that $\sigma^2_{\text{opt}}[Y] = \text{Var}[Y] = 2p$, i.e. Y is strictly sub-Gaussian.

Second, $p < \frac{1}{6}$: In this case, $N_p(0) < 0$, and N_p is increasing on $\mathbb{R}_+$, tending to $+\infty$ when $\lambda \to +\infty$. Thus, since N_p is a smooth function, there exists a unique $\lambda_0 = \operatorname{arcosh}\left(\frac{1-4p-4p^2}{2p(1-2p)}\right) > 0$ such that $N_p(\lambda_0) = 0$. Moreover N_p is strictly negative on $(0, \lambda_0)$ and strictly positive on $(\lambda_0, +\infty)$. These properties immediately extend to $g^{(3)}_{\sigma,p}$. Consequently, $g^{(2)}_{\sigma,p}$ is strictly decreasing on $(0, \lambda_0)$ and strictly increasing on $(\lambda_0, +\infty)$. In addition, we have $g^{(2)}_{\sigma,p}(0) = \sigma^2 - 2p \geq 0$ and $g^{(2)}_{\sigma,p}(\lambda_0) = \sigma^2 - \frac{(1-2p)^2}{4(1-4p)}$ and $g^{(2)}_{\sigma,p}(+\infty) = \sigma^2 > 0$. Note that if $\sigma^2 \geq \frac{(1-2p)^2}{4(1-4p)}$ then $g^{(2)}_{\sigma,p}$ is positive on $\mathbb{R}_+$ and then it is straightforward to prove that σ is a variance proxy using the same final steps as the case $p \geq \frac{1}{6}$. Thus, we obtain the following upper bound for the optimal variance proxy of Y:

$$\sigma_1^2(p) := \frac{(1-2p)^2}{4(1-4p)}. \tag{4}$$

Consequently $\sigma^2 = 2p$ is no longer a variance proxy because $g'_{\sigma,p}$ would become strictly decreasing on $(0, \lambda_0)$ and thus strictly negative on this interval (because $g'_{\sigma,p}(0) = 0$) and so $g_{\sigma,p}$ would be negative on $(0, \lambda_0)$ (again because $g_{\sigma,p}(0) = 0$). This proves that for $p < \frac{1}{6}$, Y is not strictly sub-Gaussian.

Let us be more precise in the case $p < \frac{1}{6}$. As mentioned above, we shall now only consider $2p < \sigma^2 < \frac{(1-2p)^2}{4(1-4p)}$. The previous analysis implies that there exists a unique pair (λ_1, λ_2) such that $0 < \lambda_1 < \lambda_0 < \lambda_2$ and $g^{(2)}_{\sigma,p}(\lambda_1) = g^{(2)}_{\sigma,p}(\lambda_2) = 0$. Moreover, $g^{(2)}_{\sigma,p}$ is strictly positive on $(0, \lambda_1) \cup (\lambda_2, +\infty)$ and strictly negative on (λ_1, λ_2). Note that we have explicitly:

$$\lambda_1(\sigma) := \operatorname{arcosh}\left(\frac{(1-2p)(1-2\sigma^2) - \sqrt{(1-2p)^2 - 4(1-4p)\sigma^2}}{4p\sigma^2}\right)$$
$$\lambda_2(\sigma) := \operatorname{arcosh}\left(\frac{(1-2p)(1-2\sigma^2) + \sqrt{(1-2p)^2 - 4(1-4p)\sigma^2}}{4p\sigma^2}\right).$$

This implies that $g'_{\sigma,p}$ is increasing on $(0, \lambda_1)$, then decreasing on (λ_1, λ_2) and finally increasing on $(\lambda_2, +\infty)$. Since $g'_{\sigma,p}(0) = 0$ and $g'_{\sigma,p}(+\infty) = +\infty$, the sign of $g'_{\sigma,p}$ is determined by the sign of $g'_{\sigma,p}(\lambda_2)$. Note also that a straightforward computation implies that $g'_{\sigma,p}(\lambda_0) = \sigma^2\lambda_0\sqrt{(1-6p)(1+2p)(1-4p)} > 0$. There are only two cases that we now study.

Case when $g'_{\sigma,p}(\lambda_2) \geq 0$: then $g'_{\sigma,p}$ is positive on $\mathbb{R}_+$ so that $g_{\sigma,p}$ is positive on $\mathbb{R}_+$ and thus σ is a variance proxy. Thus an upper bound is σ_2^2 that is the unique solution in $\left(2p, \frac{(1-2p)^2}{4(1-4p)}\right)$ of

$$\sigma_2^2(p) = \frac{2p\sinh\lambda_2(\sigma_2(p))}{\lambda_2(\sigma_2(p))(1 + 2p\cosh\lambda_2(\sigma_2(p)) - 2p)}. \tag{5}$$

Case when $g'_{\sigma,p}(\lambda_2) < 0$: then there exists a unique pair (λ_3, λ_4) such that $\lambda_0 < \lambda_3 < \lambda_2 < \lambda_4$ and $g'_{\sigma,p}(\lambda_3) = g'_{\sigma,p}(\lambda_4) = 0$. Moreover, $g'_{\sigma,p}$ is positive on $(0, \lambda_3) \cup (\lambda_4, +\infty)$ and negative on (λ_3, λ_4). This implies that $g_{\sigma,p}$ increases on $(0, \lambda_3)$, then decreases on (λ_3, λ_4) and finally increases on $(\lambda_4, +\infty)$. In particular, since $g_{\sigma,p}(0) = 0$ and $g_{\sigma,p}(+\infty) = +\infty$, $g_{\sigma,p}$ has only one local maximum λ_3 and one local minimum λ_4 on $(0, +\infty)$. Moreover, these local extremum satisfy $0 < \lambda_3 < \lambda_0 < \lambda_4$ and $g_{\sigma,p}(\lambda_3) > 0$. Eventually the sign of $g_{\sigma,p}(\lambda_4)$ determines if σ is a variance proxy or not. Since $g_{\sigma,p}(\lambda_4)$ is a smooth function of σ, the critical case corresponding to σ_{opt} may happen only when $g_{\sigma_{\mathrm{opt}},p}(\lambda_4) = 0$. Since we know that this critical case is achieved on $\left(2p, \frac{1-4p+4p^2}{4(1-4p)}\right)$ we conclude with the following statement.

Theorem 3.1 (Optimal variance proxy for symmetric 3-mass distribution.)**.** *Let X be a discrete random variable on the set $\{-a, 0, a\}$ where $a > 0$ and $\mathbb{P}(X = -a) = p$, $\mathbb{P}(X = 0) = 1 - 2p$, $\mathbb{P}(X = a) = p$ where $p \in \left(0, \frac{1}{2}\right)$. Then we have two regimes:*

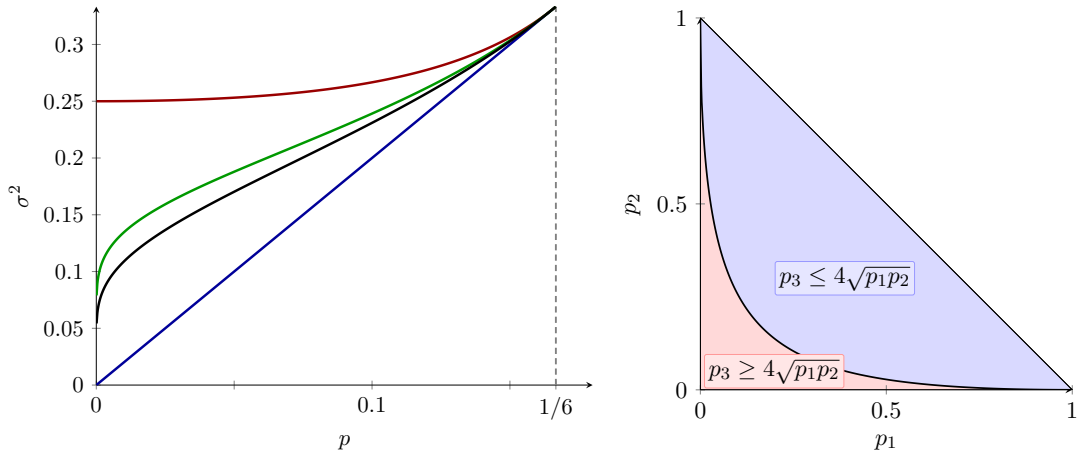


Figure 3: Left: Symmetric 3-mass case (Section 3.1). Black: Optimal variance proxy σ^2_{opt} for $0 < p < \frac{1}{6}$. Blue: variance $\text{Var}[Y] = 2p$. Red: upper bound $\sigma_1^2(p)$ of Equation (4). Green: upper bound $\sigma_2^2(p)$ of Equation (5). Right: Asymmetric 3-mass case (Section 3.2). Two different regimes depending on the relative weight of the intermediate mass.

(i) Strictly sub-Gaussian regime. *If $p \in \left[\frac{1}{6}, \frac{1}{2}\right)$, then X is strictly sub-Gaussian, i.e.,*

$$\sigma^2_{\text{opt}} = \text{Var}[X] = 2p.$$

(ii) Non-strictly sub-Gaussian regime. *If $p \in (0, \frac{1}{6})$, then X is not strictly sub-Gaussian. In this case, the optimal variance proxy is characterized by the system*

$$\begin{cases} 0 = g_{\sigma_{\text{opt}},p}(\lambda_c) = \frac{\lambda_c^2 \sigma^2_{\text{opt}}}{2} - \ln\big(2p\cosh\lambda_c + 1 - 2p\big), \\ 0 = g'_{\sigma_{\text{opt}},p}(\lambda_c) = \lambda_c \sigma^2_{\text{opt}} - \frac{2p\sinh\lambda_c}{2p\cosh\lambda_c + 1 - 2p}, \end{cases}$$

where $\lambda_c \in (\lambda_0, +\infty)$ is the unique solution with

$$\lambda_0 = \text{arcosh}\left(\frac{1 - 4p - 4p^2}{2p(1-2p)}\right) > 0.$$

Equivalently, the optimal variance proxy of the non-strictly sub-Gaussian regime admits the closed-form

$$\sigma^2_{\text{opt}} = \frac{2p\,\sinh(\lambda_c)}{\lambda_c\left(2p\cosh(\lambda_c) + 1 - 2p\right)},$$

where λ_c is the unique solution of

$$p\,\lambda_c\,\sinh(\lambda_c) - \big(1 - 2p + 2p\cosh(\lambda_c)\big)\ln\big(1 - 2p + 2p\cosh(\lambda_c)\big) = 0.$$

In this regime, $\text{Var}[X] = 2p < \sigma^2_{\text{opt}} < \sigma_1^2 = \frac{(1-2p)^2}{4(1-4p)}$.

3.2 Asymmetric 3-mass distribution

Let X be a random variable supported on $\{-a, 0, a\}$, $a > 0$, and $\mathbb{P}(X = -a) = p_1$, $\mathbb{P}(X = 0) = p_3 = 1 - p_1 - p_2$, $\mathbb{P}(X = a) = p_2$, where $p_1, p_2, p_3 \in (0, 1)$, see Appendix 1. As before, we may define $Y = \frac{1}{a}X$ and

restrict to $a = 1$. We also may assume $p_2 \geq p_1$ without loss of generality. We have $\mu := \mathbb{E}[Y] = -p_1 + p_2 \geq 0$ and $\mathrm{Var}[Y] = p_1 + p_2 - (-p_1 + p_2)^2$. We have that σ^2 is a variance proxy of Y if and only if

$$\mathbb{E}\left[e^{\lambda Y}\right] = p_1 e^{-\lambda} + p_2 e^{\lambda} + 1 - p_1 - p_2 \leq e^{(\frac{\lambda^2\sigma^2}{2} + \lambda\mu)} \ , \ \forall\, \lambda \in \mathbb{R}.$$

This inequality is equivalent to

$$g_{\sigma,p_1,p_2}(\lambda) := \frac{\lambda^2\sigma^2}{2} - \ln(u_{0;p_1,p_2}(\lambda)) + \lambda\mu \geq 0 \ , \ \forall\, \lambda \in \mathbb{R}, \tag{6}$$

with $u_{0;p_1,p_2}(\lambda) := p_1 e^{-\lambda} + p_2 e^{\lambda} + 1 - p_1 - p_2$ a smooth function and $u_{0;p_1,p_2}(\lambda) > 0$, $\forall \lambda \in \mathbb{R}$. For convenience, we drop the dependence in (p_1, p_2) in the notation, as per $u_0(\lambda)$, and we define $u_i(\lambda) = u_{i-1}'(\lambda),\ i \in \{1, 2, 3\}$, $\forall\, \lambda \in \mathbb{R}$. We have $u_1(\lambda) = -p_1 e^{-\lambda} + p_2 e^{\lambda}$ and we observe that $u_2(\lambda) = u_0(\lambda) - p_3$, $u_3(\lambda) = u_1(\lambda)$ and $u_1^2(\lambda) = (u_0(\lambda) - p_3)^2 - 4p_1p_2$ for all $\lambda \in \mathbb{R}$. The general theory implies that $\mathrm{Var}[Y] = p_1 + p_2 - (-p_1 + p_2)^2$ is a lower bound for the optimal variance proxy, thus we shall only consider larger or equal values for σ^2.

The function g_{σ,p_1,p_2} is a smooth function of $(\lambda, \sigma, p_1, p_2)$ and its first derivatives can be expressed as:

$$\begin{aligned} g'_{\sigma,p_1,p_2}(\lambda) &= \lambda\sigma^2 - \frac{u_1(\lambda)}{u_0(\lambda)} + \mu \ , \ g^{(2)}_{\sigma,p_1,p_2}(\lambda) = \sigma^2 - \frac{u_0(\lambda)p_3 - p_3^2 + 4p_1p_2}{u_0(\lambda)^2} \\ g^{(3)}_{\sigma,p_1,p_2}(\lambda) &= u_1(\lambda)\frac{p_3u_0(\lambda) + 8p_1p_2 - 2p_3^2}{u_0(\lambda)^3}. \end{aligned} \tag{7}$$

Note in particular that $g^{(3)}_{\sigma,p_1,p_2}$ does not depend on σ. Moreover, $g^{(3)}_{\sigma,p_1,p_2}$ can be written as

$$g^{(3)}_{\sigma,p_1,p_2}(\lambda) = \frac{u_1(\lambda)N_{p_1,p_2}(\lambda)}{u_0(\lambda)^3}, \ \text{where } N_{p_1,p_2}(\lambda) := p_3u_0(\lambda) + 8p_1p_2 - 2p_3^2. \tag{8}$$

Observe that $u_0(\lambda)$ is strictly positive on $\mathbb{R}$. Therefore, the function $g^{(3)}_{\sigma,p_1,p_2}$ and $u_1N_{p_1,p_2}$ share the same sign. We know that u_1 changes sign at $\lambda_0 := -\frac{1}{2}\ln(\frac{p_2}{p_1})$ assuming that $p_2 \geq p_1$ we have $\lambda_0 \leq 0$. Hence, to determine the sign of $g^{(3)}_{\sigma,p_1,p_2}$, it remains to analyze the sign of N_{p_1,p_2}. For this purpose, it is sufficient to study the sign of the polynomial

$$P(X) := p_2p_3X^2 + (8p_1p_2 - p_3^2)X + p_1p_3, \text{where } X = e^{\lambda} > 0. \tag{9}$$

To determine the sign of $P(X)$, we compute its discriminant:

$$\Delta = (p_3^2)^2 - 20p_1p_2(p_3)^2 + 64p_1^2p_2^2 = (p_3^2 - 4p_1p_2)(p_3^2 - 16p_1p_2),$$

and it gives two different regimes, separated as illustrated on the right panel of Figure 3, that we shall study separately.

3.2.1 Case $p_3 \leq 4\sqrt{p_1p_2}$

In this case, we have the following theorem.

Theorem 3.2 (Optimal variance proxy for $p_3 \leq 4\sqrt{p_1p_2}$, closed-form expression.)**.** *When $p_3 \leq 4\sqrt{p_1p_2}$, the optimal variance proxy is given by*

$$\sigma^2_{\mathrm{opt}} = \frac{2(p_2 - p_1)}{\ln p_2 - \ln p_1}.$$

The proof is deferred to the Appendix. Theorem 3.2 implies the following corollary.

Corollary 3.3. *For $p_3 \leq 4\sqrt{p_1p_2}$, the random variable Y is strictly sub-Gaussian if and only if $p_1 = p_2 = p$. In this symmetric case, the condition $p_3 \leq 4\sqrt{p_1p_2}$ is equivalent to $p \geq \frac{1}{6}$.*

Proof. It is obvious from the fact that in this case $\sigma^2_{\mathrm{opt}} = \frac{2(p_2-p_1)}{\ln(p_2/p_1)} \geq \mathrm{Var}[Y] = p_1 + p_2 - (p_2 - p_1)^2$ with equality if and only if $p_1 = p_2$. □

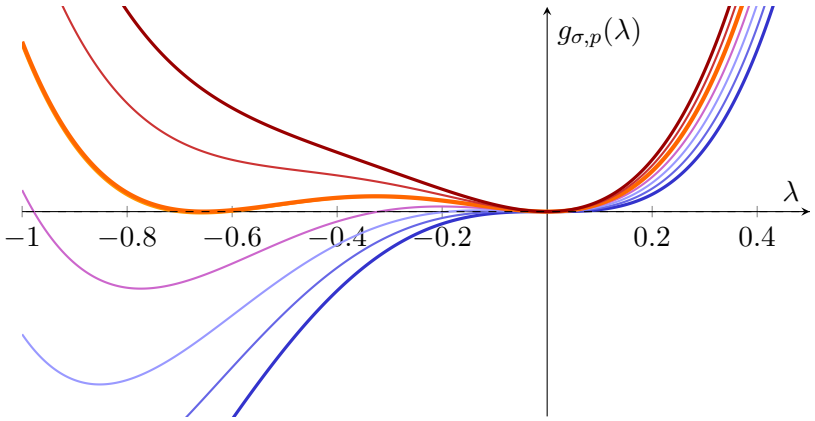


Figure 4: Functions $\lambda \mapsto g_{\sigma,p_1,p_2}(\lambda)$ of Equation (6) for $(p_1, p_2) = (0.13, 0.25)$, with σ^2 varying from the upper bound $\frac{2\sqrt{p_1p_2}}{p_3+2\sqrt{p_1p_2}}$ in dark red down to the variance $\mathrm{Var}[Y]$ in blue (the function g then becomes locally negative around $\lambda = 0_-$), while the orange curve stands for the optimal proxy variance $\sigma^2_{\mathrm{opt}} = \frac{2(p_2-p_1)}{\ln p_2 - \ln p_1}$. The intermediate curves illustrate the progressive transition from convex behavior around 0 to oscillating behavior.

3.2.2 Case $p_3 > 4\sqrt{p_1p_2}$

In this case, the discriminant Δ defined in Equation (3.2) is positive and the polynomial P of Equation (9) has two distinct positive roots so that N_{p_1,p_2} defined in Equation (8) has exactly two roots. The following lemma provides the location of the latter roots.

Lemma 3.4. *Let $p_1, p_2 \in (0,1)$ such that $p_2 \geq p_1$ and $(1-p_1-p_2)^2 > 16p_1p_2$. Then N_{p_1,p_2} defined in Equation (8) has exactly two roots which are* positive.

We now analyze the sign of the function $g_{\sigma,p_1,p_2}(.;\sigma^2)$ separately on $\mathbb{R}^*_-$ and $\mathbb{R}^*_+$. Let us first observe that the discussion made for $\mathbb{R}^*_-$ in the proof of Theorem 3.2 is still valid. Thus, $g_{\sigma,p_1,p_2}(.;\sigma^2)$ is non-negative on $\mathbb{R}^*_-$ if and only if $\sigma^2 \geq \frac{2(p_2-p_1)}{\ln(p_2/p_1)}$ (see Figure 4). Let us now discuss the situation on $\mathbb{R}^*_+$. From Theorem 3.4 and (8) we get that $g^{(3)}$ is negative on $(-\infty, \lambda_0) \cup (\lambda_-(\sigma), \lambda_+(\sigma))$ and positive on $(\lambda_0, \lambda_-(\sigma)) \cup (\lambda_+(\sigma), +\infty)$. Consequently, $g^{(2)}_{\sigma,p_1,p_2}(.;\sigma^2)$ is increasing on $(0, \lambda_-)$ and since $g^{(2)}_{\sigma,p_1,p_2}(0;\sigma^2) = 0$ is it thus positive on this interval. Then, $g^{(2)}_{\sigma,p_1,p_2}(.;\sigma^2)$ is decreasing on $(\lambda_-(\sigma), \lambda_+(\sigma))$ and then increasing on $(\lambda_+(\sigma), +\infty)$ with $g^{(2)}_{\sigma,p_1,p_2}(+\infty;\sigma^2) = \sigma^2 > 0$. Thus, there are only two possible cases: either $g^{(2)}_{\sigma,p_1,p_2}(\lambda_+(\sigma);\sigma^2) \geq 0$ and then $g_{\sigma,p_1,p_2}(.;\sigma^2)$ is strictly convex and positive on $\mathbb{R}^*_+$ or $g^{(2)}_{\sigma,p_1,p_2}(\lambda_+(\sigma);\sigma^2) < 0$ in which case $g^{(2)}_{\sigma,p_1,p_2}(.;\sigma^2)$ has exactly two zeros $(\lambda_1(\sigma), \lambda_2(\sigma))$ on $\mathbb{R}^*_+$ such that $\lambda_-(\sigma) < \lambda_1(\sigma) < \lambda_+(\sigma) < \lambda_2(\sigma)$. We may thus apply Theorem 2.6 whose conclusion depends on the existence of a solution $(\lambda, \sigma^2) \in (\lambda_-, +\infty) \times (\mathrm{Var}[Y], +\infty)$ of the equation $g_{\sigma,p_1,p_2}(\lambda;\sigma^2) = 0 = g'_{\sigma,p_1,p_2}(\lambda;\sigma^2)$. This set of equations is equivalent to

$$\sigma^2 = \frac{1}{\lambda}\left(\frac{u_1(\lambda)}{u_0(\lambda)} - \mu\right) = \frac{1}{\lambda}\left(p_1 - p_2 + \frac{p_2e^{\lambda} - p_1e^{-\lambda}}{p_1e^{-\lambda} + p_2e^{\lambda} + 1 - p_1 - p_2}\right),$$

and λ positive solution of

$$\lambda\frac{u_1(\lambda)}{u_0(\lambda)} - 2\ln(u_0(\lambda)) + \lambda\mu = 0,$$

i.e.

$$F(\lambda) := \lambda u_1(\lambda) - 2u_0(\lambda)\ln u_0(\lambda) + \lambda u_0(\lambda)(p_2 - p_1) = 0. \tag{10}$$

From Theorem 2.6, $F(\lambda) = 0$ has no solution on $\mathbb{R}^*_+$ such that $\sigma^2 = \frac{1}{\lambda}\left(\frac{u_1(\lambda)}{u_0(\lambda)} - \mu\right) < \mathrm{Var}[Y]$. Thus, we have two cases detailed in the following theorem.

Theorem 3.5 (Optimal variance proxy for $p_3 > 4\sqrt{p_1 p_2}$.)**.** *When $p_3 > 4\sqrt{p_1 p_2}$, the optimal variance proxy depends on the zero of F in Equation (10):*

- *If F has a positive zero $\lambda_c > 0$ then Theorem 2.6 implies that this zero is unique on $\mathbb{R}_+^*$ and we get that $\sigma_c^2 = \frac{1}{\lambda_c}\left(\frac{u_1(\lambda_c)}{u_0(\lambda_c)} - \mu\right) \geq \mathrm{Var}[Y]$. Theorem 2.2 implies that the optimal variance proxy is given by*

$$\sigma_{\mathrm{opt}}^2 = \max\left(\frac{2(p_2-p_1)}{\ln(p_2/p_1)}, \frac{1}{\lambda_c}\left(\frac{u_1(\lambda_c)}{u_0(\lambda_c)} - (p_2-p_1)\right)\right).$$

- *If F has no zero on $\mathbb{R}_+^*$ then $g_{\sigma,p_1,p_2}(.;\sigma^2)$ is positive on $\mathbb{R}_+^*$ and thus the optimal variance proxy is given by $\sigma_{opt}^2 = \frac{2(p_2-p_1)}{\ln(p_2/p_1)}$.*

Remark 3.6. *Note that $\lim_{\lambda\to+\infty} F(\lambda) = -\infty$ and $F(\lambda) = \frac{1}{6}M_3[Y]\lambda^3 + \frac{1}{4}\left(\frac{1}{3}M_4[Y] - \mathrm{Var}[Y]^2\right)\lambda^4 + O(\lambda^5)$, where $M_3[Y] := \mathbb{E}[(Y-\mu)^3]$ and $M_4[Y] := \mathbb{E}[(Y-\mu)^4]$. Thus, a sufficient condition for F to admit a positive zero is that $M_3[Y] > 0$ or $\left[M_3[Y] = 0 \text{ and } \frac{1}{3}M_4[Y]\,\mathrm{Var}[Y]^2 > 0\right]$. In particular for $p_1 = p_2 = p$, we have $F(\lambda) = \frac{1}{6}p(1-6p)\lambda^4 + O(\lambda^5)$ so that for $p > \frac{1}{6}$ (which is the equivalent condition to $p_3 > 4\sqrt{p_1 p_2}$ in this case) F always has a unique positive zero λ_c and the corresponding variance $\sigma_c^2 = \frac{1}{\lambda_c}\frac{u_1(\lambda_c)}{u_0(\lambda_c)}$ is always greater than $2p = \mathrm{Var}[Y]$ and hence the optimal variance proxy is always $\sigma_{\mathrm{opt}}^2 = \sigma_c^2 = \frac{1}{\lambda_c}\frac{u_1(\lambda_c)}{u_0(\lambda_c)}$ in the symmetric case $p_1 = p_2 = p$.*

4 Optimal variance proxy for the discrete uniform distribution

In this section, we briefly present the discrete, equally spaced uniform distribution, also known as the comb distribution, see Appendix 1. Our main result establishes that this law is strictly sub-Gaussian, with an optimal variance proxy that coincides with its variance.

Theorem 4.1 (Optimal variance proxy for the uniform discrete distribution.)**.** *Let X be uniformly distributed on $\{ka + b\,,\, k \in [\![1, N]\!]\}$ with $N > 1$, $a \in \mathbb{R}\setminus\{0\}$ and $b \in \mathbb{R}$. Then X is strictly sub-Gaussian, i.e. its optimal variance proxy equals its variance:*

$$\sigma_{\mathrm{opt}}^2[X] = \mathrm{Var}[X] = a^2\frac{N^2-1}{12}.$$

The proof of Theorem 4.1 is postponed to Appendix C.

5 Software implementation

To support reproducibility and facilitate the use of both state-of-the-art and our theoretical results, we developed a Python package [1] to compute the optimal sub-Gaussian variance proxy for a wide range of probability distributions. The package handles both discrete and continuous distributions, including truncated Gaussian and Exponential laws.

The implementation follows a unified principle. Whenever a closed-form expression is available, it is returned directly; this is the case for Bernoulli and Binomial distributions, uniform and discrete uniform distributions (see Theorem 4.1), as well as certain symmetric families such as Beta, Kumaraswamy, Triangular (see Arbel et al., 2020), or 3-mass distributions (see Theorem 3.1). When no closed-form can be derived, the computation relies on the general characterizations of the optimal variance proxy given in Section 2 (see also (Arbel et al., 2020)), and proceeds via numerical methods. In particular, an adaptive grid search is combined with robust root-finding algorithms such as Brent's method. This hybrid methodology ensures that both analytically tractable and intractable cases are encompassed within a single coherent framework.

The package primarily implements the characterization of the optimal variance proxy based on function g_Y defined in Equation (2) and based on the cumulant-generating function of $Y - \mu$. Computing σ^2_{opt} then reduces to solving the coupled system

$$g_Y(\lambda; \sigma^2) = 0, \qquad g'_Y(\lambda; \sigma^2) = 0,$$

as stated in Theorem 3.5. This characterization is particularly effective for discrete distributions with few support points, such as the 3-mass distribution, where it provides a tractable criterion for identifying candidate values of the variance proxy.

We also used this implementation to validate our theoretical results: the specialized method for the symmetric 3-mass case (see Theorem 3.1) was checked against the asymmetric 3-mass case (see Theorem 3.5) for the special scenario where $p_1 = p_2$, and both approaches produced identical results. This provides an additional consistency check between the theoretical framework and the numerical implementation.

The package provides utilities for visualization of objective functions and supports batch analysis across multiple parameter settings, making it a practical companion for applied research in Bayesian inference, variational methods, and concentration bounds.

6 Discussion

In this work, we advance the understanding of the sub-Gaussian property for discrete distributions by deriving the optimal sub-Gaussian variance proxy for certain 3-mass distributions, in particular those with equally spaced support such as $\{-1, 0, 1\}$. We further extend the analysis to the uniform discrete distribution on $\{1, \ldots, N\}$ with equally spaced support.

Generalizing beyond these settings to distributions with non-equidistant support or with more than 3 mass points, appears essentially intractable in full generality for N-mass categorical laws. Nonetheless, other discrete families remain of substantial interest. In Bayesian nonparametrics, for instance, one often encounters discrete distributions, e.g. those arising from the Dirichlet process (Ferguson, 1973), the Pitman–Yor process (Pitman and Yor, 1997), or Gibbs-type processes (De Blasi et al., 2015). While concentration properties for such processes have been studied in the context of large deviations (e.g., Doss and Sellke, 1982; Feng, 2007), a more refined analysis of their tails via optimal variance proxies represents a promising direction for future research.

Acknowledgment

Olivier Marchal used part of his IUF junior grant G752IUFMAR for this research, and Julyan Arbel was partially supported by ANR-21-JSTM-0001 grant.

References

Julyan Arbel, Olivier Marchal, and Hien D Nguyen. On strict sub-Gaussianity, optimal proxy variance and symmetry for bounded random variables. *ESAIM: P-S*, 24:39–55, 2020.

Mathias Barreto, Olivier Marchal, and Julyan Arbel. Optimal sub-Gaussian variance proxy for truncated Gaussian and exponential random variables. *Statistics and Probability Letters*, 2025.

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

Stéphane Boucheron, Gábor Lugosi, and Pascal Massart. *Concentration Inequalities: A Nonasymptotic Theory of Independence*. Oxford University Press, 2013.

Sébastien Bubeck, Nicolo Cesa-Bianchi, et al. Regret analysis of stochastic and nonstochastic multi-armed bandit problems. *Foundations and Trends® in Machine Learning*, 5(1):1–122, 2012.

Valerii Buldygin and Yu Kozachenko. Sub-Gaussian random variables. *Ukrainian Mathematical Journal*, 32: 483–489, 1980.

Olivier Catoni. *PAC-Bayesian Supervised Classification: The Thermodynamics of Statistical Learning*, volume 56 of *IMS Lecture Notes*. IMS, 2007.

Pierpaolo De Blasi, Stefano Favaro, Antonio Lijoi, Ramsés H Mena, Igor Prünster, and Matteo Ruggiero. Are Gibbs-type priors the most natural generalization of the Dirichlet process? *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 37(2):212–229, 2015.

Hani Doss and Thomas Sellke. The tails of probabilities chosen from a Dirichlet prior. *The Annals of Statistics*, 10(4):1302–1305, 1982.

Shui Feng. Large deviations for dirichlet processes and poisson-dirichlet distribution with two parameters. *Electron. J. Probab.*, 2007.

T.S. Ferguson. A Bayesian analysis of some nonparametric problems. *The Annals of Statistics*, 1(2):209–230, 1973. ISSN 0090-5364.

Wassily Hoeffding. Probability Inequalities for Sums of Bounded Random Variables. *Journal of the American Statistical Association*, 58(301):13–30, 1963.

Jean-Pierre Kahane. Propriétés locales des fonctions à séries de Fourier aléatoires. *Studia Mathematica*, 19: 1–25, 1960.

Michael Kearns and Lawrence K. Saul. Large deviation methods for approximate probabilistic inference. *UAI*, 1998.

Olivier Marchal and Julyan Arbel. On the sub-Gaussianity of the Beta and Dirichlet distributions. *Electronic Communications in Probability*, 22(54):1–14, 2017.

Jim Pitman and Marc Yor. The two-parameter Poisson-Dirichlet distribution derived from a stable subordinator. *The Annals of Probability*, 25(2):855–900, 1997.

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

Mark Rudelson and Roman Vershynin. Non-asymptotic theory of random matrices: extreme singular values. In *Proceedings of the ICM*, 2010.

Mariia Vladimirova, Jakob Verbeek, Pablo Mesejo, and Julyan Arbel. Understanding priors in Bayesian neural networks at the unit level. In *International Conference on Machine Learning*, 2019.

Mariia Vladimirova, Stéphane Girard, Hien Nguyen, and Julyan Arbel. Sub-Weibull distributions: Generalizing sub-Gaussian and sub-Exponential properties to heavier tailed distributions. *Stat*, 9(1), 2020.

A Proofs for Section 2

In this section we prove Theorem 2.1, Theorem 2.2, and Theorem 2.5. We also establish two lemmas, Theorem A.1 and Theorem A.2, which are useful to prove Theorem 2.6.

Proof of Theorem 2.1. It is obvious from the definition of $g_Y(\lambda;\sigma^2) = \frac{1}{2}\lambda^2\sigma^2 - M_Y(\lambda)$. The sufficient condition is also immediate since if M is a bound for Y then $|M_Y(\lambda)| \leq \lambda|M+\mu| = o(\lambda^2)$. The fact that Y is sub-Gaussian follows from that there exists $C > 0$ such that $|M_Y(\lambda)| \leq C\lambda^2$ for all $\lambda \in \mathbb{R}$. Thus, taking $\frac{C}{2}$ immediately gives that σ^2 is a variance proxy so that Y is sub-Gaussian. □

Proof of Theorem 2.2. Let us first mention that Theorem 2.1 implies that the optimal variance proxy is well-defined. Then let us consider $\sigma^2 > \max\{\mathrm{Var}[Y], \sup \mathcal{S}_c^*\}$. Since $\sigma^2 > \mathrm{Var}[Y]$, we get that $g_Y(.;\sigma^2)$ is locally convex and non-negative around $\lambda = 0$. Let us consider $\lambda_m(\sigma) \neq 0$ a local minimum of $g_Y(.,\sigma^2)$. For simplicity we shall assume that $\lambda_m(\sigma) > 0$ but a similar argument is valid if $\lambda_m(\sigma) < 0$. Then we have $\partial_\sigma[g_Y(\lambda_m(\sigma);\sigma^2)] = g'_Y(\lambda_m(\sigma);\sigma^2)\partial_\sigma\lambda_m(\sigma) + \lambda_m(\sigma)^2\sigma = \lambda_m(\sigma)^2\sigma > 0$. Assume by contradiction that $g_Y(\lambda_m(\sigma);\sigma^2) < 0$ then increasing σ would increase the value of $g_Y(\lambda_m(\sigma);\sigma^2)$. Since $\partial_\sigma[g_Y(\lambda_m(\sigma);\sigma^2)] = \lambda_m(\sigma)^2\sigma > 0$ there are only two cases:

First, $\lambda_m(\sigma)$ remains outside a positive neighborhood of 0 denoted $(0,\epsilon)$ when we increase σ and thus since $\partial_\sigma[g_Y(\lambda_m(\sigma);\sigma^2)] > \epsilon^2\sigma$ and by assumption $g_Y(\lambda_m(\sigma);\sigma^2) < 0$, there exists a value $\sigma_1 > \sigma$ for which $g_Y(\lambda_m(\sigma_1);\sigma_1^2) = 0$ which is a contradiction because $\sigma_1^2 \in \mathcal{S}_c^*$ so that we should have $\sigma \geq \sigma_1$.

Second, $\lambda_m(\sigma) \to 0_+$ when we increase σ. In this case this is a contradiction because $g_Y(.,\sigma^2)$ is locally positive and convex in a positive neighborhood of 0. Indeed, $g_Y(.,\sigma^2)$ must reach a positive local maximum on $(0,\lambda_m(\sigma))$ that we denote $\lambda_{\max}(\sigma)$. By Rolle's theorem on $(0,\lambda_{\max}(\sigma))$ and $(\lambda_{\max}(\sigma),\lambda_m(\sigma))$, there exist at least two distinct values $(\lambda_1(\sigma),\lambda_2(\sigma))$ with $0 < \lambda_1(\sigma) < \lambda_{\max}(\sigma) < \lambda_2(\sigma) < \lambda_m(\sigma)$ such that $g''_Y(\lambda_1(\sigma);\sigma^2) = g''_Y(\lambda_2(\sigma);\sigma^2) = 0$. Since $\lambda_m(\sigma) \to 0_+$, we must also have $\lambda_1(\sigma),\lambda_2(\sigma) \to 0_+$ and hence by continuity $g''_Y(0,\sigma^2) \to 0$. But this is impossible since $g''_Y(0;\sigma^2) = (\sigma^2 - \mathrm{Var}[Y]) > 0$ is a positive, increasing function of σ.

Thus we conclude that for any $\sigma^2 > \max\{\mathrm{Var}[Y], \sup \mathcal{S}_c^*\}$, all local minima of $g_Y(.;\sigma^2)$ are non-negative so that since $\lim_{\lambda\to\pm\infty} g_Y(\lambda;\sigma^2) = +\infty$, $g_Y(.;\sigma^2)$ is non-negative on $\mathbb{R}$. Hence σ^2 is a variance proxy and thus $\sigma^2_{\mathrm{opt}} \leq \max\{\mathrm{Var}[Y], \sup \mathcal{S}_c^*\}$.

Let us prove the converse inequality and assume that $\sigma^2_{\mathrm{opt}} < \max\{\mathrm{Var}[Y], \sup \mathcal{S}_c^*\}$. It is well-known that $\mathrm{Var}[Y]$ is always a lower bound for the optimal variance proxy, thus the last inequality is only possible if $\mathrm{Var}[Y] < \sigma^2_{\mathrm{opt}} < \sup \mathcal{S}_c^*$. Thus, there exists $s_c \in \mathcal{S}_c^*$ such that $\sigma^2_{\mathrm{opt}} < s_c$, i.e. there exists $\lambda_c \in \mathbb{R}^*$ such that $g_Y(\lambda_c;s_c) = g'_Y(\lambda_c;s_c) = 0$ and λ_c is a local minimum of $g_Y(.;s_c)$. We have from the fact that the dependence of g_Y relatively to σ is quadratic that:

$$g_Y(\lambda_c,\sigma^2) = g_Y(\lambda_c,s_c) + \frac{1}{2}\lambda_c^2(\sigma^2 - s_c) = \frac{1}{2}\lambda_c^2(\sigma^2 - s_c). \tag{11}$$

so that $g_Y(\lambda_c,\sigma^2) < 0$ when $\sigma^2 < s_c$. This implies that $g_Y(.;\sigma^2)$ is no longer non-negative when $\sigma^2 < s_c$ so that σ^2 is not a variance proxy. This contradicts the fact that $\sigma^2_{\mathrm{opt}} < s_c$ is an optimal variance proxy. □

Proof of Theorem 2.5. If $g''_Y(.;\sigma^2)$ has no zero on $\mathbb{R}_+^*$, then it is strictly convex on $\mathbb{R}_+^*$ and since $g_Y(0;\sigma^2) = g'_Y(0;\sigma^2) = 0$, $g_Y(.;\sigma^2)$ is positive on $\mathbb{R}_+^*$. Similarly, if $g''_Y(.;\sigma^2)$ has a unique zero on $\mathbb{R}_+^*$, then it cannot change sign at this zero, because $g''_Y(0;\sigma^2) = \sigma^2 - \mathrm{Var}[Y] \geq 0$ and $\lim_{\lambda\to+\infty} g''_Y(\lambda;\sigma^2) = \sigma^2 > 0$. Hence, we conclude similarly that $g_Y(.;\sigma^2)$ is positive on $\mathbb{R}_+^*$. □

Lemma A.1 (Local minimum when $g''_Y(.;\sigma^2)$ has two positive zeros and $\sigma^2 > \mathrm{Var}[Y]$)**.** *Assume that M_Y is a smooth function and that $M_Y(\lambda) \overset{\lambda\to\pm\infty}{=} o(\lambda^2)$. Moreover, assume that for any $\sigma^2 \in (\mathrm{Var}[Y], +\infty)$, $g''_Y(.;\sigma^2)$ has exactly two positive zeros $(\lambda_1(\sigma), \lambda_2(\sigma))$ such that $0 < \lambda_1(\sigma) < \lambda_2(\sigma)$. Then, $g_Y(.;\sigma^2)$ has at most one local minimum $\lambda_m(\sigma)$ on $\mathbb{R}_+^*$ and it is necessarily located in $\lambda_m(\sigma) \in (\lambda_1(\sigma), \lambda_2(\sigma))$. Consequently, the equations $g_Y(\lambda, \sigma^2) = 0 = g'_Y(\lambda, \sigma^2)$ have at most one solution on $\mathbb{R}_+^* \times (\mathrm{Var}[Y], +\infty)$ and we have that:*

- *if they have no solution, then $g_Y(.;\sigma^2)$ is non-negative on $\mathbb{R}_+$ for any $\sigma^2 > \mathrm{Var}[Y]$.*
- *if they have one solution (λ_c, σ_c^2), then $\lambda_c > 0$ is necessarily a local minimum and $g_Y(.;\sigma^2)$ is non-negative on $\mathbb{R}_+$ if and only if $\sigma^2 \geq \sigma_c^2$.*

A similar result is valid on $\mathbb{R}_-^*$.

Proof of Theorem A.1. The proof follows from a precise study of variations. Indeed, let us first notice that for any $\sigma^2 > \mathrm{Var}[Y]$, we have that $g_Y(.;\sigma^2)$ is locally convex and positive around $\lambda = 0$ since $g''_Y(0;\sigma^2) = \sigma^2 - \mathrm{Var}[Y] > 0$. We also remind that $\lim_{\lambda\to\pm\infty} g''_Y(\lambda;\sigma^2) = \sigma^2 > 0$ so that $g_Y(.;\sigma^2)$ is also convex at infinity. Thus, from the assumption that $g''_Y(.;\sigma^2)$ has exactly two zeros $(\lambda_1(\sigma), \lambda_2(\sigma))$ on $\mathbb{R}_+^*$, we conclude that either $g''_Y(.;\sigma^2)$ does not change sign at its zeros and in this case, $g_Y(.;\sigma^2)$ is strictly convex on $\mathbb{R}_+^*$ with $g_Y(0;\sigma^2) = 0$ so that it is positive and increasing on $\mathbb{R}_+$, thus there is no solution of $g'_Y(.;\sigma^2) = 0$ on $\mathbb{R}_+^*$. Or that $g''_Y(.;\sigma^2)$ is necessarily positive on $(0, \lambda_1(\sigma))$, negative on $(\lambda_1(\sigma), \lambda_2(\sigma))$ and positive on $(\lambda_2(\sigma), +\infty)$. Thus, $g'_Y(.;\sigma^2)$ is increasing on $(0, \lambda_1(\sigma))$ and since $g'_Y(0;\sigma^2) = 0$ it is positive on $(0, \lambda_1(\sigma))$. Then, $g'_Y(.;\sigma^2)$ is decreasing on $(\lambda_1(\sigma), \lambda_2(\sigma))$ and then increasing on $(\lambda_2(\sigma), +\infty)$. In particular, $g'_Y(.;\sigma^2)$ admits a unique local minimum at $\lambda = \lambda_2(\sigma)$ on $\mathbb{R}_+^*$. Consequently, if $g'_Y(\lambda_2(\sigma);\sigma^2) \geq 0$, then $g_Y(.;\sigma^2)$ is non-negative on $\mathbb{R}_+$ and thus since $g_Y(0;\sigma^2) = 0$, we get that $g_Y(.;\sigma^2)$ is non-negative on $\mathbb{R}_+$. Alternatively, if $g'_Y(\lambda_2(\sigma);\sigma^2) < 0$, then $g'_Y(.;\sigma^2)$ has exactly two zeros $\lambda_1(\sigma) < \lambda_l(\sigma) < \lambda_r(\sigma) < \lambda_2(\sigma)$ and it is negative on $(\lambda_l(\sigma), \lambda_r(\sigma))$ and positive on $(0, \lambda_l(\sigma)) \cup (\lambda_r(\sigma), +\infty)$. Consequently, $g_Y(.;\sigma^2)$ has exactly one local minimum on $\mathbb{R}_+^*$ denoted $\lambda_m(\sigma)$ which is necessarily located in $(\lambda_1(\sigma), \lambda_2(\sigma))$. Next, let us observe that:

$$\partial_\sigma[g'_Y(\lambda_2(\sigma);\sigma^2)] = g''_Y(\lambda_2(\sigma);\sigma^2)\partial_\sigma[\lambda_2(\sigma)] + 2\sigma\lambda_2(\sigma) = 2\sigma\lambda_2(\sigma) > 0. \tag{12}$$

Thus, the local minimum of $g'_Y(.;\sigma^2)$ is an increasing function of σ, so that if it is null for a value σ_1, then it is positive for $\sigma > \sigma_1$ and negative for $\sigma < \sigma_1$. Finally, let us observe that

$$\partial_\sigma[g_Y(\lambda_m(\sigma);\sigma^2)] = g'_Y(\lambda_m(\sigma);\sigma^2)\partial_\sigma[\lambda_m(\sigma)] + \sigma\lambda_m(\sigma)^2 = \sigma\lambda_m(\sigma)^2 > 0. \tag{13}$$

so that $\lambda_m(\sigma)$ is an increasing function of σ. Let us denote $(\lambda_c, \sigma_c^2) \in \mathbb{R}_+^* \times (\mathrm{Var}[Y], +\infty)$ a solution of $g_Y(\lambda, \sigma^2) = 0 = g'_Y(\lambda, \sigma^2)$, then for $\sigma < \sigma_c$, we have $g_Y(\lambda_m(\sigma);\sigma^2) < 0$ while for $\sigma > \sigma_c$, we have $g_Y(\lambda_m(\sigma);\sigma^2) > 0$. Since $g_Y(.;\sigma^2)$ has only a unique local minimum $\lambda_m(\sigma)$ and a local maximum on $\mathbb{R}_+^*$ which is always strictly positive, we conclude that we cannot have another solution of $g_Y(\lambda, \sigma^2) = 0 = g'_Y(\lambda, \sigma^2)$ with $\lambda \in \mathbb{R}_+^*$. When there is no solution, we have $g_Y(\lambda_m(\sigma);\sigma^2) > 0$ so that $g_Y(.;\sigma^2)$ is non-negative on $\mathbb{R}_+^*$. When we have a unique solution (λ_c, σ_c^2), then $g_Y(\lambda_m(\sigma);\sigma^2) > 0$ for $\sigma > \sigma_c$, $g_Y(\lambda_m(\sigma_c);\sigma_c^2) = 0$ and $g_Y(\lambda_m(\sigma);\sigma^2) < 0$ for $\sigma < \sigma_c$ concluding the proof. □

We complement Theorem A.1 with another lemma regarding the situation when $\sigma < \mathrm{Var}[Y]$.

Lemma A.2 (Local minimum when $g''_Y(.;\sigma^2)$ has two positive zeros and $\sigma^2 < \mathrm{Var}[Y]$)**.** *Assume that M_Y is a smooth function and that $M_Y(\lambda) \overset{\lambda\to\pm\infty}{=} o(\lambda^2)$ and that for any $\sigma^2 \in (0, \mathrm{Var}[Y])$, $g''_Y(.;\sigma^2)$ has exactly two zeros $(\lambda_1(\sigma), \lambda_2(\sigma))$ such that $0 < \lambda_1(\sigma) < \lambda_2(\sigma)$. Then, there is no solution to the set of equations $g_Y(\lambda;\sigma^2) = 0 = g'_Y(\lambda;\sigma^2)$ with $\lambda > 0$ and $\sigma^2 \in (0, \mathrm{Var}[Y])$.*
A similar result holds on $\mathbb{R}_-^*$.

Proof of Theorem A.2. For any $\sigma^2 < \mathrm{Var}[Y]$, we have that $g_Y(\lambda\sigma^2) = (\sigma^2 - \mathrm{Var}[Y])\lambda^2 + o(\lambda^2)$ when $\lambda \to 0_+$. Thus, $g_Y(.;\sigma^2)$ is locally concave and negative around $\lambda = 0_+$. Since $M_Y(\lambda) \overset{\lambda\to\pm\infty}{=} o(\lambda^2)$, we know that $g_Y(.;\sigma^2)$ is locally convex and positive when $\lambda \to +\infty$. Since $g''_Y(.;\sigma^2)$ is assumed to have exactly two positive zeros $\lambda_1(\sigma) < \lambda_2(\sigma)$, we may only have the following cases:

- $g''_Y(.;\sigma^2)$ is negative on $(0, \lambda_1(\sigma))$ and changes sign at $\lambda_1(\sigma)$. Thus, it is positive on $(\lambda_1(\sigma), \lambda_2(\sigma))$ and cannot change sign at $\lambda_2(\sigma)$ to remain positive at $+\infty$. Hence, $g''_Y(.;\sigma^2)$ is positive on $(\lambda_1(\sigma), \lambda_2(\sigma)) \cup (\lambda_2(\sigma), +\infty)$. Consequently, $g'_Y(.;\sigma^2)$ is decreasing on $(0, \lambda_1(\sigma))$ and increasing on $(\lambda_1(\sigma), +\infty)$. Since $g'_Y(0;\sigma^2)$, we have $g'_Y(\lambda_1(\sigma);\sigma^2) < 0$ and since $g'_Y(+\infty;\sigma^2) = +\infty$, g'_Y has only one zero $\lambda_0(\sigma)$ on $\mathbb{R}_+^*$ that satisfies $\lambda_0(\sigma) > \lambda_1(\sigma)$. Moreover, $g'_Y(.;\sigma^2)$ is negative on $(0, \lambda_0(\sigma))$ and positive on $(\lambda_0(\sigma), +\infty)$. Hence, $g_Y(.;\sigma^2)$ is decreasing on $(0, \lambda_0(\sigma))$ and increasing on $(\lambda_1(\sigma), +\infty)$. Since $g_Y(0;\sigma^2) = 0$ and $g_Y(+\infty;\sigma^2) = +\infty$, g_Y admits a unique zero $\lambda_*(\sigma)$ on $\mathbb{R}_+^*$ and we have $\lambda_*(\sigma) > \lambda_0(\sigma)$. Hence, there are no simultaneous solutions to $g_Y(\lambda;\sigma^2) = 0 = g'_Y(\lambda;\sigma^2)$ with $\lambda > 0$.
- $g''_Y(.;\sigma^2)$ is negative on $(0, \lambda_1(\sigma))$ and does not change sign at $\lambda_1(\sigma)$ so it is negative on $(0, \lambda_2(\sigma))$. In order to be positive at $\lambda \to +\infty$, $g''_Y(.;\sigma^2)$ must change sign at $\lambda = \lambda_2(\sigma)$. Thus, $g'_Y(.;\sigma^2)$ is decreasing on $(0, \lambda_2(\sigma))$ and increasing on $(\lambda_2(\sigma), +\infty)$. Since $g'_Y(0;\sigma^2) = 0$ and $g'_Y(+\infty;\sigma^2) = +\infty$, $g'_Y(.;\sigma^2)$ admits a unique zero $\lambda_0(\sigma)$ on $\mathbb{R}_+^*$ and it satisfies $\lambda_0(\sigma) > \lambda_2(\sigma)$. Moreover, $g'_Y(.;\sigma^2)$ is negative on $(0, \lambda_0(\sigma))$ and positive on $(\lambda_0(\sigma), +\infty)$. Since $g_Y(0;\sigma^2) = 0$ and $g_Y(+\infty;\sigma^2) = +\infty$, $g_Y(.,\sigma^2)$ is decreasing and negative on $(0, \lambda_0(\sigma))$ and increasing on $(\lambda_0(\sigma), +\infty)$. Hence, $g_Y(.;\sigma^2)$ admits a unique zero $\lambda_*(\sigma)$ on $\mathbb{R}_+^*$ and we have $\lambda_*(\sigma) > \lambda_0(\sigma)$. Hence, there are no simultaneous solutions to $g_Y(\lambda;\sigma^2) = 0 = g'_Y(\lambda;\sigma^2)$ with $\lambda > 0$.

□

B Proofs for Section 3

In this section we prove Theorem 3.2 and Theorem 3.4.

Proof of Theorem 3.2. Let us first observe that $\left(2\lambda_0, \sigma^2_{\mathrm{opt}}\right) = \left(-\ln\frac{p_2}{p_1}, \sigma^2_{\mathrm{opt}}\right)$ is a solution (λ, σ^2) of the system of equations

$$g_{\sigma,p_1,p_2}(\lambda) = 0 \text{ and } g'_{\sigma,p_1,p_2}(\lambda) = 0.$$

Moreover, the case $p_3 \leq 4\sqrt{p_1p_2}$ is equivalent to the fact that $\Delta \leq 0$ or $\Delta > 0$ with P admitting two strictly negative roots (the sum of roots (X_1, X_2) is $X_1 + X_2 = \frac{p_3^2 - 8p_1p_2}{p_2p_3} < 0$ and the product of roots $X_1X_2 = \frac{p_1}{p_2} > 0$). Thus, the function N_{p_1,p_2} is strictly positive on $\mathbb{R}$. Consequently, $g^{(3)}_{\sigma,p_1,p_2}$ and $u_1(\lambda)$ share the same sign and hence $g^{(3)}_{\sigma,p_1,p_2}$ is strictly negative on $(-\infty, \lambda_0)$ and strictly positive on $(\lambda_0, +\infty)$. Furthermore, since $\lim_{\lambda\to\pm\infty} g^{(2)}_{\sigma,p_1,p_2}(\lambda) = \sigma^2$, it follows that λ_0 is a global minimum of $g^{(2)}_{\sigma,p_1,p_2}$ with value $g^{(2)}_{\sigma,p_1,p_2}(\lambda_0) = \sigma^2 - \frac{2\sqrt{p_1p_2}}{p_3+2\sqrt{p_1p_2}}$. If $\sigma^2 \geq \frac{2\sqrt{p_1p_2}}{p_3+2\sqrt{p_1p_2}}$, then $g^{(2)}_{\sigma,p_1,p_2}$ is non-negative on $\mathbb{R}$ and so g'_{σ,p_1,p_2} is a strictly increasing function. Since $g'_{\sigma,p_1,p_2}(0) = 0$, g'_{σ,p_1,p_2} is negative on $\mathbb{R}_-$ and positive on $\mathbb{R}_+$ and finally g_{σ,p_1,p_2} has a global minimum at $\lambda = 0$ which is precisely null so it is positive and σ^2 is a variance proxy. This gives that $\frac{2\sqrt{p_1p_2}}{p_3+2\sqrt{p_1p_2}}$ is an upper bound for the optimal variance proxy.

On the contrary, if $\sigma^2 < \frac{2\sqrt{p_1p_2}}{p_3+2\sqrt{p_1p_2}}$, then $g^{(2)}_{\sigma,p_1,p_2}$ has two distinct zeros $(\lambda_1(\sigma), \lambda_2(\sigma))$ such that $\lambda_1(\sigma) < \lambda_0 < \lambda_2(\sigma) \leq 0$. Moreover, since $g''_{\sigma,p_1,p_2}(.;\sigma^2)$ is positive on $\mathbb{R}_+^*$, we get that $g_{\sigma,p_1,p_2}(.;\sigma^2)$ is always positive on $\mathbb{R}_+^*$ for any $\sigma^2 \in \left(\mathrm{Var}[Y], \frac{2\sqrt{p_1p_2}}{p_3+2\sqrt{p_1p_2}}\right)$. Since, we have observed that the equations $g_{\sigma,p_1,p_2}(\lambda,\sigma^2) = 0 = g'_{\sigma,p_1,p_2}(\lambda,\sigma^2)$ admits $\left(2\lambda_0, \frac{2(p_2-p_1)}{\ln(p_2/p_1)}\right) \in \mathbb{R}_-^* \times (\mathrm{Var}[Y], +\infty)$ as solution, application of Theorem 2.6 on $\mathbb{R}_-^*$

implies that $g_{\sigma,p_1,p_2}(.;\sigma^2)$ is non-negative on $\mathbb{R}_-$ if and only if $\sigma^2 \geq \frac{2(p_2-p_1)}{\ln(p_2/p_1)}$. Thus the optimal variance proxy in this case is $\sigma^2_{\text{opt}} = \frac{2(p_2-p_1)}{\ln(p_2/p_1)}$. □

Proof of Theorem 3.4. Let us first observe that we have:

$$\lambda_\pm = \ln\left(\frac{p_3^2 - 8p_1p_2 \pm \sqrt{(p_3^2-4p_1p_2)(p_3^2-16p_1p_2)}}{2p_1p_2}\right).$$

We shall denote for compactness $x := p_3^2 > 0$, $y := p_1p_2 > 0$ so that we have the condition $x > 16y$. Since $\lambda_0 < 0$, we only need to prove that $\lambda_- > 0$. We will now prove that $\lambda_- > 0$ under the condition $x > 16y > 0$. To establish this result, we proceed through a chain of equivalent inequalities, beginning with the definition of λ_-

$$\begin{aligned}
\lambda_- > 0 &\iff \frac{x - 8y - \sqrt{(x-4y)(x-16y)}}{2y} > 1 \\
&\iff x - 8y - \sqrt{(x-4y)(x-16y)} > 2y \\
&\iff (x-10y)^2 > (x-4y)(x-16y) \quad \text{(because } x > 16y \Rightarrow x - 10y > 6y > 0\text{)} \\
&\iff x^2 - 20xy + 100y^2 > x^2 - 20xy + 64y^2 \\
&\iff 100y^2 > 64y^2 \\
&\iff 36y^2 > 0 \quad \text{(always valid for } y \neq 0\text{)}.
\end{aligned}$$

Thus we get $\lambda_0 < 0 < \lambda_- < \lambda_+$ under the condition $p_3^2 > 16p_1p_2$ ending the proof of the lemma. □

C Proofs for Section 4

By linearity of the log-MGF and the scaling property of variance proxies, it is convenient to normalize X. Define

$$Y := \frac{X-b}{a}.$$

Then Y is uniformly distributed on $[\![1, N]\!]$ and

$$\sigma_{\text{opt}}[X] = |a|\,\sigma_{\text{opt}}[Y].$$

Hence, without loss of generality, we assume $a = 1$ and $b = 0$. Under this assumption, the variable Y uniformly distributed on the integer set $\{1, 2, \ldots, N\}$ with moments:

$$\mu := \mathbb{E}[Y] = \frac{N+1}{2}, \quad \sigma^2 := \text{Var}[Y] = \frac{N^2-1}{12}, \quad \kappa_3[Y] := \mathbb{E}[(Y - \mathbb{E}[Y])^3] = 0. \tag{14}$$

By definition, $\sigma > 0$ is a variance proxy of Y if and only if

$$\mathbb{E}\left[e^{\lambda Y}\right] = \frac{1}{N}\sum_{k=1}^{N} e^{\lambda k} \;\leq\; \exp\left(\tfrac{\lambda^2\sigma^2}{2} + \lambda\mu\right), \quad \forall\, \lambda \in \mathbb{R}. \tag{15}$$

Equivalently, defining the log-partition function

$$u(\lambda) := \ln\left(\frac{1}{N}\sum_{k=1}^{N} e^{\lambda k}\right),$$

this condition becomes equivalent to the non-negativity of

$$g_{\sigma,N}(\lambda) := \tfrac{\lambda^2\sigma^2}{2} - u(\lambda) + \lambda\mu \ \geq 0, \qquad \forall\, \lambda \in \mathbb{R}.$$

It is known that the variance is a universal lower bound for variance proxies. Hence in our analysis we consider only $\sigma^2 \geq \mathrm{Var}[Y] = \frac{N^2-1}{12}$.

To characterize the optimal variance proxy, it is essential to study the properties of the function $g_{\sigma,N}$. Observe that $g_{\sigma,N}$ is a smooth function of $(\sigma, \lambda) \in \mathbb{R}^2$. Its first three derivatives with respect to λ are explicitly computed as:

$$g'_{\sigma,N}(\lambda) = \lambda\sigma^2 - u'(\lambda) + \mu\,,\ g^{(2)}_{\sigma,N}(\lambda) = \sigma^2 - u^{(2)}(\lambda)\,,\ g^{(3)}_{\sigma,N}(\lambda) = -u^{(3)}(\lambda). \tag{16}$$

To analyze the log-partition function u, and consequently those of $g_{\sigma,N}$ it is convenient to introduce an auxiliary family of probability distributions. For every real λ, define the probability distribution P_λ on $\{1, \ldots, N\}$ by

$$P_\lambda(k) = \frac{e^{\lambda k}}{\sum_{j=1}^{N} e^{\lambda j}}, \qquad \forall\, k \in [\![1, N]\!].$$

Let $Z_\lambda \sim P_\lambda$ denote the associated random variable. Note that $Z_0 = Y$ coincides with the uniform distribution. This family of probability distributions satisfies several fundamental identities, including symmetry properties and a moment derivative formula.

Symmetry identities. For all $\lambda \in \mathbb{R}$:

$$P_{-\lambda}(k) = P_\lambda(N - k + 1), \quad \mathbb{E}[Z_{-\lambda}] = N + 1 - \mathbb{E}[Z_\lambda], \quad \mathrm{Var}[Z_{-\lambda}] = \mathrm{Var}[Z_\lambda]. \tag{17}$$

Moment derivative identity. For all integers $m \geq 1$:

$$\frac{d}{d\lambda}\mathbb{E}[Z_\lambda^m] = \mathbb{E}[Z_\lambda^{m+1}] - \mathbb{E}[Z_\lambda]\,\mathbb{E}[Z_\lambda^m]. \tag{18}$$

Lemma C.1 (Derivatives of log-partition function as moments)**.** *The derivatives of $\lambda \mapsto u(\lambda)$ give the moments of the associated random variables:*

$$u'(\lambda) = \mathbb{E}[Z_\lambda], \quad u^{(2)}(\lambda) = \mathrm{Var}[Z_\lambda], \quad u^{(3)}(\lambda) = \mathbb{E}[(Z_\lambda - \mathbb{E}[Z_\lambda])^3] = \kappa_3[Z_\lambda]. \tag{19}$$

Proof. These identities follow directly from the moment derivative formula combined with the expressions for the first moments. Applying the derivative identity recursively yields the expressions for $u'(\lambda)$, $u^{(2)}(\lambda)$, and $u^{(3)}(\lambda)$, which correspond to the mean, variance, and third centered moment of Z_λ. □

From the symmetry identities, we know that $\mathrm{Var}[Z_\lambda]$ is an even function of λ. This symmetry allows us to restrict our analysis to $\mathbb{R}_+$. Moreover, given the derivative relation $\kappa_3[Z_\lambda] = \frac{d}{d\lambda}\mathrm{Var}[Z_\lambda]$ along with the evenness of the variance, it follows that the third central moment is an odd function of λ. The main technical task is then to prove that the third derivative $\lambda \mapsto u^{(3)}(\lambda)$ is negative on $\mathbb{R}_+$.

Lemma C.2 (Negativity of the third derivative of the log-partition function)**.** *Let $N \geq 2$ be an integer and let $\lambda > 0$ be a real number. Then the third derivative of the log-partition function is strictly negative. In other words,*

$$u^{(3)}(\lambda) = -\frac{N^3 e^{\lambda N}(1 + e^{\lambda N})}{(1 - e^{\lambda N})^3} + \frac{e^{\lambda}(1 + e^{\lambda})}{(1 - e^{\lambda})^3} < 0, \quad \forall\, \lambda \in \mathbb{R}_+^*. \tag{20}$$

Consequently,

$$\kappa_3[Z_\lambda] < 0 \,\textit{for all}\, \lambda \in \mathbb{R}_+^*. \tag{21}$$

Proof. The explicit expression of $u^{(3)}(\lambda)$ follows by direct differentiation of $u(\lambda) = \ln\left(\frac{1}{N}\sum_{k=1}^{N} e^{\lambda k}\right)$ and using the closed-form expression for the geometric sum. Then, consider the auxiliary function

$$f(t) = \frac{t(t+1)}{(1-t)^3}, \quad \text{for } t = e^{\lambda} > 1. \tag{22}$$

Its derivative is:

$$f'(t) = \frac{1 + 4t + t^2}{(1-t)^4}.$$

Since both the numerator and the denominator are strictly positive for all $t > 1$, we conclude that $f'(t) > 0$. Hence, f is strictly increasing on the interval $(1, \infty)$. Now fix $\lambda > 0$, so that $t = e^{\lambda} > 1$. Since $N \geq 2$, we have $t^N > t$. By the monotonicity of f, it follows that

$$f(t^N) > f(t),$$

which yields the inequality

$$\frac{e^{\lambda N}(1 + e^{\lambda N})}{(1 - e^{\lambda N})^3} > \frac{e^{\lambda}(1 + e^{\lambda})}{(1 - e^{\lambda})^3}.$$

Furthermore, since $N^3 \geq 8$ for all $N \geq 2$, we obtain

$$\frac{N^3 e^{\lambda N}(1 + e^{\lambda N})}{(1 - e^{\lambda N})^3} > \frac{e^{\lambda}(1 + e^{\lambda})}{(1 - e^{\lambda})^3}.$$

Therefore, the third derivative of the log-partition function satisfies

$$u^{(3)}(\lambda) = -\frac{N^3 e^{\lambda N}(1 + e^{\lambda N})}{(1 - e^{\lambda N})^3} + \frac{e^{\lambda}(1 + e^{\lambda})}{(1 - e^{\lambda})^3} < 0 \ , \ \forall\, \lambda > 0$$

which completes the proof of the lemma. □

The end of the proof of Theorem 4.1 is now straightforward.

Variance proxy optimality. Theorem C.1 and Theorem C.2 show that

$$u^{(3)}(\lambda) = \kappa_3[Z_\lambda] = \frac{d}{d\lambda}\operatorname{Var}[Z_\lambda] < 0, \quad \forall\, \lambda > 0.$$

As $\operatorname{Var}[Z_\lambda]$ is an even function of λ, it follows that it is strictly increasing on $\mathbb{R}_-$, strictly decreasing on $\mathbb{R}_+$ and thus achieves a unique global maximum at $\lambda = 0$. In particular,

$$\operatorname{Var}[Z_\lambda] \leq \operatorname{Var}[Z_0] = \operatorname{Var}[Y], \quad \forall\, \lambda \in \mathbb{R}.$$

Hence, for any $\sigma^2 \geq \operatorname{Var}[Y]$, we have $g^{(2)}_{\sigma,N}(\lambda) = \sigma^2 - \operatorname{Var}[Z_\lambda] \geq 0$, with equality if and only if $\lambda = 0$. Finally for any $\sigma^2 \geq \operatorname{Var}[Y]$, given that $g^{(2)}_{\sigma,N}(\lambda)$ has no solution on $\mathbb{R}^*_+$ and $\mathbb{R}^*_-$, it follows from Theorem 2.5 that $g_{\sigma,N}$ is non-negative on $\mathbb{R}$. This shows that Y is strictly sub-Gaussian ending the proof.