Constructive approaches to concentration inequalities with independent random variables

Céline Moucer, Adrien Taylor, Francis Bach

Constructive approaches to concentration inequalities with independent random variables

Céline Moucer
Inria, Département d'Informatique de l'Ecole Normale Supérieure, PSL Research University, Paris, France. celine.moucer@inria.fr
Ecole Nationale des Ponts et Chaussées, Marne-la-Vallée, France.

Adrien Taylor
Inria, Département d'Informatique de l'Ecole Normale Supérieure, PSL Research University, Paris, France. adrien.taylor@inria.fr

Francis Bach
Inria, Département d'Informatique de l'Ecole Normale Supérieure, PSL Research University, Paris, France. francis.bach@inria.fr

Abstract. Concentration inequalities, a major tool in probability theory, quantify how much a random variable deviates from a certain quantity. This paper proposes a systematic convex optimization approach to studying and generating concentration inequalities with independent random variables. Specifically, we extend the generalized problem of moments to independent random variables.

We first introduce a variational approach that extends classical moment-generating functions, focusing particularly on first-order moment conditions. Second, we develop a polynomial approach, based on a hierarchy of sum-of-square approximations, to extend these techniques to higher-moment conditions. Building on these advancements, we refine Hoeffding's, Bennett's and Bernstein's inequalities, providing improved worst-case guarantees compared to existing results.

Key words: concentration inequalities, semidefinite programming, sum-of-square, generalized problem of moments, convex optimization

Subject classifications: 60E15, 90C22, 90C25, 60F10

Introduction.

Concentration inequalities have emerged as a major tool in probability theory, finding applications in learning theory (Devroye et al. 1996), in random matrix theory (Tao 2011) or statistical physics or mechanics (Dembo and Zeitouni 1998). These inequalities quantify how much a random variable deviates from a certain quantity, usually its mean. Classical examples include Markov's inequality for bounding probabilities of deviations from zero or Chebyshev's inequalities for deviation from the mean. In machine learning and statistics, where the data are often assumed to be independent and identically distributed (i.i.d.), basic inequalities such as Hoeffding's inequality, Bennett's inequality or Bernstein's inequality (Boucheron et al. 2013, Chapter 6) are extensively used, e.g., for characterizing generalization properties of machine learning algorithms (Bach 2024). For instance, Hoeffding's inequality (Hoeffding 1963) states that for $X_1, \dots, X_n$ i.i.d. random variables taking their values almost surely in $[0, 1]$, the sum $\sum_{i=1}^n X_i$ has a subgaussian tail with deviation $t \geqslant 0$:

$$\mathbb{P}\left(\sum_{i=1}^{n}(X_i - \mathbb{E}[X_i]) \geqslant nt\right) \leqslant \exp\left(-2nt^2\right).$$

In this work, we provide a principled approach to concentration inequalities for independent univariate random variables with finite moments. Let $\mathcal{X}$ be a subset of $\mathbb{R}$ and $\mathcal{P}(\mathcal{X})$ be the set of distributions

on $\mathcal{X}$. Given $X_1, \dots, X_n$ independent random variables generated from distributions $p_1, \dots, p_n \in \mathcal{P}(\mathcal{X})$, we formulate the generalized problem of moments for independent random variables as follows:

$$\begin{aligned}\rho_n &= \sup_{p_1,\dots,p_n \in \mathcal{P}(\mathcal{X})} \mathbb{E}_{p_1,\dots,p_n}[F(X_1,\dots,X_n)] \text{ such that } \forall i, \mathbb{E}_{p_i}[g_i(X_i)] = \mu_i, \\ &= \sup_{p_1,\dots,p_n \in \mathcal{P}(\mathcal{X})} \int_{\mathcal{X}^n} F(x_1,\dots,x_n) dp_1(x_1) \cdots dp_n(x_n) \text{ such that } \forall i, \int_{\mathcal{X}} g_i(x_i) dp_i(x_i) = \mu_i,\end{aligned} \tag{1}$$

for some functions $F : x \in \mathcal{X}^n \mapsto \mathbb{R}^+$ and $g_i : x \in \mathcal{X} \mapsto \mathbb{R}^m$. For instance, Hoeffding's inequality involves $F(x) = \mathbf{1}_{\sum_{i=1}^n x_i \geqslant nt + \sum_{i=1}^n \mathbb{E}[X_i]}$, and therefore $\mathbb{E}_{p_1,\dots,p_n}[F(x)] = \mathbb{P}(\sum_{i=1}^n (X_i - \mathbb{E}[X_i]) \geqslant nt)$. Problem (1) is an infinite-dimensional non-convex problem. Without further assumptions on F and g, minimizing with respect to distributions p_i is often intractable.

This problem is closely related to the generalized problem of moments formalized by Lasserre (2008), which extends the traditional problem of moments (Landau 1998) that seeks a measure matching a given set of moments. The search for optimal multivariate Chebyshev's inequalities began in the 1960s (Marshall and Olkin 1960, Isii 1962). Isii (1962, 1964), along with Karlin and Studden (1966), formalized the pursuit of sharp inequalities by framing it as an optimization problem. Let $\mathcal{Y} \subset \mathbb{R}^n$ and $\mathcal{P}(\mathcal{Y})$ be the set of probability distributions on $\mathcal{Y}$. The generalized problem of moments takes the form:

$$\begin{aligned}\rho &= \sup_{p \in \mathcal{P}(\mathcal{Y})} \mathbb{E}_p[F(X)], \text{ such that } \mathbb{E}_p[g(X)] = \mu \\ &= \sup_{p \in \mathcal{P}(\mathcal{Y})} \int_{\mathcal{Y}} F(x) dp(x) \text{ such that } \int_{\mathcal{Y}} g(x) dp(x) = \mu,\end{aligned} \tag{2}$$

where $g : \mathcal{Y} \mapsto \mathbb{R}^m$. Compared to Problem (1), the generalized problem of moments optimizes over distributions $p \in \mathcal{P}(\mathcal{Y})$ that are not necessarily products of their marginals. Under mild assumptions on the moment vector μ, strong duality holds (Isii 1964, Theorem 3.1). Its Lagrangian relaxation was first formulated by Isii (1964) as follows:

$$\begin{aligned}\rho &= \inf_{\alpha \in \mathbb{R}, \beta \in \mathbb{R}^m} \alpha + \beta\mu \text{ such that } \forall x \in \mathcal{Y}, F(x) \leqslant \alpha + \beta^\top g(x), \\ &= \inf_{\alpha \in \mathbb{R}, \beta \in \mathbb{R}^m} \alpha + \beta\mu + \sup_{x \in \mathcal{Y}} \{F(x) - (\alpha + \beta^\top g(x))\}.\end{aligned} \tag{3}$$

where $\alpha \in \mathbb{R}$ corresponds to the dual variable associated to the constraint $\int_{\mathcal{Y}} dp(x) = 1$, and $\beta \in \mathbb{R}^m$ to the constraint $\int_{\mathcal{Y}} g(x) dp(x) = \mu$. The dual Problem (3) is convex as it is expressed as the pointwise supremum of affine functions. Yet, it is unclear how to deal with the constraint "$\forall x \in \mathcal{Y}, F(x) \leqslant \alpha + \beta^\top g(x)$", because it corresponds to infinitly many linear constraints in (α, β).

Those problems are now traditionally approached via convex reformulations or approximations using semidefinite programming (SDP). Bertsimas and Popescu (2005) first investigated optimal bounds for $\mathbb{E}_p[F(X)] = \mathbb{P}(X \in S)$ assuming $\mathcal{X}$ and S to be semi-algebraic sets. For univariate random variables, they efficiently solved it using a single SDP, allowing them to derive tight bounds. For multivariate random variables, they proposed a series of semidefinite relaxations using sum-of-square representations (SoS).

More generally, Lasserre (2008) investigated the generalized problem of moments and derived a hierarchy of SDPs converging to the optimal value (extending the methodology developed for approximating global optimization problems (Lasserre 2001)). Simultaneously, Vandenberghe et al. (2007), Comanor et al. (2006) reformulated the generalized Chebyshev inequality as linear matrix inequalities (LMI) using an S-procedure.

These SDP-based approaches, along with tight guarantees, often allow reconstructing corresponding worst-case distributions (Bertsimas and Popescu 2005, Section 5.1) (Vandenberghe et al. 2007, Section 2.2). These extremal distributions turns out to be discrete (Rogosinski 1958, Theorem 1) and may even be specified with $m+2$ Dirac, where m is the number of constraints in (2). When the distributions in the problem under consideration are continuous with additional properties like symmetry or unimodality, these bounds may no longer be sharp. Therefore, Popescu (2005) generalized Chebyshev's inequality to convex classes of distributions generated by an appropriate parametric family of distributions. Building on Choquet's theory and conic duality, they provided a SDP reformulation (resp. approximation) of the generalized problem of moments for such univariate (resp. multivariate) distributions. From this framework, Van Parys et al. (2015) extended Gauss inequalities to multivariate unimodal distributions and outlined a methodology for computing worst-case unimodal distributions related to this problem.

Research questions and assumptions. Throughout this work, we assume $X_1, \dots, X_n$ to be independent random variables in $\mathcal{X} \subset \mathbb{R}$ with finite moments $\forall i, \mathbb{E}_{p_i}[g(X_i)] = \mu_i \in \mathbb{R}^m$. For all i, we define $\mathcal{P}_{\mu_i}(\mathcal{X}) = \{p_i \in \mathcal{P}(\mathcal{X}), \int_{\mathcal{X}} g_i(x_i) dp_i(x_i) = \mu_i\}$ the set of (univariate) distributions on $\mathcal{X}$ with moment μ_i. If in addition, $X_1, \dots, X_n$ follows the same distributions with moments $\mu_1 = \mu_2 = \cdots = \mu_n$, they are said to be independent and identically distributed (i.i.d.). We assume that F is an indicator function, that is for S an appropriately selected compact semi-algebraic subset of $\mathcal{X}^n$ that emerges from the problem under consideration, $\forall x \in \mathcal{X}^n, F(x) = \mathbf{1}_{x \in S}$. The generalized moment problem for independent variables takes the form:

$$\begin{aligned} \rho_n &= \sup_{\forall i, p_i \in \mathcal{P}(\mathcal{X})} \int_{x \in \mathcal{X}^n} \mathbf{1}_{x \in S} dp_1(x_1) \cdots dp_n(x_n) \ \text{ such that } \forall i, \ \int_{\mathcal{X}} g(x_i) dp_i(x_i) = \mu_i, \\ &= \sup_{\forall i, p_i \in \mathcal{P}_{\mu_i}(\mathcal{X})} \int_{x \in \mathcal{X}^n} \mathbf{1}_{x \in S} dp_1(x_1) \cdots dp_n(x_n). \end{aligned} \tag{4}$$

In this formulation, independence is encoded by explicitly accounting for the constraint "$\forall x \in \mathcal{X}^n, p(x) = p_1(x_1) \cdots p_n(x_n)$". Classical concentration inequalities correspond to upper bounds to these problems for specific choices of g_i's. It is natural to wonder how tight such bounds are. To this end, we propose constructive and tractable approaches to upper bounding Problem (4), along with a comparison to existing bounds. Then, we raise the issue of reconstructing worst-case distributions that can potentially match these bounds in some scenarios.

Contributions. We propose constructive approaches for computing concentration inequalities of independent variables given a set of moments. To this end, we start by considering variational formulations to Problem (4):

$$\rho_n^{\mathcal{H}} = \inf_{H \in \mathcal{H}} \sup_{\forall i, p_i \in \mathcal{P}(\mu_i)} \int_{\mathcal{X}^n} H(x) dp_1(x) \cdots dp_n(x) \text{ such that } \forall x \in \mathcal{X}^n, F(x) \leqslant H(x), \tag{5}$$

where $\mathcal{H}$ is a well-chosen set of functions. It is straightforward that Problem (5) yields an upper bound to the generalized problem of moments for independent random variables, that is:

$$\rho_n \leqslant \rho_n^{\mathcal{H}}.$$

Efficiently leveraging the variational formulation (5) requires strategies to enforce the constraint $\forall x \in \mathcal{X}^n, F(x) \leqslant H(x)$ and to bound the objective. The choice of the function families $\mathcal{H}$ ensure these two requirements are satisfied. We propose two natural and complementary strategies based on convex optimization, each relying on different choices for $\mathcal{H}$.

1. The first strategy revolves around a family of product-functions $\forall x \in \mathcal{X}^n, U(x) = \prod_{i=1}^n u_i(x_i)$ inspired from classical probability proofs and variational probabilistic inference (Jaakkola and Jordan 1999). Given such a function, we define a *separable approach* by formulating Problem (5) as n univariate subproblems. We then develop a *variational approach* by optimizing over the family of product-funtions. When F is log-convex and first-order moments are finite with $\forall i, \mathbb{E}[g_i(X_i)] = \mathbb{E}[X_i] = \mu_i$, the variational approach takes the form of a finite-dimensional convex optimization reformulation, that can be efficiently solved. This strategy significantly improves Hoeffding's inequality for small values of n and accurately meets the asymptotic large deviations for large n. Moreover, this approach allows the reconstruction of distributions involved at the optimum of the upper bound Problem (5). However, when it comes to higher-order finite moments, the variational approach formulates as a nonconvex problem that cannot be solved efficiently anymore.

2. The second strategy relies on a family of polynomial upper bounds, which is particularly suited to higher-moment conditions. Referred to as a *polynomial approach*, this strategy formulates as a non-convex optimization problem that can be effectively approximated by a series of sum-of-square formulations (Lasserre 2008). This approach contributes to refining Bernstein's and Bennett's inequality, which are fundamental probabilistic bounds.

3. Finally, we extend the polynomial approach to a *feature-based approach* relying on broader families of upper bounds $\mathcal{H}$. Compared to the variational approach, this method refines Hoeffding's inequality using higher-order polynomials when applied to two random variables. While this methodology introduces finer approximations $\rho_n \leqslant \rho_n^{\mathcal{H}}$, it often requires well-chosen relaxations to approximate $\rho_n^{\mathcal{H}}$.

Outline of the paper. Section 1 focuses on computing Problem (5) derived from a family of product-functions, which is particularly suited to first-order moment assumptions. We formally compare these bounds to existing results. In Section 2, we study in depth the variational and separable approaches associated with Hoeffding's inequality. Furthermore, we propose a methodology for reconstructing distributions that match the separable or variational optimization problem in the worst-case scenario. Section 3 considers a family of polynomial upper bounds and approximate the resulting upper optimization Problem (5) using sum-of-square formulations. Thereby, we derive numerical evaluations of Bennett's and Bernstein's inequality. Finally, we introduce a feature-based framework that expands the scope of upper bounds families, providing a comprehensive approach to Hoeffding's inequality that includes both the polynomial and variational methodologies.

Codes. All codes are provided at `https://github.com/CMoucer/ConcentrationInequalities`. We use standard solvers SCS (O'Donoghue et al. 2016) and MOSEK (ApS 2022).

1. A variational approach based on product-functions.

Classical functions $F(\cdot)$ usually correspond to the probability tails of a sum of independent random variables, specifically, $F(X) = \mathbf{1}_{\sum_{i=1}^n X_i \in S}$ representing $\mathbb{P}(\sum_{i=1}^n X_i \in S)$. This applies, among others, to Hoeffding's, Bennett's, and Bernstein's inequalities (see, e.g., (Boucheron et al. 2013, Bach 2024, Vershynin 2018) and references therein). Their proofs typically rely on the Cramér-Chernoff technique, which essentially combines the exponential Chernoff's inequality with the independence of random variables. Specifically, they use moment-generating functions as follows:

$$\mathbb{P}\left(\sum_{i=1}^n (X_i - \mu_i) \geqslant nt\right) \leqslant \inf_{\lambda \geqslant 0} e^{-\lambda nt}\mathbb{E}[e^{\lambda \sum_{i=1}^n (X_i-\mu_i)}] = \inf_{\lambda \geqslant 0} \prod_{i=1}^n e^{-\lambda(\mu_i+t)}\mathbb{E}[e^{\lambda X_i}],$$

and then optimize over $\lambda \geqslant 0$ for obtaining the smallest possible valid upper bound within this family. A natural, but richer, family of inequalities for obtaining concentration bounds involves constructing upper bounding functions as products of univariate functions, which are classical in probabilistic variational inference:

$$\mathcal{U} = \left\{ U : \mathcal{X}^n \mapsto \mathbb{R}^+ \text{ such that } \forall x \in \mathcal{X}^n, U(x) = \prod_{i=1}^n u_i(x_i) \text{ and } \forall i, u_i : \mathcal{X} \mapsto \mathbb{R}^+ \right\}. \tag{6}$$

In the Cramér-Chernoff technique, the functions u_i correspond to moment-generating functions with $\forall x_i \in \mathcal{X},\ u_i(x_i) = e^{\lambda(x_i-\mu_i-t)}$ and serve as natural upper bounds to the indicator function $\mathbf{1}_{\sum_{i=1}(x_i-\mu_i)\geqslant nt} \leqslant \prod_{i=1}^n e^{\lambda(x_i-\mu_i-t)}$.

Throughout this section, we assume the existence of a product-function $U \in \mathcal{U}$ (6) such that $\forall x \in \mathcal{X}^n, F(x) \leqslant U(x)$, from which we define two strategies. First, we introduce a separable approach that generalizes classical probability proofs. Second, we optimize over the family of product-functions (6) and show how it formulates as a convex optimization problem. Depending on the moments and product-functions under consideration, we demonstrate that these approaches yield tractable upper bounds.

1.1. The separable approach.

Let $U \in \mathcal{U}$ be a product-function verifying $\forall x \in \mathcal{X}^n,\ F(x) \leqslant U(x)$. This section studies the properties of the optimization problem:

$$\rho_n^U = \sup_{\forall i, p_i \in \mathcal{P}_{\mu_i}(\mathcal{X})} \int_{\mathcal{X}^n} U(x_1, \dots x_n) dp_1(x_1) \cdots dp_n(x_n).$$

This problem naturally provides an upper bound to the generalized problem of moments for independent random variables (4), specifically $\rho_n \leqslant \rho_n^U$. By definition, the function $U \in \mathcal{U}$ has a product structure $\forall x \in \mathcal{X}^n, U(x) = \prod_{i=1}^n u_i(x_i)$. This allows the separation of the integral over $\mathcal{X}^n$ into n integrals over $\mathcal{X}$, thus decoupling the optimization problem into n independent optimization problems over $p_i \in \mathcal{P}(\mathcal{X})$. In other words, the family $\mathcal{U}$ aligns with the structure imposed by independence:

$$\rho_n^U = \sup_{\forall i, p_i \in \mathcal{P}_{\mu_i}(\mathcal{X})} \int_{\mathcal{X}} \cdots \int_{\mathcal{X}} \prod_{i=1}^n u_i(x_i) dp_1(x_1) \cdots dp_n(x_n) = \prod_{i=1}^n \sup_{p_i \in \mathcal{P}_{\mu_i}(\mathcal{X})} \int_{\mathcal{X}} u_i(x_i) dp_i(x_i).$$

For univariate distributions with μ_i in the interior of $\mathcal{X}$, strong duality holds. Then,

$$\begin{aligned} \rho_n^U &= \prod_{i=1}^n \inf_{\alpha_i \in \mathbb{R}, \beta_i \in \mathbb{R}^m} \{\alpha_i + \beta_i^\top \mu_i\} \text{ such that } \forall\ x_i \in \mathcal{X}, u_i(x_i) \leqslant \alpha_i + \beta_i^\top g_i(x_i), \\ &= \prod_{i=1}^n \inf_{\alpha_i \in \mathbb{R}, \beta_i \in \mathbb{R}^m} \sup_{x_i \in \mathcal{X}} \{u_i(x_i) - \beta_i^\top (g_i(x_i) - \mu_i)\}. \end{aligned} \tag{7}$$

Problem (7) is finite-dimensional and convex, as it is the pointwise supremum of affine functions (Boyd and Vandenberghe 2004, Section 3.2.3). In the case of i.i.d. random variables, the problem simplifies significantly to a single optimization problem with $\rho_n^U = (\rho_1^{\exp})^n$. However, computing $\sup_{x_i \in \mathcal{X}} \{u_i(x_i) - \beta_i^\top (g_i(x_i) - \mu_i)\}$ often remains numerically intractable. Under strong assumptions, such as the finiteness of the support or the convexity of the objective function on a compact set, it reduces to a finite number of constraints. Proposition 1 outlines a useful tractable setting that will be used in Section 2 for improving Hoeffding's inequality.

PROPOSITION 1. *Let $\mathcal{X}$ be compact and $x_i \mapsto u_i(x_i) - \beta^\top (g_i(x_i) - \mu_i)$ be convex. Then, Problem* (7) *formulates as a tractable convex optimization problem with a finite number of constraints:*

$$\rho_n^U = \prod_{i=1}^n \inf_{\alpha_i \in \mathbb{R}, \beta_i \in \mathbb{R}^m} \{\alpha_i + \beta_i^\top \mu_i\} \text{ such that } \forall x \in \text{Extremal}(\mathcal{X}),\ \forall i,\ u_i(x_i) \leqslant \alpha_i + \beta_i^\top g_i(x_i).$$

Proof. Under the assumptions of Proposition 1, the maximization of the convex function $u_i(x_i) - \beta_i^\top (g_i(x_i) - \mu_i)$ over the compact set $\mathcal{X}$ is achieved at extremal points of $\mathcal{X}$ (Boyd and Vandenberghe 2004, Section 3.2.3). □

This technique faces two major challenges: first, constructing a valid upper bound $U \in \mathcal{U}$ can be difficult; second, even with a suitable upper bound, the resulting optimization Problem (7) may be numerically intractable. The next section focuses on more accurate approximations to the generalized problem of moments by optimizing over the family of upper bounds (instead of keeping one such upper bound fixed).

1.2. Optimizing over $\mathcal{U}$.

This section explores Problem (5) for the class of product-functions (6). Combined with the dual formulation (7), we define:

$$\rho_n^{\text{var}} = \inf_{u_i \geqslant 0} \inf_{\alpha \in \mathbb{R}^n, \beta \in \mathbb{R}^{n \times m}} \prod_{i=1}^n (\alpha_i + \beta_i^\top \mu_i) \text{ such that } \forall i, \forall x_i \in \mathcal{X}, u_i(x_i) \leqslant \alpha_i + \beta_i^\top (x_i), \\ \forall x \in \mathcal{X}^n, F(x_1, \dots, x_n) \leqslant \prod_{i=1}^n u_i(x_i). \tag{8}$$

Optimizing with respect to $u_1, \dots, u_n \geqslant 0$, it holds that:

$$\rho_n^{\text{var}} = \inf_{\alpha \in \mathbb{R}^n, \beta \in \mathbb{R}^{n \times m}} \prod_{i=1}^n (\alpha_i + \beta_i^\top \mu_i) \text{ such that } \forall x_i \in \mathcal{X}, \forall i, \alpha_i + \beta_i^\top g(x_i) \geqslant 0, \\ \forall x \in \mathcal{X}^n, F(x_1, \dots, x_n) \leqslant \prod_{i=1}^n (\alpha_i + \beta_i^\top g(x_i)). \tag{9}$$

As expected, Problem (9) shows no dependence on the u_i's. In Proposition 2, we formally compare ρ_n (4) to ρ_n^{var} (9) and to ρ_n^U (7) for any function $U \in \mathcal{U}$.

PROPOSITION 2. *Let $U \in \mathcal{U}$, ρ_n^U be defined in* (7)*, ρ_n in* (4) *and ρ_n^{var} in* (9)*. Then it holds that*

$$\rho_n \leqslant \rho_n^{\text{var}} \leqslant \rho_n^U.$$

In addition, the equality $\rho_n^U = \rho_n^{\text{var}}$ holds for optimal values $(u_\star, \alpha_\star, \beta_\star)$ such that $\forall x_i \in \mathcal{X}, u_{i,\star}(x_i) = \alpha_{i,\star} + (\beta_{i,\star})^\top g_i(x_i)$.

Proposition 2 provides optimal product-functions $U \in \mathcal{U}$, as affine functions of the moments. This specific structure indicates that moment-generating functions, used in the Cramér-Chernoff method, are not optimal. Computing $(\alpha_{i,\star}, \beta_{i,\star})$ often remains difficult (that is, solving Problem (9) which is finite-dimensional but nonconvex). Proposition 3 ensures a convex reformulation of Problem (9).

PROPOSITION 3. *Let ρ_n^{var} be defined in* (8)*. Then, it holds that*

$$\log(\rho_n^{\text{var}}) = \inf_{\alpha \in \mathbb{R}^n, \beta \in \mathbb{R}^{n \times m}} \sum_{i=1}^n \{\alpha_i + \beta_i^\top \mu_i - 1\} + \sup_{x \in \mathcal{X}^n} \left\{ \log(F(x)) - \sum_{i=1}^n \log(\alpha_i + \beta_i^\top g_i(x_i)) \right\}. \tag{10}$$

In addition, if $x \mapsto \log(F(x)) - \sum_{i=1}^n \log(\alpha_i + \beta_i^\top g_i(x_i))$ is convex and $\mathcal{X}$ is compact, then,

$$\sup_{x \in \mathcal{X}^n} \left\{ \log(F(x)) - \sum_{i=1}^n \log(\alpha_i + \beta_i^\top g_i(x_i)) \right\} = \sup_{x \in \text{Extremal}(\mathcal{X}^n)} \left\{ \log(F(x) - \sum_{i=1}^n \log(\alpha_i + \beta_i^\top g_i(x_i)) \right\}.$$

Proof. First, let us consider the logarithm $\log(\rho_n^{\text{var}}) = \inf_{\alpha \in \mathbb{R}^n, \beta \in \mathbb{R}^{n \times m}} \sum_{i=1}^n \log(\alpha_i + \beta_i^\top \mu_i) + \sup_{x \in \mathcal{X}^n} \{\log(F(x) - \sum_{i=1}^n \log(\alpha_i + \beta_i^\top g_i(x_i))\}$. Noticing that $\inf_{t \geqslant 0} \{\big(t\alpha_i + t\beta_i^\top \mu_i - 1\big) - \log(t\alpha_i + t\beta_i g(x_i))\} = \log(\alpha_i + \beta_i^\top \mu_i) - \log(\alpha_i + \beta_i g(x_i))$, we conclude the reformulation in (10). The second assertion follows by maximization of a convex function over a compact set. □

Proposition 3 details a set of assumptions on $\mathcal{X}$, F and g, under which the constraints "$\log(F(x_1,\dots,x_n)) \leqslant \sum_{i=1}^n \log(\alpha_i + \beta_i^\top g_i(x_i))$" reduces to a finite number of points. For instance, these assumptions are satisfied for finite first-order moments $g_i(x_i) = x_i$ together with log-convex objectives F (such as exponentials or indicator functions $\mathbf{1}_S$, with compact sets $S \subset \mathcal{X}^n$, see (Boyd and Vandenberghe 2004, Section 3.5)). An alternative convexification proof is achieved via optimal transport in Appendix A, which also includes a formulation of the gap to the generalized problem of moments.

We have established two approaches for deriving upper bounds to the generalized problem of moments for independent random variables (4) using a family of product-functions (6). First, we introduced a separable approach (7) that formulates as a product of n convex optimization problems. However, constructing product functions may not be straightforward. Then, we formulate a variational Problem (9) emerging from (7) by optimizing with respect to product-functions. It turns out that Problem (9) benefits from a convex reformulation which does not require a priori upper bounds to F (but constructs such bounds in the process). Without further assumptions on the probability support $\mathcal{X}$, the objective F or moments g_i, both approaches are intractable. We will see next how they effectively apply in the context of Hoeffding's inequality.

2. Revisiting Hoeffding's inequality.

Hoeffding's inequality establishes a subgaussian tail for the sum of independent random variables taking their values in a bounded set with finite means. This section is devoted to refining Hoeffding's inequality, applying the separable and variational frameworks developed in Section 1. First, let us recall Hoeffding's inequality as stated in Theorem 1.

THEOREM 1. *(Hoeffding 1963) Let $X_1, ..., X_n$ be independent random variables taking their values in $[a_1, b_i]$ almost surely. Then, for every $t \geqslant 0$,*

$$\mathbb{P}\left(\frac{1}{n}\sum_{i=1}^n (X_n - \mathbb{E}[X_n]) \geqslant t\right) \leqslant \exp\left(-\frac{2n^2t^2}{\sum_{i=1}^n (b_i - a_i)^2}\right) = \rho_n^{\text{Hoeffding}}. \tag{11}$$

Throughout this section, $X_1, \dots, X_n$ are independent random variables with mean $\mu_1, \dots, \mu_n$. Without loss of generality, we assume that they all take their values in $\mathcal{X} = [0,1]$ (that is, $b_i = 1$ and $a_i = 0$) and thereby, have finite mean $\mathbb{E}[X_i] = \mu_i$ for all i. Our goal is to approximate the probability $\mathbb{P}(\sum_{i=1}^n X_i \geqslant nt + \sum_{i=1}^n \mu_i)$ for $t \geqslant 0$. In our framework, it translates to functions where for all $x \in [0,1]^n$, $F(x) = \mathbf{1}_{\sum_{i=1}^n X_i - \sum_{i=1}^n \mu_i - nt}$ and $\forall i, \forall x_i \in [0,1], g_i(x_i) = x_i$.

As a reference, we first consider one random variable, that benefits from an exact analytical bound (4). We compare it to the separable technique (7) on the moment-generating function. Then, we extend the analysis to n random variables, comparing bounds in the separable and variational approaches to Hoeffding's inequality. More precisely, we examine cases where random variables are i.i.d., and where random variables are divided into two blocks with different means $\mu_1 = \mu_2 = \cdots = \mu_m$ and $\mu_{m+1} = \cdots = \mu_n$ (with $1 \leqslant m \leqslant n-1$). For the case $\mu_1 = \mu_2 = \cdots = \mu_n$, our results asymptotically correspond to large deviations. Finally, we propose a methodology to reconstruct a distribution in the worst-case scenario.

2.1. One random variable: comparison of the exact and exponential bounds

Computing optimal bounds for univariate random variables has been extensively studied in past years and is encompassed in the multivariate analyses proposed by Isii (1962), Bertsimas et al. (2000), Vandenberghe et al. (2007). Let us compute the exact closed-form solution (4) and the separable scenario for moment-generating functions (7).

Exact optimization problem. Let X_1 be a random variable in $\mathcal{X} = [0,1]$, with $\mathbb{E}[X_1] = \mu$. Recall the exact optimization problem (4) for every $t \geqslant 0$,

$$\begin{aligned}\rho_1(t) &= \sup_{p_1 \in \mathcal{P}([0,1])} \int_0^1 \mathbf{1}_{x_1 \geqslant \mu + t} dp_1(x_1) \text{ such that } \int_0^1 x_1 dp_1(x_1) = \mu, \\ &= \inf_{\alpha,\beta \in \mathbb{R}} \alpha + \beta\mu \text{ such that } \forall\, x_1 \in [0,1], \mathbf{1}_{x_1 \geqslant \mu + t} \leqslant \alpha + \beta x_1.\end{aligned} \tag{12}$$

Problem 12 defines a function $\rho_1(\cdot)$ as a solution to a linear program for every $t \geqslant 0$ and verifies $\rho_1(t) = \mathbb{P}(X_1 \geqslant t + \mu)$. As stated by Bertsimas and Popescu (2005, Theorem 2.2), strong duality holds for $\mu \in]0,1[$. It turns out that Problem (12) is a particular case of both the generalized moment problem (2) for univariate distributions and of the generalized problem of moments for independent random variables (1). It admits a closed-form solution, detailed in Proposition 4.

PROPOSITION 4. *Let $0 \leqslant t \leqslant 1 - \mu$. Then, $\rho_1(t)$ as defined in* (12) *verifies:*

$$\rho_1(t) = \frac{\mu}{\mu + t}. \tag{13}$$

Proof. Functions $x_1 \mapsto -(\alpha + \beta x_1)$ is convex on $[0,1]$. Thus the constraint in (12) can be reduced to two constraints : $\alpha \geqslant 0$ and $\beta(\mu + t) \geqslant 1$. It follows that $\alpha = 0$ and $\beta = \frac{1}{\mu + t}$. □

Separable approach. Now, let us compute the bound defined in (7) considering moment-generating functions $u_\lambda(x) = e^{\lambda(x - (\mu + t))}$. By construction, $\forall x \in \mathbb{R}, \mathbf{1}_{x \geqslant \mu + t} \leqslant e^{\lambda(x - (\mu + t))}$. We define the family of upper bounds $\rho_1^{\exp}$. For every $\lambda \in \mathbb{R}$ and for every $t \geqslant 0$,

$$\rho_1^{\exp}(\lambda, t) = \inf_{\alpha,\beta} \alpha + \beta\mu, \text{ such that } \forall x_1 \in [0,1], e^{\lambda(x_1 - (\mu + t))} \leqslant \alpha + \beta x_1. \tag{14}$$

Problem (14) is a convex optimization problem, that is well-defined for every $t \geqslant 0$ and $\lambda \in \mathbb{R}$. For every $t \geqslant 0$, it admits an optimal moment-generating function analytically given in Proposition 5.

PROPOSITION 5. *Let $t \geqslant 0$, and $\rho_1^{\exp}(\lambda)$ be defined in* (14)*. Then, for every $t \geqslant 0$,*

$$\rho_1^{\exp}(t) = \left(\frac{\mu}{\nu}\right)^{\nu} \left(\frac{1 - \mu}{1 - \nu}\right)^{1 - \nu}, \tag{15}$$

where $\nu = \mu + t$, and $\rho_{1,\star}^{\exp}$ is the optimal value when optimizing (14) *with respect to λ.*

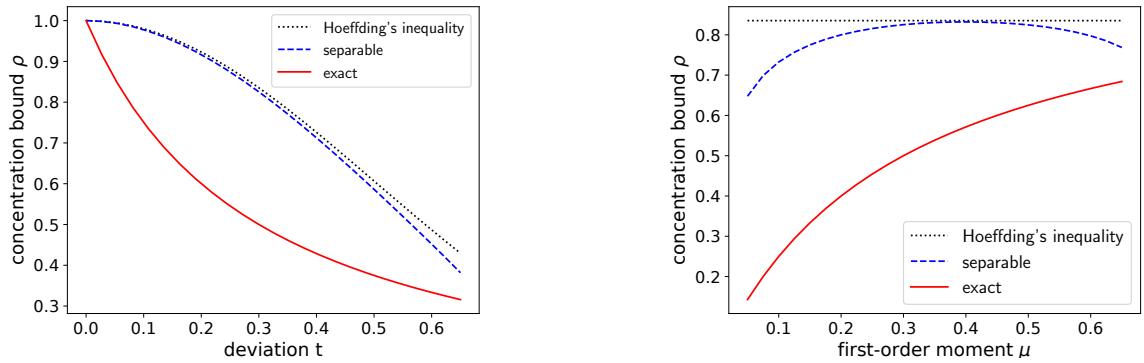


Figure 1 **Comparison of the bound in the exact** (13) **and separable** (15) **approaches to Hoeffding's inequality** (11)**. On the left, bounds are plot as a function of** t **with** $\mu = 0.3$**, and on the right as a function of** μ **with** $t = 0.3$**.**

Proof. See Appendix B.1.1. □

Proposition 5 yields exactly the Chernoff-Bound for a Bernoulli random variable on the support $\{0, 1\}$ with mean μ (Boucheron et al. 2013, Section 2.2). This result was originally stated in the seminal work of Hoeffding (1963), but our approach ensures tightness for this family of moment-generating functions. It can also be directly derived from Kullback's inequality, as detailed in Appendix B.1.2.

In Figure 1, we observe that the exact bound ρ_1 (12) significantly improves upon Hoeffding's bound, whereas the bound $\rho_1^{\exp}$ (14) in the separable approach only shows improvement for large values of t. Both approaches benefit from a dependence in the first-order moment μ, that does not appear in Hoeffding's inequality (11). The next section extends beyond the univariate case.

2.2. Generalization to n independent random variables.

We consider $X_1, \ldots, X_n$ independent random variables with finite means μ_i. We first examine the case of i.i.d. random variables, where $\mu_1 = \cdots = \mu_n$. The bound obtained in the variational approach (9) asymptotically matches the large deviations, and cannot therefore be much improved for a large number of variables. Next, we consider the case of independent random variables with different means. Specifically, we formulate a tractable upper bound using the separable treatment and discuss the computational limits of the variational approach for large n due to an exponential number of constraints. In the special case of two blocks of variables with different means, the number of constraints involved is $O(n^2)$.

2.2.1. Independent and identically distributed random variables (equal means).

This section focuses on i.i.d. random variables, that is with $\mu_1 = \mu_2 = \cdots = \mu_n$. Thanks to symmetry properties, both the separable and variational approaches yield tractable solutions for any number of variables n.

Separable approach. Let us consider the moment generating function of $X_1 + \cdots + X_n$: $\forall \lambda \in \mathbb{R}, \forall x \in [0,1]^n, u_\lambda(x) = e^{\lambda(\sum_{i=1}^n \{x_i - \mu - t\})} = \prod_{i=1}^n e^{\lambda(x_i-\mu-t)}$. This is a product-function verifying for all $x \in [0,1]^n, F(x) = \mathbf{1}_{\sum_{i=1}^n x_i \geqslant n(t+\mu)} \leqslant u_\lambda(x)$. We thus define the separable Problem (7) for $\lambda \in \mathbb{R}$ and $t \geqslant 0$:

$$\rho_n^{\exp}(\lambda, t) = \prod_{i=1}^n \inf_{\alpha_i \in \mathbb{R}, \beta_i \in \mathbb{R}} (\alpha_i + \beta_i \mu) \text{ such that } \forall x_i \in [0,1]^n, e^{\lambda(x_i-\mu-t)} \leqslant \alpha_i + \beta_i \mu. \tag{16}$$

Optimizing with respect to λ, Proposition 6 provides a closed-form as a function of t.

PROPOSITION 6. *Let $\mu_1 = \cdots = \mu_n$ and let $\rho_n^{\exp}$ be defined in* (16)*. Then, it holds for all $t \geqslant 0$:*

$$\rho_{n,\star}^{\exp}(t) = \inf_{\lambda \in \mathbb{R}} \rho_n^{\exp}(\lambda, t) = \left(\frac{\mu}{\nu}\right)^{n\nu} \left(\frac{1-\mu}{1-\nu}\right)^{n(1-\nu)} = (\rho_{1,\star}^{\exp}(t))^n. \tag{17}$$

Proof. Since $\mu_1 = \cdots = \mu_n$, Problem (16) benefits from a symmetry property and simplifies into $\forall t \geqslant 0,\ \forall \lambda \in \mathbb{R},\ \rho_n^{\exp}(\lambda, t) = \prod_{i=1}^n \inf_{\alpha,\beta \in \mathbb{R}} (\alpha + \beta\mu),\ \text{s.t. } \forall x_i \in [0,1]^n, e^{\lambda(x_i-\mu-t)} \leqslant \alpha + \beta\mu$, that is $\rho_n^{\exp}(\lambda, t) = (\rho_1^{\exp}(\lambda))^n$. We then optimize over λ as in Proposition 5. □

The symmetry properties induced by the i.i.d. assumption allow simplifying (16) into a single univariate convex optimization problem, which can be solved efficiently. Again, this is exactly the Chernoff bound for n i.i.d. Bernoulli variables taking their values in $\{0,1\}$ with mean n. We conclude that the separable approach improves Hoeffding's inequality for large deviations t (as for a unique univariate variable). Let us now explore how the variational approach might offer further improvements.

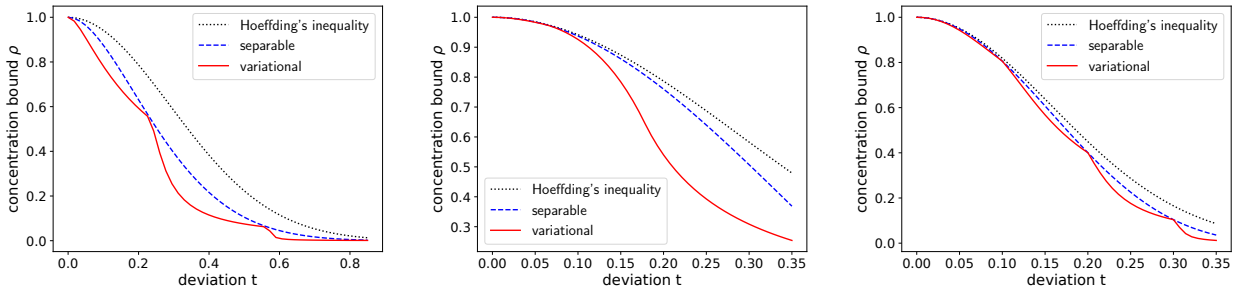
Variational approach. In the context of Hoeffding's inequality and given symmetry properties induced by $\mu_1 = \cdots = \mu_n = \mu$, the bound ρ_n^{var} defined in (9) takes the form for all $t \geqslant 0$:

$$\rho_n^{\text{var}}(t) = \inf_{\alpha \in \mathbb{R}, \beta \in \mathbb{R}} \prod_{i=1}^n (\alpha + \beta^\top \mu) \ \text{ such that } \forall x \in [0,1]^n,\ \mathbf{1}_{x_1+\cdots+x_n \geqslant n(\mu+t)} \leqslant \prod_{i=1}^n (\alpha + \beta x_i),$$
$$\forall x_i \in [0,1],\ \alpha + \beta x_i \geqslant 0.$$

This problem has a convex reformulation with a finite number of constraints, as proven in Proposition 3,

$$\log \rho_n^{\exp}(t) = \inf_{\alpha,\beta,t \geqslant 0} n(\alpha + \beta\mu - 1) \text{ such that } \ -\sum_{i=1}^n \log(\alpha + \beta x_i) \leqslant 0,\ x \in \text{extremal}(\bar{\mathcal{X}}_n), \tag{18}$$

where $\bar{\mathcal{X}}_n = \{(x_1, \ldots, x_n) \in [0,1]^n, x_1 + \cdots + x_n \geqslant n(\mu + t)\}$ and where the constraint "$\forall x \in [0,1], \alpha + \beta x \geqslant 0$" is implied by the logarithm. The set $\bar{\mathcal{X}}_n \subset [0,1]^n$ is compact and symmetric in $(x_1, \ldots, x_n)$. At first sight, the set $\text{extremal}(\mathcal{X})_n$ appears to grow exponentially with n. However, due to symmetry properties and the structure of the constraints, it reduces to $O(n)$ constraints (see Appendix B.2, via computation of extremal points). Therefore, Problem (18) can be efficiently addressed using standard solvers for convex optimization. In addition, it is possible to derive closed-form solutions for $(\alpha_\star, \beta_\star)$, by enumerating all extremal points given $\mu + t$ and solving the KKT condition for Problem (18). We provide an example for $n = 2$ in Appendix B.3.



(a) $\mu = 0.1$ and $n = 3$. (b) $\mu = 0.6$ and $n = 3$. (c) $\mu = 0.6$ and $n = 10$.

Figure 2 Comparison of bounds derived in the separable (16) **and variational approaches** (18) **to Hoeffding's inequality** (11)**, as a function of the deviation** t**.**

Figure 2 illustrates the comparison between the separable and variational approaches to the Hoeffding's bound for different numbers of i.i.d. random variables. It appears that $\rho^{\exp}_{n,\star}$ (16) closely tracks Hoeffding's bound. In contrast, ρ^{var}_{n} (18) provides significant numerical improvements over Hoeffding's inequality when a small number of random random variables are in play. Furthermore, as the number of variables increases, the variational approach asymptotically matches the separable treatment, as expected from the large deviations theory.

In probability theory, the study of the asymptotic behavior of tails of random variables is known as the large deviations theory, introduced by Varadhan (1988). In particular, the large deviation principle provides a guarantee on rare events, as outlined in Theorem 2 for the sum of i.i.d. random variables.

THEOREM 2 **(Cramér (1938): Large Deviations)**. *Let* $X_1, \dots, X_n$ *be i.i.d. random variables with finite moment-generating functions, and let* $\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$*. Then, for all* $x \in \mathbb{R}$*,*

$$\lim_{n\to\infty} \frac{1}{n} \log(\mathbb{P}(\bar{X}_n \geqslant x)) = -\Gamma^{\star}(x),$$

where $\Gamma^{\star}(x) = \sup_{t\geqslant 0}(tx - \Gamma(t))$ *and* $\Gamma(t) = \log(\mathbb{E}[\exp(tX_1)])$.

The moment-generating function of a univariate random variable formulates as an optimization problem as in (14). Corollary 1 provides the large deviations asymptotic for i.i.d. random variables.

COROLLARY 1. Let $X_1, \dots, X_n$ be i.i.d. random variables with mean $\mathbb{E}[X_1] = \mu$, and taking their value in $[0, 1]$ almost surely, and let $\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$. It holds that, for all $t \geqslant 0$,

$$\lim_{n\mapsto\infty} \frac{1}{n} \log(\mathbb{P}(\bar{X}_n \geqslant \mu + t)) = -(\mu+t)\log\left(\frac{\mu}{\mu+t}\right) - (1-(\mu+t))\log\left(\frac{1-\mu}{1-(\mu+t)}\right).$$

Proof. From the separable approach, $\Gamma(t) = \sup_{p\in\mathcal{P}(\mathcal{X})} \int_0^1 e^{tx} dp(x)$ such that $\int_0^1 x dp(x) = \mu$. Strong duality holds and $\Gamma(t) = \inf_{\alpha,\beta\in\mathbb{R}} \alpha + \beta\mu,$ such that $\forall x \in [0,1],\ e^{tx} \leqslant \alpha + \beta x$. Thus, $\Gamma(t) = 1 + \mu(e^t - 1)$. The desired statement is obtained by computing the Fenchel conjugate. □

Corollary 1 demonstrates that the probability of $\bar{X}_n$ deviating from the mean μ converges exactly to the bound in the separable approach (17): for all $t \geqslant 0$, $\mathbb{P}(X_1 + \cdots + X_n \geqslant n(\mu + t)) \approx (\rho_1^{\exp})^n$. Numerical results presented in Figure 2 are thus consistent with these large deviation estimates for relatively large n.

As a conclusion, when random variables are i.i.d., optimization problems in the separable and variational approaches benefit from tractable formulations. The bound in the variational treatment (18) shows significant improvements over Hoeffding's inequality when a small number n of random variables are in play, but suffers an increasing number of constraints. As n increases, the separable approach (16) provides a close estimate of the generalized problem of moments at a lower computational cost.

2.2.2. Two blocks of random variables with different means. Hoeffding's inequality, as presented in Theorem 1, applies generally to independent random variables without specific assumptions on their means. However, formulating the separable and variational approaches in a generic setting can be computationally challenging for a large number of variables n. To simplify this, we focus on the scenario where the random variables are divided into two blocks with different means.

Separable approach. Consider the optimization problem (7) with different means in the context of Hoeffding's inequality. After solving each subproblem in (α_i, β_i) as in Proposition 5, the bound takes the form, for any $t \geqslant 0$:

$$\rho_{n,\star}^{\exp}(t) = \inf_{\lambda \in \mathbb{R}} e^{-n\lambda(\bar{\mu}_n + t)} \prod_{i=1}^{n} \left(1 + \mu_i(e^{\lambda} - 1)\right). \tag{19}$$

Optimizing over $\lambda \in \mathbb{R}$ for different means μ_i cannot be achieved in closed-form as in the case of i.i.d. variables. Note that $\log(\rho_{n,\star}^{\exp})$ could be computed as the minimum of a convex objective in λ. We rather explicit an analytical upper bound in Proposition 7.

PROPOSITION 7. *Let $\mu_1, \ldots, \mu_n$ be in $]0,1[$ and $\mu_i \neq \frac{1}{2}$. Then it holds for $t \geqslant 0$ that:*

$$\rho_{n,\star}^{\exp}(t) \leqslant \exp\left(\frac{nt^2/2}{\frac{1}{n}\sum_{i=1}^{n} \log(\frac{\mu_i}{1-\mu_i})}\right).$$

In addition, $\rho_{n,\star}^{\exp}(t) \leqslant \rho_n^{\text{Hoeffding}}$.

Proof. Let us consider $f_i(\lambda) = \log(1 + \mu_i(e^{\lambda} - 1))$. A quadratic upper bound for f was derived by (Jaakkola and Jordan 2000, Section 2.2), such that $\forall \lambda \in \mathbb{R}, \log(1 + \mu_i(e^{\lambda} - 1)) \leqslant \lambda\mu_i + \frac{\lambda^2}{4}\frac{2\mu_i - 1}{\log(\frac{\mu_i}{1-\mu_i})}$. In addition, $\lambda\mu_i + \frac{\lambda^2}{4}\frac{2\mu_i - 1}{\log(\frac{\mu_i}{1-\mu_i})} \leqslant \lambda\mu_i + \frac{\lambda^2}{8}$, leading to the final assertion. □

REMARK 1. In the proof for Proposition 7, considering the naive upper function $\log(1 + \mu_i(e^{\lambda} - 1)) \leqslant \lambda\mu_i + \frac{\lambda^2}{8}$ would have led to Hoeffding's inequality.

Given random variables with different means, Proposition 7 shows a control of $\rho_{n,\star}^{\exp}(t)$ by an upper bound depending on $(\mu_i)_{i=1,\ldots,n}$ and improving Hoeffding's inequality. Therefore, it appears to be suited to different means as well as to two blocks of random variables.

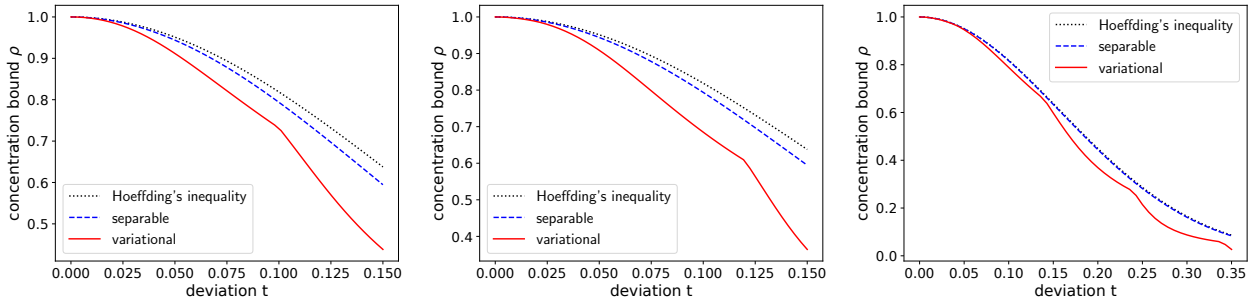
Variational approach. Recall the optimization problem defining the variational approach for random variables in $[0,1]$ having different means $\mu_1,\dots,\mu_n$. For $t \geqslant 0$, we define:

$$\log(\rho_n^{\text{var}}(t)) = \inf_{\alpha\in\mathbb{R}^n,\beta\in\mathbb{R}^n} \sum_{i=1}^n \{\alpha_i + \beta_i\mu_i - 1\}, \text{ such that } -\sum_{i=1}^n \log(\alpha_i + \beta_i x_i) \leqslant 0, \forall x \in \text{ extremal}(\bar{\mathcal{X}}_n), \tag{20}$$

where $\bar{\mathcal{X}}_n = \{(x_1,\dots,x_n) \in [0,1]^n, \sum_{i=1}^n x_i \geqslant nt + \sum_{i=1}^n \mu_i\}$. Without further assumptions on the μ_i's, this optimization problem may have up to $O(n!)$ constraints. To simplify the computations, we consider a first group of variables $X_1,\dots X_m$ with mean μ_1, and a second group $X_{m+1},\dots,X_n$ with mean μ_2, with $1 \leqslant m \leqslant n$. Then, the number of constraints under consideration can be reduced to $O(n)$, as shown in Lemma 1.

LEMMA 1. *Let $\mu_1 = \dots = \mu_m$ and $\mu_{m+1} = \dots = \mu_n$ in* (20)*. Then, the number of constraints in Problem* (20) *reduces to $O(n^2)$.*

Proof. See Appendix B.4. □



(a) $n = 10$, $m = 5$, $\mu_1 = 0.2$, $\mu_2 = 0.8$ (b) $n = 10$, $m = 2$, $\mu_1 = 0.2$, $\mu_2 = 0.8$ (c) $n = 10$, $m = 2$, $\mu_1 = 0.4$ and $\mu_2 = 0.6$

Figure 3 Comparison of the variational (20) **and separable** (19) **approaches to Hoeffding's inequality** (11)**, in the context of independent random variables divided into two blocks of size m and $n-m$, with mean μ_1 and μ_2,**

In Figure 3, we observe that the variational approach (20) improves largely upon the separable scenario (19), as soon as the means of the two blocks differs significantly.

Given n random variables divided into two subgroups with different means, we have formulated tractable formulations in the variational (20) and separable (19) approaches. On the one hand, we provide an upper bound to the separable approach in Proposition 7 using a quadratic upper bound, which still offers improvements over Hoeffding's inequality. On the other hand, the variational approach can be expressed as a manageable convex optimization problem with $O(n)$ constraints, as stated in Lemma 1. Numerical comparisons show that both approaches behave similarly for a large number of variables n, but tend to differ for small n. In the following, we explore the reconstruction of extremal distributions involved at the optimum.

2.3. Reconstructing extremal distributions.

A bound is said to be tight if a distribution satisfies the bound with equality, as defined by Bertsimas and Popescu (2005, Section 2.). Such a distribution is referred to as an *extremal distribution*. We first restrict our attention to one random variable, for which the exact optimization formulation (12) and the separable approach (14) admit analytical solutions. Building on these results, we propose a strategy to construct an extremal distribution for n random variables in the variational (20) and separable approaches (19).

2.3.1. Dirac distributions for one univariate random variable. When it comes to one univariate random variables, we derived in Section 2.1 closed form of the exact (12) and separable (14) approaches. In the following, we construct their corresponding extremal distributions.

Exact optimization problem. Recall the exact optimization problem (12) for one variable, for all $t \geqslant 0$:

$$\rho_1(t) = \sup_{p_1 \in \mathcal{P}([0,1])} \int_0^1 \mathbf{1}_{x_1 \geqslant \mu + t} dp_1(x_1) \text{ such that } \int_0^1 x_1 dp_1(x_1) = \mu. \tag{21}$$

Its exact value is provided in Proposition 4, for all $t \geqslant 0, \rho_1(t) = \frac{\mu}{\mu+t}$. Dualizing twice the optimization problem (21), we propose in Proposition 8 a strategy for reconstructing an extremal distribution.

PROPOSITION 8. *Let $t \geqslant 0$. The following distribution is an optimal solution to* (12)*:*

$$\forall x \in [0,1], p(x) = \frac{t}{\mu + t}\delta_{x=0} + \frac{\mu}{\mu + t}\delta_{x=\mu+t}. \tag{22}$$

Proof. The result is obtained by computing directly $\int_0^1 \mathbf{1}_{x \geqslant \mu+t} dp(x) = \frac{\mu}{\mu+t}$. We rather propose a constructive approach. First, recall the dual of Problem (21): for all $t \geqslant 0, \rho_1(t) = \inf_{\alpha,\beta \in \mathbb{R}} \alpha + \beta\mu, \text{ such that } \forall x \in [0,1], \mathbf{1}_{x \geqslant \mu+t} \leqslant \alpha + \beta x$. By convexity of $x \in [0,1] \mapsto \mathbf{1}_{x \geqslant \mu+t} \leqslant \alpha + \beta x$, the problem reduces for all $t \geqslant 0$ to $\rho_1(t) = \inf_{\alpha,\beta} \alpha + \beta\mu, \text{ such that } \alpha \geqslant 0, 1 \geqslant \alpha + \beta(\mu + t)$. At optimality, $(\alpha_\star, \beta_\star) = (0, \frac{1}{\mu+t})$ meaning that the points $x = 0$ and $x = \mu + t$ are active. We compute again the dual, which reformulates as a problem in the probability space by strong duality: for all $t \geqslant 0$ $\rho_1(t) = \sup_{\lambda_1,\lambda_2 \geqslant 0} \lambda_2, \text{ such that } 1 = \lambda_1 + \lambda_2, \mu = \lambda_2(\mu + t)$. The constraint "$1 = \lambda_1 + \lambda_2$" corresponds to initial constraint $\int_0^1 dp(x) = 1$, and the second one to the first-order moment condition $\int_0^1 x dp(x) = \mu$. At optimality, $(\lambda_{1,\star}, \lambda_{2,\star}) = (\frac{t}{\mu+t}, \frac{\mu}{\mu+t})$. We conclude by identification. □

REMARK 2. Other distributions achieve the optimal bound, such as $p(x) = (1-\mu)\delta_0 + \mu\delta_1$.

Proposition 8 offers a strategy for reconstructing an extremal distribution, mostly by reducing the constraints involved at the optimum to some active points and by dualizing twice.

Separable approach. We derive the same technique described in the proof for Proposition 8 to the separable approach. Recall the separable approach applied to moment-generating functions (14) is defined for $\lambda \in \mathbb{R}$ and $t \in \geqslant 0$,

$$\rho_1^{\exp}(\lambda, t) = \inf_{\alpha,\beta} \alpha + \beta\mu \text{ such that } \forall\, x_1 \in [0,1], e^{\lambda(x_1 - (\mu+t))} \leqslant \alpha + \beta x_1.$$

In the proof for Proposition 5, we showed that $\rho_1^{\exp}(\lambda, t) = e^{-\lambda(\mu+t)}\left((1 + \mu(e^\lambda - 1)\right)$. We compute an example of an extremal distribution in Proposition 9, that is independent of t and λ.

PROPOSITION 9. *Let $t, \lambda \in \mathbb{R}$. The following distribution is an optimal solution to* (14)*:*

$$p(x) = (1-\mu)\delta_{x=0} + \mu\delta_{x=1}.$$

Proof. Following the same approach as in Proposition 8 for determining the extremal distribution of the exact optimization problem. We prove in Appendix B.1.1 that $(\alpha_\star, \beta_\star) = (e^{-\lambda(\mu+t)}, e^{-\lambda(\mu+t)}(e^\lambda - 1))$. After redualizing, it leads to the optimization problem $\sup_{\lambda_1,\lambda_2 \geqslant 0} \lambda_1 e^{-\lambda(\mu+t)} + \lambda_2 e^{\lambda(1-\mu-t)}$, such that $1 = \lambda_1 + \lambda_2, \mu = \lambda_2$. We conclude by identification. □

When studying concentration inequalities applied to a one (univariate) random variable, we derived examples of extremal distributions. This strategy can be decomposed into two steps. First, the convexity properties of the constraints $\forall x \in [0,1], \mathbf{1}_{x \geqslant \mu+t} \leqslant \alpha + \beta x$ and $\forall x \in [0,1], e^{\lambda(x-\mu-t)} \leqslant \alpha + \beta x$ allow determining active constraints and thus identifying the Dirac delta functions involved at optimality. Second, leveraging strong duality, we dualize the dual, leading to an optimization problem with respect to a simplified space of distributions. We then conclude by identification. In what follows, we show that this technique extends well to n univariate random variables.

2.3.2. Generalization to n random variables. We now turn to the problem of deriving extremal distribution for n independent random variables. The strategy developed for one random variable extends well to multiple random variables in the separable approach. In the variational approach however, this process often involves computing analytically the solution in the dual before obtaining the extremal distribution.

Separable approach. The separable approach (7) benefits from a decoupling into n independent optimization problems on one random variables. In the context of Hoeffding's inequality, recall its dual formulation (19) below, for $t \geqslant 0$ and $\lambda \in \mathbb{R}$:

$$\rho_n^{\exp}(\lambda, t) = \prod_{i=1}^n \inf_{\alpha_i,\beta_i} (\alpha_i + \beta_i \mu_i) \text{ such that } \forall x_i \in [0,1]^n, e^{\lambda(x_i - \mu_i - t)} \leqslant \alpha_i + \beta_i \mu_i.$$

Proposition 9 applies on each subproblem i, leading to Corollary 2.

COROLLARY 2. The following distribution is an optimal solution to (19) $p(x) = \prod_{i=1}^n p_i(x_i)$, with $\forall x_i \in [0,1], p_i(x_i) = (1-\mu_i)\delta_{x_i=0} + \mu_i \delta_{x_i=1}$.

Variational approach. At first glance, the variational approach reintroduces coupling between variables in the optimization problem. Let us recall its dual formulation (20) in the context of Hoeffding's inequality, for $t \geqslant 0$,

$$\rho_n^{\text{var}}(t) = \inf_{\alpha \in \mathbb{R}^n, \beta \in \mathbb{R}^n} \prod_{i=1}^n (\alpha_i + \beta_i \mu_i) \text{ such that } \forall\, x \in [0,1]^n, \mathbf{1}_{\sum_{i=1}^n x_i \geqslant \sum_{i=1}^n \mu_i + nt} \leqslant \prod_{i=1}^n (\alpha_i + \beta_i x_i), \quad \forall\, i, \forall\, x_i \in [0,1], 0 \leqslant \alpha_i + \beta_i x_i. \tag{23}$$

The Lagrangian dual of Problem (23) cannot be formulated in closed form due to the product form in the constraints. Let us revisit the original optimization problem from which this problem is derived:

$$\rho_n^{\text{var}} = \inf_{\forall i, u_i \geqslant 0} \inf_{\alpha \in \mathbb{R}^n, \beta \in \mathbb{R}^n} \prod_{i=1}^n (\alpha_i + \beta_i \mu_i), \text{ such that } \forall x \in [0,1]^n, \mathbf{1}_{\sum_{i=1}^n x_i \geqslant \sum_{i=1}^n \mu_i + nt} \leqslant \prod_{i=1}^n u_i(x_i), \quad (24)$$
$$\forall i, \forall x_i \in [0,1], u_i(x_i) \leqslant \alpha_i + \beta_i x_i.$$

Using this formulation, Proposition 10 computes explicitly an extremal distribution.

PROPOSITION 10. *Let* $(u_\star, \alpha_\star, \beta_\star)$ *be a solution to* (24)*. Then,* $\forall i, \forall x_i \in [0,1], u_{i,\star}(x_i) = \alpha_{i,\star} + \beta_{i,\star} x_i$*. In addition, an extremal distribution is given by:*

$$p(x) = \prod_{i=1}^n \left((1-\mu_i)\delta_{x_i=0} + \mu_i \delta_{x_i=1} \right).$$

Proof. Let $(u_\star, \alpha_\star, \beta_\star)$ be optimal solutions in (24). Then, Proposition 2 provides the form of the optimal product-function upper bounding F, that we recall: $\forall i, \forall x_i \in [0,1], u_{i,\star}(x_i) = \alpha_{i,\star} + \beta_{i,\star} x_i$. Thus, considering this specific product-function, we have

$$\begin{aligned}
\rho_{relax}^n &= \prod_{i=1}^n \sup_{p_i \in \mathcal{P}(\mathcal{X})} \int_0^1 u(x_i) dp_i(x_i) \text{ such that } \int_0^1 x_i dp_i(x_i) = \mu_i, \\
&= \prod_{i=1}^n \inf_{\lambda_i, \nu_i} (\lambda_i \mu_i + \nu_i) \text{ such that } \alpha_{i,\star} - \nu_i + (\beta_{i,\star} - \lambda_i) x_i \leqslant 0, \forall x_i \in [0,1],
\end{aligned}$$

where the constraint "$\alpha_{i,\star} - \nu_i + (\beta_{i,\star} - \lambda_i) x_i \leqslant 0, \forall x_i \in [0,1]$" reduces to $\alpha_{i,\star} - \nu_i \leqslant 0$ for $x_i = 0$ and $\alpha_{i,\star} - \nu_i + \beta_{i,\star} - \lambda_i \leqslant 0$ for $x_i = 1$. We conclude by identification that $p(x) = (1-\mu)\delta_0 + \mu \delta_1$. □.

The extremal distribution derived in Proposition 10 is exactly equal to the extremal distribution in the separable treatment approach moment-generating functions in Proposition 2. Again, it is independent of the deviation t. In both cases, it turns out that the dependence in t is only supported in the optimal upper function $U_\star$, either in the moment generating function or in the linear function parametrized by $(\alpha_\star, \beta_\star)$, as detailed in Appendix B.3 for $n = 2$.

In this section, we have thus refined Hoeffding's inequality through two different approaches. When the random variables are i.i.d., the separable approach applied to moment-generating functions aligns with the large deviation principle and slightly improves the traditional Hoeffding's inequality. In contrast, the variational approach yields significantly smaller bounds for a small number of variables but requires $O(n)$ constraints. The case of distinct means $\mu_1, \ldots, \mu_n$ is more advanced and we only explicitly attacked the scenario of two blocks with distinct means. Finally, we proposed a strategy for reconstructing an extremal distribution. As a natural extension, these strategies could potentially be applied to Bennett or Bernstein's inequalities, which assume first and second-order conditions. However, extending to higher-order assumptions reveals challenges where both the separable and variational approaches fail to provide computable solutions. In what follows, we propose a new family of upper bounds adapted to such scenarios.

3. A polynomial approach based on sum-of-square decomposition.

The separable and variational approaches are constructed from the family of product-functions (6) that upper bound F on $\mathcal{X}$. They turn out to be effective for finite first-order moments, but remain computationally out of reach when assuming fixed higher-order moments. Even a simple second-order assumption in Hoeffding's inequality (or an assumption on the variance) results in challenging constraints in the separable approach:

$$\rho_n^{\exp}(\lambda) = \prod_{i=1}^{n} \inf_{\alpha_i,\beta_i} (\alpha_i + \beta_i^{(1)} x_i + \beta_i^{(2)}) \text{ s.t. } \forall x_i \in [0,1]^n, e^{\lambda(x_i-\mu_i-t)} \leqslant \alpha_i + \beta_i^{(1)} x_i + \beta_i^{(2)} x_i^2,$$

where $\mu_i^{(1)}$ (resp. $\mu_i^{(2)}$) represents the first (resp.second) order moment condition. The constraint "$\forall x_i \in [0,1]^n, e^{\lambda(x_i-\mu_i-t)} - (\alpha_i + \beta_i^{(1)} x_i + \beta_i^{(2)} x_i^2)$ admits no closed-form solution. This issue also prevents from computing the large deviations for i.i.d. random variables, which involves computing $\Gamma(t) = \log(\mathbb{E}[\exp(tX_1)])$ as presented in Theorem 2. Similarly, the variational approach involves polynomial constraints with a product structure, making the optimization problem more complex to solve:

$$\rho_n^{\text{var}} = \inf_{\alpha\in\mathbb{R}^n, \beta\in\mathbb{R}^{n\times m}} \prod_{i=1}^{n} (\alpha_i + \beta_i^{(1)}\mu^{(1)} + \beta_i^{(2)}\mu^{(2)}) \text{ such that } \forall x_i \in \mathcal{X}, \forall i, \alpha_i + \beta_i^{(1)} x_i + \beta_i^{(2)} x_i^2 \geqslant 0,$$
$$\forall x \in \mathcal{X}^n, F(x_1,\ldots,x_n) \leqslant \prod_{i=1}^{n} (\alpha_i + \beta_i^{(1)} x_i + \beta_i^{(2)} x_i^2).$$

In this section, we explore polynomial families of upper bounds adapted to such scenarios. First, we propose to analyze Problem (5) using the family of linear upper bounds, for which we derive closed-form upper bounds. This *linear approach* offers already an improvement to Hoeffding's inequality in comparison with the variational approach for small values of t. We then extend this approach to a polynomial family of upper bounds, whose degree equals the number of variables. This results in an optimization problem with an infinite number of polynomial constraints. This so-called *polynomial approach* is closely related to the work of Bertsimas and Popescu (2005), who introduced a series of SDPs to approximate the generalized problem of moments for multivariate random variables. This approach numerically improves Bernstein's and partially Bennett's inequality. Finally, we introduce a *feature-based approach* generalizing the polynomial approach to a broader family of upper bounds. This allows in particular analyzing Hoeffding's inequality using second-order polynomials.

3.1. A simple linear upper bound for Hoeffding's concentration.

Before studying polynomial upper bounds, we start by exploring the simpler family of linear upper bounds and applying it in the context of Hoeffding's inequality. This scenario outlines the key concepts that inspire the polynomial approach and results in an optimization problem that can be solved in closed-form.

Let $X_1,\ldots,X_n$ be i.i.d. random variables with mean $\mathbb{E}[X_i] = \mu$ and taking their values in $[0,1]$ and let us introduce the family of linear functions $\mathcal{L} = \{a^\top x + b, a \in \mathbb{R}^n, b \in \mathbb{R}\}$.

REMARK 3. For $n \geqslant 2$, notice that linear functions cannot be formulated as product-functions. Indeed, for two i.i.d. random variables, product-functions takes the form $U(x) = (\alpha + \beta x_1)(\alpha + \beta x_2) = \alpha^2 + \alpha\beta(x_1 + x_2) + \beta^2 x_1 x_2$.

The associated optimization problem in the linear approach takes the form, for $t \geqslant 0$:

$$\begin{aligned} \rho_n^{\text{lin}}(t) = \inf_{a\in\mathbb{R}^n, b\in\mathbb{R}} \sup_{\forall i, p_i \in \mathcal{P}_{\mu_i}(\mathcal{X})} \int_{\mathcal{X}^n} \left(\sum_{i=1}^n a_i x_i + b \right) dp_1(x_1) \cdots dp_n(x_n), \\ \text{such that } \forall x \in \mathcal{X}^n, F(x) \leqslant \sum_{i=1}^n a_i x_i + b. \end{aligned} \tag{25}$$

By construction, $\rho_n \leqslant \rho_n^{\text{lin}}$. We solve Problem (25) analytically in Proposition 11.

PROPOSITION 11. *Let $X_1, \ldots, X_n$ be i.i.d. random variables taking their values in $[0,1]$ and with finite mean $\mathbb{E}[X_i] = \mu$. Then, it holds for all $t \in [0, 1-\mu]$ that:*

$$\rho_n^{\text{lin}}(t) = \frac{\mu}{\mu + t}.$$

Proof. Under the assumptions of Hoeffding's inequality and by symmetry,it holds for $t \geqslant 0$:

$$\rho_n^{\text{lin}}(t) = \inf_{a\in\mathbb{R}, b\in\mathbb{R}} an\mu + b \text{ such that } \forall x \in [0,1]^n, \mathbf{1}_{x_1+\cdots+x_n \geqslant n(\mu+t)} \leqslant a \sum_{i=1}^n x_i + b.$$

It follows that $b = 0$ by considering $x = 0$ and minimizing with respect to b. By construction, $\sum_{i=1}^n x_i \geqslant n(\mu + t)$ implies $1 \leqslant a \sum_{i=1}^n x_i \leqslant an(\mu + t)$. We conclude that $a = \frac{1}{n(\mu+t)}$ and the desired result. □

In Proposition 11, we show that ρ_n^{lin} is exactly equal to the exact bound for one univariate random variable as defined in (12), that is $\rho_n^{\text{lin}} = \rho_1$. In addition, ρ_n^{lin} shows a dependence on the mean μ, but no dependence on the number of variables under consideration.

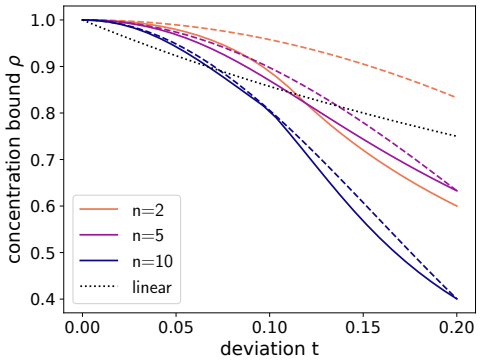


Figure 4 **Comparison of the bound obtained in the linear approach** (25) **to the variational** (18) **and separable** (16) **(in dashed lines) approaches, for $\mu = 0.6$ and several values for n.**

Surprisingly in Figure 4, it happens to improve the variational (18) and separable (16) approach for small values of the deviation t, even for a large number of variables n. However, for larger values of the deviation,

the variational and separable approaches significantly outperform the linear approach. In the following, we introduce a family of polynomials encompassing linear functions, that comes at the cost of computationally more expensive approximations.

3.2. A polynomial upper bound: a SoS approximation for polynomial moment assumptions.

This section studies an alternate approach to (1) using a family of polynomial upper bounds. Under some assumptions on the degree of these polynomials, tt turns out that this family contains the linear and variational approaches, as feasible points. Its associated Problem (5) is therefore guaranteed to provide better approximations to the generalized problem of moments.

We assume $X_1, \ldots, X_n$ to be independent random variables with monomials moments of degree at most $a \in \mathrm{N}^\star$, that is $\forall i, \mathbb{E}[g_i(X_i)] = (\mathbb{E}[X_i^k])_{k=1,\ldots,a} = (\mu_i^{(k)})_{k=1,\ldots,a}$. Denote $J_d = \{(k_1, \ldots k_n) \in \mathrm{N}^n,\ k_i \in \mathrm{N},\ k_1 + \ldots + k_n \leqslant d\}$ and $\forall \kappa \in J_d$, $\bar{x}^\kappa = \prod_{i=1}^n x_i^{k_i}$ monomials of degree d. The family of (multivariate) polynomials under consideration is

$$\mathcal{Q}_d^n = \left\{ Q(x) = \sum_{\kappa \in J_d} q_\kappa \bar{x}^\kappa, q_\kappa \in \mathbb{R}^{|J_d|} \right\}, \tag{26}$$

with a certain degree $d \in \mathrm{N}^+$. In particular, linear functions belongs to this set $\mathcal{L} \subset \mathcal{Q}_d^n$ along with product functions for $d \geqslant n$. Finally, recall that $F(x) = \mathbf{1}_{x \in S}$, where S has a structure specified in Assumption 1.

ASSUMPTION 1. *Let $S = \{x \in \mathbb{R}^n, h_1(x) \geqslant 0, \ldots, h_m(x) \geqslant 0\}$ and assume that there exists $s(x)$ a polynomial such that $s(x) = s_0(x) + \sum_{j=1}^m s_j(x) h_j(x)$, with $\{x \in \mathbb{R}^n, h(x) \geqslant 0\}$ a compact set and where $h_i(x)$ are polynomial that admits a sum-of-square decomposition.*

In particular, any compact polyhedron verifies Assumption 1 (Bertsimas and Popescu 2005, Theorem 4.1). Then, Problem (5) associated with polynomials in $\mathcal{Q}_d^n$ takes the form,

$$\begin{aligned} \rho_n^{\text{polynomial},d} = \inf_{Q \in \mathcal{Q}_d^n} \sup_{\forall i, p_i \in \mathcal{P}(\mathcal{X})} & \int_{\mathcal{X}} \cdots \int_{\mathcal{X}} Q(x_1, \ldots, x_n) p_1(x_1) \cdots dp_n(x_n), \\ \text{such that } \forall i, \forall k \in 1, \ldots, a, & \int_{\mathcal{X}} x_i^k dp_i(x_i) = \mu_i^{(k)}, \text{ and } \forall x \in \mathcal{X}^n, \mathbf{1}_{x \in S} \leqslant Q(x). \end{aligned} \tag{27}$$

By construction, $\rho_n \leqslant \rho_n^{\text{polynomial,a}}$.

PROPOSITION 12. *Let $d \leqslant a$, and let the covariance matrix take the form $\sigma_\kappa = \prod_{i=1}^n \mu_i^{(k_i)}$ for $\kappa \in J_d$. Then Problem* (27) *reformulates as:*

$$\rho_n^{\text{polynomial},d} = \inf_{q_\kappa \in \mathbb{R}^{|J_d|}} \sum_{\kappa \in J_d} q_\kappa \sigma_\kappa \text{ such that } \forall x \in \mathcal{X}^n, \mathbf{1}_{x \in S} \leqslant \sum_{\kappa \in J_d} \bar{x}_\kappa q_\kappa. \tag{28}$$

Proof. By definition $Q \in \mathcal{Q}_d^n$, $\forall x \in \mathcal{X}^d, Q(x) = \sum_{\kappa \in J_d} q_\kappa \bar{x}^\kappa$. Since $d \leqslant a$ and $\forall i, \forall k \in 1, \ldots, a, \int_{\mathcal{X}} x_i^k dp_i(x_i) = \mu_i^{(k)}$, the objective function $\int_{x \in \mathcal{X}^n} Q(x) dp_1(x_1) \cdots dp_n(x_n)$ for $Q \in \mathcal{Q}_d^n$ can be decomposed as the weighted sum of moment μ_i^k. □

By construction, Proposition 12 reveals a condition on the degree d, so that $\mathbb{E}[Q]$ for $Q \in \mathcal{Q}_d$ formulates as a product combination of moments $(\mu_i^{(k)})_{i,k}$. In addition, Problem (28) happens to be exactly the Lagrangian dual of the generalized problem of moments applied to multivariate random variable in $\mathbb{R}^n$ (Bertsimas and Popescu 2005, Equation 2.2). The major difference lies in the structure of the matrix $\forall \kappa \in J_d, \sigma_\kappa = \mathbb{E}[\prod_{i=1}^n X_i^{k_i}] = \prod_{i=1}^n E[X_i^{k_i}]$ due to independence.

Problem (28) suffers from an infinite number of constraints "$\forall x \in \mathcal{X}^n, \mathbf{1}_{x\in S} \leqslant \sum_{\kappa\in J_d} \bar{x}_\kappa q_\kappa$". Connections between such nonnegative polynomial and sum-of-square decomposition have been hightlighted by several authors (Lasserre 2002, 2008, de Klerk and Laurent 2019). They were later formulated as SDPs by Parrilo (2003) and Lasserre (2001). In their study of optimal bounds for the generalized problem of moments, Bertsimas and Popescu (2005) constructed a sequence of SDPs approximating $\rho_n^{\text{polynomial},d}$, that is recalled below.

THEOREM 3. *(Bertsimas and Popescu 2005, Theorem 4.3) Let $S = \{h_j(x) \geqslant 0, j = 1, \dots, m\}$ and $\mathcal{X} = \{\omega_j(x) \geqslant 0, j = 1, \dots, l\}$ verify Assumption 1. For every $\epsilon > 0$, there exists a nonnegative integer $r \in \mathrm{N}$ such that $|\rho_n^{\text{polynomial},d} - \tilde{\rho}_n^{\text{polynomial,d}}(r)| \leqslant \epsilon$, where $\tilde{\rho}_n^{\text{polynomial,d}}(r)$ is the value of the following SDP:*

$$
\begin{aligned}
\tilde{\rho}_n^{\text{polynomial,d}}(r) = \inf_{q_\kappa, S\succcurlyeq 0, P\succcurlyeq 0} & \sum_{\kappa\in J_d} q_\kappa \sigma_\kappa, \\
\text{such that } & q_\kappa - \delta_{\kappa=0} = s_\kappa^0 + \sum_{i=1}^m \sum_{\eta,\theta\in J_r, \eta+\theta=\kappa} s_\eta^i h_\theta^i, \ \forall \kappa \in J_d, \\
& 0 = \sum_{i=1}^m \sum_{\eta,\theta\in J_r, \eta+\theta=\kappa} s_\eta^i h_\theta^i, \ \forall \kappa \in J_r \backslash J_d, \\
& q_\kappa = p_\kappa^0 + \sum_{i=1}^l \sum_{\eta,\theta\in J_r, \eta+\theta=\kappa} p_\eta^i \omega_\theta^i, \ \forall \kappa \in J_d, \\
& 0 = p_\kappa^0 + \sum_{i=1}^l \sum_{\eta,\theta\in J_d, \eta+\theta=\kappa} p_\eta^i \omega_\theta^i, \ \forall \kappa \in J_r \backslash J_d, \\
& s_\kappa^i = \sum_{\eta,\theta\in J_r, \eta+\theta=\kappa} s_{\eta,\theta}^i, \ S^i = [s_{\eta,\theta}]_{\eta,\theta\in J_r} \succcurlyeq 0, \ \forall \kappa \in J_r, \ i = 1, \dots, m, \\
& p_\kappa^i = \sum_{\eta,\theta\in J_r, \eta+\theta=\kappa} p_{\eta,\theta}^i, \ P^i = [p_{\eta,\theta}]_{\eta,\theta\in J_r} \succcurlyeq 0, \ \forall \kappa \in J_r, \ i = 1, \dots, l.
\end{aligned}
\tag{29}
$$

Theorem 3 provides a sequence of SDPs approximation Problem (28), but does not specify at which degree r a certain precision level ϵ is attained. By construction, it is clear that their degree r must be larger than the degree d of the polynomials under consideration. Bertsimas and Popescu (2005) highlighted a hierarchy between these SDP approximations, which is explicited in Corollary 3.

COROLLARY 3. (Bertsimas and Popescu 2005, Corollary 4.4) Let $\rho_n^{\text{polynomial},d}$ be defined as in (28) and $\tilde{\rho}_n^{\text{polynomial,d}}(r)$ as in (29). Then, it follows that:

$$
\rho_n \leqslant \rho_n^{\text{polynomial},d} \leqslant \tilde{\rho}_n^{\text{polynomial,d}}(r) \leqslant \cdots \leqslant \tilde{\rho}_n^{\text{polynomial,d}}(1). \tag{30}
$$

From a numerical perspective, the size of SDPs in (29) grows exponentially with degrees r and d, limiting both the precision of the approximation problem (28). In the next section however, we apply these approximations to two basic concentration inequalities and improve the existing bounds for a small number of variables n and a small degree r.

3.3. Applications to Bernstein's and Bennett's inequalities

The variational and separable approaches failed to provide tractable optimization problem for second-order conditions, that appear for example in Bennett's and Bernstein's inequalities. For two random variables, we compute refined bounds using the polynomial approach, and more precisely the SDP aproximations defined in (29).

Bernstein's inequality. Bernstein's inequality controls the deviation of the sum of independent random variables to their mean, given an appropriate control of moments. There exists different versions for Bernstein's inequality, and we consider a convenient version requiring finite second-order moments, as provided for example in (Boucheron et al. 2013, Corollary 2.11).

COROLLARY 1. *Let $X_1, \dots, X_n$ be independent random variables in $[-c, c]$ with means $\mathbb{E}[X_i] = \mu_i$, and variance $v = \sum_{i=1}^n \mathbb{E}[X_i^2]$. Then, for all $t > 0$,*

$$\mathbb{P}\left(\sum_{i=1}^n X_i \geqslant nt + \sum_{i=1}^n \mu_i\right) \leqslant \exp\left(-\frac{n^2 \frac{t^2}{2}}{v + c\frac{nt}{3}}\right). \tag{31}$$

Let X_1, X_2 be two i.i.d. random variables taking their values almost surely in $[-1, 1]$ with $\mathbb{E}[X_1] = \mathbb{E}[X_2] = \mu^{(1)}$, and $\mathbb{E}[X_1^2] = \mathbb{E}[X_2^2] = \mu^{(2)}$. We consider the polynomial family $\mathcal{Q}_2 = \{x \in \mathbb{R}^2 \mapsto x_{(2)}^\top Q x_{(2)}, Q \in \mathbb{R}^{(6\times 6)}\}$ where $x_{(2)} = (1, x_1, x_2, x_1x_2, x_1^2, x_2^2)$. Then, Problem (28) takes the form:

$$\begin{aligned} \forall t \in \geqslant 0,\ \rho_2^{\text{polynomial},2}(t) = \inf_{Q\in\mathbb{R}^{(6\times 6)}} \operatorname{Tr}(Q\Sigma) \text{ such that } \forall x \in \mathcal{X}^2,\ 1 \leqslant x_{(2)}^\top Q x_{(2)}, \\ \forall x \in [-1,1]^2, 0 \leqslant x_{(2)}^\top Q x_{(2)}, \end{aligned} \tag{32}$$

where $\bar{\mathcal{X}_2} = \{(x_1, x_2) \in [-1, 1], x_1 + x_2 \geqslant 2(\mu^{(1)} + t)\}$, $\Sigma = \sigma\sigma^\top$ and $\sigma = (1, \mu^{(1)}, \mu^{(1)}, (\mu^{(1)})^2, \mu^{(2)}, \mu^{(2)})$. It results from Theorem 3 that Problem (32) can be approximated by a hierarchy of sum-of-square (29).

Figure 5 shows that the SDP approximation of the polynomial approach (32) improves upon Bernstein's inequality (31) for a small degree in the SoS hierarchy (here $r = 2$). This improvement is not straightforward, since proofs for Bernstein's inequality often require a specific control of moments $\mathbb{E}[X_i^k]$ for all $k \in \mathrm{N}$ (see, e.g.,(Boucheron et al. 2013, proof for Theorem 2.9)). Therefore, considering higher-degree d in the polynomial approximations would probably produce a lower value for $\rho_2^{\text{polynomial,d}}$. However, increasing d requires to increase the degree $r \in \mathrm{R}$ of the sum-of-square approximations (29) and thereby, leads to very large SDPs that cannot be handled by our solvers.

Bennett's inequality. Bennett's inequality is similar with Bernstein's inequality, but applies to upper bounded random variables, as recalled in Theorem 4.

(a) $\mu_1 = -0.3$, $\mu_2 = 0.1$ (b) $\mu_1 = -0.3$, $\mu_2 = 0.5$

Figure 5 **Comparison of Bernstein's inequality** (31) **to an SDP approximation** (29) **to the polynomial approach** (32)**, as a function of** t**, for** $n = 2$**,** $d = 2$ **and** $r = 2$**.**

THEOREM 4. *Bennett (1968) Let* $X_1, \cdots, X_n$ *be i.i.d. random variables such that* $X_i \leqslant a$ *almost surely and* $\sigma^2 = \sum_{i=1}^n \mathbb{E}[(X_i - \mathbb{E}X_i)^2]$*, then for any* $t \geqslant 0$*,*

$$\mathbb{P}\left(\sum_{i=1}^{n}\{X_i - \mathbb{E}[X_i]\} \geqslant nt\right) \leqslant \exp\left(-\frac{\sigma^2}{a^2}h(\frac{atn}{\sigma^2})\right), \tag{33}$$

where $h(t) = (1+t)\log(1+t) - t$*. It implies that* $\mathbb{P}(\sum_{i=1}^n \{X_i - \mathbb{E}[X_i]\} \geqslant nt) \leqslant \exp\left(-\frac{t^2n^2}{2(\sigma^2 + atn/3)}\right)$.

Let X_1, X_2 be two i.i.d. random variables taking their values almost surely in $\mathcal{Z} = [-\infty, 1]$ with $\mathbb{E}[X_1] = \mathbb{E}[X_2] = \mu^{(1)}$, and $\mathbb{E}[X_1^2] = \mathbb{E}[X_2^2] = \mu^{(2)}$. Then, Problem (28) takes the form:

$$\begin{aligned}\rho_2^{\text{polynomial},2}(t) = \inf_{Q \in \mathbb{R}^{6\times 6}} \ \mathrm{Tr}(Q\Sigma) \text{ such that } \forall x \in \bar{\mathcal{Z}}_2,\ 1 \leqslant x_{(2)}^\top Q x_{(2)},\\ \forall x \in [-\infty, 1]^2, 0 \leqslant x_{(2)}^\top Q x_{(2)},\end{aligned} \tag{34}$$

where $\bar{\mathcal{Z}}_2 = \{(x_1, x_2) \in [-\infty, 1], x_1 + x_2 \geqslant 2(\mu^{(1)} + t)\}$, $\Sigma = \sigma\sigma^\top$ and $\sigma = (1, \mu^{(1)}, \mu^{(1)}, (\mu^{(1)})^2, \mu^{(2)}, \mu^{(2)})$.

(a) $\mu_1 = -0.3$, $\mu_2 = 1$ (b) $\mu_1 = -0.3$, $\mu_2 = 0.2$

Figure 6 **Comparison of Bennett's inequality** (33) **to an SDP approximation** (29) **to the polynomial approach** (34)**, as a function of** t **for** $n = 2$**,** $d = 2$ **and** $r = 2$**.**

Figure 6 reveals the limit of the polynomial approach, which fails to improve Bennett's inequality for all values of $t \geqslant 0$. The proof relies indeed on additional arguments, such as Jensen's inequality, or Taylor approximations such as $u \mapsto \log(1+u) \leqslant u$, which are not exploited in this approach.

To conclude, the polynomial approach (28) effectively refines Bernstein and Bennett's inequalities for some range of values $t \geqslant 0$. However it is hindered by the necessity of large SDP approximations: handling higher-order polynomials and a large number of random variables increases the size of the SDPs under consideration, making them difficult to solve. A review of the complexity of semidefinite optimization and interior point methods can be found in Wolkowicz et al. (2000), Vandenberghe and Boyd (1996).

3.4. Higher-order polynomial approximations : a variational reformulation.

The variational, separable and linear approaches allowed refining Hoeffding's inequality. In the polynomial approach developed above, Proposition 12 entails that the degree of the polynomials under consideration is controlled by the highest-order moment. In the context of Hoeffding's inequality, where the first-order moment is finite, it implies that the best polynomial representation is actually linear ($d \leqslant 1$).

In the following, we introduce a numerical procedure to incorporate higher-order polynomials for studying Hoeffding's inequality. We then lay the foundation for a feature-based approach that generalizes the polynomial approach to broader families of upper functions. Finally, we derive a SDP relaxation for the case of two independent random variables.

3.4.1. Studying Hoeffding's inequality with second-order polynomials. Hoeffding's inequality requires finite first-order moments, together with almost surely bounded variables. In what follows, we show how it affects second-order moments. Based on this, we integrate second-order moment conditions into a polynomial approach for studying Hoeffding's inequality.

Let $X_1, \dots, X_n$ be independent random variables taking their values in $[0,1]$ almost surely with $\mathbb{E}[X_i] = \mu_i^{(1)}$ for all $i = 1, \dots, n$. Lemma (C) ensures the existence and a control on the second-order moment (see proof in Appendix C).

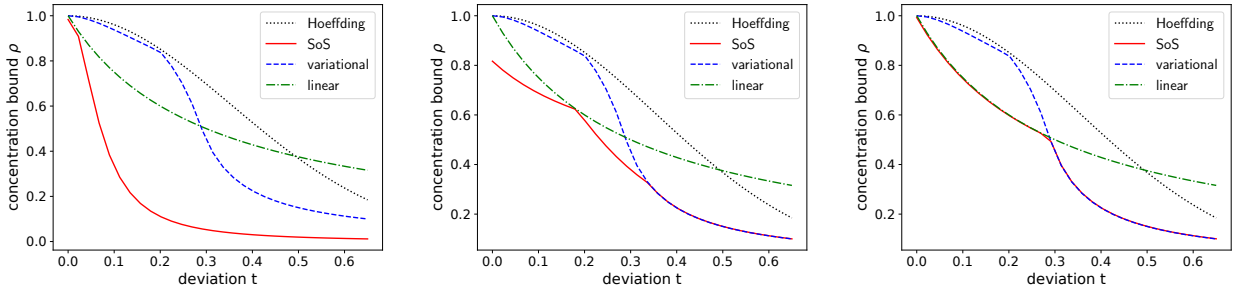
LEMMA 2. *Let X be a random variable almost surely on $[0,1]$ with mean $\mathbb{E}[X] = \mu^{(1)}$. Then X admits a finite second-order moment, denoted $\mathbb{E}[X^2] = \mu^{(2)}$. In addition, it holds that $(\mu^{(1)})^2 \leqslant \mu^{(2)} \leqslant \mu^{(1)}$.*

Lemma 2 provides a control of the second-order moment by the first-order moment for bounded random variables. From that, we define an optimization problem based on the polynomial approach (28):

$$\begin{aligned}\tilde{\rho}_n^{\text{polynomial},2} = \inf_{\mu^{(2)} \in [(\mu^{(1)})^2, \mu^{(1)}]} \inf_{Q \in \mathcal{Q}_2^n} \sup_{p_1, \dots, p_n \in \mathcal{P}(\mathcal{X})} & \int_{\mathcal{X}} \dots \int_{\mathcal{X}} Q(x_1, \dots, x_n) dp_1(x_1) \cdots dp_n(x_n), \\ & \text{such that } \forall i, \int_{\mathcal{X}} x_i dp_i(x_i) = \mu_i^{(1)}, \\ & \forall i, \int_{\mathcal{X}} x_i^2 dp_i(x_i) = \mu_i^{(2)}, \\ & \forall x \in \mathcal{X}^n, \mathbf{1}_{x \in S} \leqslant Q(x),\end{aligned} \tag{35}$$

$$\tilde{\rho}_n^{\text{polynomial},2} = \inf_{\mu^{(2)} \in [(\mu^{(1)})^2, \mu^{(1)}]} \rho_n^{\text{polynomial},2}(\mu^{(2)}).$$

Given a value of $\mu^{(2)} \in [(\mu^{(1)})^2, \mu^{(1)}]$, the inner optimization problem defining $\rho_n^{\text{polynomial},2}$ can be approached with a hierarchy of sum-of-square defining $\rho_n^{\text{polynomial},2}(r, \mu^{(2)})$, as stated in Proposition 12 and Theorem 3. In Figure 7, we use a gridsearch procedure on $\mu^{(2)}$ for approximating $\rho_n^{\text{polynomial},2}(\mu^{(2)})$. It turns out that optimizing $\rho_n^{\text{polynomial},2}(\mu^{(2)})$ with respect to $\mu^{(2)}$ aligns exactly with the minimum bound achieved by the linear (25) and variational (18) approaches.



(a) $\mu_1 = 0.3$ and $\mu_2 = 0.1$ (b) $\mu_1 = 0.3$ and $\mu_2 = 0.2$ (c) Optimized version for $\mu_1 = 0.3$.

Figure 7 Comparison of the bound in polynomial approach (28) **(with an SDP approximation of degree $r = 3$), to the variational** (18) **and linear** (25) **approaches, and to Hoeffding's inequality** (11)**, as a function of the deviation parameter t for $n = 2$.**

In conclusion, linear functions and product-functions appear to be the minimal polynomial families suitable for studying Hoeffding's inequality. Although the polynomial approach (35) incurs high computational costs, the linear approach offers a closed-form solution and the variational approach provides tractable reformulations for i.i.d. random variables and structured independent variables.

3.4.2. Generalization to tighter bounds for F: a variational feature-based formulation We have previously explored the family of product-functions, linear functions as well as polynomials. These approaches were limited by the power of representation allowed by moments (see Proposition 12). In the context of Hoeffding's inequality, we addressed higher-order polynomials using finite lower-order moments through a numerical procedure. We now extend this approach by laying the foundations for a feature-based approach.

Let us start by recalling the variational formulation (5):

$$\inf_{H \in \mathcal{H}} \sup_{\forall i, p_i \in \mathcal{P}_{\mu_i}(\mathcal{X})} \int_{\mathcal{X}^n} H(x) dp_1(x_1) \cdots dp_n(x_n) \text{ such that } \forall x \in \mathcal{X}^n, F(x) \leqslant H(x),$$

where $\mathcal{H}$ is a family of upper bounds. We introduce a feature vector $\phi : \mathcal{X} \mapsto \mathbb{R}^l, l \in \mathrm{N}$ such that functions have the following representation:

$$F(x) = \left\langle \bar{F}, \bigotimes_{i=1}^{n} \phi(x_i) \right\rangle = \sum_{i_1, \ldots, i_n = 1}^{l} \bar{F}_{i_1, \ldots, i_n} \prod_{k=1}^{n} \phi(x_k)_{i_k}.$$

In the polynomial (resp. variational) approach for examples, features ϕ were monomials (resp. linear functions). Let us consider a family of functions H decomposing with respect to these features:

$$\sup_{\forall i, p_i \in \mathcal{P}_{\mu_i}(\mathcal{X})} \int_{\mathcal{X}^n} H(x) dp_1(x_1) \cdots dp_n(x_n) = \sup_{\forall i, p_i \in \mathcal{P}_{\mu_i}(\mathcal{X})} \left\langle \bar{H}, \bigotimes_{i=1}^n \int_{\mathcal{X}} \phi(x_i) dp_i(x_i) \right\rangle = \sup_{\forall i, \sigma_i \in \mathcal{K}(\mu_i)} \left\langle \bar{H}, \bigotimes_{i=1}^n \sigma_i \right\rangle,$$

where $\mathcal{K}(\mu_i)$ is the set of achievable moments $\mathbb{E}_{p_i \in \mathcal{P}_{\mu_i}(\mathcal{X})}[\phi(X_i)] = \int_{\mathcal{X}} \phi(x_i) dp_i(x_i)$ such that $p_i \in \mathcal{P}_{\mu_i}(\mathcal{X})$. In the context of Hoeffding's inequality, the set of achievable moments $\mathcal{K}(\mu_i)$ corresponds to the intuitive idea that there is no assumption on second-order moments, but that they are related to lower-order moments. Then, the overall optimization problem takes the form:

$$\inf_{H \in \mathcal{H}} \sup_{\forall i, \sigma_i \in \mathcal{K}(\mu_i)} \left\langle \bar{H}, \bigotimes_{i=1}^n \sigma_i \right\rangle \text{ such that } \forall x \in \mathcal{X}^n, \left\langle \bar{F} - \bar{H}, \bigotimes_{i=1}^n \phi(x_i) \right\rangle \leqslant 0. \tag{36}$$

Compared to the polynomial approach developed above (28), Problem (36) provides a more generic formulation for any family of features. The family of features must satisfy two key components : a (tighter) relaxation of the constraint $\forall x \in \mathcal{X}^n, \langle \bar{F} - \bar{H}, \bigotimes_{i=1}^n \phi(x_i) \rangle \leqslant 0$, which can be managed using sum-of-square for polynomial features, and a relaxation for $\sup_{\forall i, \sigma_i \in \mathcal{K}(\mu_i)} \langle \bar{H}, \bigotimes_{i=1}^n \sigma_i \rangle$. In the following proposition, we analyze a simple relaxation of $\sup_{\forall i, \sigma_i \in \mathcal{K}(\mu_i)} \langle \bar{H}, \bigotimes_{i=1}^n \sigma_i \rangle$ for two i.i.d. random variables in the context of Hoeffding's inequality (i.e., $n = 2$). For this case, the tensor formulation simplifies into matrices and admits an SDP relaxation.

PROPOSITION 13. *Let X_1, X_2 be i.i.d. random variables taking their values almost surely in $[0,1]$ with finite mean $\mu^{(1)}$ and let $\phi(x) = (1, x, x^2)$ be the feature vector. Then,*

$$\sup_{\sigma \in \mathcal{K}(\mu^{(1)})} \operatorname{Tr}\left(\tilde{H} \sigma \sigma^\top\right) \leqslant \sup_{M \succcurlyeq 0} \operatorname{Tr}\left(\tilde{H} M\right) \text{ such that } 0 \leqslant \operatorname{Tr}(M E_{1,1}) \leqslant \operatorname{Tr}(M E_{0,2}) \leqslant \operatorname{Tr}(M E_{0,1}) \leqslant 1,$$
$$\operatorname{Tr}(M E_{0,0}) = 1.$$

Proof. Let $\tilde{H} \in \mathbb{R}^{3 \times 3}$ be a symmetric matrix representation of vector $\bar{H} \in \mathbb{R}^6$, such that $\sup_{\sigma \in \mathcal{K}(\mu^{(1)})} \operatorname{Tr}(\tilde{H} \sigma \sigma^\top) = \sup_{M \succcurlyeq 0} \operatorname{Tr}(\tilde{H} M)$ such that $M \in \operatorname{hull}\{\sigma \sigma^\top, \sigma \in \mathcal{K}(\mu)\}$. By definition, $\sigma = (1, \mu^{(1)}, \mu^{(2)})$ holds for the first and second-order moments. We relax this problem by optimizing over $M \succcurlyeq 0$ and incorporating additional constraints on M to ensure it accurately incorporates relationships between the first and second-order moments (Lemma 2, namely $(\mu^{(1)})^2 \leqslant \mu^{(2)} \leqslant \mu^{(1)}$ and $\int_0^1 dp(x) = 1$.). $\square$.

Proposition 13 provides an SDP relaxation for approximating Hoeffding's inequality for two random variables and given a quadratic features. Despite the simple relaxation, solving this problem requires to solve efficiently a saddle-point problem and to quantify how far such a relaxation is from the generalized problem of moments (1), which are left to future work.

4. Conclusion and future works.

Conclusion. In this work, we introduced two families of upper bounds for approximating the generalized problem of moments for independent random variables. First, we studied a separable approach, leveraging specific upper functions adapted to finite first-order moments. This approach is complemented by a convex variational method, optimizing over the entire family of product-functions. When studying Hoeffding's inequality, these formulations are particularly effective for both i.i.d. random variables and cases where variables are divided into groups with different means, facilitating the reconstruction of associated extremal distributions. Due to the computational limitations of the product-functions based approaches, we broadened our scope by introducing a polynomial family of upper bounds. Here, carefully selected polynomials are employed, resulting in non-negative polynomials that can be approximated using sum-of-square techniques, at a higher computational cost. This framework enables exploration of concentration properties concerning Bennett's and Bernstein's inequalities, although it does not universally improve theoretical bounds. We finally extended the polynomial approach into a feature-based approach, using polynomials independently of the order of moment assumptions. While not focused on computational efficiency, this method offers a more flexible and comprehensive way to study concentration inequalities. In summary, our methodologies each address different complexities inherent in the problem of moments and independence.

Future works. Throughout this work, we have highlighted several limitations related to these approaches. The separable approach could benefit from exploring and constructing new product-functions that lead to closed-form solutions, going beyond moment-generating functions. Meanwhile, the polynomial approach, when approximated by SDPs of increasing sizes, converges slowly even for a reasonable number of variables and low-degree polynomials. Studying the connection between the family of polynomials, the linear functions and product-functions would simplify the underlying optimization problems. In addition, numerical experiments have shown the limitations of the polynomial approach, which does not account for some analytical arguments, such as convexity or inequalities derived from Taylor expansion. Introducing key components of these analyses would probably help improve (or match) known bounds. Furthermore, the feature-based approach could still benefit from exploring efficient relaxations to improve known bounds.

Finally, the generalized problem of moments requires few assumptions on the distribution under consideration, such as bounded moments or random variables lying in bounded sets (e.g., intervals). Other paths to improvements of concentration bounds without independence were explored, by exploiting additional structural properties of the distributions. For instance, Popescu (2005) extended the generalized problem of moments to convex classes of distributions, such as unimodal or bimodal distributions. Given unimodal distributions, Van Parys et al. (2015) reconstructed non-discrete extremal distributions, which are thereby more representative of the effective behaviors of random variables. Extending their work to the case of independent variables would probably help refine inequalities that can be adapted to different problem structures. In particular, introducing subgaussian assumptions, which appear in many probability proofs in machine learning, would be of interest, thus improving probabilistic bounds for such problems.

A. Exact formulation of generalized problem of moments for independent random variables using optimal transport:

This proof is an alternative to the convexification of the variational approach from Proposition 3. Using optimal transport, we compute an exact formulation for the generalized problem of moments (1), and recover the convex reformulation for the variational approach as result of weak duality.

Let us introduce for $x \in \mathcal{X}^n$, $G(x) = \log(F(x_1, ..., x_n))$. Then, the exact formulation takes the form for the KL divergence $D(q\left|\right|p) = \int_{\mathcal{X}} \log\left(\frac{p(x)}{q(x)}\right) dq(x)$:

$$\begin{aligned}\log(\rho_n) &= \sup_{\forall i, p_i \in \mathcal{P}_{\mu_i}(\mathcal{X})} \log\left(\int_{\mathcal{X}^n} e^{G(x)} dp_1(x_1) \cdots dp_n(x_n)\right), \\ &= \sup_{q \in \mathcal{P}(\mathcal{X}^n)} \sup_{\forall i, p_i \in \mathcal{P}_{\mu_i}(\mathcal{X})} \int_{\mathcal{X}^n} G(x) dq(x) - D\left(q\left|\right|\prod_{i=1}^n p_i\right),\end{aligned}$$

By Donsker-Varadhan's inequality. By Pythagorean theorem for the KL divergence and mutual information, we have:

$$\log(\rho_n) = \sup_{q \in \mathcal{P}(\mathcal{X}^n)} \sup_{\forall i, p_i i \mathcal{P}_{\mu_i}(\mathcal{X})} \int_{\mathcal{X}^n} G(x) dq(x) - D\left(q\left|\right|\prod_{i=1}^n q_i\right) - \sum_{i=1}^n D(p_i\left|\right|q_i).$$

The variational relaxation corresponds in fact to using the fact that $D(q\left|\right|\prod_{i=1}^n q_i) \geqslant 0$. In addition, we recover the convex formulation from the Lagrangian relaxation of:

$$\begin{aligned}\log(\rho_n) = \sup_{q \in \mathcal{P}(\mathcal{X}^n)} \inf_{\forall i, \alpha_i \in \mathbb{R}, \beta_i \in \mathbb{R}^m} &\int_{\mathcal{X}^n} G(x) dq(x) - D\left(q\left|\right|\prod_{i=1}^n q_i\right) + \sum_{i=1}^n \{\alpha_i + \beta_i^\top \mu_i - 1\} \\ &+ \sup_{x \in \mathcal{X}^n} \{G(x) - \sum_{i=1}^n \log(\alpha_i + \beta_i^\top g(x_i))\}, \\ \leqslant \inf_{\forall i, \alpha_i \in \mathbb{R}, \beta_i \in \mathbb{R}^m} &\sum_{i=1}^n \{\alpha_i + \beta_i^\top \mu_i - 1\} + \sup_{x \in \mathcal{X}^n} \{G(x) - \sum_{i=1}^n \log(\alpha_i + \beta_i^\top g(x_i))\} \\ &+ \sup_{q \in \mathcal{P}(\mathcal{X}^n)} \int_{\mathcal{X}^n} G(x) dq(x) - D\left(q\left|\right|\prod_{i=1}^n q_i\right),\end{aligned}$$

by weak duality. Ignoring the term, $D\left(q\left|\right|\prod_{i=1}^n q_i\right)$, the maximum of $\sup_{\mathcal{X}^n} G(x) dq(x) = \max_{x \in \mathcal{X}^n} G(x)$ with respect to $q \in \mathcal{P}(\mathcal{X}^n)$ is attained at a Dirac, at which $D\left(q\left|\right|\prod_{i=1}^n q_i\right) = 0$. Thus, the fact that $\rho_n \leqslant \rho_n^{\text{var}}$ holds by weak duality.

B. Proofs for Hoeffding's inequality

B.1. Separable approach for univariate distributions

B.1.1. Proof for Proposition 5 Let us introduce a family of functions $u_\lambda(x) = e^{\lambda(x-(\mu+t))}$. Then, for every $t \geqslant 0, \lambda \in \mathbb{R}$,

$$\rho_1^{\exp}(\lambda, t) = \inf_{\alpha \in R, \beta \in \mathbb{R}} \alpha + \beta\mu \text{ such that } \forall x \in X, e^{\lambda(x-(\mu+t))} \leqslant \alpha + \beta x.$$

The function $x \mapsto e^{\lambda(x-(\mu+t))} - (\alpha + \beta x)$ is convex on $[0,1]$. Applying Proposition 1, we have:

$$\rho_1^{\exp}(\lambda, t) = \inf_{\alpha \in R, \beta \in \mathbb{R}} \alpha + \beta\mu \text{ such that } e^{-\lambda(\mu+t)} \leqslant \alpha \text{ and } e^{s(1-(\mu+t))} \leqslant \alpha + \beta.$$

The solution is given by $\alpha = e^{-\lambda(\mu+t)}$ and $\beta = e^{-\lambda(\mu+t)}(e^{\lambda} - 1)$, and thus we have $\rho_1^{\exp}(\lambda, t) = e^{-\lambda(\mu+t)}\left(1 + \mu(e^{\lambda} - 1)\right)$. By optimizing over λ, we have $\lambda_\star = \log\left(\frac{(\mu+t)(1-\mu)}{\mu(1-(\mu+t))}\right)$ and it holds for all $t \in [0, 1-(\mu)]$ with $\nu = \mu + t$:

$$\rho_{1,\star}^{\exp}(t) = \left(\frac{\mu}{\nu}\right)^{\nu} \left(\frac{1-\mu}{1-\nu}\right)^{1-\nu}.$$

B.1.2. Alternative proof for Proposition 5 Let X be a random variable taking its value almost surely in $[0,1]$, with mean $\mathbb{E}[X] = \mu$ and associated with the distribution $p \in \mathcal{P}([0,1])$. Then, for $KL(p,q) = \int_X \log(\frac{p(x)}{q(x)}) dp(x)$ and $kl(\mu, \nu) = \log(\frac{\mu}{\nu})\mu$, the Kullback-Leibler divergence, it holds that

$$\begin{aligned}
\log \mathbb{P}(X \geqslant t) &\leqslant \inf_s -st + \log(\mathbb{E}[e^{sX}]) \text{ by Markov's exponential inequality,} \\
&\leqslant \inf_s -st + \sup_q s\mathbb{E}_q[X] - KL(q,p), \text{ by Donsker} - \text{Varadhan's variational formula,} \\
&= \inf_s -st + \sup_\nu s\nu - kl(\nu, \mu), \\
&= -kl(t, \mu).
\end{aligned}$$

B.2. Variational approach with equal means: computing the extremal points for $\bar{X}_n$

Recall the optimization problem in the variational approach:

$$\begin{aligned}
\rho_n^{\exp} = \inf_{\alpha,\beta,t \geqslant 0} n(\alpha + \beta\mu - 1) \text{ such that } & -\sum_{i=1}^n \log(\alpha + \beta x_i^j) \leqslant 0,\ x^j \in \text{extremal}(\bar{\mathcal{X}}_n), \\
& \alpha + \beta x \geqslant 0,\ x \in [0,1],
\end{aligned}$$

where $\bar{\mathcal{X}}_n = \{(x_1, \ldots, x_n) \in [0,1]^n, x_1 + \cdots + x_n \geqslant n(\mu + t)\}$. Constraints $-\sum_{i=1}^n \log(\alpha + \beta x_i^j) \leqslant 0$ and $x_1 + \cdots + x_n \geqslant n(\mu + t)$ are symmetric in the coordinates, meaning that they have the same value when permuting $x_1^j, \ldots, x_n^j$. Therefore, the optimization problem requires to formulate only $O(n)$ extremal points, depending on the value for $n(\mu + t) \in [0,1]$:

- If $n \leqslant n(\mu+t) > n-1$: Extremal$(\mathcal{X}^n) = \{(1, \ldots, 1); (n(\mu+t) - (n-1), 1, \ldots, 1)\}$,
- If $n - 2 < n(\mu + t) \leqslant n - 1$: Extremal$(\mathcal{X}^n) = \{(1, \ldots, 1); (0, 1, \ldots, 1); (0, n(\mu + t) - (n - 2), 1, \ldots, 1)\}$,
- If $n - 3 < n(\mu + t) \leqslant n - 2$: Extremal$(\mathcal{X}^n) = \{(1, \ldots, 1); (0, 1, \ldots, 1); (0, 0, 1, \ldots, 1); (0, 0, n(\mu + t) - (n-3), 1, \ldots, 1)\}$,
- $\ldots$,
- If $n(\mu + t) \leqslant 1$: Extremal$(\mathcal{X}^n) = \{(1, \ldots, 1); (0, 1, \ldots, 1); \ldots; (0, \ldots, 0, 1); (0, \ldots, 0, n(\mu + t))\}$.

Thus, there are at most $\lfloor n(\mu + t) \rfloor + 1$ extremal points to consider.

B.3. Closed-form solution to the variational reformulation for $n = 2$

Recall the optimization problem under consideration for $n = 2$, for two i.i.d. random variables taking their values in $[0, 1]$ with mean $\mu \in \mathbb{R}$,

$$\log \rho_n^{\exp}(t) = \inf_{\alpha,\beta,t\geqslant 0} 2(\alpha + \beta\mu - 1) \text{ such that } \; -\log(\alpha + \beta x_1^j) - \log(\alpha + \beta x_2^j) \leqslant 0, \; x^j \in \text{extremal}(\bar{\mathcal{X}}_2), \tag{37}$$

where $\bar{\mathcal{X}}_2 = \{(x_1, x_2) \in [0, 1]^2, x_1 + x_2 \geqslant 2(\mu + t)\}$. Extremal points of $\bar{\mathcal{X}}_2$ depends on the value of $\mu + t$:

- If $1/2 \leqslant \mu + t \leqslant 1$, $\text{extremal}(\bar{\mathcal{X}}_2) = \{(1, 1), (1, 2(\mu + t)1)\}$;
- If $0 \leqslant \mu + t \leqslant 1/2$, $\text{extremal}(\bar{\mathcal{X}}_2) = \{(1, 1), (0, 1), (0, 2(\mu + t)\}$.

We derive an optimal solution to problem (37) together with optimal values $(\alpha_\star, \beta_\star)$ in Proposition

PROPOSITION 14. *Let $\mu \in [0, 1]$, and $t \in [0, 1 - \mu]$. Then, optimal solutions $(\alpha_\star, \beta_\star)$ to* (37) *verify:*

$$\alpha_\star = \begin{cases} \sqrt{\frac{\mu}{\mu+t}} & \text{if } 0 \leqslant t \leqslant \frac{1}{2} - \mu, \\ \sqrt{\frac{1-2t-\mu}{1-\mu}} - \frac{t(2(\mu+t)-1)}{\sqrt{(1-\mu)(1-2t-\mu)}(1-(\mu+t))} & \text{if } \frac{1}{2} - \mu \leqslant t \leqslant \frac{(1-\mu)^2}{2-\mu}, \\ 0 & \text{if } t \geqslant \frac{(1-\mu)^2}{2-\mu}, \end{cases}$$

$$\beta_\star = \begin{cases} \frac{t}{(\mu+t)\sqrt{\mu(\mu+2t)}} & \text{if } 0 \leqslant t \leqslant \frac{1}{2} - \mu, \\ \frac{t}{(1-(\mu+t))\sqrt{(1-\mu)(1-\mu-2t)}} & \text{if } \frac{1}{2} - \mu \leqslant t \leqslant \frac{(1-\mu)^2}{2-\mu}, \\ \frac{\mu}{\sqrt{2(\mu+t)-1}} & \text{if } t \geqslant \frac{(1-\mu)^2}{2-\mu}. \end{cases}$$

Proof. The proof is divided into two parts, depending on the value for $\mu + t$.

1. Let us first assume that $\mu + t \in]1/2, 1]$. Then, the optimization problem (37) takes the form:

$$\begin{aligned} \inf_{\alpha,\beta} \; & 2(\alpha + \beta\mu - 1), \text{ such that } \alpha \geqslant 0, && (\times\lambda_1) \\ & -2\log(\alpha + \beta) \leqslant 0, && (\times\lambda_2) \\ & -\log(\alpha + \beta) - \log(\alpha + (2(\mu + t) - 1)\beta) \leqslant 0 && (\times\lambda_3), \end{aligned}$$

where $\lambda_1, \lambda_2, \lambda_3 \geqslant 0$ are dual variables. Reformulating the second constraint into "$\alpha + \beta \geqslant 1$", it still holds that this problem is a convex. We compute the KKT conditions:

$$\begin{aligned} 2 &= \lambda_1 + \lambda_2 + \lambda_3 \left(\frac{1}{\alpha + \beta} + \frac{1}{\alpha + (2(\mu + t) - 1)\beta} \right), \\ 2\mu &= \lambda_2 + \lambda_3 \left(\frac{1}{\alpha + \beta} + \frac{(2(\mu + t) - 1)}{\alpha + (2(\mu + t) - 1)\beta} \right), \\ 0 &= \lambda_1 \alpha, \\ 0 &= \lambda_2(1 - (\alpha + \beta)), \\ 0 &= \lambda_3(-\log(\alpha + \beta) - \log(\alpha + (2(\mu + t) - 1)\beta)), \\ 0 &\leqslant \lambda_1, \lambda_2, \lambda_3. \end{aligned}$$

We proceed by a distinction of cases:

- Assume first that $\lambda_1 \neq 0$, then $\alpha = 0$. If $\lambda_2 \neq 0$, then $\beta = 1$ and $\lambda_3 = 0$, $\lambda_2 = 2 = 2\mu$, which is false. We conclude that $\lambda_2 = 0$, and that $\lambda_3 \neq 0$. This fixes the value for β, by solving $\log(\beta) + \log(\beta(2(\mu+t)-1)) = 0$ and injecting this relationship into the two first KKT conditions above.
- Assume now that $\alpha \neq 0$. Then, $\lambda_1 = 0$. At least λ_2 or λ_3 must be nonzero, and both cannot be nonzero if $t + \mu \neq 1/2$. Assuming $\log(\alpha+\beta) + \log(\alpha + (2(\mu+t)-1)\beta) \neq 0$ entails $\lambda_3 = 0$ and $\lambda_2 \neq 0$, that is $\alpha + \beta = 1$. Then, $\log(\alpha + (2(\mu+t)-1)\beta) < 0$ which is false. We conclude that $\log(\alpha+\beta) + \log(\alpha + (2(\mu+t)-1)\beta) = 0$, and thus, that $\lambda_3 \neq 0$ and $\lambda_2 = 0$. This leads to the solutions :

$$\alpha_\star = \sqrt{\frac{1-2t-\mu}{1-\mu}} - \frac{t(2(\mu+t)-1)}{\sqrt{(1-\mu)(1-2t-\mu)}(1-(\mu+t))},$$
$$\beta_\star = \frac{t}{(1-(\mu+t))\sqrt{(1-\mu)(1-\mu-2t)}},$$

for which $\alpha_\star > 0$ for $t \leqslant \frac{(1-\mu)^2}{2-\mu}$.

2. We now consider the case where $\mu + t \in [0, 1/2]$. The optimization problem under consideration takes the form:

$$\begin{aligned}
\inf_{\alpha,\beta}\ 2(\alpha+\beta\mu-1),\ \text{such that } \alpha \geqslant 0, & \qquad (\times\lambda_1)\\
\alpha+\beta \geqslant 1, & \qquad (\times\lambda_2)\\
-\log(\alpha) - \log(\alpha+\beta) \leqslant 0 & \qquad (\times\lambda_3),\\
-\log(\alpha) - \log(\alpha+2(\mu+t)\beta) \leqslant 0 & \qquad (\times\lambda_4).
\end{aligned}$$

We compute the KKT conditions, leading to

$$\begin{aligned}
2 &= \lambda_1 + \lambda_2 + \lambda_3\left(\frac{1}{\alpha} + \frac{1}{\alpha+\beta}\right) + \lambda_4\left(\frac{1}{\alpha} + \frac{1}{\alpha+\beta 2(\mu+t)}\right),\\
2\mu &= \lambda_2 + \lambda_3\frac{1}{\alpha+\beta} + \lambda_4\frac{2(\mu+t)}{\alpha+\beta 2(\mu+t)},\\
0 &= \lambda_1\alpha,\\
0 &= \lambda_2(1-(\alpha+\beta)),\\
0 &= \lambda_3(-\log(\alpha) - \log(\alpha+\beta)),\\
0 &= \lambda_4(-\log(\alpha) - \log(\alpha+\beta 2(\mu+t))),\\
0 &\leqslant \lambda_1, \lambda_2, \lambda_3, \lambda_4.
\end{aligned}$$

First, note that $\alpha > 0$, and thus, $\lambda_1 = 0$. In addition, $\lambda_2 = 0$, otherwise, it implies that $\lambda_3 = \lambda_4 = 0$ and $\mu = 1$, which is impossible. In addition, if $\lambda_3 \neq 0$, then $\alpha(\alpha+\beta) = 1$. Yet, since $\alpha > 0$, $\beta \geqslant 0$. The function $t \mapsto \alpha(\alpha + t\beta)$ is nondecreasing, and thus $\alpha(\alpha + 2(\mu+t)\beta) > 1$, which is impossible. We conclude that $\lambda_3 = 0$. By construction, $\lambda_4 \neq 0$, which implies that $\alpha(\alpha + 2(\mu+t)\beta) = 1$, and thus,

$$\begin{aligned}
2 &= \lambda_4(\frac{1}{\alpha} + \alpha),\\
2\mu &= \lambda_4\alpha 2(\mu+t),
\end{aligned}$$

from which we conclude the solutions $\alpha = \sqrt{\frac{\mu}{\mu+t}}$ and $\beta = \frac{t}{(\mu+t)\sqrt{\mu(\mu+2t)}}$. □

B.4. Proof for Lemma 1.

Problem (20) simplifies into

$$\inf_{\alpha\in\mathbb{R}^2,\beta\in\mathbb{R}^2} m(\alpha_1+\beta_1\mu_1)+(n-m)(\alpha_2+\beta_2\mu_n),$$
$$\text{such that } \forall x\in \text{extremal}(\bar{\mathcal{X}}_n),\ -\sum_{i=1}^{m}\log(\alpha_1+\beta_1 x_i)-\sum_{i=m+1}^{n}\log(\alpha_2+\beta_2 x_i)\leqslant 0,$$

where $\bar{\mathcal{X}}_n = \{(x_1,\ldots,x_n)\in[0,1]^n, x_1+\ldots+x_n \geqslant nt+m\mu_1+(n-m)\mu_n\}$. We define $\bar{\mu}_n = \frac{1}{n}\sum_{i=1}^n \mu_i$ and denote $q = n(\bar{\mu}_n+t) - \lfloor n(\bar{\mu}_n+t)\rfloor$. We denote the points with the notation $(x_1,\ldots x_m|x_{m+1},\ldots,x_n)$. Let $n(\bar{\mu}_n+t)\in[n-k+1,n-k[$. The following assertions are true:

- $(1,\ldots,1)\in\text{extremal}(\bar{\mathcal{X}}_n)$
- If $(1,\ldots,1|0,1,\ldots,1)\in\text{extremal}(\bar{\mathcal{X}}_n)$, all its permutations are in $\text{extremal}(\bar{\mathcal{X}}_n)$. By symmetry, it is sufficient to notice that the point $(1,\ldots,1,0|1,\ldots,1)$ is in $\text{extremal}(\bar{\mathcal{X}}_n)$.
- If $(1,\ldots,1|0,0,1,\ldots,1)\in\text{extremal}(\bar{\mathcal{X}}_n)$, all its permutations are in $\text{extremal}(\bar{\mathcal{X}}_n)$. By symmetry, it is sufficient to notice that the point $(1,\ldots,1,0|0,1,\ldots,1)$ and $(1,\ldots,1,0,0|1,\ldots,1)$ are in $\text{extremal}(\bar{\mathcal{X}}_n)$.
- If $(1,\ldots,1|0,\ldots 0,1,\ldots,1)\in\text{extremal}(\bar{\mathcal{X}}_n)$, all its permutations are in $\text{extremal}(\bar{\mathcal{X}}_n)$. By symmetry, it is sufficient to notice that the point $(1,\ldots,1,0|0,\ldots,0,1,\ldots,1),\ldots,(1,\ldots,1,0,\ldots,0|1,\ldots,1)$ are in $\text{extremal}(\bar{\mathcal{X}}_n)$ (as long as $k\geqslant m$). That is about $O(|k-m|)$ points.
- If $(1,\ldots,1|0,\ldots,0,q,1,\ldots,1)\in\text{extremal}(\bar{\mathcal{X}}_n)$, all its permutations are in $\text{extremal}(\bar{\mathcal{X}}_n)$. By symmetry, it is sufficient to notice that points $(1,\ldots,1,q|0,\ldots,0,1,\ldots,1)$, $(1,\ldots,1,q,0|0,\ldots,0,1,\ldots,1)$, …, $(1,\ldots,1,q,0,\ldots,0|1,\ldots,1)$ are in $\text{extremal}(\bar{\mathcal{X}}_n)$. That is, about $O(2|k-m|)$ points.

If they are k zeros elements, there are $O(\sum_{i=1}^{k-m} i) = O(k^2)\leqslant O(n^2)$ points.

C. Connecting the second-order to the first-order moment: proof for Lemma 2

Let X be a random variable almost surely in $[0,1]$ with mean $\mu^{(1)}$. First, its second-order moment exists. In addition, we have,

$$\begin{aligned}
\mu^{(2)} &\leqslant \sup_{p\in\mathcal{P}([0,1])}\int_0^1 x^2 dp(x),\ \text{such that}\ \int_0^1 x dp(x)=\mu^{(1)},\\
&= \inf_{\alpha,\beta}\alpha+\beta\mu^{(1)}\ \text{such that}\ \forall x\in[0,1], x^2\leqslant\alpha+\beta x,\\
&= \inf_{\alpha,\beta}\alpha+\beta\mu^{(1)}\ \text{such that}\ \alpha\geqslant 0, 1\leqslant\alpha+\beta\ \ (\text{by convexity}),\\
&= \mu^{(1)},
\end{aligned}$$

with $\alpha=0,\beta=1$. In addition, by Cauchy-Schwarz, $\int_0^1 x dp(x)\leqslant\sqrt{\left(\int_0^1 x^2 dp(x)\right)}$ that is $\mu^{(1)}\leqslant\sqrt{\mu^{(2)}}$.

Acknowledgments

This work was funded by MTE and the Agence Nationale de la Recherche as part of the "Investissements d'avenir" program, reference ANR-19-P3IA-0001 (PRAIRIE 3IA Institute). We also acknowledge support from the European Research Council (grant SEQUOIA 724063).

References

ApS, MOSEK. 2022. *The MOSEK optimization toolbox for MATLAB manual. Version 10.0.*. URL `http://docs.mosek.com/9.0/toolbox/index.html`.

Bach, Francis. 2024. *Learning Theory from First Principles*. MIT Press (to appear).

Bennett, George. 1968. A one-sided probability inequality for the sum of independent, bounded random variables. *Biometrika* **55**(3) 565–569.

Bertsimas, Dimitris, Ioana Popescu. 2005. Optimal inequalities in probability theory: A convex optimization approach. *SIAM Journal on Optimization* **15**(3) 780–804.

Bertsimas, Dimitris, Ioana Popescu, Jay Sethuraman. 2000. Moment problems and semidefinite optimization. *Handbook of Semidefinite Programming* 311–339.

Boucheron, Stéphane, Gabor Lugosi, Pascal Massart. 2013. *Concentration Inequalities : A non Asymptotic Theory of Independence*. Oxford University Press.

Boyd, Stephen, Lieven Vandenberghe. 2004. *Convex Optimization*. Cambridge University Press.

Comanor, Katherine, Lieven Vandenberghe, Stephen Boyd. 2006. Semidefinite programming and multivariate Chebyshev bounds. *IFAC Proceedings Volumes* **39**(9) 597–601.

Cramér, Harald. 1938. Sur un nouveau théorème limite de la théorie des probabilités. *Actualites Scientifiques et Industrielles* **736** 2–23.

de Klerk, Etienne, Monique Laurent. 2019. A Survey of Semidefinite Programming Approaches to the Generalized Problem of Moments and Their Error Analysis. *World Women in Mathematics 2018*. 17–56.

Dembo, Amir, Ofer Zeitouni. 1998. *Large Deviations Techniques and Applications*. Springer.

Devroye, Luc, László Györfi, Gábor Lugosi. 1996. *A Probabilistic Theory of Pattern Recognition*, *Stochastic Modelling and Applied Probability*, vol. 31. Springer.

Hoeffding, Wassily. 1963. Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association* **58**(301) 13–30.

Isii, Keiiti. 1962. On sharpness of Tchebycheff-type inequalities. *Annals of the Institute of Statistical Mathematics* **14** 185–197.

Isii, Keiiti. 1964. Inequalities of the types of Chebyshev and Cramér-rao and mathematical programming. *Annals of the Institute of Statistical Mathematics* **16** 277–293.

Jaakkola, Tommi S., Michael I. Jordan. 1999. Variational probabilistic inference and the QMR-DT network. *Journal of Artificial Intelligence Research* **10**(1) 291–322.

Jaakkola, Tommi S., Michael I. Jordan. 2000. Bayesian parameter estimation via variational methods. *Statistics and Computing* **10** 25–37.

Karlin, Samuel, William J. Studden. 1966. Tchebycheff systems: With applications in analysis and statistics. *Pure and Applied Mathematics, A Series of Texts and Monographs* .

Landau, Henry J. 1998. *Moments in Mathematics*. American Mathematical Society.

Lasserre, Jean Bernard. 2001. Global optimization with polynomials and the problem of moments. *SIAM Journal on Optimization* **11**(3) 796–817.

Lasserre, Jean-Bernard. 2002. Bounds on measures satisfying moment conditions. *The Annals of Applied Probability* **12**(3) 1114 – 1137.

Lasserre, Jean-Bernard. 2008. A semidefinite programming approach to the generalized problem of moments. *Mathematical Programming* **112** 65–92.

Marshall, Albert W., Ingram Olkin. 1960. Multivariate Chebyshev inequalities. *The Annals of Mathematical Statistics* **31**(4) 1001–1014.

O'Donoghue, Brendan, Eric Chu, Neal Parikh, Stephen Boyd. 2016. Conic optimization via operator splitting and homogeneous self-dual embedding. *Journal of Optimization Theory and Applications* **169**(3) 1042–1068.

Parrilo, Pablo. 2003. Semidefinite programming relaxations for semialgebraic problems. *Mathematical Programming, Series B* **96** 293–320.

Popescu, Ioana. 2005. A semidefinite programming approach to optimal-moment bounds for convex classes of distributions. *Mathematics of Operation Researche* **30**(3) 632–657.

Rogosinski, Werner W. 1958. Moments of non-negative mass. *Proceedings of the Royal Society of London. Series A, Mathematical and Physical Sciences* **245**(1240) 1–27.

Tao, Terence. 2011. *Topics in random matrix theory*. American Mathematical Society.

Van Parys, Bart, Paul Goulart, Daniel Kuhn. 2015. Generalized gauss inequalities via semidefinite programming. *Mathematical Programming* **156** 1–32.

Vandenberghe, Lieven, Stephen Boyd. 1996. Semidefinite programming. *SIAM Review* **38**(1) 49–95.

Vandenberghe, Lieven, Stephen Boyd, Katherine Comanor. 2007. Generalized Chebyshev bounds via semidefinite programming. *SIAM Review* **49**(1) 52–64.

Varadhan, Srinivasa. R. S. 1988. *Large deviations and applications*. École d'Été de Probabilités de Saint-Flour XV–XVII, 1985–87.

Vershynin, Roman. 2018. *High-Dimensional Probability: An Introduction with Applications in Data Science*. Cambridge Series in Statistical and Probabilistic Mathematics, Cambridge University Press.

Wolkowicz, Henry, Romesh Saigal, Lieven Vandenberghe. 2000. *Handbook of Semidefinite Programming*. Kluwer Academic Publishers.