Tight Bound for Sum of Heterogeneous Random Variables: Application to Chance Constrained Programming

Quentin Jacquet, Riadh Zorgati

Tight Bound for Sum of Heterogeneous Random Variables: Application to Chance Constrained Programming

Quentin Jacquet[1,2], Riadh Zorgati[2]

[1] INRIA, CMAP, Ecole Polytechnique, Palaiseau, France

[2] EDF R&D Saclay, Palaiseau, France

{quentin.jacquet,riadh.zorgati}@edf.fr

Abstract

We study a tight Bennett-type concentration inequality for sums of heterogeneous and independent variables, defined as a one-dimensional minimization. We show that this refinement, which outperforms the standard known bounds, remains computationally tractable: we develop a polynomial-time algorithm to compute confidence bounds, proved to terminate with an ϵ-solution. From the proposed inequality, we deduce tight distributionally robust bounds to Chance-Constrained Programming problems. To illustrate the efficiency of our approach, we consider two use cases. First, we study the chance-constrained binary knapsack problem and highlight the efficiency of our cutting-plane approach by obtaining stronger solution than classical inequalities (such as Chebyshev-Cantelli or Hoeffding). Second, we deal with the Support Vector Machine problem, where the convex conservative approximation we obtain improves the robustness of the separation hyperplane, while staying computationally tractable.

Keywords: Concentration inequalities, Chance-constrained programming, Confidence bounds, Knapsack problem, Support Vector Machine

1 Introduction

Concentration inequalities – such as Hoeffding [19], Bennett [5] or McDiarmid [25] to cite a few – were originally introduced to quantify how a random variable deviates from their expectation. The crux of the matter is the imperfect knowledge of a random process: varying between the inequalities, the only available information are about the two first moments (mean and variance) or the length of the support of the distribution. These inequalities have now a wide variety of applications, see e.g. [8], including chance constrained programming or machine learning [27, 28, 39, 23].

Many refinements of Hoeffding and Bennett's inequalities have been proposed: all these works exploit Chernoff's inequality but differ in the estimation of the moment-generating function $t \mapsto \mathbb{E}[e^{tX}]$. Figure 1 proposes a schematic classification of the literature. From and Swift [15] and Zheng [41] both use a linear approximation of $x \mapsto e^{tx}$, that is tighter than Hoeffding' bound [19] for variables in $[0, 1]$. They differ in the use of the arithmetic-geometric mean inequality. Jebara [21] exploits an inequality from [5, (b)] to derive an analytic one-sided bound for sum of heterogeneous random variables. Finally, Cheng and Li [11] insert a multipoint approximation of e^{tX} and compare their results with [41]. We

emphazise that the classification we made – which is a contribution on its own – focuses on the crucial approximation done while tackling with Chernoff's inequalities, and does not directly compare the final bounds obtained in each work.

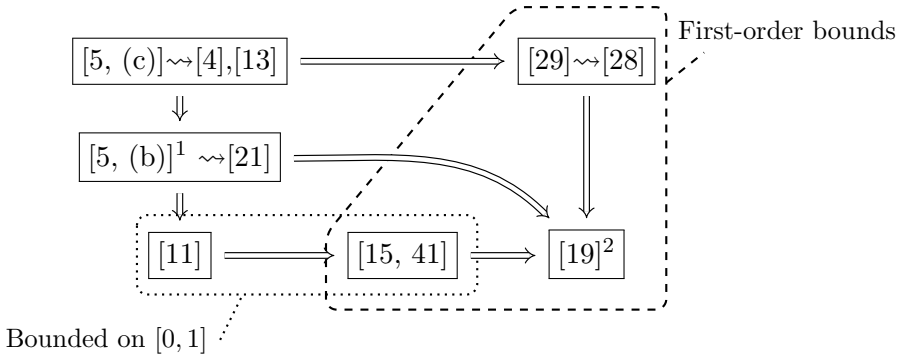


Figure 1: Classification of $t \mapsto \mathbb{E}[e^{tX}]$ estimations
[1] Bennett's inequality
[2] Hoeffding's inequality

If [a] $\Rightarrow$ [b], the upper estimator of the moment-generating function in [a] is tighter than in [b]. If [a] $\rightsquigarrow$ [b], [b] uses the same moment-generating estimator but improves / extends the results. Proofs of the different implications are provided in Appendix A.1.

In this paper, we focus on the moment-generating estimator introduced in [5, (c)]: for a random variable X with mean μ, variance σ^2 and such that $|X - \mu| \leq b$, we have the following inequality:

$$\forall t \in \mathbb{R},\ \mathbb{E}\left[e^{tX}\right] \leq e^{t\mu} \frac{\sigma^2 e^{|t|b} + b^2 e^{-\frac{|t|\sigma^2}{b}}}{b^2 + \sigma^2} \quad . \tag{1}$$

This estimator is as tight as possible (knowing only μ, σ and b), since it has been proved to be exact for a particular Bernoulli distribution, see e.g. [5]. Dembo and Zeitouni [13] exploit this inequality but limit the study to identically distributed variables to obtain a closed-form expression involving a Kullback–Leibler divergence. Bennett [4] extends the results to non identically distributed variables, but, in order to obtain explicit formula, further approximations have been made, leaving room for possible improvements. In contrast, we do not make additional approximations and directly construct the Chernoff bound using (1). Even if an analytic solution is not known in the heterogeneous setting, we prove that this bound can be used in many applications.

We first focus on the computation of confidence bound and introduce a double bisection algorithm (Algorithm 1). We prove that this algorithm computes a bound with arbitrary precision in polynomial time (Theorem 3.2). This algorithm belongs to the class of Probabilistic Bisection Algorithms (PBA), see e.g. [20, 38], but instead of having a zero-mean noise, the error is bounded and controlled by a parameter.

We then apply this result on Chance-Constrained Programming (CCP) [10, 26, 31, 18, 36], a very attractive tool for dealing with uncertainty in optimization problems in addition to stochastic [6, 22, 34] and robust [1, 2] optimization approaches. This approach relies upon the characterization of uncertainty by means of probabilistic information and tries to find a good solution in a probabilistic sense. CCP aims at finding the best solution which satisfies uncertain constraints of the form $g(x, \xi) \geq$

0 with a given probability p, typically close to 1. A general CCP is expressed as:

$$\begin{aligned} \min \quad & f[c(x,\xi)] \\ s.t. \quad & \mathbb{P}[g_i(x,\xi) \geq 0, (i=1,...,m)] \geq p \\ & x \in X, \end{aligned} \tag{2}$$

where $x \in \mathbb{R}^n$ denote a decision vector, f is some general risk (for example $f(.) = \mathbb{E}.$) applied to the cost function c that is impacted by uncertainty $\xi \in \mathbb{R}^m$ a general random vector/process, $\mathbb{P}$ is the probability measure associated to the probability space $(\Omega, \mathcal{F}, \mathbb{P})$ on which is defined ξ applied to a whole system of m stochastic inequalities $g_i(x,\xi) \geq 0, (i=1,...,m)$, $p \in (0,1)$ is a given confidence level and X models the deterministic feasible set for the decision vector x.

Although the underlying mathematical difficulties lead to very challenging tasks in the general case, CCP may lead to tractable algorithms. It is in particular the case for individual chance-constrained optimization $\mathbb{P}[g_i(x,\xi) \geq 0] \geq p_i, \ (i=1,...,m)$ in Gaussian setting leading to a SOCP [17], a convex optimization problem. More generally, for specific families of distributions, it is known that the set of a probabilistic constraint is convex, making possible the use of nonlinear methods, see e.g. [32, 24, 37]. However, the distributions are commonly unavailable in many applications, and even the evaluation of the constraint is not easy. A classical method is then to find a conservative approximation of the problem that is distributionally robust. In this case, the chance constraint are satisfied for any distribution and an optimal solution of the approximate problem gives a feasible solution of the original chance-constrained problem. Concentration inequalities have already been used in this context: to the best of our knowledge, Pinter [29] was the first to use concentration inequalities in optimization problem. Nemirovski and Shapiro [27] proved that the use of Chernoff's bounds provides tractable conservative approximation of chance-constrained problems. In particular, they detailed the convex approximation for several families of univariate distributions. Peng, Maggioni and Lisser [28] focuses on SOCP conservative approximations of two types: distributionnaly robust formulations based on Hoeffding and Chebyshev's inequalities, and models that assumes a normally distributed uncertainty. In particular, they deal with joint independent chance constraints, which is known to be of a high complexity, see e.g. [27].

Here, we compare various formulations on knapsack problems [9, 35] by specializing the study to second-order bounds (knowledge of means and variances). To this purpose, we introduce a new convex conservative approximation based on Bernstein's inequality (Proposition 4.1) and derive from the tight Bennett's inequality a strong approximation (Proposition 4.2). We show that the two formulations can be efficiently solved by a cutting-plane approach and lead to solution improvement on instances from the literature (Table 2). In particular, for a given budget, we improve the objective value compared to the SOCP formulations [28].

Finally, we focus on the Support Vector Machine (SVM) problem with uncertainties, where the difficulty lies in the large number of probabilistic constraints. The distributionally robust version of the problem has been addressed in [3, 39, 23]. In particular, Ben-Tal et al. [3] first consider the same moment-generating estimator (1), but make additional approximations in order to obtain SOCP formulations. Using convex optimization tools, we numerically highlight that our approach increases further the quality of the separation hyperplane while staying tractable for instances of substantial size.

This paper is organized as follows. In Section 2, we first derive properties of the proposed inequal-

ity, and numerically observe its asymptotic behavior. Then, we introduce in Section 3 an algorithm to compute confidence bounds. Finally, in Section 4, we apply the inequality to chance-constrained programming, focusing on knapsack problems (Section 4.1) and on Support Vector Machine problem (Section 4.2).

Notations. For two vectors a, b of $\mathbb{R}^N$, we denote by $\langle a, b\rangle$ the Euclidean scalar product. Moreover, $a \wedge b$ (resp. $a \vee b$) stands for the component-wise maximum (resp. the minimum) between a and b. Besides, for a discrete set X, $\mathrm{Conv}(X)$ will read as the convex envelope of X.

2 On the tightest Cramér-Chernoff bound

We first recall Hoeffding's [19] and Bennett's [5] inequalities:

Proposition 2.1 (Hoeffding)**.** *Let $X_1, \ldots, X_N$ be N independent random variables such that $\mathbb{P}[a_k \leq X_k - \mathbb{E}[X_k] \leq b_k] = 1$ for all $k \in \{1, \ldots, N\}$. Then, for all $d \geq 0$,*

$$\ln\ \mathbb{P}\left[\sum_{k=1}^{N} X_k - \mathbb{E}[X_k] \geq d\right] \leq -\frac{2d^2}{\sum_{k=1}^{N}(b_k - a_k)^2}\ . \tag{3}$$

As a consequence, for all $\tau \in]0,1[$, $\mathbb{P}\left[\sum_{k=1}^{N} X_k - \mathbb{E}[X_k] \geq d_\tau\right] \leq \tau$ where $d_\tau = \|b-a\|_2\sqrt{-\ln(\sqrt{\tau})}$.

Proposition 2.2 (Bennett)**.** *Let $X_1, \ldots, X_N$ be N independent random variables such that*

(i) $\mathbb{P}[X_k - \mathbb{E}[X_k] \leq b] = 1$, $k \in \{1, \ldots, N\}$,

(ii) $\sum_{k=1}^{N} \mathbb{E}[X_k^2] \leq \sigma^2$.

Then, with $g : u \mapsto (1+u)\ln(1+u) - u$, we get for all $d \geq 0$,

$$\ln\ \mathbb{P}\left[\sum_{k=1}^{N} X_k - \mathbb{E}[X_k] \geq d\right] \leq -\frac{\sigma^2}{b^2} g\left(\frac{bd}{\sigma^2}\right)\ . \tag{4}$$

As a consequence, for all $\tau \in]0,1[$, $\mathbb{P}\left[\sum_{k=1}^{N} X_k - \mathbb{E}[X_k] \geq d_\tau\right] \leq \tau$ where $d_\tau = \frac{\sigma^2}{b} g^{-1}\left(\frac{b^2}{\sigma^2}\ln\left(\frac{1}{\tau}\right)\right)$.

Proposition 2.1 and Proposition 2.2 does not suppose the same *a priori* knowledge on the random variables: in the latter, information in second-moment is supposed whereas the former only needs knowledge on the mean of each random variable. We now focus on the tightest second-order Cramér-Chernoff bound, firstly introduced in [5], and based on (1):

Theorem 2.3 (Refined Bennett's inequality [27],Table 2)**.** *Let $X_1, \ldots, X_N$ be N independent random variables such that*

(i) $\mathbb{P}[X_k - \mathbb{E}[X_k] \leq b_k] = 1$, $k \in \{1, \ldots, N\}$,

(ii) $\mathrm{Var}(X_k) \leq \sigma_k^2$, $k \in \{1, \ldots, N\}$.

Then, introducing $\gamma_k := \frac{\sigma_k^2}{b_k^2}$, for all $d \geq 0$

$$\forall \lambda \in \mathbb{R}_+^N, \quad \ln\ \mathbb{P}\left[\langle \lambda, X - \mathbb{E}[X]\rangle \geq d\right] \leq \inf_{t>0}\left\{-td + \sum_{k=1}^{N} \ln\left(\frac{\gamma_k e^{t\lambda_k b_k} + e^{-t\lambda_k b_k \gamma_k}}{1+\gamma_k}\right)\right\}\ . \tag{5}$$

In addition, if $\mathbb{P}[X_k - \mathbb{E}[X_k] \geq -b_k] = 1$,

$$\forall \lambda \in \mathbb{R}^N, \quad \ln\ \mathbb{P}\left[\langle \lambda, X - \mathbb{E}[X]\rangle \geq d\right] \leq \inf_{t>0} \left\{ -td + \sum_{k=1}^{N} \ln\left(\frac{\gamma_k e^{t|\lambda_k| b_k} + e^{-t|\lambda_k| b_k \gamma_k}}{1+\gamma_k} \right) \right\} \ . \tag{6}$$

Proof. Using the Chernoff' inequality on the variable $\langle \lambda, X - \mathbb{E}[X]\rangle$, we obtain

$$\mathbb{P}\left[\langle \lambda, X - \mathbb{E}[X]\rangle \geq d\right] \leq e^{-t(d + \langle \lambda, \mathbb{E}[X]\rangle)}\ \mathbb{E}\left[e^{t\langle \lambda, X\rangle}\right] \ .$$

By the independence of the variables X_k, we have $\mathbb{E}\left[e^{t\langle \lambda, X\rangle}\right] = \prod_{k=1}^{N} \mathbb{E}\left[e^{t\lambda_k X_k}\right]$. Finally, using (1), we obtain for all $t \geq 0$:

$$\mathbb{P}[\langle \lambda, X - \mathbb{E}[X]\rangle \geq d] \leq e^{-td} \prod_{k=1}^{N} \left(\frac{\gamma_k e^{t|\lambda_k| b_k} + e^{-t|\lambda_k|\gamma_k b_k}}{1+\gamma_k} \right) \ .$$

We conclude by applying the logarithm and by rearranging the terms. □

The right-hand sides of (5) and (6) correspond to the Cramér transform [13, Section 2.2] of the Bernoulli distribution that achieves the equality in (1). The scope of Theorem 2.3 is slightly more general than Hoeffding and Bennett inequality since we allow to have sum of weighted heterogeneous random variables (positive or negative weights).

Under the assumptions of Theorem 2.3, and introducing $\tau^- := \prod_{k=1}^{N} \frac{\gamma_k}{1+\gamma_k}$, we get as an immediate corollary :

$$\ln\ \mathbb{P}\left[\sum_{k=1}^{N} X_k - \mathbb{E}[X_k] \geq \alpha N \right] \leq \varphi_\alpha^* := \inf_{t \geq 0} \varphi_\alpha(t) \ , \tag{7}$$

where $\overline{b} = \frac{1}{N}\sum_{k=1}^{N} b_k$ and

$$\varphi_\alpha : t \geq 0 \mapsto \ln(\tau^-) + Nt\left(\overline{b} - \alpha\right) + \sum_{k=1}^{N} \ln\left(1 + \gamma_k^{-1} e^{-t b_k (1+\gamma_k)}\right) \ . \tag{8}$$

The expression of φ_α is derived from (5) with $\lambda_k = 1$. In the specific case where the coefficient $\mathbb{E}[X_k], \sigma_k$ and b_k are identical for all $k \in \{1 \dots N\}$, the minimization in t that appears in (5) has an analytic solution (using Kullback-Leibler divergence), see e.g. [13, 33]. In the framework of this paper, we allow heterogeneous parameters, and therefore the minimum is no longer analytically known. Nonetheless, the following properties show that the one-dimensional minimization is well defined:

Proposition 2.4 (Study of φ_α)**.** *Let* $\alpha \geq 0$*, then* $\varphi_\alpha(0) = 0$*. Moreover, the mapping* φ_α *is twice differentiable and their respective derivatives are*

(i) $$\frac{d}{dt}\varphi_\alpha(t) = N\left(\overline{b} - \alpha\right) - \sum_{k=1}^{N} \frac{b_k(1+\gamma_k)}{1 + \gamma_k e^{t b_k (1+\gamma_k)}} \ ,$$

(ii) $$\frac{d^2}{dt^2}\varphi_\alpha(t) = \sum_{k=1}^{N} b_k^2 (1+\gamma_k)^2 \frac{\gamma_k e^{t b_k (1+\gamma_k)}}{\left(1 + \gamma_k e^{t b_k (1+\gamma_k)}\right)^2} \ .$$
Moreover, $0 \leq \frac{d^2}{dt^2}\varphi_\alpha(t) \leq M := \frac{1}{2}\sum_{k=1}^{N} b_k^2 (1+\gamma_k)^2$.

Proof. Let us introduce for all $\gamma, d \in \mathbb{R}_+^*$, $f(t) = \ln\left(1 + \gamma^{-1} e^{-td}\right)$. It follows that $f'(t) = \frac{-d}{1+\gamma e^{td}}$ and $f''(t) = d^2 \frac{\gamma e^{td}}{\left(1+\gamma e^{td}\right)^2} \in \left[0, \frac{d^2}{2}\right]$. To recover the result, note that $\varphi_\alpha(t)$ is the sum of functions $f(\cdot)$ with

$d_k = b_k(1+\gamma_k)$. □

We immediately deduce from Proposition 2.4 that the function $\varphi_\alpha(\cdot)$ is strictly convex, and thus the position of the minimum, denoted by $t^*_\alpha \in \mathbb{R}_+ \cup \{+\infty\}$, is unique. The following lemma lists useful properties of φ^*_α that will be used in the sequel.

Lemma 2.5 (Study of φ^*_α)**.**

(i) For $\alpha \leq \overline{b} := \frac{1}{N}\sum_{k=1}^N b_k$, the function $\alpha \mapsto \varphi^*_\alpha$ is decreasing.
Moreover, $\varphi^*_0 = 0$, $\varphi^*_{\overline{b}} = \ln(\tau^-)$ and for $\alpha > \overline{b}$, $\varphi^*_\alpha = -\infty$.

(ii) For $\alpha < \min_k\{b_k\}$, $t^*_\alpha \in \mathrm{Conv}\left(\left\{t^{(k)}_\alpha\right\}_k\right)$, where $t^{(k)}_\alpha = \frac{1}{b_k(1+\gamma_k)}\ln\left(\frac{\alpha+b_k\gamma_k}{\gamma_k(b_k-\alpha)}\right)$.
Moreover, $t^*_\alpha \leq -\frac{1}{N(\overline{b}-\alpha)}\ln(\tau^-)$.

(iii) For $\alpha_1, \alpha_2 < \overline{b}$, $\left|\varphi^*_{\alpha_2} - \varphi^*_{\alpha_1}\right| \geq N\min\{t^*_{\alpha_1}, t^*_{\alpha_2}\}\,|\alpha_2 - \alpha_1|$.

Proof. (i) Let $\alpha < \beta$. There exits t^*_α such that $\varphi^*_\alpha = \varphi_\alpha(t^*_\alpha)$. Besides, $\varphi_\alpha(t^*_\alpha) = Nt^*_\alpha(\beta-\alpha) + \varphi_\beta(t^*_\alpha) > \varphi_\beta(t^*_\alpha)$. As $\varphi^*_\beta \leq \varphi_\beta(t^*_\alpha)$ by optimality, we easily conclude.
Then, if $\alpha = \overline{b}$, the infimum is reached for $t \to \infty$ and is equal to $\ln(\tau^-)$. If now $\alpha = 0$, then the minimum is attained at $t = 0$ ($\frac{d\varphi_\alpha}{dt}(t) \geq 0$). Finally, if $\alpha > \overline{b}$, then all the terms that appear in φ^*_α are decreasing, and the function diverges to $-\infty$.

(ii) $t^{(k)}_\alpha$ would the minimum if there were only the k-term in φ. As the minimum of a sum of convex functions lies in the convex envelope of the set of minimizers of each term, we get the property. Besides, by (i), $\varphi_\alpha(t^*_\alpha) \leq \varphi_\alpha(0) = 0$. Therefore, $\ln(\tau^-) + Nt^*_\alpha(\overline{b}-\alpha) \leq 0$, and so $Nt^*_\alpha \leq \frac{-1}{\overline{b}-\alpha}\ln(\tau^-)$.

(iii) Let $\alpha_1, \alpha_2 \leq \overline{b}$. Then, by definition,

$$\varphi^*_{\alpha_1} = \varphi_{\alpha_1}(t^*_{\alpha_1}) = Nt^*_{\alpha_1}(\alpha_2-\alpha_1) + \varphi_{\alpha_2}(t^*_{\alpha_1})$$
$$\varphi^*_{\alpha_2} = \varphi_{\alpha_2}(t^*_{\alpha_2}) = Nt^*_{\alpha_2}(\alpha_1-\alpha_2) + \varphi_{\alpha_1}(t^*_{\alpha_2})$$

from which we deduce by optimality of $t^*_{\alpha_1}$ and $t^*_{\alpha_2}$:

$$\varphi^*_{\alpha_1} \geq Nt^*_{\alpha_1}(\alpha_2-\alpha_1) + \varphi^*_{\alpha_2}$$
$$\varphi^*_{\alpha_2} \geq Nt^*_{\alpha_2}(\alpha_1-\alpha_2) + \varphi^*_{\alpha_1}$$

As a consequence, $\left|\varphi^*_{\alpha_2} - \varphi^*_{\alpha_1}\right| \geq N\min\{t^*_{\alpha_1}, t^*_{\alpha_2}\}\,|\alpha_2-\alpha_1|$.

□

The next theorem can be directly derived from Lemma 2.5 and provides an alternative confidence bound $\alpha_\tau N$ to the bound d_τ provided in Proposition 2.1 and Proposition 2.2.

Theorem 2.6. *For all $\tau \in [\tau^-, 1[$, there exists a unique α_τ such that $\varphi^*_{\alpha_\tau} = \ln(\tau)$. As a consequence, $\mathbb{P}\left[\sum_{k=1}^N X_k - \mathbb{E}[X_k] \geq \alpha_\tau N\right] \leq \tau$.*

Numerical experiments. We aim to numerically compare the bounds developed in Section 2 with four inequalities: Hoeffding (3), Bennett (4), Cantelli (a one-sided improvement of Chebyshev's inequality, see e.g. [8]) and the bound introduced by Jebara [21]. To this purpose, we follow the

$\mathbb{E}[X_k]$	$\mathcal{U}(0,1)$
a_k	$\mathcal{U}(-1,0)$
b_k	$\mathcal{U}(0,1)$
σ_k	$\mathcal{U}\left(0,\left(b_k-a_k\right)/2\right)$
α	$\mathcal{U}\left(0,\bar{b}\right)$

Table 1: Definition of the random variables

methodology of [21]: we search to bound $\ln\ \mathbb{P}\left[\sum_{k=1}^{N} X_k - \mathbb{E}[X_k] \geq \alpha N\right]$, where the parameters $\mathbb{E}[X_k], \sigma_k, a_k, b_k$ and α are randomly generated following the rules described in Table 1.

In order to have a fast implementation of φ_α^*, we introduce a bisection algorithm, see Algorithm 1. Note that this bisection method is only valid because we have shown that φ_α is convex and t_α^* is bounded, see Lemma 2.5. The four other bounds are immediate to compute as they are analytically known.

Algorithm 1 Bisection Search to compute φ_α^*

Require: N, α, b_k, σ_k, ϵ_t
  $t^-, t^+ \leftarrow 0, -\frac{1}{N(\bar{b}-\alpha)}\ln(\tau^-)$
  while $t^+ - t^- > \epsilon_t$ **do**
    $\hat{t} \leftarrow \frac{1}{2}(t^- + t^+)$
    $g \leftarrow \frac{d}{dt}\varphi_\alpha(\hat{t})$
    if $g \geq 0$ **then** $t^+ \leftarrow \hat{t}$ **else** $t^- \leftarrow \hat{t}$
  return $\hat{\varphi}$

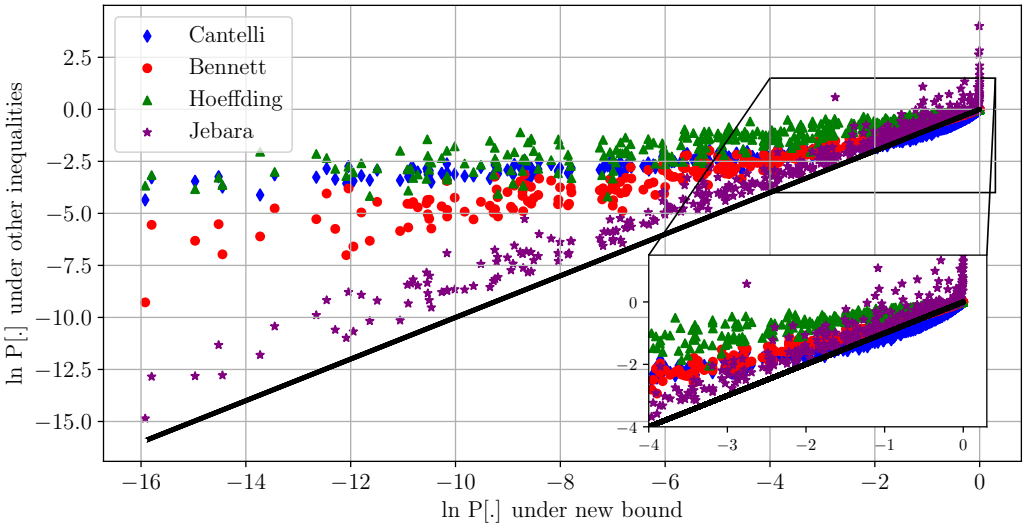
The result are depicted in Figure 2 for 500 realizations and are performed on a laptop `Intel Core i7 @2.20GHz` $\times$ `12`. The log-probability of the four methods are represented on the y-axis for each value of φ_α^*, represented on the x-axis. We recover the results proved in Appendix A.1: φ_α^* always outperforms Bennett, Hoeffding and Jebara's inequalities. We observe that Cantelli's bound is better for large probability error (small α) – typically exp $\varphi_\alpha^* \geq 20\%$ – but becomes rapidly dominated by the four other Chernoff's inequalities. In fact, Chebyshev's inequality has a quadratic decay in α when the Chernoff's bounds has exponential behaviors . For $\varphi_\alpha^* \geq -5$, the bound from [21] may be less efficient. Possibly, this bound can exceed 1, because the minimizer that is used in the Chernoff's inequality has no guarantee to be optimal.

The computation of the Bennett's and Hoeffding's bounds is almost immediate. The refined version of [21] takes around 1ms per instance for $N = 100$ (due to the computation of Lambert function), and φ_α^* takes around 5ms per instance for $N = 100$ for precision $\epsilon_t =$ 1e-6.

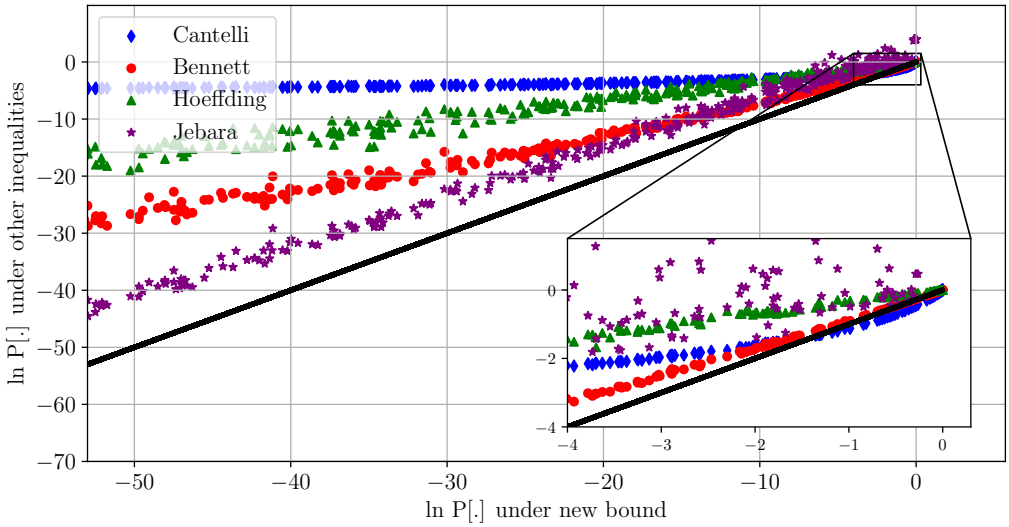
Remark 2.1. *We did not display the results for Bernstein's bound, as it is known that this inequality is strictly looser than Bennett's inequality [5], see e.g. [21] for a proof.*

3 Computing confidence bounds

In this section, we aim to derive a confidence bound α_τ with a given maximum probability error τ such that $\varphi_{\alpha_\tau}^* = \ln(\tau)$. We first prove that the confidence bound we obtain always provides useful information, as α_τ is strictly lower than $\bar{b}$ for $\tau \geq \tau^-$ (for probability error less than τ^-, we can only certify a confidence bound of $\bar{b}$) :



(a) Random instances, made of $N = 10$ heterogeneous variables



(b) Random instances, made of $N = 100$ heterogeneous variables

Figure 2: Comparison of four bounds with φ_α^*.
If the marker is above the line, then the method gives looser bound compared to φ_α^*.

Lemma 3.1. *Suppose that* $\tau \in]\tau^-, 1]$, *then*

$$\alpha_\tau \leq \bar{b} - \sqrt{\frac{1}{N\Gamma} \ln(\tau/\tau^-)} \ ,$$

where $\Gamma = 1 + (\min\{\gamma_k\} \min\{b_k(\gamma_k + 1)\})^{-1}$.

Proof. Let $\alpha_\kappa = \overline{b} - \kappa$ et $t = \frac{1}{\sqrt{\kappa}}$. Then,

$$\begin{aligned}
\varphi_{\alpha_\kappa}(t) &= \ln(\tau^-) + N\sqrt{\kappa} + \sum_{k=1}^{N} \ln\left(1 + \gamma_k^{-1} e^{-\frac{1}{\sqrt{\kappa}} b_k(\gamma_k+1)}\right) \\
&\leq \ln(\tau^-) + N\sqrt{\kappa} + \sum_{k=1}^{N} \gamma_k^{-1} e^{-\frac{1}{\sqrt{\kappa}} b_k(\gamma_k+1)} \\
&\leq \ln(\tau^-) + N\left[\sqrt{\kappa} + (\min\{\gamma_k\})^{-1} e^{-\frac{1}{\sqrt{\kappa}} \min\{b_k(\gamma_k+1)\}}\right] \\
&\leq \ln(\tau^-) + N\left[\sqrt{\kappa} + \frac{(\min\{\gamma_k\})^{-1}\sqrt{\kappa}}{\sqrt{\kappa} + \min\{b_k(\gamma_k+1)\}}\right] \leq \ln(\tau^-) + N\Gamma\sqrt{\kappa} \ .
\end{aligned}$$

The last inequality is obtained using $e^{-x} \leq (1+x)^{-1}$ for $x > 0$. Therefore, for $\kappa^2 \leq \frac{1}{N\Gamma}\ln(\tau/\tau^-)$, $\varphi_{\alpha_\kappa}(t) \leq \ln(\tau)$ and as a consequence $\varphi^*_{\alpha_\kappa} \leq \ln(\tau) = \varphi^*_{\alpha_\tau}$. As $\alpha \mapsto \varphi^*_\alpha$ is decreasing, we obtain that $\alpha_\tau \leq \alpha_\kappa$. □

Note that, under its apparent simplicity, this property does not hold for other bound such that Hoeffding or Bennett.

Double bisection search algorithm. We now present a fast algorithm to compute α_τ introduced in Theorem 2.6. Algorithm 2 consists of two nested bisection searches. The inner one is dedicated to find the minimum in t – see Algorithm 1 – to an arbitrary precision ϵ_t and, as a consequence, to compute φ^* to a precision $M\epsilon_t^2$. This estimation of φ^* constitutes the oracle for the outer bisection search. Therefore, the test is more elaborated as it checks whether the decision is sure or not: we only reduce the space by half when the oracle returns value far enough from the target $\ln(\tau)$ i.e., with a distance greater than $M\epsilon_t^2$. If not, then, it means that we obtain at a certain iteration an estimation close enough to $\ln(\tau)$, so we stop at this point. This outer bisection search is a particular case of Probabilistic Bisection Algorithm (PBA) [20, 38], where the error term is not necessarily of zero mean but takes values in a small bounded interval.

Algorithm 2 Double Bisection Search for confidence bound's computation

Require: τ, N, b_k, σ_k, ϵ_t, ϵ_α
  $\alpha^-,\ \alpha^+ \leftarrow 0,\ \overline{b} - \sqrt{\ln(\tau/\tau^-)\,/\,(N\Gamma)}$ ▷ Init α-bisection
  $tol \leftarrow$ `false`
  while $\alpha^+ - \alpha^- > \epsilon_\alpha$ or $tol =$ `false` **do**
    $\hat{\alpha} \leftarrow \frac{1}{2}(\alpha^- + \alpha^+)$
    $t^-, t^+ \leftarrow 0, -\frac{1}{N(\overline{b}-\hat{\alpha})}\ln(\tau^-)$ ▷ Init t-bisection
    while $t^+ - t^- > \epsilon_t$ **do**
      $\hat{t} \leftarrow \frac{1}{2}(t^- + t^+)$
      $g \leftarrow \frac{d}{dt}\varphi_\alpha(\hat{t})$
      if $g \geq 0$ **then** $t^+ \leftarrow \hat{t}$ **else** $t^- \leftarrow \hat{t}$
    $\hat{\varphi} \leftarrow \varphi_{\hat{\alpha}}(\hat{t})$
    if $\hat{\varphi} > \ln(\tau) + M\epsilon_t^2$ **then** $\alpha^- \leftarrow \hat{\alpha}$
    else if $\hat{\varphi} < \ln(\tau) - M\epsilon_t^2$ **then** $\alpha^+ \leftarrow \hat{\alpha}$
    else $tol \leftarrow$ `true`
  return $\hat{\alpha}$

Termination guarantees. The following proposition proves that this algorithm is fast (log convergence) and provides a solution arbitrary close to the optimal solution.

Theorem 3.2. *Let $\tau \in]\tau^-, 1]$. Algorithm 2 ends with a value $\hat{\alpha}$ such that*

$$|\hat{\alpha} - \alpha_\tau| \leq \epsilon_\alpha \wedge \sqrt{\frac{2M}{N \min\{m_k\}}}\epsilon_t \wedge \frac{2M}{N \min\{b_k m_k\}}\epsilon_t^2 \ , \tag{9}$$

where $m_k := \frac{\ln(2+\gamma_k^{-1})}{b_k^2(1+\gamma_k)}$. Moreover, the total number of iterations I_τ is bounded:

$$I_\tau \leq \left\lceil \log_2\left(\frac{\bar{b}}{\epsilon_\alpha}\right)\right\rceil \left\lceil \log_2\left(\frac{\sqrt{\Gamma}\ln(1/\tau^-)}{\epsilon_t\sqrt{N\ln(\tau/\tau^-)}}\right)\right\rceil \ .$$

Proof. At the end of the algorithm, one obtain from the inner bisection that $t^- \leq t^*_{\hat{\alpha}}, \hat{t} \leq t^+$ and $|t^+ - t^-| \leq \epsilon_t$. Suppose that the algorithm ends with a value $\hat{\alpha}$ and $\hat{\varphi} = \varphi_{\hat{\alpha}}(\hat{t})$. Then, from the mean-value theorem, there exists $t \in [t^-, t^+]$ such that

$$\left|\frac{d}{dt}\varphi_\alpha(t^-) - \frac{d}{dt}\varphi_\alpha(t^+)\right| = \left|\frac{d^2}{dt^2}\varphi_\alpha(t)\right|(t^+ - t^-) \leq M\epsilon_t \ .$$

As the derivative of φ_α is decreasing and positive (resp. negative) in t^- (resp. t^+), $|\frac{d}{dt}\varphi_\alpha(t)| \leq M\epsilon_t$ for all $t \in [t^-, t^+]$, and using once again the mean-value theorem,

$$|\hat{\varphi} - \varphi^*_{\hat{\alpha}}| \leq M\epsilon_t^2 \ .$$

1st case: the algorithm ends with $tol = \texttt{true}$.
Therefore (criteria), $\left|\hat{\varphi} - \varphi^*_{\alpha_\tau}\right| \leq M\epsilon_t^2$, and so $\left|\varphi^*_{\hat{\alpha}} - \varphi^*_{\alpha_\tau}\right| \leq 2M\epsilon_t^2$. Using Lemma 2.5, item (iii), we obtain

$$|\hat{\alpha} - \alpha_\tau| \leq \frac{2M\epsilon_t^2}{N \min\{t^*_{\alpha_\tau}, t^*_{\hat{\alpha}}\}} \ .$$

Then, using Lemma 2.5, item (ii), for all α,

$$t^*_\alpha \geq \min_{k|b_k>\alpha}\left\{\frac{1}{b_k(1+\gamma_k)}\ln\left(\frac{\alpha + b_k\gamma_k}{\gamma_k(b_k - \alpha)}\right)\right\} \ .$$

By concavity, $\ln(1+x) \geq \ln(1+z)\min\{x/z, 1\}$ for all $x, z \geq 0$. Therefore, for all k such that $b_k > \alpha$ (it exists otherwise $\alpha > \bar{b}$), we obtain:

$$\ln\left(\frac{\alpha + b_k\gamma_k}{\gamma_k(b_k - \alpha)}\right) = \ln\left(1 + \frac{\alpha(1+\gamma_k)}{\gamma_k(b_k - \alpha)}\right) \geq \ln\left(2 + \frac{1}{\gamma_k}\right)\min\{\frac{\alpha}{b_k - \alpha}, 1\} \geq \frac{1}{b_k}\ln\left(2 + \frac{1}{\gamma_k}\right)\min\{\alpha, b_k\} \ .$$

Then, $t^*_\alpha \geq \alpha \min_k\{m_k\} \vee \min_k\{b_k m_k\}$. Therefore, as $\hat{\alpha}$ and α_τ are positive quantities,

$$\begin{aligned} t^*_{\alpha_\tau} \vee t^*_{\hat{\alpha}} &\geq \alpha_\tau \min_k\{m_k\} \vee \hat{\alpha}\min_k\{m_k\} \vee \min_k\{b_k m_k\} \\ &\geq |\hat{\alpha} - \alpha_\tau| \min_k\{m_k\} \vee \min_k\{b_k m_k\} \ . \end{aligned}$$

Finally,

$$|\hat{\alpha} - \alpha_\tau| \leq \max\left\{\sqrt{\frac{2M}{N\min\{m_k\}}}\epsilon_t, \frac{2M}{N\min\{b_k m_k\}}\epsilon_t^2\right\} \ .$$

2nd case: the algorithm ends with $|\alpha^+ - \alpha^-| \leq \epsilon_\alpha$.
Then, as *tol* = `false`, at each iteration, $|\hat{\varphi} - \ln(\tau)| \geq M\epsilon_t^2$, and so α_τ lies in $[\alpha^-, \alpha^+]$. Therefore, $|\hat{\alpha} - \alpha_\tau| \leq |\alpha^+ - \alpha^-| \leq \epsilon_\alpha$.

Besides, denoting by I_t the number of iterations for the inner bisection search, we have

$$I_t \leq \left\lceil \log_2 \left(\frac{-\ln(\tau^-)}{N(\bar{b} - \alpha)\epsilon_t} \right) \right\rceil \quad .$$

Then, as $\bar{b} - \alpha \geq \sqrt{\frac{1}{N\Gamma} \ln(\tau/\tau^-)}$ (see Lemma 3.1), $I_t \leq \left\lceil \log_2 \left(\frac{\sqrt{\Gamma}\ln(1/\tau^-)}{\sqrt{N \ln(\tau/\tau^-)}\epsilon_t} \right) \right\rceil$. Furthermore, denoting by I_α the number of iterations for the outer bisection search, we have

$$I_\alpha \leq \left\lceil \log_2 \left(\frac{\bar{b}}{\epsilon_\alpha} \right) \right\rceil \quad .$$

□

Theorem 3.2 proves that Algorithm 2 is fast (log convergence), and provides a solution with an arbitrary precision. Note that the number of iterations is impacted by the distance of τ from the minimal value τ^-. In fact, very close to τ^-, the minimizer t_α^* tends to $+\infty$, and therefore the width of the bisection search space becomes large. Nonetheless, for reasonable error tolerance τ, the algorithm takes very few iterations. Meanwhile, the precision of $\hat{\alpha}$ does not depend on τ.

Numerical experiments. We use the instances developed in Table 1. The results are depicted in Figure 3 for 1000 realizations, and are fast to obtain (few seconds in total). Of course, the confidence bounds we obtain are greater than the value computed with normal distributions, and so all values are greater than 1. We recover the superiority of the studied bound compared to the standard Bennett's inequality. Besides, Chebyshev-Cantelli's bound is only valuable for low probability level, and becomes inefficient for probabilities close to 1.

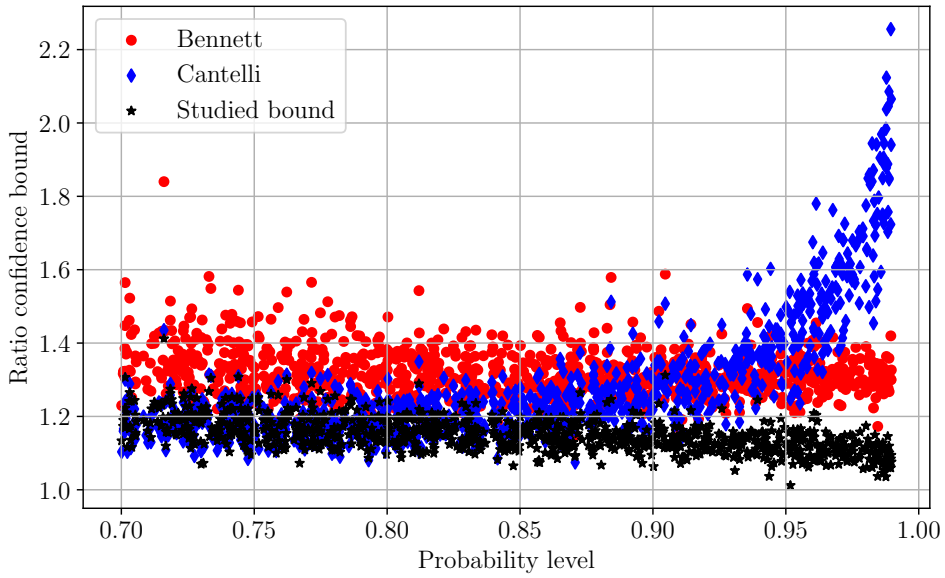


Figure 3: Random instances, made of $N = 10$ heterogeneous variables.
For different probability levels $(1 - \tau)$ and different inequalities, we display the value of the confidence bound normalized by the normal case, i.e., the (exact) value for normal distributions.

4 Application to Chance-Constrained Programming

In the two next subsections, we will use the proposed concentration inequality (5)-(6) into CCP problems of the form (2) with *individual* bilinear constraints of the form $g(x, \xi) = \xi^T x$ - b, with $b \in \mathbb{R}$. The first application (knapsack problems) is of a combinatorial nature and contains only one chance constraint whereas the second one (Support Vector Machine problems) is a continuous problem but contains as many chance constraints as training points.

4.1 Chance-constrained binary Knapsack problem

Let us consider the Knapsack problem with random weights, stated as a chance-constrained problem:

$$\begin{aligned} \max_{y \in \{0,1\}^N} \ & \pi^T y \\ \text{s.t} \quad & \mathbb{P}\left[\omega^T y \geq C\right] \leq \tau \end{aligned} \tag{CKP}$$

where $\pi \in \mathbb{R}^N$ denotes the utility of each item, $\omega \in \mathbb{R}^N$ denotes the random weights of each item, and $C \in \mathbb{R}$ is the maximum budget.

If ω_k follows a normal distribution $\mathcal{N}(\overline{\omega}_k, \sigma_k)$, the chance constraint can be computed *exactly* by the following problem, see e.g. [32, Theorem 10.4.1] :

$$\begin{aligned} \max_{y \in \{0,1\}^N} \ & \pi^T y \\ \text{s.t} \quad & \Phi^{-1}(1-\tau)\sqrt{y^T \Sigma y} + \overline{\omega}^T y \leq C \end{aligned} \tag{CKP-N}$$

where Φ is the cumulative distribution for the standard normal distribution. In this specific setting, Han et al. [16] provides efficient algorithms to obtain robust solutions.

Here, we focus on random weights whose distributions are unknown, but where the two first moments are available, as well as upper bounds. This distributionally robust approach has been firstly studied by Calafiore and El Ghaoui [9], where they focused on Hoeffding-type approximations. Recently, Ryu and Park [35] proposed to repeatedly solve ordinary binary knapsack subproblems to deduce bounds on SOCP approximations (such as Chebyshev-Cantelli). As we will compare the different approximations in the sequel, we first recall the next two classical results: let us define $B = \text{diag}(b^2)$ and $\Sigma = \text{diag}(\sigma^2)$, then

(i) (Hoeffding) the problem (CKP-H) is a valid conservative approximation of (CKP)

$$\begin{aligned} \max_{y \in \{0,1\}^N} \ & \pi^T y \\ \text{s.t} \quad & \sqrt{2\ln(1/\tau)}\sqrt{y^T B y} + \overline{\omega}^T y \leq C \end{aligned} \tag{CKP-H}$$

(ii) (Chebyshev-Cantelli) the problem (CKP-C) is a valid conservative approximation of (CKP)

$$\begin{aligned} \max_{y \in \{0,1\}^N} \ & \pi^T y \\ \text{s.t} \quad & \sqrt{\tfrac{1}{\tau} - 1}\sqrt{y^T \Sigma y} + \overline{\omega}^T y \leq C \end{aligned} \tag{CKP-C}$$

This comparison is inspired by the work of Peng, Maggioni and Lisser [28] where first-order bounds (Hoeffding and an approximation of Bernstein bound) are used in the continuous knapsack problem,

and compared to exact SOCP relaxation for normal variables.

Using Theorem 2.3 with $X_k := \omega_k$ and $d := C - \overline{\omega}^T y$, we obtain (tighter) conservative approximation. A lower bound of (CKP) is then obtained by solving:

$$\begin{array}{ll} \max\limits_{\substack{y\in\{0,1\}^N \\ t\geq 0}} & \pi^T y \\ \text{s.t} & t\left[\overline{\omega}^T y - C\right] + \sum\limits_{k=1}^{N} \ln\left(\dfrac{\gamma_k e^{y_k b_k} + e^{-t y_k b_k \gamma_k}}{1+\gamma_k}\right) \leq \ln(\tau) \end{array} \tag{$\overline{\text{CKP}}$}$$

The constraint contains bilinear terms ty_k. A naïve approach could be to consider a Fortet linearization of the bilinear terms, see e.g. [14]. Here, using the structure of the constraint, we succeed in reformulating the constraint. Let us consider the change of variable $z := 1/t$ and divide the constraint by t:

$$(\overline{\text{CKP}}) \iff \begin{cases} \max\limits_{\substack{y\in\{0,1\}^N \\ z\geq 0}} & \pi^T y \\ \text{s.t} & \overline{\omega}^T y + \sum\limits_{k=1}^{N} z \ln\left(\dfrac{\gamma_k e^{\frac{y_k}{z} b_k} + e^{-\frac{y_k}{z} b_k \gamma_k}}{1+\gamma_k}\right) \leq C + z\ln(\tau) \end{cases} \tag{10}$$

Proposition 4.1. *For every $\gamma, b \geq 0$, the function $\Psi^+_{\gamma,b} : \mathbb{R}\times\mathbb{R}_+ \to \mathbb{R}_+$, defined as,*

$$\Psi^+_{\gamma,b}(y,z) := z\ln\left(\frac{\gamma e^{\frac{y}{z}b} + e^{-\frac{y}{z}b\gamma}}{1+\gamma}\right) \ ,$$

is jointly convex. Therefore, the problem ($\overline{\text{CKP}}$) – without the integrity condition – is convex.

Proof. From elementary calculation, the Jacobian and Hessian of Ψ^+ are respectively

$$J_{\Psi^+_{\gamma,b}}(y,z) = \left[b - \frac{b(1+\gamma)}{1+\gamma e^{\frac{y}{z}b(1+\gamma)}} \quad \ln\left(\frac{\gamma + e^{-\frac{y}{z}b(1+\gamma)}}{1+\gamma}\right) + \frac{b(1+\gamma)y}{z\left(1+\gamma e^{\frac{y}{z}b(1+\gamma)}\right)}\right]$$

$$H_{\Psi^+_{\gamma,b}}(y,z) = \gamma e^{\frac{y}{z}b(1+\gamma)}\left(\frac{b(1+\gamma)}{1+\gamma e^{\frac{y}{z}b(1+\gamma)}}\right)^2 \begin{bmatrix} \frac{1}{z} & -\frac{y}{z^2} \\ -\frac{y}{z^2} & \frac{y^2}{z^3} \end{bmatrix}$$

As $\text{Tr}\left(H_{\Psi_{\gamma,b}}(y,z)\right) \geq 0$ and $\det\left(H_{\Psi_{\gamma,b}}(y,z)\right) = 0$, the Hessian is always a positive semi-definite matrix, and the function is jointly convex. □

Remark 4.1. *We can directly obtain the convexity of the function by noting that $\Psi^+_{\gamma,b}$ is the composition of the perspective function [12] with a particular log-sum-exp function. This proposition is generalized in [27] to a class of moment-generating function's estimators. Nonetheless, we make explicit the Jacobian and the Hessian of the function, since it will be necessary for numerical optimization.*

By Proposition 4.1, Problem (CKP) reduces to a convex Mixed-Integer Non-Linear Programming (MINLP) problem. In fact, there is a unique nonlinear constraint (the budget constraint). For such problems with a moderate degree of nonlinearity, cutting-plane methods [40] are known to be efficient. In this approach, the nonlinear constraint is described by a set of linear constraints, incrementally built by adding at each iteration a new cutting-plane of the original constraint. The subproblem at

iteration j is then expressed as:

$$\left(y^{(j+1)}, z^{(j+1)}\right) = \underset{\substack{y\in\{0,1\}^N \\ z\geq 0}}{\arg\max} \left\{ \pi^T y \;\middle|\; \overline{\omega}^T y + \sum_{k=1}^{N} \left\langle s_k^{(i)}, \begin{pmatrix} y_k \\ z \end{pmatrix} \right\rangle \leq C + z\ln(\tau),\ 1 \leq i \leq j \right\} \quad , \tag{11}$$

where $s_k^{(i)} := \nabla\Psi^+_{\gamma_k,b_k}\left(y_k^{(i)}, z^{(i)}\right)$. The convergence of this approach has been proved, see e.g. [7, Theorem 9.6], and is achieved in a finite number of steps for this particular problem (there is a finite number of knapsack-filling scenarios). Note that the cuts are dynamically added in the branch-and-bound at each node where an integer solution is found (*Lazy constraint*), so that the solver does not need to perform a complete MILP solving at each iteration.

As an alternative to the model ($\overline{\text{CKP}}$), we also introduce a convex reformulation of the problem under Bernstein's inequality: let suppose that $|\omega_k - \overline{\omega}_k| \leq b_k$ (and not only $\omega_k - \overline{\omega}_k \leq b_k$), then a valid conservative estimation of (CKP) can be obtained by replacing the probabilistic constraint by Bernstein's inequality (see e.g. [8] for more details on this inequality). We obtain the following formulation:

$$\begin{aligned} \max_{y\in\{0,1\}^N} \ & \pi^T y \\ \text{s.t} \quad & \exp\left(-\frac{\frac{1}{2}(C-\overline{\omega}^T y)^2}{\sum_{i=1}^N y_k^2\sigma_k^2 + \frac{1}{3}z(C-\overline{\omega}^T y)}\right) \leq \tau \\ & z \geq b_k y_k,\ 1 \leq k \leq N \end{aligned} \tag{CKP-B}$$

Proposition 4.2. *Problem* (CKP-B) *is equivalent with the following problem:*

$$\begin{aligned} \max_{y\in\{0,1\}^N, z\geq 0} \ & \pi^T y \\ s.t \quad & \tfrac{1}{3}\ln(1/\tau)z + \sqrt{(y^T\ z)\,\Gamma\begin{pmatrix} y \\ z \end{pmatrix}} + \overline{\omega}^T y \leq C \\ & z \geq b_k y_k,\ 1 \leq k \leq N \end{aligned} \tag{12}$$

where $\Gamma = diag\begin{bmatrix} (2\ln(1/\tau)\sigma_k^2)_{1\leq k\leq N} \\ \frac{1}{9}\ln(1/\tau)^2 \end{bmatrix}$. *Therefore, problem* (CKP-B) *– without the integrity condition – is convex.*

Proof. We reformulate the constraint so that we end up with a convex reformulation:

$$\begin{aligned} & \exp\left(-\frac{\frac{1}{2}t^2}{\sum_{i=1}^N y_k^2\sigma_k^2 + \frac{1}{3}zt}\right) \leq \tau, \quad t = C - \overline{\omega}^T y,\ z = \max_k\{b_k y_k\} \\ \iff & \ln(1/\tau)\left[\sum_{i=1}^N y_k^2\sigma_k^2 + \tfrac{1}{3}zt\right] \leq \frac{1}{2}t^2 \\ \iff & \ln(1/\tau)\sum_{i=1}^N y_k^2\sigma_k^2 \leq \frac{1}{2}\left[t - \tfrac{1}{3}\ln(1/\tau)z\right]^2 - \tfrac{1}{18}\ln(1/\tau)^2 z^2 \\ \iff & \sqrt{2\ln(1/\tau)\sum_{i=1}^n y_k^2\sigma_k^2 + \tfrac{1}{9}\ln(1/\tau)^2 z^2} \leq C - \overline{\omega}^T y - \tfrac{1}{3}\ln(1/\tau)z \\ \iff & \sqrt{\begin{pmatrix} y^T & z \end{pmatrix}\Gamma\begin{pmatrix} y \\ z \end{pmatrix}} \leq C - \overline{\omega}^T y - \tfrac{1}{3}\ln(1/\tau)z \end{aligned}$$

Note that, in the optimization problem, it is sufficient to consider $z \geq \max_k\{b_k y_k\}$ as the optimization will search for the lowest value possible (z only appears on the constraint above), ans so the constraint will be naturally saturated. □

The SOCP formulation introduced in Proposition 4.2 provides an alternative conservative approximation, which can be directly compared to the classical Chebyshev approximation, as they both belongs to the same class of problem. In contrast, the formulation ($\overline{\text{CKP}}$) is not expressed as a cone programming, but we provide in Appendix B a reformulation where constraints are expressed via exponential cones.

Remark 4.2. *We already know (proved theoretically above and highlighted by Figure 2) that* ($\overline{\text{CKP}}$) *gives better solution than all other formulations (apart from exact one in the case of normally distributed weights), as the set of admissible solutions is larger. Note also that we did not provide optimization model for the bound developed in [21] and [5], as a convex expression of the chance constraint is all but immediate to obtain (if it exists).*

Numerical results. In order to obtain chance-constrained instances, we adapted deterministic instances from the literature[1], see [30], by adding a maximum standard deviation of 5% of the original weight (taken as mean value), and setting the maximum value to $b = 5\sigma$. Note that for normal distribution, the probability of exceeding $\overline{\omega} + 3\sigma$ is 0.997. Finally, the maximum probability error τ is taken to 3%.

We use `Cplex v12.10` as a MILP solver and the tests are performed on a laptop `Intel Core i7 @2.20GHz` $\times$ `12`. The MIP gap tolerance is taken to 0.001% and the Integrity tolerance to 1e-8. The tests shows that the cutting-plane method adds very few cuts. For instance, the solver added 190 cuts for the instance `1_10000`.

Instance	KP	(CKP-N)	($\overline{\text{CKP}}$)	Prob.	Time	(CKP-B)	(CKP-C)	(CKP-H)
1_100	9147	8842	8817	0.19	0.1	8719	8817	8150
1_200	11238	11227	10962	0.81	0.1	10682	10832	10353
1_500	28857	28606	28405	2.11	0.4	28152	28127	27924
1_1000	54503	54105	53836	1.58	0.65	53617	53267	52109
1_2000	110625	110130	109779	2.95	1.8	109621	109148	107228
1_5000	276457	275685	275220	2.99	33.4	275068	274151	*271160*
1_10000	563647	562560	561968	3.00	97.4	561809	560387	*556126*
2_100	1514	1513	1512	0.82	0.1	1456	1476	1395
2_200	1634	1619	1594	0.69	0.2	1558	1592	1508
2_500	4566	4537	4504	2.31	0.5	4472	4472	4348
2_1000	9052	9008	8970	2.87	1.52	8951	8927	8761
2_2000	18051	17991	17946	2.85	4.0	17925	17872	*17635*
2_5000	44356	44262	44201	2.86	32.7	44184	44073	*43696*
2_10000	90204	90071	89996	2.99	84.2	89975	89807	*89265*

Table 2: Results for knapsack instances. We compare the two new formulations ($\overline{\text{CKP}}$) and (CKP-B) to the existing methods (CKP-H) and (CKP-C). The method KP corresponds to the deterministic case, and (CKP-N) corresponds to a normally-distributed uncertainty. For ($\overline{\text{CKP}}$), we also provide the probability error and the computational time. When the objective is in italic, the solver does not succeed to prove the optimality in the given time.

[1]The instances are extracted from the website `http://artemisa.unicauca.edu.co/~johnyortega/instances_01_KP/` We use the set of instances `knapPI_{X}_1000_1` where `X` goes from `1_100` to `2_10000` (the second number stands for the number of items in the instances), see Table 2.

The numerical tests show the efficiency of the proposed relaxation: the use of the second-order information leads to a substantial improvement of the optimal objective-function value, compared to the classical Hoeffding bound. Besides, this method appears to be easy tractable, as we were able to solve instances of 10000 items in less than two minutes. Note that the relaxation seems to be a bit more tractable than the Hoeffding bound, as the solver cannot prove the optimality of the solution with the desired precision in less than 10 minutes.

Remark 4.3. *We present here the results for mixed-integer problems, but the results and the methodology does not exploit the integrity condition of the variables, and so the results and the methodology are still applicable on the (simpler) continuous problem. In particular, a cutting-plane approach still converges.*

4.2 Distributionally Robust Support Vector Machine problem

Let us consider a dataset of M points $\{x_i, l_i\}$ where each point x_i is a vector of $\mathbb{R}^N$. The points lies into two classes, indexed by labels $l_i \in \{-1, 1\}$. The chance-constrained formulation of the Support Vector Machine (SVM) problem (with soft margin) is defined as follows:

$$\begin{aligned} \min_{w\in\mathbb{R}^N, w_0\in\mathbb{R}, \xi\in\mathbb{R}^M_+} \quad & \frac{1}{2}\|w\|_2^2 + C\sum_{i=1}^M \xi_i \\ \text{s.t.} \quad & \mathbb{P}\left[l_i(w^T x_i + w_0) \le 1 - \xi_i\right] \le \tau_i \\ & \xi_i \ge 0,\ i = 1, \dots, M \end{aligned} \tag{SVM – CCP}$$

This robust version has been recently studied, see e.g. [39, 23]. Here, we focus on independent noises for each points as in [3]. In contrast with the knapsack problem (CKP), the random training points are multiplied by w, which can be either positive or negative. Therefore, we can no longer apply (5) and must use (6) which contains absolute values. Moreover, each training feature (point) defines a (nonlinear) chance constraint.

Proposition 4.3. *Let* $(\overline{\text{SVM}})$ *be defined as*

$$\begin{aligned} \min_{w\in\mathbb{R}^N, w_0\in\mathbb{R}, \xi\in\mathbb{R}^M_+} \quad & \frac{1}{2}\|w\|_2^2 + C\sum_{i=1}^M \xi_i \\ s.t \quad & -l_i(w_0 + w^T\overline{x}_i) + \sum_{k=1}^N \Psi_{\gamma_{ik},b_{ik}}(w_k, z_i) \le \xi_i - 1 + z_i\ln(\tau_i),\ 1 \le i \le M \end{aligned} \tag{$\overline{\text{SVM}}$}$$

where $\Psi_{\gamma,b} : \mathbb{R}\times\mathbb{R}_+ \to \mathbb{R}_+$ *is defined as* $\Psi_{\gamma,b}(y,z) = \Psi^+_{\gamma,b}(|y|,z)$. *This problem is a valid conservative approximation of* (SVM – CCP).

Proof. We follow the similar steps as for the knapsack case. In particular, we use Theorem 2.3 for each chance constraint $i \in \{1, \dots, M\}$ with $X = x$, $\lambda = -l_i w$ and $d = \xi_i - 1 + l - i(w_0 + \overline{x})$. Then, we apply the chance of variable $z = 1/t$. □

The following proposition shows that the function Ψ keeps the same regularity as Ψ^+:

Proposition 4.4. *The function $\Psi_{\gamma,b}$ is convex and twice continuously differentiable. Moreover,*

$$\Psi_{\gamma,b}(y,z) = \begin{cases} \Psi^+_{\gamma,b}(y,z), & y \geq 0 \\ \Psi^+_{\frac{1}{\gamma},b\gamma}(y,z), & y \leq 0 \end{cases} \tag{13}$$

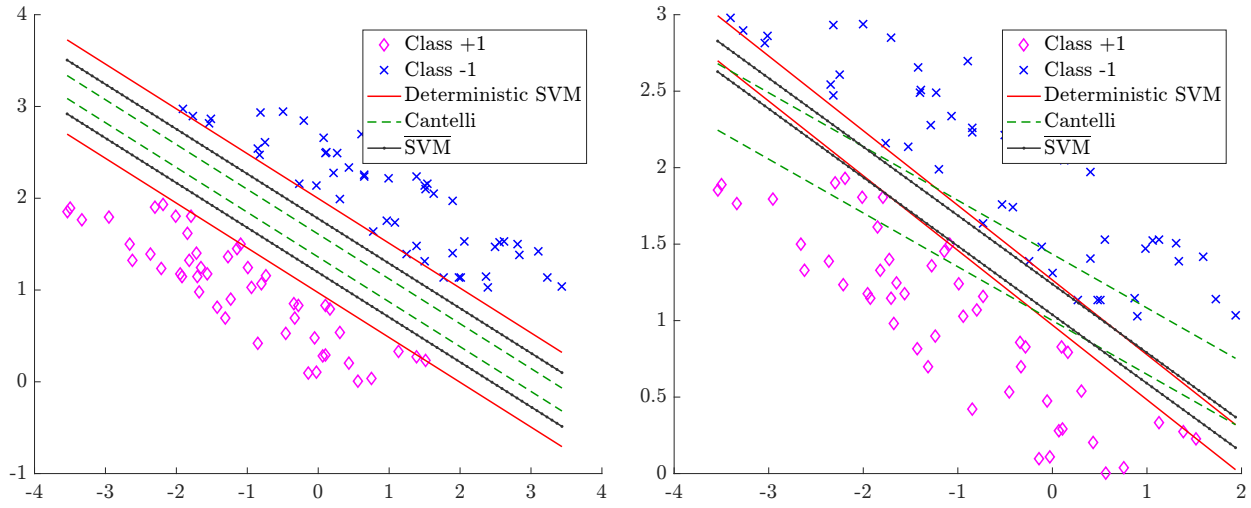
As a consequence, problem ($\overline{\text{SVM}}$) *is a* convex *conservative approximation of* (SVM – CCP).

Proof. We have the following direct equalities:

$$\Psi^+_{\gamma,b}(-y,z) = z \ln\left(\frac{e^{-\frac{y}{z}b} + \gamma^{-1}e^{\frac{y}{z}b\gamma}}{1+\gamma^{-1}}\right) = z\ln\left(\frac{e^{-\frac{y}{z}(b\gamma)\gamma^{-1}} + \gamma^{-1}e^{\frac{y}{z}(b\gamma)}}{1+\gamma^{-1}}\right) = \Psi^+_{\gamma^{-1},b\gamma}(y,z) \ .$$

Furthermore, to check the regularity property, it suffices to verify the condition in $y=0$. $\nabla\Psi^+_{\gamma,b}(0,z) = 0$ for all γ and b, so $\nabla\Psi_{\gamma,b}(0^-,z) = \nabla\Psi_{\gamma,b}(0^+,z) = 0$. Moreover, $H_{\Psi^+_{\gamma,b}}(0,z) = \begin{bmatrix} b^2\gamma & 0 \\ 0 & 0 \end{bmatrix} = H_{\Psi^+_{\gamma^{-1},b\gamma}}(0,z)$. Therefore, $\Psi_{\gamma,b}$ is twice continuously differentiable in $y=0$. □

Numerical results. In the tests, ($\overline{\text{SVM}}$) is implemented using the interior-point nonlinear solver `IPOPT`[2]. The solver always returns the optimal solution as the problem has been proved to be convex, see Proposition 4.4. For comparison, we also implement the robust SVM approximation using Chebyshev-Cantelli inequality – see e.g. [39] – which can be efficiently solved by any SOCP solver.



(a) Robust separation
(no slack variables used)

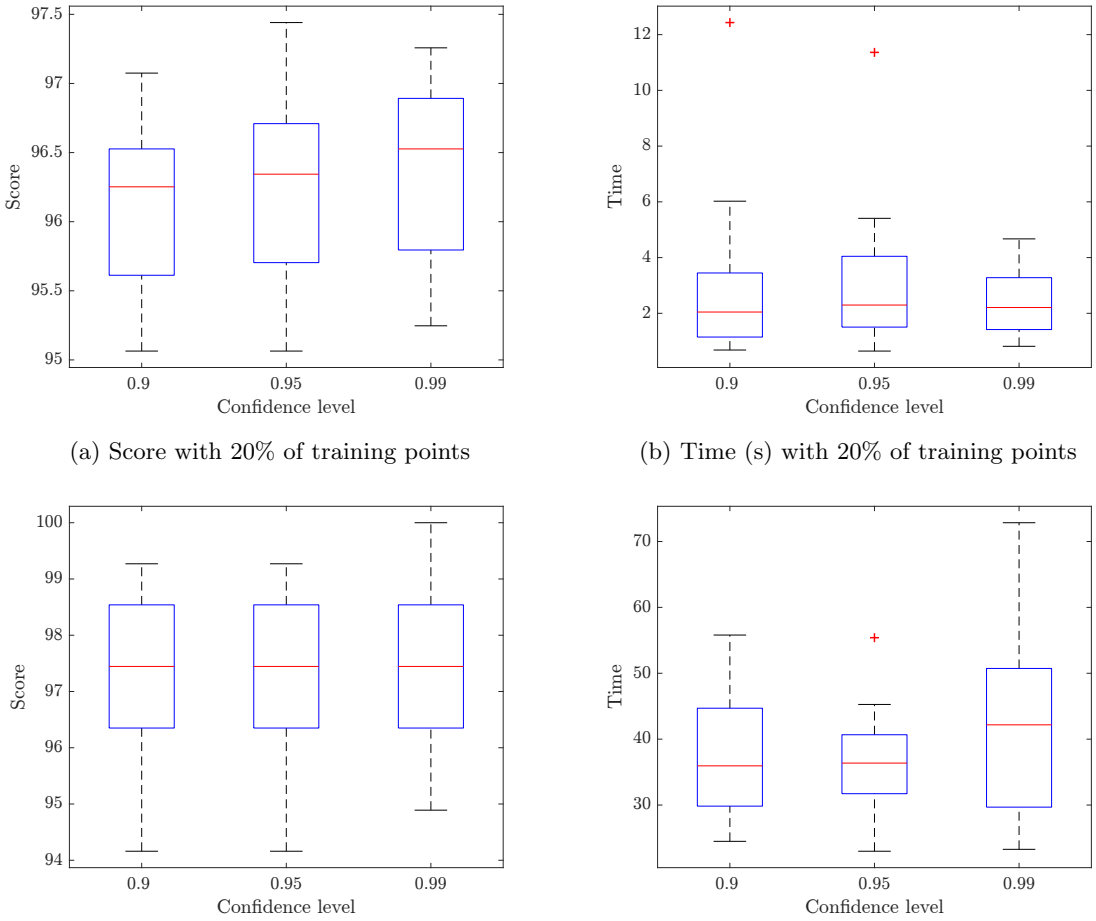
(b) Robust separation
(active slack variables for some points)

Figure 4: Two-dimensions SVM with $\tau = 0.02$, $M = 100$, $C = 100$.
We directly represent the margin around the hyperplane (centered between the two lines)

First, we construct 2D instances with linearly separable classes. To compute the standard deviation of each point, we follow the method of [39] by calculating the standard deviation of the training points for each class and then divide by 10. This appears to be reasonable as an uncertainty set for each data point. The results are displayed on Figure 4. On the left, the classes are sufficiently distant so that it is not necessary to activate slack variables ξ_i. We observe that all the methods find the same hyperplane, but differ on the size of the margin width. As expected, the Chebyshev-Cantelli'

[2] `https://coin-or.github.io/Ipopt/`

inequality is more conservative on this example. On Figure 4b, we reduce the space between the two classes. The points are still linearly separable in the deterministic setting, but are not robustly separable both for Chebyshev and for the proposed method. Nonetheless, we numerically observe that Cantelli relaxation needs to activate more slack variables, and so the optimal value is greater than the one found by the proposed method.



(a) Score with 20% of training points

(b) Time (s) with 20% of training points

(c) Score with 80% of training points

(d) Time (s) with 80% of training points

Figure 5: Two-dimensions SVM with $\tau = 0.02$, $M = 100$, $C = 100$

We then use the proposed method on instances from the literature. In particular, we use data on Breast Cancer in the Wisconsin[3]. This data set contains 683 samples of dimension 10, see e.g. [39, 23] for more information on the dataset. Figure 5 displays the time and the score (the percentage of test data that satisfies the classification obtained with the training set) for two configurations. We observe that the mean score is always higher than 96% (same order as in [39, 23]), and the time stays reasonable even for a substantial number of features (more than 500 training points, see Figures 5b and 5d).

5 Conclusion and perspectives

In this paper, we studied a refined Bennett-type inequality, originally developed in the homogeneous setting and extended here to the heterogeneous case. We have shown that this concentration inequality can be used in a wide range of applications. First, we introduce a double bisection search which

[3]https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+%28Diagnostic%29

computes (in logarithmic time) confidence bounds proved to be tighter than other classical approaches. In particular, it outperforms the standard Chebyshev's approach for high probability precision. Besides, we obtained tight distributionally robust bounds to individual Chance-constrained Programming which can be formulated as convex problem. In particular, we highlighted that the inequality can be inserted into CCP binary knapsack problem while staying tractable (instances of 10 000 binary variables). Tests on SVM problems have also been performed, obtaining a better separability of the data on instances from the literature (containing up to 500 points).

Future works will be dedicated to the extension of the results to the independent joint probability constraint case. Moreover, we think that this inequality can be helpful in many concrete applications to estimate more precisely error bounds, especially we will focus on electricity bill estimates.

References

[1] Aharon Ben-Tal, Laurent El Ghaoui, and Arkadi Nemirovski. *Robust optimization*. Vol. 28. Princeton university press, 2009.

[2] Aharon Ben-Tal and Arkadi Nemirovski. "Robust solutions of linear programming problems contaminated with uncertain data". In: *Mathematical programming* 88.3 (2000), pp. 411–424.

[3] Aharon Ben-Tal et al. "Chance constrained uncertain classification via robust optimization". In: *Mathematical Programming* 127.1 (Oct. 2010), pp. 145–173. DOI: 10.1007/s10107-010-0415-1.

[4] George Bennett. "A one-sided probability inequality for the sum of independent, bounded random variables". In: *Biometrika* 55.3 (1968), pp. 565–569. DOI: 10.1093/biomet/55.3.565.

[5] George Bennett. "Probability Inequalities for the Sum of Independent Random Variables". In: *Journal of the American Statistical Association* 57.297 (Mar. 1962), pp. 33–45. DOI: 10.1080/01621459.1962.10482149.

[6] John R. Birge and François Louveaux. *Introduction to Stochastic Programming*. Springer New York, 2011. DOI: 10.1007/978-1-4614-0237-4.

[7] Joseph-Frédéric Bonnans et al. *Numerical optimization: theoretical and practical aspects*. Springer Science and Business Media, 2006.

[8] Stéphane Boucheron, Gábor Lugosi, and Olivier Bousquet. "Concentration Inequalities". In: *Advanced Lectures on Machine Learning*. Springer Berlin Heidelberg, 2004, pp. 208–240. DOI: 10.1007/978-3-540-28650-9_9.

[9] G. C. Calafiore and L. El Ghaoui. "On Distributionally Robust Chance-Constrained Linear Programs". In: *Journal of Optimization Theory and Applications* 130.1 (Dec. 2006), pp. 1–22. DOI: 10.1007/s10957-006-9084-x.

[10] A. Charnes and W. Cooper. "Chance-Constrained Programming". In: *Management Science* 6 (Oct. 1959), pp. 73–79.

[11] Xueqin Cheng and Yanpeng Li. "An improved Hoeffding's inequality for sum of independent random variables". In: *Statistics and Probability Letters* 183 (Apr. 2022), p. 109349.

[12] Patrick L. Combettes. "Perspective Functions: Properties, Constructions, and Examples". In: *Set-Valued and Variational Analysis* 26.2 (Apr. 2017), pp. 247–264. DOI: 10.1007/s11228-017-0407-x.

[13] Amir Dembo and Ofer Zeitouni. *Large Deviations Techniques and Applications*. Springer Berlin Heidelberg, 2010. DOI: 10.1007/978-3-642-03311-7.

[14] R. Fortet. "Applications de l'algèbre de boole en recherche opérationelle". In: *Revue Français de Recherche Opérationelle* (1960).

[15] Steven G. From and Andrew W. Swift. "A refinement of Hoeffding's inequality". In: *Journal of Statistical Computation and Simulation* 83.5 (May 2013), pp. 977–983. DOI: 10.1080/00949655.2011.644290.

[16] Jinil Han et al. "Robust optimization approach for a chance-constrained binary knapsack problem". In: *Mathematical Programming* 157.1 (July 2015), pp. 277–296. DOI: 10.1007/s10107-015-0931-0.

[17] R. Henrion. "Structural properties of linear probabilistic constraints". In: *Optimization* 56.4 (Aug. 2007), pp. 425–440. DOI: 10.1080/02331930701421046.

[18] René Henrion. "Perturbation Analysis of Chance-constrained Programs under Variation of all Constraint Data". In: *Lecture Notes in Economics and Mathematical Systems*. Springer Berlin Heidelberg, 2004, pp. 257–274. DOI: 10.1007/978-3-642-55884-9_13.

[19] Wassily Hoeffding. "Probability Inequalities for Sums of Bounded Random Variables". In: *Journal of the American Statistical Association* 58.301 (Mar. 1963), pp. 13–30. DOI: 10.1080/01621459.1963.10500830.

[20] M. Horstein. "Sequential transmission using noiseless feedback". In: *IEEE Transactions on Information Theory* 9.3 (1963), pp. 136–143. DOI: 10.1109/TIT.1963.1057832.

[21] Tony Jebara. *A refinement of Bennett's inequality with applications to portfolio optimization*. 2018. DOI: 10.48550/ARXIV.1804.05454.

[22] Peter Kall and Stein Wallace. *Stochastic Programming*. Vol. 46. Jan. 1994. DOI: 10.2307/2584504.

[23] Rashed Khanjani-Shiraz et al. "Distributionally robust joint chance-constrained support vector machines". In: *Optimization Letters* (Mar. 2022). DOI: 10.1007/s11590-022-01873-x.

[24] Constantino M. Lagoa, Xiang Li, and Mario Sznaier. "Probabilistically Constrained Linear Programs and Risk-Adjusted Controller Design". In: *SIAM Journal on Optimization* 15.3 (Jan. 2005), pp. 938–951. DOI: 10.1137/s1052623403430099.

[25] Colin McDiarmid. "On the method of bounded differences". In: *Surveys in Combinatorics, 1989*. Cambridge University Press, Aug. 1989, pp. 148–188. DOI: 10.1017/cbo9781107359949.008.

[26] Bruce L. Miller and Harvey M. Wagner. "Chance Constrained Programming with Joint Constraints". In: *Operations Research* 13.6 (Dec. 1965), pp. 930–945. DOI: 10.1287/opre.13.6.930.

[27] Arkadi Nemirovski and Alexander Shapiro. "Convex Approximations of Chance Constrained Programs". In: *SIAM Journal on Optimization* 17.4 (Jan. 2007), pp. 969–996. DOI: 10.1137/050622328.

[28] Shen Peng, Francesca Maggioni, and Abdel Lisser. "Bounds for probabilistic programming with application to a blend planning problem". In: *European Journal of Operational Research* 297.3 (Mar. 2022), pp. 964–976. DOI: 10.1016/j.ejor.2021.09.023.

[29] Janos Pinter. "Deterministic approximations of probability inequalities". In: *Zeitschrift für Operations-Research* 33.4 (1989), pp. 219–239.

[30] David Pisinger. "Where are the hard knapsack problems?" In: *Computers and Operations Research* 32.9 (Sept. 2005), pp. 2271–2284. DOI: 10.1016/j.cor.2004.03.002.

[31] András Prékopa. "On probabilistic constrained programming". In: *Proceedings of the Princeton Symposium on Mathematical Programming*. Princeton University Press, 1970, pp. 113–138. DOI: 10.1515/9781400869930-009.

[32] András Prékopa. *Stochastic Programming*. Springer Netherlands, 1995. DOI: 10.1007/978-94-017-3087-7.

[33] Maxim Raginsky and Igal Sason. "Concentration of Measure Inequalities in Information Theory, Communications, and Coding". In: *Foundations and Trends in Communications and Information Theory* 10.1-2 (2013), pp. 1–246. DOI: 10.1561/0100000064.

[34] Andrzej Ruszczyński and Alexander Shapiro. "Stochastic Programming Models". In: *Handbooks in Operations Research and Management Science*. Elsevier, 2003, pp. 1–64. DOI: 10.1016/s0927-0507(03)10001-1.

[35] Jaehyeon Ryu and Sungsoo Park. *Robust solutions for stochastic and distributionally robust chance-constrained binary knapsack problems*. 2021. DOI: 10.48550/ARXIV.2105.11875.

[36] W. van Ackooij. "A discussion of probability functions and constraints from a variational perspective". In: *Set-Valued and Variational Analysis* 28.4 (2020), pp. 585–609. DOI: 10.1007/s11228-020-00552-2.

[37] W. van Ackooij and J. Malick. "Eventual convexity of probability constraints with elliptical distributions". In: *Mathematical Programming* 175.1 (2019), pp. 1–27. DOI: 10.1007/s10107-018-1230-3.

[38] Rolf Waeber, Peter I. Frazier, and Shane G. Henderson. "Bisection search with noisy responses". In: *SIAM Journal on Control and Optimization* 51.3 (Jan. 2013), pp. 2261–2279. DOI: 10.1137/120861898.

[39] Ximing Wang, Neng Fan, and Panos M. Pardalos. "Robust chance-constrained support vector machines with second-order moment information". In: *Annals of Operations Research* 263.1-2 (Oct. 2015), pp. 45–68. DOI: 10.1007/s10479-015-2039-6.

[40] Tapio Westerlund and Frank Pettersson. "An extended cutting plane method for solving convex MINLP problems". In: *Computers and Chemical Engineering* 19 (June 1995), pp. 131–136. DOI: 10.1016/0098-1354(95)87027-x.

[41] Songfeng Zheng. "A refined Hoeffding's upper tail probability bound for sum of independent random variables". In: *Statistics and Probability Letters* 131 (Dec. 2017), pp. 87–92. DOI: 10.1016/j.spl.2017.08.012.

A Proofs

A.1 Comparison of $\mathbb{E}[e^{tX}]$ estimations from the litterature

[5, (c)]⇒[5, (b)]. Suppose that $X - \mathbb{E}[X] \leq b$ and $\mathrm{Var}(X) \leq \sigma^2$. Then, in [21], the upper estimator of the moment-generating function is $J(t) = 1 + \gamma\left(e^{tb} - 1 - tb\right)$, where $\gamma = (\sigma/b)^2$. Besides, in [13], the upper estimator is

$$D(t) := \frac{\gamma e^{tb} + e^{-t\gamma b}}{1+\gamma} \ .$$

If now we consider D as a function of γ, i.e., $D(t,\gamma) = D(t)$, then, the second partial derivative w.r.t. γ is $\partial^2_\gamma D(t,\gamma) = \frac{2}{(1+\gamma)^3}\left[e^{\gamma t} - e^t\right] \leq 0$. Therefore, $D(t,\cdot)$ is concave for any fixed $t \geq 0$ and

$$D(t,\gamma) \leq D(t,0) + \gamma\partial_\gamma D(t,0) = J(t) \ .$$

[5, (c)]⇒[29]⇒[19]. As $\gamma \mapsto D(t,\gamma)$ is increasing, then $D(t,\gamma) \geq D(t,1) = \cosh(tb)$, which is exactly the bound obtained by Pinter with $a = b$. As $\cosh(x) \leq \exp(x^2/2)$, we have $D(t,1) \leq e^{(tb)^2/2}$, which is exactly the Hoeffding's estimator.

[15, 41]⇒[19]. Now, until the end of the proof, let us suppose that $X \in [0,1]$, i.e., $a = -\mathbb{E}[X]$ and $b = 1 - \mathbb{E}[X]$. We denote by $p = \mathbb{E}[X]$ the mean value and by σ^2 the variance. Then, in [19], the upper estimator of the moment-generating function is $H(t) = e^{tp+t^2/8}$. In [41] and [15], the upper estimator of the moment-generating function $\mathbb{E}[e^{t(X-p)}]$ is $Z(t) := 1 + p(e^t - 1)$. By basic algebra, $H'(t) - Z'(t) = \left(p + \frac{t}{4}\right)e^{tp+t^2/8} - pe^t$. Then

$$H'(t) - Z'(t) \geq 0 \iff \ln\left(1 + \frac{t}{4p}\right) + t(p-1) + t^2/8 \geq 0 \ .$$

As $\ln(1+x) \geq \frac{x}{1+\frac{1}{2}x}$ for $x \geq 0$, $H'(t) - Z'(t) \geq 0$ if

$$\left[\frac{1}{4p} + p - 1\right] + t\left[\frac{1}{8} + \frac{p-1}{8p}\right] + t^2\left[\frac{1}{8^2 p}\right] \geq 0 \ .$$

The above condition holds since the discriminant of this second-order equation is zero. Therefore, $H'(t) - Z'(t) \geq 0$, and since $H(0) = Z(0)$, we finally conclude that $H(t) \geq Z(t)$ for $t > 0$.

[11]⇒[15, 41]. In [11], the upper estimator is a family of function C_k such that

$$C_k(t) := 1 + k\left(e^{t/k} - 1\right)(p - \sigma^2 - p^2) + (\sigma^2 + p^2)(e^t - 1) \ ,$$

One can prove that $\{C_k(t)\}_k$ is decreasing $\forall t \in \mathbb{R}_+$, and

$$\lim_{k\to\infty} C_k(t) = C_\infty(t) := 1 + t(p-q) + q(e^t - 1) \ ,$$

where $q := \sigma^2 + p^2 \leq p$. Hence, $C_\infty(t) - Z(t) = (p-q)\left(1 + t - e^t\right) \leq 0$.

[5, (b)]⇒[11]. For $X \in [0,1]$, the upper estimator of [21] is $J(t) = 1 + \gamma\left(e^{t(1-p)} - 1 - t(1-p)\right)$. Using the notation $\theta = (1-p)^2 \in [0,1]$, we express C_∞ and J in (θ, γ)-coordinates:

$$\begin{cases} C_\infty(t,\gamma,\theta) = 1 + t(1\text{-}\sqrt{\theta}) + \left[\gamma\theta + (1\text{-}\sqrt{\theta})^2\right]\left[e^t - 1 - t\right] \\ J(t,\gamma,\theta) = 1 + \gamma\left(e^{t\sqrt{\theta}} - 1 - t\sqrt{\theta}\right) \end{cases}$$

Now, the partial derivatives w.r.t γ are

$$\begin{cases} \partial_\gamma C_\infty(t,\gamma,\theta) = \theta\left[e^t - 1 - t\right] \\ \partial_\gamma J(t,\gamma,\theta) = e^{t\sqrt{\theta}} - 1 - t\sqrt{\theta} \end{cases}$$

The function $[0,1] \ni \theta \mapsto e^{t\sqrt{\theta}} - 1 - t\sqrt{\theta}$ is convex for any $t \geq 0$, and so $\partial_\gamma J \leq \partial_\gamma C_\infty$. As $J(t,0,\theta) = 1 \leq C_\infty(t,0,\theta)$, we conclude that $J(t) \leq C_\infty(t)$.

B Conic reformulation

First, $\Psi_{\gamma,b}$ can also be written $\Psi_{\gamma,b}(y,z) = \max\left\{\Psi^+_{\gamma,b}(y,z), \Psi^+_{\gamma^{-1},b\gamma}(y,z)\right\}$. Therefore,

$$\sum_{k=1}^N \Psi_{\gamma_k,b_k}(y_k,z) \leq u \iff \begin{cases} \sum_{k=1}^N v_k \leq u \\ \Psi^+_{\gamma_k,b_k}(y_k,z) \leq v_k \\ \Psi^+_{\gamma_k^{-1},b_k\gamma_k}(y_k,z) \leq v_k \end{cases}$$

Now, denoting the exponential cone by $\mathcal{K}_{exp} = \{(x, 1, x_2, x_3) : x_1 \geq x_2 e^{x_3/x_2}\}$,

$$\begin{aligned}
\Psi^{+}_{\gamma_k, b_k}(y_k, z) \leq v_k &\iff \gamma_k e^{\frac{y_k}{z} b_k} + e^{-\frac{y_k}{z} b_k \gamma_k} \leq e^{\frac{v_k}{z}} (1 + \gamma_k) \\
&\iff \gamma_k e^{\frac{y_k b_k - v_k}{z} b_k} + e^{\frac{-y_k b_k \gamma_k - v_k}{z}} \leq 1 + \gamma_k \\
&\iff \begin{cases} \gamma_k \eta_k + \nu_k \leq (1 + \gamma_k) z \\ (\eta_k, z, y_k b_k - v_k) \in \mathcal{K}_{exp} \\ (\nu_k, z, -y_k b_k \gamma_k - v_k) \in \mathcal{K}_{exp} \end{cases}
\end{aligned}$$

This formulation has a number of variables and (conic) constrains of order $O(NM)$. Therefore, this conic reformulation is only valuable for small to medium instance sizes.