Optimization Methods and Software

ISSN: 1055-6788 (Print) 1029-4937 (Online) Journal homepage: https://www.tandfonline.com/loi/goms20

Non-asymptotic confidence bounds for the optimal value of a stochastic program

Vincent Guigues, Anatoli Juditsky & Arkadi Nemirovski

To cite this article: Vincent Guigues, Anatoli Juditsky & Arkadi Nemirovski (2017) Non-asymptotic confidence bounds for the optimal value of a stochastic program, Optimization Methods and Software, 32:5, 1033-1058, DOI: 10.1080/10556788.2017.1350177

To link to this article: https://doi.org/10.1080/10556788.2017.1350177

Optimization Methods & Software, 2017
Vol. 32, No. 5, 1033–1058, https://doi.org/10.1080/10556788.2017.1350177

Non-asymptotic confidence bounds for the optimal value of a stochastic program

Vincent Guigues[a]*, Anatoli Juditsky[b] and Arkadi Nemirovski[c]

[a]*FGV/EMAp, Rio de Janeiro, Brazil;* [b]*LJK, Université Grenoble Alpes, Grenoble Cedex 9, France;* [c]*Georgia Institute of Technology, Atlanta, GA, USA*

(*Received 31 January 2016; accepted 27 June 2017*)

We discuss a general approach to building non-asymptotic confidence bounds for Stochastic Optimization problems. Our principal contribution is the observation that a Sample Average Approximation of a problem supplies upper and lower bounds for the optimal value of the problem which are essentially better than the quality of the corresponding optimal solutions. At the same time, such bounds are more reliable than 'standard' confidence bounds obtained through the asymptotic approach. We also discuss bounding the optimal value of MinMax Stochastic Optimization and stochastically constrained problems. We conclude with a simulation study illustrating the numerical behaviour of the proposed bounds.

Keywords: Sample Average Approximation; confidence interval; MinMax Stochastic Optimization; stochastically constrained problems

AMS Subject Classification: 90C15; 90C90; 90C30

1. Introduction

Consider the following Stochastic Programming (SP) problem:

$$\mathrm{Opt} = \min_x [f(x) = \mathbf{E}\{F(x,\xi)\},\ x \in X], \tag{1}$$

where X is a nonempty bounded closed convex set of a Euclidean space E, ξ is a random vector with probability distribution P on $\Xi \subset \mathbf{R}^k$ and $F:\ X \times \Xi \to \mathbf{R}$. There are two competing approaches for solving (1) when a sample $\xi^N = (\xi_1, \ldots, \xi_N)$ of realizations of ξ (or a device to sample from the distribution P) is available – Sample Average Approximation (SAA) and the Stochastic Approximation (SA). The basic idea of the SAA method is to build an approximation of the 'true' problem (1) by replacing the expectation $f(x)$ with its SAA

$$f_N(x, \xi^N) = \frac{1}{N}\sum_{t=1}^{N} F(x, \xi_t), \quad x \in X.$$

*Corresponding author. Email: vguigues@fgv.br

The resulting optimization problem has been extensively studied theoretically and numerically (see, e.g. [9,12,13,26,27,30], among many others). In particular, it was shown that the SAA method (coupled with a deterministic algorithm for minimizing the SAA) is often efficient for solving large classes of stochastic programs. The alternative SA approach was also extensively studied since the pioneering work by Robbins and Monro [21]. Though possessing better theoretical accuracy estimates, SA was long time believed to underperform numerically. It was recently demonstrated (cf. [2,14,28]) that a proper modification of the SA approach, based on the ideas behind the Mirror Descent algorithm [15] can be competitive and can even significantly outperform the SAA method on a large class of convex stochastic programs.

Note that in order to qualify the accuracy of approximate solutions (e.g. to build efficient stopping criteria) delivered by the stochastic algorithm of choice, one needs to construct lower and upper bounds for the optimal value Opt of problem (1) from stochastically sampled observations. Furthermore, the question of computing reliable upper and, especially, lower bounds for the optimal value is of interest in many applications. Such bounds allow statistical decisions (e.g. computing confidence intervals, testing statistical hypotheses) about the optimal value. For instance, using the approach to regret minimization, developed in [3,17], they may be used to construct *risk-averse strategies* for multi-armed bandits, and so on.

An important methodological feature of the SAA approach is its asymptotic framework which explains how to provide asymptotic estimates of the accuracy of the obtained solution by computing asymptotic upper and lower bounds for the optimal value of the 'true' problem (see, e.g. [4,8,13,18–20,24], and references therein).

However, as is always the case with techniques which are validated asymptotically, some important questions, such as 'true' reliability of bounds, cannot be answered by the asymptotic analysis. Note that the non-asymptotic accuracy of optimal solutions of the SAA problem was recently analysed (see, e.g. [7,19,20,23,26,27]), yet, to the best of our knowledge, the literature does not provide any *non-asymptotic* construction of lower and upper bounds for the optimal value of (1) by SAA. On the other hand, non-asymptotic lower and upper bounds for the objective value by SA method were built in [5,11].

Our objective in this work is to fill this gap, by building reliable finite-time evaluations of the optimal value of (1), which are also good enough to be of practical interest. Our basic methodological observation is Proposition 1 which states that the SAA of problem (1) comes with a 'built-in' non-asymptotic lower and upper estimation of the 'true' objective value. The accuracy of these estimations is essentially higher than the available theoretical estimation of the *quality of the optimal solution* of the SAA. Indeed, when solving a high-dimensional SAA problem, the (theoretical bound of) inaccuracy of the optimal solution becomes a function of dimension. In particular, when the set X is a unit Euclidean ball of $\mathbf{R}^n$, the accuracy of the SAA *optimal solution* may be by factor $O(n)$ worse than the corresponding accuracy of the SA solution [14]. In contrast to this, the *optimal value* of the SAA problem supplies an approximation of the 'true' optimal value of accuracy which is (almost) independent of problem's dimension and may be used to construct non-trivial non-asymptotic confidence bounds for the true optimal value. This fact is surprising, because the bad theoretical accuracy bound for optimal solutions of SAA reflects their actual behaviour on some problem instances (see Proposition 2 and the discussion in Section 2.1.3).

The paper is organized as follows.

We present the construction of lower and upper confidence bounds for the optimal value of a stochastic problem in Section 2. Specifically, in Section 2.1, we develop confidence bounds for the optimal value of problem (1). Then in Section 2.2, we build lower and upper bounds for the optimal value of MinMax Stochastic Optimization and show how the confidence bounds can be constructed for an ϵ-*underestimation* of the optimal value of a (stochastically) constrained Stochastic Optimization problem.

Finally, several simulation experiments illustrating the properties of the bounds built in Section 2 are presented in Section 3. Proofs of theoretical statements are collected in the appendix.

2. Confidence bounds via Sample Average Approximation

2.1 *Problem without stochastic constraints*

2.1.1 *Situation*

In the sequel, we fix a Euclidean space E. We denote by $B_{\|\cdot\|}$ the closed unit ball of the norm $\|\cdot\|$, and by $\|\cdot\|_*$ the norm conjugate to $\|\cdot\|$:

$$\|y\|_* = \max_{\|x\|\le 1} \langle x, y\rangle.$$

Let us now assume that we are given a function $\omega(\cdot)$ which is continuously differentiable on $B_{\|\cdot\|}$ and strongly convex with respect to $\|\cdot\|$, with parameter of strong convexity equal to one, i.e. such that and for every $x, y \in B_{\|\cdot\|}$

$$(\nabla\omega(x) - \nabla\omega(y))^{\mathrm{T}}(x - y) \ge \|x - y\|^2,$$

with $\omega(0) = 0$ and $\nabla\omega(0) = 0$ (in other words, $\omega(\cdot)$ is a *distance-generating function compatible with* $\|\cdot\|$). We denote

$$\Omega = \max_{x:\|x\|\le 1} \sqrt{2\omega(x)}. \tag{2}$$

Let, further,

- X be a convex compact subset of E,
- $R = R_{\|\cdot\|}[X]$ be the smallest radius of a $\|\cdot\|$-ball containing X,
- P be a Borel probability distribution on $\mathbf{R}^k$, Ξ be the support of P, and

  $$F(x, y) : X \times \Xi \to \mathbf{R}$$

  be a Borel function which is convex in $x \in X$ for P-almost all $y \in \Xi$, and is P-summable for every $x \in X$, so that the function

  $$f(x) = \mathbf{E}\{F(x, \xi)\} : X \to \mathbf{R}$$

  is well defined and convex.

  We denote

  $$L(x, \xi) = \max\{\|g - h\|_* : g \in \partial_x F(x, \xi), h \in \partial f(x)\}.$$

The outlined data give rise to the stochastic program

$$\mathrm{Opt} = \min_{x\in X}[f(x) = \mathbf{E}\{F(x, \xi)\}]$$

and its Sample Average Approximation

$$\mathrm{Opt}_N(\xi^N) = \min_{x\in X}\left[f_N(x, \xi^N) := \frac{1}{N}\sum_{t=1}^{N} F(x, \xi_t)\right], \tag{3}$$

where $\xi^N = (\xi_1, \ldots, \xi_N)$, and $\xi_1, \xi_2, \ldots$ are drawn independently from P. Our immediate goal is to understand how well the optimal value $\mathrm{Opt}_N(\xi^N)$ of SAA approximates the true optimal value Opt.

2.1.2 *Confidence bounds*

Our main result is as follows.

PROPOSITION 1 *In the situation of Section 2.1.1, let us assume that f is differentiable on X and that for some positive M_1, M_2 and all $x \in X$ one has*

$$\text{(a)} \quad \mathbf{E}[\mathrm{e}^{(F(x,\xi)-f(x))^2/M_1^2}] \le \mathrm{e}, \quad \text{(b)} \quad \mathbf{E}[\mathrm{e}^{L^2(x,\xi)/M_2^2}] \le \mathrm{e}. \tag{4}$$

Define

$$a(\mu, N) = \frac{\mu M_1}{\sqrt{N}} \quad \text{and} \quad b(\mu, s, \lambda, N) = \frac{\mu M_1 + [\Omega[1+s^2] + 2\lambda]M_2 R}{\sqrt{N}},$$

where Ω is as in (2), and let $\tau_* = 0.557409\ldots$ be the smallest positive real such that $\mathrm{e}^t \le t + \mathrm{e}^{\tau_* t^2}$ for all $t \in \mathbf{R}$. Then for all $N \in \mathbf{Z}_+$ and $\mu \in [0, 2\sqrt{\tau_* N}]$

$$\mathrm{Prob}\{\mathrm{Opt}_N(\xi^N) > \mathrm{Opt} + a(\mu, N)\} \le \mathrm{e}^{-\mu^2/4\tau_*}; \tag{5}$$

and for all $N \in \mathbf{Z}_+$, $\mu \in [0, 2\sqrt{\tau_* N}]$, $s > 1$ and $\lambda \ge 0$,

$$\mathrm{Prob}\{\mathrm{Opt}_N(\xi^N) < \mathrm{Opt} - b(\mu, s, \lambda, N)\} \le \mathrm{e}^{-N(s^2-1)} + \mathrm{e}^{-\mu^2/4\tau_*} + \mathrm{e}^{-\lambda^2/4\tau_*}. \tag{6}$$

We have the following obvious corollary to this result.

COROLLARY 1 *Under the assumptions of Proposition 1, let*

$$\mathtt{Low}^{\mathrm{SAA}}(\mu_1, N) = \mathrm{Opt}_N(\xi^N) - a(\mu_1, N),$$
$$\mathtt{Up}^{\mathrm{SAA}}(\mu_2, s, \lambda, N) = \mathrm{Opt}_N(\xi^N) + b(\mu_2, s, \lambda, N).$$

Then for all $N \in \mathbf{Z}_+$, $s > 1$, $\lambda \ge 0$, $\mu_1, \mu_2 \in [0, 2\sqrt{\tau_* N}]$

$$\mathrm{Prob}\{\mathrm{Opt} \in [\mathtt{Low}^{\mathrm{SAA}}(\mu_1, N),\ \mathtt{Up}^{\mathrm{SAA}}(\mu_2, s, \lambda, N)]\} \ge 1 - \beta,$$

where $\beta[= \beta(\mu_1, \mu_2, s, \lambda, N)] = \mathrm{e}^{-\mu_1^2/4\tau_*} + \mathrm{e}^{-\mu_2^2/4\tau_*} + \mathrm{e}^{-N(s^2-1)} + \mathrm{e}^{-\lambda^2/4\tau_*}$. In other words, for the choice of μ_1, μ_2, s, λ and N such that $0 < \beta < 1$, the segment $[\mathtt{Low}^{\mathrm{SAA}}, \mathtt{Up}^{\mathrm{SAA}}]$ is the confidence interval for Opt *of level $1 - \beta$.*

2.1.3 *Discussion*

The result of Proposition 1 merits some comments.

(1) Confidence bounds of Proposition 1 and Corollary 1 involve constants M_1 and M_2, defined in (4). Valid upper bounds on these constants are crucial to obtain sound confidence bounds. To the best of our knowledge, there is no generic procedure which allows us to construct such estimates. Nevertheless, it is possible to build 'reasonably good' bounds for M_1 and M_2 in specific problem settings. For instance, we provide such bounds for the examples used to illustrate the results of this section in the numerical experiments of Section 3 (see Appendix 2 for details of the calculations).

(2) 'As is', Proposition 1 requires $f(\cdot)$ to be differentiable. This purely technical assumption is in fact not restrictive at all. Indeed, we can associate with (1) its 'smoothed' approximation

$$\min_{x\in X}\left[f_\epsilon(x) := \int_{\Xi\times E} F_\epsilon(x,[\upsilon;\xi])P(d\xi)p(\upsilon)\,\mathrm{d}\upsilon\right], \quad F_\epsilon(x,[\upsilon;\xi]) = F(x+\epsilon\upsilon,\xi),$$

where $p(\cdot)$ is, say, the density of the uniform distribution U on the unit ball $B_{\|\cdot\|}$ in E. Assuming that bounds (4(a)) and (4(b)) hold for all x from an open set X^+ containing X, it is immediately seen that f_ϵ is, for values of ϵ small enough, a continuously differentiable function on X which converges, uniformly on X, to f as $\epsilon \to +0$. Given a possibility to sample from the distribution P, we can sample from the distribution $P^+ := P \times U$ on $\Xi^+ = \Xi \times E$, and thus can build the SAA of the problem $\min_{x\in X} f_\epsilon(x)$. When ϵ is small, this smoothed problem satisfies the assumptions of Proposition 1, the parameters M_1, M_2 remaining unchanged, and its optimal value can be made as close to Opt as we wish by an appropriate choice of ϵ. As a result, by passing from the SAA of the original problem to the SAA of the smoothed one, ϵ being small, we ensure, 'at no cost,' smoothness of the objective, and thus – applicability of the large deviation bounds stated in Proposition 1.

(3) The standard theoretical results on the SAA of a Stochastic Optimization problem (1), see, e.g. [14,25] and references therein, are aimed at quantifying the sample size $N = N(\epsilon, n)$ which, with overwhelming probability, ensures that an *optimal solution* $x(\xi^N)$ to the SAA of the problem of interest satisfies the relation $f(x(\xi^N)) \le \mathrm{Opt} + \epsilon$, for a given $\epsilon > 0$. The corresponding bounds on N are similar, but not identical, to the bounds in Proposition 1. Let us consider, for instance, the simplest case of 'Euclidean geometry' where $\|x\| = \|x\|_2 = \sqrt{\langle x, x\rangle}$, $\omega(x) = \frac{1}{2}\|x\|^2$, and X is the unit $\|\cdot\|_2$-ball. In this case, Proposition 1 states that for a given $\epsilon > 0$, the sample size N for which $\mathrm{Opt}(\xi^N)$ is, with probability at least $1-\alpha$, ϵ-close to Opt, can be upper bounded for small enough ϵ and α by

$$N_\epsilon := C\frac{[M_1 + M_2]^2 \ln(1/\alpha)}{\epsilon^2}$$

(here C is a positive absolute constant).[1] It should be stressed that both the bound itself and the range of 'small enough' values of ϵ, α for which this bound is valid are independent of the dimension n of the decision vector x. In contrast to this, available estimation of the complexity $N(\epsilon, n)$ relies upon uniform convergence arguments and is affected by problem's dimension: up to logarithmic terms, $N(\epsilon, n) = nN_\epsilon$ (cf. the discussion in [23,27]). This phenomenon – linear dependence on the problem's dimension n of the SAA sample size yielding, with high probability, an ϵ-optimal solution to a stochastic problem – is not an artefact stemming from an imperfect theoretical analysis of the SAA but reflects the actual performance of SAA on some instances. Indeed, we have the following.

PROPOSITION 2 *For any $n \ge 3$, and $R, L > 0$ one can point out a convex Lipschitz continuous function f with Lipschitz constant L on the Euclidean ball $B_{\|\cdot\|_2}(R)$ of radius R for $\|\cdot\|_2$, and an integrand $F(x,\xi)$ convex in x such that $\mathbf{E}_\xi\{F(x,\xi)\} = f(x)$, $\|F'(x,\xi) - f'(x)\|_2^2 \le L$ a.s., for all $x \in B_{\|\cdot\|_2}(R)$, and such that with probability at least $1 - \mathrm{e}^{-1}$ there is an optimal solution $x(\xi^N)$ to the SAA*

$$\min\left[f_N(x,\xi^N) = \frac{1}{N}\sum_{i=1}^N F(x,\xi_i) :\ x \in B_{\|\cdot\|_2}(R)\right],$$

sampled over $N \le n$ i.i.d. realizations of ξ, satisfying

$$f(x(\xi^N)) - \mathrm{Opt} \ge c_0 LR, \tag{7}$$

where c_0 is a positive absolute constant.

Note that for large-scale problems, the presence of the factor n in the sample size bound is a definite and serious drawback of SAA. A nice fact about the SAA approach, as expressed by Proposition 1, is that as far as reliable ϵ-approximation of the optimal value *(rather than building an ϵ-solution)* is concerned, the performance of the SAA approach, at least in the case of favourable geometry, is not affected by the problem's dimension. It should be stressed that the crucial role in Proposition 1 is played by convexity which allows us to express the quality to which the SAA reproduces the optimal value in (1) in terms of how well $f_N(x, \xi^N)$ reproduces the first-order information on f at a single point $x_* \in \mathrm{Argmin}_X f$, see the proof of Proposition 1. In a 'favourable geometry' situation, e.g. in the Euclidean geometry case, the corresponding sample size is not affected by problem's dimension. In contrast to this, to yield reliably an ϵ-solution, the SAA requires, in general, $f_N(x, \xi^N)$ to be ϵ-close to f uniformly on X with overwhelming probability; and the corresponding sample size, even in the case of Euclidean geometry, grows with problem's dimension.

(4) Note that (at least in the case of Euclidean geometry) without additional, as compared to those in Proposition 1, restrictions on F and/or the distribution P, the quality of the SAA estimate $\mathrm{Opt}_N(\xi^N)$ of Opt (and thus, the quality of the confidence interval for it provided by Corollary 1) is, within an absolute constant factor, the best allowed by the laws of Statistics. Namely, we have the following lower bound for the widths of the confidence intervals for the optimal value valid already for a class of *linear* stochastic problems.

PROPOSITION 3 *For any $n \geq 1, M_1 \geq M_2 > 0$, one can point out a family of linear Stochastic Optimization problems, i.e. linear functions f on the unit Euclidean ball $B_{\|\cdot\|_2}$ of $\mathbf{R}^n$ and corresponding integrands $F(x, \xi)$ linear in x such that $\mathbf{E}_\xi\{F(x, \xi)\} = f(x)$, satisfying the assumptions of Proposition* 1 *and Corollary* 1, *and such that the width of the confidence interval for* $\mathrm{Opt} = \min_{x \in B_{\|\cdot\|_2}} f(x)$ *of confidence level $\geq 1 - \alpha$ cannot be less than*

$$\underline{W} = 2\gamma q_{\mathcal{N}}(1 - \alpha)\frac{M_1}{\sqrt{N}}, \tag{8}$$

where $q_{\mathcal{N}}(\beta)$ is the β-quantile of the standard normal distribution, and $\gamma > 0$ is given by the relation

$$\mathbf{E}_{\zeta \sim \mathcal{N}(0,1)}\{\exp\{\gamma^2\zeta^2\}\} = \exp\{1\},$$

or, equivalently, $\gamma^2 = \frac{1}{2}(1 - \exp\{-2\})$.

In Table 1, we provide the ratios R_{W} of the widths of the confidence intervals, as given by Corollary 1 and their lower bounds for some combinations of risks α and parameters M_1, M_2 and N.

2.2 *Constrained case*

Now consider a convex stochastic problem of the form

$$\mathrm{Opt} = \min_{x \in X}\left[f_0(x) := \int_\Xi F_0(x, \xi)P(\mathrm{d}\xi) : f_i(x) := \int_\Xi F_i(x, \xi)P(\mathrm{d}\xi) \leq 0,\ 1 \leq i \leq m\right], \tag{9}$$

where, similarly to the above, X is a convex compact set in a Euclidean space E, P is a Borel probability distribution on $\mathbf{R}^k$, Ξ is the support of P, and

$$F_i(x, \xi) : X \times \Xi \to \mathbf{R}, \quad 0 \leq i \leq m,$$

Table 1. Ratios R_{W} of the widths of the confidence intervals as given by Corollary 1 and their lower bounds from Proposition 3.

$\alpha = 0.1$	$M_1 = M_2 = 1$			$M_1 = 10,\ M_2 = 1$			$M_1 = 100,\ M_2 = 1$		
N	10	100	1000	10	100	1000	10	100	1000
R_{W}	8.086	7.803	7.775	3.772	3.744	3.741	3.341	3.338	3.337
$\alpha = 0.01$									
R_{W}	5.586	5.362	5.340	2.666	2.644	2.642	2.374	2.372	2.372
$\alpha = 0.001$									
R_{W}	4.908	4.689	4.667	2.368	2.346	2.344	2.114	2.112	2.112

are Borel functions convex in x and P-summable in ξ for every x, implying that the functions f_i, $0 \le i \le m$, are convex. As in the previous section, we assume that E is equipped with a norm $\|\cdot\|$, the conjugate norm being $\|\cdot\|_*$, and a compatible with $\|\cdot\|$ distance-generating function for the unit ball $B_{\|\cdot\|}$ of the norm.

We put

$$L(x,\xi) = \max_{0\le i\le m} \{\|g-h\|_* : g \in \partial_x F_i(x,\xi), h \in \partial f_i(x)\}.$$

Assuming that we can sample from the distribution P, and given a sample size N, we can build Sample Average Approximations (SAA's) of functions f_i, $0 \le i \le m$:

$$f_{i,N}(x,\xi^N) = \frac{1}{N}\sum_{t=1}^{N} F_i(x,\xi_t).$$

Here, as above, $\xi_1, \xi_2, \ldots$ are drawn, independently of each other from P and $\xi^N = (\xi_1, \ldots, \xi_N)$. Same as above, we want to use these SAA's of the objective and the constraints of (9) to infer conclusions on the optimal value of the problem of interest (9).

Our first observation is that in the constrained case, one can hardly expect a reliable and tight approximation to Opt to be obtainable from noisy information. The reason is that in the general constrained case, even the special one where F_i (and thus f_i) are affine in x, the optimal value is highly unstable: arbitrarily small perturbations of the data (e.g. the coefficients of affine functions F_i in the special case or parameters of distribution P) can result in large changes in the optimal value. As a result, with noisy observations of the data, one could hardly expect to get a good estimate of Opt via a sample of instance-independent size. The standard remedy is to impose an priori upper bound on the magnitude of optimal Lagrange multipliers for the problem of interest, e.g. by imposing the assumption that this problem is strictly feasible, with the level of strict feasibility

$$\varkappa := -\min_{x\in X} \max[f_1(x), \ldots, f_m(x)] \tag{10}$$

lower bounded by a known in advance positive quantity. Since in many cases an priori lower bound on $\varkappa$ is unavailable, we intend in the sequel to utilize an alternative approach, specifically, as follows. Let us associate with (9) the univariate (max-)function

$$\Phi(r) = \min_{x\in X} \max[f_0(x) - r, f_1(x), \ldots, f_m(x)].$$

Clearly, Φ is a continuous convex nonincreasing function of $r \in \mathbf{R}$ such that $\Phi(r) \to \infty$ as $r \to -\infty$. This function has a zero if and only if (9) is feasible, and Opt is nothing but the smallest zero of Φ.

DEFINITION 1 *Given $\epsilon > 0$, a real ρ ϵ-underestimates* Opt *if $\rho \leq$ Opt and $\Phi(\rho) \leq \epsilon$.*

Note that $\Phi(\rho) \leq \epsilon$ implies that

$$\rho \geq \mathrm{Opt}(\epsilon) := \min_{x \in X}[f_0(x) - \epsilon : f_i(x) \leq \epsilon,\ 1 \leq i \leq m].$$

Thus, ρ ϵ-underestimates Opt if and only if ρ is in-between the optimal value of the problem of interest (9) and the problem obtained from (9) by 'optimistic' ϵ-perturbation of the objective and the constraints.

Remark 1 Let ρ ϵ-underestimate Opt. When (9) is feasible and the magnitude (absolute value) ϑ of the left derivative of $\Phi(\cdot)$ taken at Opt is positive, from convexity of Φ it follows that

$$\mathrm{Opt} - \frac{\epsilon}{\vartheta} \leq \rho < \mathrm{Opt}.$$

Thus, unless ϑ is small, ρ is an $O(\epsilon)$-tight lower bound on Opt. Note that when (9) is strictly feasible, ϑ indeed is positive, and it can be bounded away from zero. Indeed, we have the following:

LEMMA 1 *Let ϑ be the magnitude of the left derivative of Φ at* Opt *and assume that $\varkappa$ given by* (10) *is positive. Then*

$$\vartheta \geq \frac{\varkappa}{V + \varkappa} \quad \text{where } V = \max_{x \in X} f_0(x) - \mathrm{Opt}.$$

In respect to the constrained problem (9), our main result is as follows:

PROPOSITION 4 *In the just described situation, assume that f_i, $0 \leq i \leq m$, are differentiable on X, and that for some positive M_1, M_2 one has for $i = 0, 1, \ldots, m$ and all $x \in X$:*

$$\mathbf{E}[\mathrm{e}^{(F_i(x,\xi) - f_i(x))^2/M_1^2}] \leq \mathrm{e}, \quad \mathbf{E}[\mathrm{e}^{L^2(x,\xi)/M_2^2}] \leq \mathrm{e}.$$

Assume also that (9) *is feasible, and that for $N \in \mathbf{Z}_+$, $s > 1$, and $\lambda, \mu \in [0, 2\sqrt{\tau_* N}]$, ϵ and β satisfy*

$$\begin{aligned} \epsilon &> 2N^{-1/2}\left[\mu M_1 + M_2 R\left[\frac{\Omega}{2}[1 + s^2] + \lambda\right]\right], \\ \beta &= \beta(\mu, s, \lambda, N) = \mathrm{e}^{-N(s^2-1)} + \mathrm{e}^{-\lambda^2/4\tau_*} + (m+2)\mathrm{e}^{-\mu^2/4\tau_*}, \end{aligned} \tag{11}$$

where Ω is given by (2), *and τ_* is given in Proposition* 1. *Then the random quantity*

$$\mathrm{Opt}_N(\xi^N) = \min_{x \in X}[f_{0,N}(x, \xi^N) - \mu M_1 N^{-1/2} : f_{i,N}(x, \xi^N) - \mu M_1 N^{-1/2} \leq 0,\ 1 \leq i \leq m]$$

ϵ-underestimates Opt *with probability $\geq 1 - \beta$.*

MinMax Stochastic Optimization. The proof of Proposition 4 also yields the following result which is of interest by its own right.

PROPOSITION 5 *In the notation and under assumptions of Proposition* 4, *consider the minimax problem*

$$\mathrm{Opt} = \min_{x\in X} \max[f_1(x), \ldots, f_m(x)] \tag{12}$$

along with its Sample Average Approximation

$$\mathrm{Opt}_N(\xi^N) = \min_{x\in X} \max[f_{1,N}(x, \xi^N), \ldots, f_{m,N}(x, \xi^N)].$$

Then for every $N \in \mathbf{Z}_+$, $s > 1$ *and* $\lambda, \mu \in [0, 2\sqrt{\tau_* N}]$ *one has*

$$\mathrm{Prob}\{\mathrm{Opt}_N(\xi^N) > \mathrm{Opt} + \mu M_1 N^{-1/2}\} \le m\mathrm{e}^{-\mu^2/4\tau_*} \tag{13}$$

and

$$\begin{aligned}\mathrm{Prob}\left\{\mathrm{Opt}_N(\xi^N) < \mathrm{Opt} - \left[\mu M_1 + 2M_2\left[\frac{\Omega}{2}[1+s^2] + 2\lambda\right]\right]N^{-1/2}\right\}\\ \le \mathrm{e}^{-\mu^2/4\tau_*} + 2[\mathrm{e}^{-N(s^2-1)} + \mathrm{e}^{-\lambda^2/4\tau_*}].\end{aligned} \tag{14}$$

An attractive feature of bounds (13) and (14) is that they are only weakly affected by the number m of components in the minimax problem (12).

3. Numerical experiments

The goal of the experiments of this section is to illustrate numerically the ideas developed above.

3.1 *Confidence intervals for problems without stochastic constraints*

Here we consider three risk-averse optimization problems of the form (1) and we compare the properties of three confidence intervals for Opt computed for the confidence level $1 - \alpha = 0.9$:

(1) the asymptotic confidence interval

$$\mathcal{C}_a(\alpha) = \left[\hat{f}_{2N} - q_{\mathcal{N}}\left(1 - \frac{\alpha}{2}\right)\frac{\hat{\sigma}_{2N}}{\sqrt{N}},\ \hat{f}_{2N} + q_{\mathcal{N}}\left(1 - \frac{\alpha}{2}\right)\frac{\hat{\sigma}_{2N}}{\sqrt{N}}\right]. \tag{15}$$

Here $\hat{f}_{2N}$ and $\hat{\sigma}^2_{2N}$ are estimations of expectation $f(x(\xi^N))$ of $F(x(\xi^N), \xi')$ and of its variance, taken over the distribution of independent from ξ^N random vector ξ' (here $x(\xi^N)$ is the SAA (3) optimal solution built using the N-sample ξ^N). They are computed using a second sample $\bar{\xi}^N$ of ξ of size N independent of ξ^N:

$$\hat{f}_{2N} = \frac{1}{N}\sum_{t=1}^{N} F(x(\xi^N), \bar{\xi}_t), \quad \hat{\sigma}^2_{2N} = \frac{1}{N}\sum_{t=1}^{N} F(x(\xi^N), \bar{\xi}_t)^2 - \hat{f}^2_{2N} \tag{16}$$

(for a justification, see [25]).

(2) The (non-asymptotic) confidence interval $\mathcal{C}_{\mathrm{SMD}}(\alpha)$ is built using the offline accuracy certificates for the Stochastic Mirror Descent algorithm, cf. Section 3.2 and Theorem 2 of [11]. The non-Euclidean algorithm with entropy distance-generating function provided the best results in these experiments and was used for comparison.

(3) The (non-asymptotic) confidence interval, denoted $\mathcal{C}_{\mathrm{SAA}}(\alpha)$, is based on the bounds of Proposition 1. Specifically, we use the lower $1-\alpha/2$-confidence bound $\mathtt{Low}^{\mathrm{SAA}}$ of Corollary 1. To construct the upper bound, we proceed as follows: first we compute the optimal solution $x(\xi^N)$ of the SAA using a simulation sample ξ^N of size N; then we compute an estimation $\hat{f}_{2N}$ of the objective value using the independent sample $\bar{\xi}^N$ as in (16). Finally, we build the upper confidence bound

$$\mathtt{Up}' = \hat{f}_{2N} + 2M_1\sqrt{\frac{\tau^* \ln[4\alpha^{-1}]}{N}},$$

where τ^* and M_1 are as in Proposition 1 (cf. the bound (5)). Finally, the upper bound $\overline{\mathtt{Up}}^{\mathrm{SAA}}$ computed as the minimum of $\mathtt{Up}'$ and the upper bound $\mathtt{Up}^{\mathrm{SAA}}$ by Corollary 1, tuned for the confidence level $1-\alpha/4$, was used.[2]

For the sake of completeness, we provide in Appendix 2 detailed computation of the constants involved for the three optimization problems considered in this section. SAA formulations of these problems were solved numerically using Mosek optimization toolbox [1].

3.1.1 *Quadratic risk minimization*

Consider the following instance of problem (1): let X be the standard simplex in $\mathbf{R}^n$: $X = \{x \in \mathbf{R}^n : x_i \geq 0, \sum_{i=1}^n x_i = 1\}$, Ξ is a part of the unit box $\{\xi = [\xi_1; \ldots; \xi_n] \in \mathbf{R}^n : \|\xi\|_\infty \leq 1\}$,

$$F(x,\xi) = \kappa_0 \xi^{\mathrm{T}} x + \frac{\kappa_1}{2}(\xi^{\mathrm{T}} x)^2, \quad f(x) = \kappa_0 \mu^{\mathrm{T}} x + \frac{\kappa_1}{2} x^{\mathrm{T}} V x,$$

with $\kappa_1 \geq 0$ and $\mu = \mathbf{E}\{\xi\}$, $V = \mathbf{E}\{\xi\xi^{\mathrm{T}}\}$.

In our experiments, $\kappa_0 = 0.1$, $\kappa_1 = 0.9$, and ξ has independent Bernoulli entries: $\mathrm{Prob}(\xi_i = 1) = \theta_i$, $\mathrm{Prob}(\xi_i = -1) = 1 - \theta_i$, with θ_i drawn uniformly over $[0, 1]$. This implies that

$$\mu_i = 2\theta_i - 1, \quad V_{i,j} = \begin{cases} \mathbf{E}\{\xi_i\}\mathbf{E}\{\xi_j\} = (2\theta_i - 1)(2\theta_j - 1) & \text{for } i \neq j, \\ \mathbf{E}\{\xi_i^2\} = 1 & \text{for } i = j. \end{cases}$$

For several problem and sample sizes, we present in Table 2 the empirical 'coverage probabilities' of the 'asymptotic' confidence interval $\mathcal{C}_a(\alpha)$ (i.e. the ratio of realizations for which $\mathcal{C}_a(\alpha)$ covers the true optimal value) for $\alpha = 0.1$ and 'target coverage probability' $1 - \alpha = 0.9$, computed over 500 realizations (the coverage probabilities of the two non-asymptotic confidence intervals are equal to one for all parameter combinations). We observe that empirical coverage probabilities degrade when the problem size n increases (and, as expected, they tend to increase with the sample size). For instance, these probabilities are much smaller than the target level, unless the size N of the simulation sample is much larger than problem dimension n. On the

Table 2. Quadratic risk minimization.

	Problem size n			
Sample size N	2	10	20	100
20	0.94	0.68	0.59	0.10
100	0.95	0.87	0.70	0.46
10,000	0.94	0.95	0.91	0.85

Note: Estimated coverage probabilities of the asymptotic confidence intervals $\mathcal{C}_a(0.1)$.

Table 3. Quadratic risk minimization.

	$\frac{\lvert\mathcal{C}_{\mathrm{SAA}}(\alpha)\rvert}{\lvert\mathcal{C}_a(\alpha)\rvert}$, problem size n					$\frac{\lvert\mathcal{C}_{\mathrm{SMD}}(\alpha)\rvert}{\lvert\mathcal{C}_a(\alpha)\rvert}$, problem size n				
Sample size N	2	10	20	100	200	2	10	20	100	200
100	6.37	9.18	10.18	29.50	47.43	30.57	65.87	78.5	274.63	474.68
1000	3.27	4.33	4.52	13.92	22.46	15.52	32.56	36.98	134.67	232.32
10,000	3.15	4.37	4.40	13.44	21.96	15.46	32.40	35.87	131.70	227.56

Note: Average ratio of the widths of the non-asymptotic and asymptotic confidence intervals.

other hand, not surprisingly, the non-asymptotic bounds yield confidence intervals much larger than the asymptotic confidence interval. We report in Table 3 the mean ratio of the widths of non-asymptotic – $\mathcal{C}_{\mathrm{SAA}}(\alpha)$ and $\mathcal{C}_{\mathrm{SMD}}(\alpha)$ – and asymptotic confidence intervals $\mathcal{C}_a(\alpha)$.[3] These ratios increase significantly with problem size (in part because the asymptotic interval becomes indeed too short), and we observe that the confidence interval $\mathcal{C}_{\mathrm{SAA}}(\alpha)$ based on SAA remains much smaller than the interval $\mathcal{C}_{\mathrm{SMD}}(\alpha)$ yielded by Stochastic Approximation.

3.1.2 *Gaussian VaR optimization*

We consider the instance of problem (1) where $X \subset \mathbf{R}^n$ is the standard simplex, ξ has normal distribution $\mathcal{N}(0, \Sigma)$ on $\mathbf{R}^n$ with $\Sigma_{i,i} \leq \sigma_{\max}$, and $F(x, \xi) = \kappa_0 \xi^T x + \kappa_1 |\xi^{\mathrm{T}} x|$, with $\kappa_1 \geq 0$, so that $f(x) = \kappa_1 \sqrt{2/\pi} \sqrt{x^{\mathrm{T}} \Sigma x}$. Observe that in the present situation, minimizing $f(x)$ is equivalent to maximizing the ε-quantile of the distribution of $\xi^{\mathrm{T}} x$ (Value-at-Risk $\mathrm{VaR}(\varepsilon)$) with $\varepsilon = 1 - \Psi(\kappa_1 \sqrt{2/\pi})$ where $\Psi(\cdot)$ is the standard normal CDF.

We generated instances of the problem of different sizes with $\kappa_0 = 0.9, \kappa_1 = 0.1$, and diagonal matrix Σ with diagonal entries drawn uniformly over $[1, 6]$ ($\sigma_{\max} = \sqrt{6}$).

We reproduce the experiments of the previous section in this setting, namely, for several problem and sample sizes, we compute empirical 'coverage probabilities' of the confidence intervals over 500 realizations. We report the results for the 'asymptotic' confidence interval $\mathcal{C}_a(\alpha)$ in Table 4 for 'target coverage probability' $1 - \alpha = 0.9$ (same as above, coverage probabilities of non-asymptotic intervals are equal to one for all parameter combinations). We especially observe extremely low coverage probabilities for $n = 100$ and $N = 20$ or $N = 100$.

In Table 5, the average ratios of the widths of non-asymptotic and asymptotic confidence intervals are provided for the same experiment. Same as in the experiments described in the previous section, these ratios increase with problem size, and the confidence intervals by SMD are much more conservative than those by SAA.

Table 4. Gaussian VaR optimization.

	Problem size n			
Sample size N	2	10	20	100
20	0.95	0.73	0.53	0.05
100	0.9	0.78	0.48	0.006
10,000	0.92	0.91	0.92	0.68
100,000	0.94	0.92	0.92	0.92

Note: Estimated coverage probabilities of asymptotic confidence intervals.

Table 5. Gaussian VaR optimization.

	$\frac{\lvert\mathcal{C}_{\rm SAA}(\alpha)\rvert}{\lvert\mathcal{C}_a(\alpha)\rvert}$ for problem size n					$\frac{\lvert\mathcal{C}_{\rm SMD}(\alpha)\rvert}{\lvert\mathcal{C}_a(\alpha)\rvert}$ for problem size n				
Sample size N	2	10	20	100	200	2	10	20	100	200
20	4.42	6.15	6.11	6.27	6.35	40.16	112.38	133.80	183.61	205.66
100	5.04	9.11	10.79	12.87	13.44	46.41	172.00	244.68	397.01	458.85
10,000	5.27	12.17	16.29	26.65	30.28	49.15	237.79	386.31	974.32	1088.90

Note: Average ratio of the widths of the non-asymptotic and asymptotic confidence intervals.

3.1.3 *CVaR optimization*

We consider here the following CVaR optimization problem: given $\varepsilon > 0$, find

$$\mathrm{Opt}_\varepsilon = \min_{x'} \kappa_0 \mathbf{E}\{\xi^{\mathrm{T}} x'\} + \kappa_1 \mathrm{CVaR}_\varepsilon(\xi^{\mathrm{T}} x') \quad x' \in \mathbf{R}^n \ \sum_{i=1}^n x'_i = 1, \quad x' \geq 0, \tag{17}$$

where the support Ξ of ξ is a part of the unit box $\{\xi = [\xi_1; \ldots; \xi_n] \in \mathbf{R}^n : \|\xi\|_\infty \leq 1\}$, and where

$$\mathrm{CVaR}_\varepsilon(\xi^{\mathrm{T}} x') = \min_{x_0 \in \mathbf{R}} \{x_0 + \mathbf{E}\{\varepsilon^{-1}[\xi^{\mathrm{T}} x' - x_0]_+\}\}$$

is the Conditional Value-at-Risk of level $0 < \varepsilon < 1$, see [22]. Observing that $|\xi^{\mathrm{T}} x'| \leq 1$ a.s., the above problem is clearly of the form (1) with $X = \{x = [x_0; x'_1; \ldots; x'_n] \in \mathbf{R}^{n+1} : \ |x_0| \leq 1, \ x'_1, \ldots, x'_n \geq 0, \ \sum_{i=1}^n x'_i = 1\}$ and

$$F(x, \xi) = \kappa_0 \xi^{\mathrm{T}} x' + \kappa_1 \left(x_0 + \frac{1}{\epsilon}[\xi^{\mathrm{T}} x' - x_0]_+ \right).$$

We consider random instances of the problem with $\kappa_0, \kappa_1 \in [0, 1]$, and ξ with independent Bernoulli entries: $\mathrm{Prob}(\xi_i = 1) = \theta_i$, $\mathrm{Prob}(\xi_i = -1) = 1 - \theta_i$, with θ_i, $i = 1, \ldots, n$ drawn uniformly from $[0, 1]$.

We compare the non-asymptotic confidence interval $\mathcal{C}_{\rm SAA}(\alpha)$ for Opt_ε to the asymptotic confidence interval $\mathcal{C}_a(\alpha)$ with confidence level $1 - \alpha = 0.9$. We consider two sets of problem parameters: $(\kappa_0, \kappa_1, \varepsilon) = (0, 1, 0.5)$ and $(\kappa_0, \kappa_1, \varepsilon) = (0.1, 0.9, 0.1)$. [4] The empirical coverage probabilities for the asymptotic confidence interval are reported in Table 6. As in other experiments, the coverage probability is still below the target probability $1 - \alpha = 0.9$ when the sample size is not much larger than the problem size. For SAA, the coverage probabilities are equal to one for all parameter combinations.

We report in Table 7 the average ratio of the widths of non-asymptotic and asymptotic confidence intervals. Note that the Lipschitz constant of $F(\cdot, \xi)$ is proportional to $1/\varepsilon$ when ε is small.

Table 6. CVaR optimization.

	$\varepsilon = 0.1$, problem size $n+1$				$\varepsilon = 0.5$, problem size $n+1$			
Sample size N	3	11	21	101	3	11	21	101
100	0.96	0.74	0.85	0.78	0.87	0.79	0.82	0.70
1000	0.95	0.88	0.86	0.67	0.90	0.87	0.81	0.90
10,000	0.92	0.93	0.91	0.94	0.98	0.88	0.91	0.95

Note: Estimated coverage probabilities of asymptotic confidence intervals.

Table 7. CVaR optimization.

	$\varepsilon = 0.5$, problem size $n+1$					$\varepsilon = 0.1$, problem size $n+1$				
Sample size N	3	11	21	101	201	3	11	21	101	201
100	5.89	3.95	4.59	15.54	14.74	293.47	27.61	9.14	14.32	14.44
1000	6.48	5.36	5.38	14.27	22.97	294.16	27.04	8.72	34.43	37.42
10,000	4.30	4.62	6.30	8.74	8.77	293.92	26.91	8.66	31.70	34.18

Note: Average ratio $|\mathcal{C}_{\mathrm{SAA}}(\alpha)|/|\mathcal{C}_a(\alpha)|$ of the widths of the non-asymptotic and asymptotic confidence intervals.

This explains the fact that for small values of ϵ, the ratio of the widths of the proposed non-asymptotic and asymptotic confidence intervals grows up significantly, especially for problem size $n+1=3$.

The experiments of this section show that when the sample size is not much larger than the problem dimension, the asymptotic computations fail to provide the confidence set of the prescribed risk. In such case the proposed approach, though conservative, seems to be the only option available for constructing a reliable confidence interval.

3.2 *Lower bounding the optimal value of a minimax problem*

We illustrate here the application of Proposition 5 to lower bounding the optimal value of the MinMax problem (12). To this end we consider the toy problem

$$\mathrm{Opt} = \min_x \max \left[f_i(x),\ i = 1, \ldots, 3,\ x = [u; v],\ v \in \mathbf{R},\ u \in \mathbf{R}^n,\ \sum_{i=1}^n u_i = 1,\ u \geq 0 \right], \tag{18}$$

where

$$f_1(x) = v + \mathbf{E}\{\varepsilon^{-1}[\xi^{\mathrm{T}}u - v]_+\} + \chi_1, \quad f_2(x) = \mathbf{E}\{\xi^{\mathrm{T}}u\} + \chi_2, \quad f_3(x) = \chi_3 - \mathbf{E}\{\xi^{\mathrm{T}}u\},$$

ε and χ being some given parameters. The SAA of the problem reads

$$\mathrm{Opt}(\xi^N) = \min_x \max \left[f_{i,N}(x, \xi^N),\ i = 1, \ldots, 3,\ x = [u; v]\ v \in \mathbf{R},\ u \in \mathbf{R}^n,\ \sum_{i=1}^n u_i = 1,\ u \geq 0 \right], \tag{19}$$

with

$$f_{1,N}(x, \xi^N) = v + \frac{1}{N\varepsilon} \sum_{t=1}^N [\xi_t^{\mathrm{T}}u - v]_+ + \chi_1,$$

$$f_{2,N}(x, \xi^N) = \frac{1}{N} \sum_{t=1}^N \xi_t^{\mathrm{T}}u + \chi_2, \quad f_{3,N}(x, \xi^N) = \chi_3 - \frac{1}{N} \sum_{t=1}^N \xi_t^{\mathrm{T}}u.$$

One can try to build an 'asymptotic' lower bound for Opt as follows (note that here we are not concerned with the theoretical validity of this construction): given the optimal solution $x(\xi^N)$ to SAA (19) and an independent sample $\bar{\xi}^N$, compute empirical estimations $\hat{f}_{i,2N}$ and $\hat{\sigma}^2_{i,2N}$ of expectation and variance of $F_i(x(\xi^N), \xi')$, as explained in Section 3.1, then compute the lower

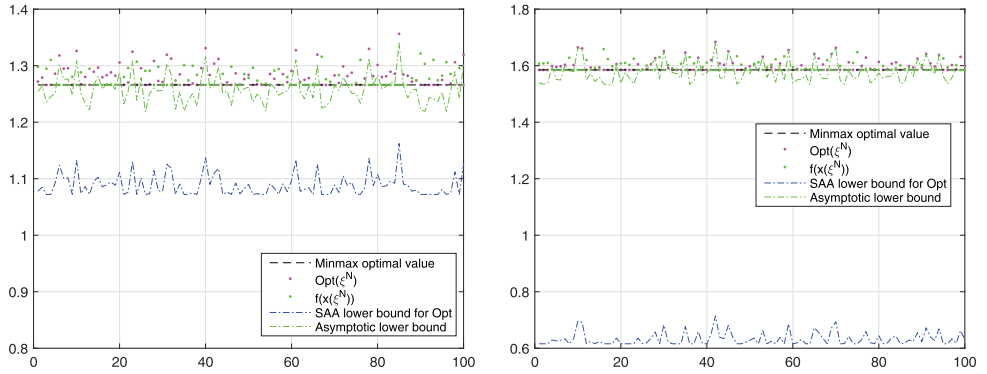

Figure 1. Optimal value Opt of the stochastic program (18) along with lower bound derived from the results of Proposition 5 and 'asymptotic' lower bound $\underline{\text{Opt}}(\xi^N)$. The results for $\varepsilon = 0.5$ on plot (a), for $\varepsilon = 0.1$ on plot (b).

bound 'of asymptotic risk α' according to

$$\underline{\text{Opt}}(\xi^N) = \max_{i=1,\ldots,3} \left\{ \hat{f}_{i,2N} - q_{\mathcal{N}} \left(1 - \frac{\alpha}{3}\right) \frac{\hat{\sigma}_{i,2N}}{\sqrt{N}} \right\}.$$

In Figure 1, we present the simulation results for the case of $\xi \in \mathbf{R}^n$ with independent Bernoulli components: $\text{Prob}(\xi_i = 1) = \theta_i$, $\text{Prob}(\xi_i = -1) = 1 - \theta_i$, with θ_i randomly drawn over $[0, 1]$. Parameters χ_i, $i = 1, 2, 3$, are chosen in such a way that f_1, f_2, and f_3 are equal at the minimizer of (18). More precisely, the results of 100 simulations of the problem with $n = 2$ and $N = 128$ are presented in Figure 1 for the value of CVaR parameter $\varepsilon = 0.5$ and $\varepsilon = 0.1$. Note that in this case the risk of the lower bound $\underline{\text{Opt}}(\xi^N)$ is significantly larger than the prescribed risk $\varepsilon = 0.1$ already for small problem dimension – the 'asymptotic' lower bound failed in 33 of 100 realizations in the experiment with $\varepsilon = 0.5$, and in 36 of 100 realizations in the experiment with $\varepsilon = 0.1$.

3.3 *Optimal value of a stochastically constrained problem*

An SAA of a stochastically constrained problem, even with a single linear constraint, can easily become unstable when the constraint is 'stiff'. As a simple illustration, let us consider a stochastically (linearly) constrained problem

$$\text{Opt}_\chi = \min_x \left[f_0(x) : f_1(x) \le 0,\ x = [u; v],\ v \in \mathbf{R},\ u \in \mathbf{R}^n,\ \sum_{i=1}^n u_i = 1,\ u \ge 0 \right], \tag{20}$$

where

$$f_0(x) = v + \mathbf{E}\{\varepsilon^{-1}[\xi^{\mathrm{T}} u - v]_+\} \quad \text{and} \quad f_1(x) = \chi - \mathbf{E}\{\xi^{\mathrm{T}} u\},$$

and ε and χ are problem parameters. The SAA of the problem is

$$\text{Opt}_\chi(\xi^N) = \min_{x=[u;v]} \left[f_{0,N}(x) : f_{1,N}(x) \le 0,\ v \in \mathbf{R},\ u \in \mathbf{R}^n,\ \sum_{i=1}^n u_i = 1,\ u \ge 0 \right], \tag{21}$$

where

$$f_{0,N}(x, \xi^N) = v + \frac{1}{N\varepsilon} \sum_{t=1}^N [\xi_t^{\mathrm{T}} u - v]_+ \quad \text{and} \quad f_{1,N}(x, \xi^N) = \chi - \frac{1}{N} \sum_{t=1}^N \xi_t^{\mathrm{T}} u.$$

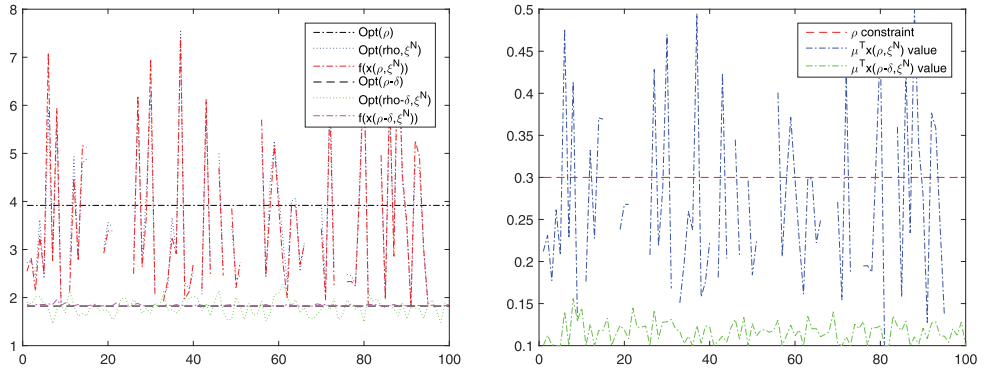

Figure 2. Plot (a): optimal value Opt of the stochastic program (20) with constraint right-hand side χ and $\chi - \delta$, along with corresponding optimal values of the SAA. Plot (b): 'true value' of the linear form $\mu^{\mathrm{T}} x(\xi^N)$ at the SAA solution.

Consider now a toy example of the problem with $u \in \mathbf{R}^2$, $\xi \sim \mathcal{N}(\mu, \Sigma)$ with $\mu = [0.1; 0.5]$ and $\Sigma = \mathrm{diag}([1; 4])$. Let $N = 128$, $\chi = 0.3$, and $\varepsilon = 0.1$. One can expect that in this case the optimal value $\mathrm{Opt}_\chi(\xi^N)$ of the SAA is unstable (in fact, problem (21) is infeasible with probability $\mathrm{Prob}\{(1/N)\sum_{t=1}^N \xi_{t,2} < \chi\} = \mathrm{Prob}\{2\mathcal{N}(0,1)/\sqrt{N} \le -0.2\} = 0.128\ldots$). We compare the solution to (21) with the SAA in which the right-hand side χ of the stochastic constraint is replaced with $\chi - \delta$ where $\delta = q_{\mathcal{N}}(1 - \varepsilon/n)(\sigma_{\max}/\sqrt{N}) = 0.5815\ldots$, $\sigma_{\max} = \max_i \Sigma_{i,i}$. In Figure 2 we present the simulation results of 100 independent realizations of the above problem. As expected, the SAA (21) is unstable; the problem turned infeasible in 22% of realizations. The SAA with the relaxed constraint exhibits much better stability.

Disclosure statement

No potential conflict of interest was reported by the authors.

Funding

Research of the first author was supported by an FGV grant, Conselho Nacional de Desenvolvimento Científico e Tecnológico [grants 307287/2013-0, 401371/2014-0], FAPERJ [grants E-26/110.313/2014, E-26/201.599/2014]. The second author was supported by the Centre National de la Recherche Scientifique CNRS-Mastodons project GARGANTUA, the LabEx PERSYVAL-Lab (ANR-11-LABX-0025), and CNPq [grant 401371/2014-0]. Research of the third author was supported by National Science Foundation [grants CMMI-1232623, CMMI-1262063, CCF-1415498], and CNPq [grant 401371/2014-0].

Notes

1. For example, for $\alpha \le \frac{1}{2}$ and $\epsilon \le M_1 + 4M_2$ we have an immediate (though rough) bound
$$N_\epsilon = \frac{4\tau_*(M_1 + 4M_2)^2 \ln(4/\alpha)}{\epsilon^2}.$$
2. It is worth to mention that in our experiments the upper bound $\mathrm{Up}^{\mathrm{SAA}}$ was too conservative and was systematically 'outperformed' by the upper bound Up'.
3. Note that asymptotic estimation $\hat{\sigma}_N$ of the noise variance often degenerates. To avoid division by zero problems, we only kept the realizations where asymptotic confidence intervals cover the true optimal value.
4. Value $\varepsilon = 0.1$ is a typical choice in risk-averse portfolio optimization setting. On the other hand, CVaR minimization for $\varepsilon = 0.5$ corresponds to *median regression*, an important special case of quantile regression model, see [10].
5. For details, see, e.g. [16, Theorem 2.1]

6. In fact, in the numerical experiments we have used a slightly better bound M_2 which can be defined as follows. Let t_n, $0 < t_n < \sigma_{\max}$ be the unique solution of the equation

$$\tilde{h}_n(t_n) = \frac{n^{2t_n^2\sigma_{\max}^2}}{1 - 2t_n^2\sigma_{\max}^2} = \mathrm{e}$$

(observe that $\tilde{h}_n(\cdot)$ is monotone on $]0, 1/\sqrt{2}\sigma_{\max}[$, so t_n can be computed using bisection). The same reasoning as in the proof of Lemma A.2 results in the bound

$$M_2 = \frac{(|\kappa_0| + \kappa_1)}{t_n} + \kappa_1\sigma_{\max}\sqrt{\frac{2}{\pi}}. \tag{22}$$

For instance, in the experiments of Section 3.1.2, for $\sigma_{\max} = \sqrt{6}$ and $n \in \{2, 10, 20, 100\}$, the values of $1/t_n$ (resp. of its upper bound $\sigma_{\max}\sqrt{2(2 + \ln n)}$) were 4.97,6.46,7.05,8.27 (resp. 5.68,7.19,7.74,8.90).

References

[1] E.D. Andersen and K.D. Andersen (2013). Available at http://docs.mosek.com/7.0/toolbox/.
[2] L. Bottou, *Large-Scale Machine Learning with Stochastic Gradient Descent*, Proceedings of COMPSTAT'2010, Springer, 2010, pp. 177–186.
[3] S. Bubeck, V. Perchet, and P. Rigollet, *Bounded regret in stochastic multi-armed bandits* (2015). Available at http://www.jmlr.org/proceedings/papers/v30/Bubeck13.pdf.
[4] J. Dupacovà and R.B. Wets, *Asymptotic behavior of statistical estimators and of optimal solutions of stochastic optimization problems*, Ann. Stat. 16 (1988), pp. 1517–1549.
[5] V. Guigues, *Multistep stochastic mirror descent for risk averse convex stochastic programs based on extended polyhedral risk measures*, Math. Program. 163 (2017), pp. 169–212.
[6] A. Juditsky and A.S. Nemirovski, *Large deviations of vector-valued martingales in 2-smooth normed spaces*, preprint (2008). Available at arXiv:0809.0813.
[7] Y.M. Kaniovski, A.J. King, and R.J.-B. Wets, *Probabilistic bounds (via large deviations) for the solutions of stochastic programming problems*, Ann. Oper. Res. 56 (1995), pp. 189–208.
[8] A.J. King and R.T. Rockafellar, *Asymptotic theory for solutions in statistical estimation and stochastic programming*, Math. Oper. Res. 18 (1993), pp. 148–162.
[9] A.J. Kleywegt, A. Shapiro, and T. Mello-de-Homem, *The sample average approximation method for stochastic discrete optimization*, SIAM J. Optim. 12 (2002), pp. 479–502.
[10] R. Koenker, *Quantile Regression*, Cambridge university press, Cambridge, UK, 2005.
[11] G. Lan, A. Nemirovski, and A. Shapiro, *Validation analysis of mirror descent stochastic approximation method*, Math. Program. 134 (2012), pp. 425–458.
[12] J. Linderoth, A. Shapiro, and S. Wright, *The empirical behavior of sampling methods for stochastic programming*, Ann. Oper. Res. 142 (2006), pp. 215–241.
[13] W.-K. Mak, D.P. Morton, and R.K. Wood, *Monte Carlo bounding techniques for determining solution quality in stochastic programs*, Oper. Res. Lett. 24 (1999), pp. 47–56.
[14] A. Nemirovski, A. Juditsky, G. Lan, and A. Shapiro, *Robust stochastic approximation approach to stochastic programming*, SIAM J. Optim. 19 (2009), pp. 1574–1609.
[15] A. Nemirovsky and D. Yudin. *Problem complexity and method efficiency in optimization*, Wiley, New York, 1983.
[16] Y. Nesterov and A. Nemirovski, *On first-order algorithms for l 1/nuclear norm minimization*, Acta Numer. 22 (2013), pp. 509–575.
[17] V. Perchet and P. Rigollet, *The multi-armed bandit problem with covariates*, Ann. Stat. 41 (2013), pp. 693–721.
[18] G.Ch. Pflug, *Asymptotic stochastic programs*, Math. Oper. Res. 20 (1995), pp. 769–789.
[19] G.Ch. Pflug, *Stochastic programs and statistical data*, Ann. Oper. Res. 85 (1999), pp. 59–78.
[20] G. Ch. Pflug, *Stochastic optimization and statistical inference*, in *Stochastic Programming: Handbooks in Operations Research and Management Science*, Vol. 10, A. Ruszczynski and A. Shapiro, eds., Elsevier, Amsterdam, 2003, pp. 427–480.
[21] H. Robbins and S. Monro, *A stochastic approximation method*, Ann. Math. Stat. 22 (1951), pp. 400–407.
[22] R.T. Rockafellar and S. Uryasev, *Conditional value-at-risk for general loss distributions*, J. Banking Finance 26 (2002), pp. 1443–1471.
[23] S. Shalev-Shwartz, O. Shamir, N. Srebro, and K. Sridharan, *Learnability and stability in the general learning setting*, in *COLT*, 2009.
[24] A. Shapiro, *Asymptotic analysis of stochastic programs*, Ann. Oper. Res. 30 (1991), pp. 169–186.
[25] A. Shapiro, D. Dentcheva, and A. Ruszczyński, Lectures on stochastic programming: modeling and theory, SIAM, Philadelphia, 2014.
[26] A. Shapiro, *Monte Carlo sampling methods*, in *Handbooks in Operations Research and Management Science*, A. Ruszczyński and A. Shapiro, eds., Vol. 10, Elsevier, 2003, pp. 353–425.
[27] A. Shapiro and A. Nemirovski, *On complexity of stochastic programming problems*, in *Continuous Optimization: Current Trends and Applications*, V. Jeyakumar and A. Rubinov, eds., Springer, 2005, pp. 111–146.
[28] N. Srebro and A. Tewari, *Stochastic optimization for machine learning*, ICML Tutorial (2010).

[29] S. van de Geer and J. Lederer, *The Bernstein–Orlicz norm and deviation inequalities*, Probab. Theory Rel. Fields 157 (2013), pp. 225–250.
[30] B. Verweij, S. Ahmed, A.J. Kleywegt, G. Nemhauser, and A. Shapiro, *The sample average approximation method applied to stochastic routing problems: A computational study*, Comput. Optim. Appl. 24 (2003), pp. 289–333.

Appendix 1. Proofs

A.1 *Preliminaries: large deviations of vector-valued martingales*

The result to follow is a slightly simplified and refined version of the bounds on probability of large deviations for vector-valued martingales developed in [6,14].

Let $\|\cdot\|$ be a norm on Euclidean space E, $\|\cdot\|_*$ be the conjugate norm, and $B_{\|\cdot\|}$ be the unit ball of the norm. Further, let ω be a continuously differentiable distance-generating function for $B_{\|\cdot\|}$ compatible with the norm $\|\cdot\|$ and attaining its minimum on $B_{\|\cdot\|}$ at the origin: $\omega'(0) = 0$, with $\omega(0) = 0$ and $\Omega = \max_{x:\|x\|\le 1} \sqrt{2[\omega(x)]}$.

LEMMA A.1 *Let $d_1, d_1, \ldots$ be a scalar martingale-difference such that for some $\sigma > 0$ it holds*

$$\mathbf{E}\{e^{d_t^2/\sigma^2} \mid d_1, \ldots, d_{t-1}\} \le e \quad \text{a.s. } t = 1, 2, \ldots$$

Then

$$\text{Prob}\left\{\underbrace{\sum_{t=1}^N d_t}_{D_N} > \lambda\sigma\sqrt{N}\right\} \le \begin{cases} e^{-\lambda^2/4\tau_*}, & 0 \le \lambda \le 2\sqrt{\tau_* N}, \\ e^{-\lambda^2/3}, & \lambda > 2\sqrt{\tau_* N}, \end{cases} \tag{A1}$$

where τ_* is defined in Proposition 1.

Proof Assuming without loss of generality that $\sigma = 1$ observe that under lemma's assumptions we have $\mathbf{E}\{e^{\tau_*\theta^2 d_t^2} \mid d_1, \ldots, d_{t-1}\} \le e^{\tau_*\theta^2}$ whenever $\tau_*\theta^2 \le 1$ where τ_* is defined in Proposition 1, and therefore for almost all $d^{t-1} = (d_1, \ldots, d_{t-1})$ we have for $0 \le \theta \le 1/\sqrt{\tau_*}$

$$\mathbf{E}\{e^{\theta d_t} \mid d^{t-1}\} \le \mathbf{E}\{\theta d_t + e^{\tau_*\theta^2 d_t^2} \mid d^{t-1}\} = \mathbf{E}\{e^{\tau_*\theta^2 d_t^2} \mid d^{t-1}\} \le e^{\tau_*\theta^2}. \tag{A2}$$

Thus, for $0 \le \theta \le 1/\sqrt{\tau_*}$, we have $\mathbf{E}\{e^{\theta D_N}\} \le e^{\tau_*\theta^2 N}$, and $\forall \lambda > 0$

$$\text{Prob}\{D_N > \lambda\sqrt{N}\} \le e^{\tau_*\theta^2 N - \lambda\theta\sqrt{N}}.$$

When minimizing the resulting probability bound over $0 \le \theta \le 1/\sqrt{\tau_*}$, we get the inequality (A1) for $\lambda \in [0, 2\sqrt{\tau_* N}]$: $\text{Prob}\{D_N > \lambda\sqrt{N}\} \le e^{-\lambda^2/4\tau_*}$. The corresponding bound for $\lambda > 2\sqrt{\tau_* N}$ is given by exactly the same reasoning as above in which (A2) is substituted with the inequality

$$\mathbf{E}\{e^{\theta d_t} \mid d^{t-1}\} \le \mathbf{E}\{e^{3\theta^2/8 + 2d_t^2/3} \mid d^{t-1}\} \le e^{3\theta^2/8 + 2/3} \le e^{3\theta^2/4}$$

when $\theta > 1/\sqrt{\tau_*}$. ■

PROPOSITION A.1 *Let $(\chi_t)_{t=1,2,\ldots}$, $\chi_t \in E$, be a martingale-difference such that for some $\sigma > 0$ it holds*

$$\mathbf{E}\{e^{\|\chi_t\|_*^2/\sigma^2} \mid \chi_1, \ldots, \chi_{t-1}\} \le e \quad \text{a.s. } t = 1, 2, \ldots \tag{A3}$$

Then for every $s > 1$, we have

$$\text{Prob}\left\{\left\|\sum_{t=1}^N \chi_t\right\|_* > \sigma\left[\frac{\Omega\sqrt{N}}{2}[1 + s^2] + \lambda\sqrt{N}\right]\right\} \le \begin{cases} e^{-N(s^2-1)} + e^{-\lambda^2/4\tau_*}, & 0 \le \lambda \le 2\sqrt{\tau_* N}, \\ e^{-N(s^2-1)} + e^{-\lambda^2/3}, & \lambda > 2\sqrt{\tau_* N}, \end{cases} \tag{A4}$$

where τ_* is defined in Proposition 1 and Ω is given by (2).

Proof By homogeneity, it suffices to consider the case when $\sigma = 1$, which we assume from now on.

1^0. Let $\gamma > 0$. We denote

$$V_x(u) = \omega(u) - \omega(x) - \langle \omega'(x), u - x\rangle \quad [u, x \in B_{\|\cdot\|}]$$

and consider the recurrence

$$x_1 = 0, \quad x_{t+1} = \operatorname*{argmin}_{y \in B_{\|\cdot\|}} [V_{x_t}(y) - \langle \gamma \chi_t, y\rangle].$$

Observe that x_t is a deterministic function of $\chi^{t-1} = (\chi_1, \ldots, \chi_{t-1})$, and that by the standard properties of proximal mapping (see. e.g. [14, Lemma 2.1]),

$$\forall (u \in B_{\|\cdot\|}) : \gamma \sum_{t=1}^{N} \langle \chi_t, u - x_t\rangle \le V_0(u) - V_{x_{N+1}}(u) + \frac{\gamma^2}{2} \sum_{t=1}^{N} \|\chi_t\|_*^2 \le \frac{1}{2}\Omega^2 + \frac{\gamma^2}{2} \sum_{t=1}^{N} \|\chi_t\|_*^2.$$

Thus

$$\max_{u \in B_{\|\cdot\|}} \left\langle \sum_{t=1}^{N} \chi_t, u \right\rangle \le \frac{\Omega^2}{2\gamma} + \frac{\gamma}{2} \underbrace{\sum_{t=1}^{N} \|\chi_t\|_*^2}_{\eta_N} + \underbrace{\sum_{t=1}^{N} \langle \chi_t, x_t\rangle}_{\zeta_N}.$$

Setting $\gamma = \Omega/\sqrt{N}$, we arrive at

$$\max_{u \in B_{\|\cdot\|}} \left\langle \sum_{t=1}^{N} \chi_t, u \right\rangle \le \frac{\Omega\sqrt{N}}{2} \left[1 + \frac{\eta_N}{N}\right] + \zeta_N. \tag{A5}$$

Invoking (A3), we get

$$\mathbf{E}\{e^{\eta_N}\} \le e^N$$

(recall that $\sigma = 1$), whence

$$\forall s > 0 : \quad \mathrm{Prob}\{\eta_N > s^2 N\} \le \min[1, e^{N(1-s^2)}]. \tag{A6}$$

2^0. When invoking (A3) and taking into account that x_t is a deterministic function of χ^{t-1} such that $\|x_t\| \le 1$ (since $x_t \in B_{\|\cdot\|}$), we get

$$\mathbf{E}\{\langle \chi_t, x_t\rangle \mid \chi^{t-1}\} = 0, \quad \mathbf{E}\{e^{\langle \chi_t, x_t\rangle^2} \mid \chi^{t-1}\} \le e. \tag{A7}$$

Applying Lemma A.1 to the random sequence $d_t = \langle \chi_t, x_t\rangle$, $t = 1, 2, \ldots$ (which is legitimate, with σ set to 1, by (A7)), we get

$$\mathrm{Prob}\{\zeta_N > \lambda\sqrt{N}\} \le \begin{cases} e^{-\lambda^2/4\tau_*}, & 0 \le \lambda \le 2\sqrt{\tau_* N}, \\ e^{-\lambda^2/3}, & \lambda > 2\sqrt{\tau_* N}. \end{cases} \tag{A8}$$

In view of (A6) and (A8), relation (A5) implies the bound (A4) of the proposition. ■

A.2 *Proof of Proposition* 1

Let x_* be an optimal solution to (SP), and let $h = \nabla f(x_*)$, so that by optimality conditions

$$\langle h, x - x_*\rangle \ge 0, \quad \forall x \in X. \tag{A9}$$

1^0. Setting $\delta(\xi) = F(x_*, \xi) - f(x_*)$, invoking (4.*a*) and applying Lemma A.1 to the random sequence $d_t = \delta(\xi_t)$ and $\sigma = M_1$ (which is legitimate by (4.*a*)), we get

$$\forall (N \in \mathbf{Z}_+,\ \mu \in [0, 2\sqrt{\tau_* N}]) : \ \mathrm{Prob}\left\{\frac{1}{N}\sum_{t=1}^{N} \delta(\xi_t) > \mu M_1 N^{-1/2}\right\} \le e^{-\mu^2/4\tau_*}. \tag{A10}$$

Since clearly

$$\mathrm{Opt}_N(\xi^N) \le f_N(x_*, \xi^N) = \mathrm{Opt} + \frac{1}{N}\sum_{t=1}^{N} \delta(\xi_t),$$

we get

$$\mathrm{Prob}\{\mathrm{Opt}_N(\xi^N) > \mathrm{Opt} + \mu M_1 N^{-1/2}\} \le e^{-\mu^2/4\tau_*}. \tag{A11}$$

2^0. It is immediately seen that under the assumptions of Proposition 1, for every measurable vector-valued function $g(\xi) \in \partial_x F(x_*, \xi)$ we have

$$h = \int_{\Xi} g(\xi) P(d\xi). \tag{A12}$$

Observe that $h_N(\xi^N) = (1/N)\sum_{t=1}^N g(\xi_t)$ is a subgradient of $f_N(x,\xi^N)$ at the point x_*. Consequently, for all $x \in X$,

$$\begin{aligned} f_N(x,\xi^N) &\geq f_N(x_*,\xi^N) + \langle h_N(\xi^N), x - x_*\rangle \\ &\geq \underbrace{[f(x_*) + \langle h, x - x_*\rangle]}_{\geq \text{Opt by}(A9)} + [[f_N(x_*,\xi^N) - f(x_*)] + \langle h_N(\xi^N) - h, x - x_*\rangle] \\ &\geq \text{Opt} + \frac{1}{N}\sum_{t=1}^N \delta(\xi_t) - \|h - h_N(\xi^N)\|_*\|x - x_*\| \geq \text{Opt} + \frac{1}{N}\sum_{t=1}^N \delta(\xi_t) - 2\|h - h_N(\xi^N)\|_* R \end{aligned}$$

(the concluding inequality is due to $x, x_* \in X$ and thus $\|x - x_*\| \leq 2R$ by definition of R). It follows that

$$\text{Opt}_N(\xi^N) \geq \text{Opt} + \frac{1}{N}\sum_{t=1}^N \delta(\xi_t) - 2\|h - h_N(\xi^N)\|_* R. \tag{A13}$$

Applying Lemma A.1 to the random sequence $d_t = -\delta(\xi_t)$ we, similarly to the above, get

$$\forall (N, \mu \in [0, 2\sqrt{\tau_* N}]) : \text{Prob}\left\{\frac{1}{N}\sum_{t=1}^N \delta(\xi_t) < -\mu M_1 N^{-1/2}\right\} \leq \mathrm{e}^{-\mu^2/4\tau_*}. \tag{A14}$$

Further, setting $\Delta(\xi) = g(\xi) - \nabla f(x_*)$, the random vectors $\chi_t = \Delta(\xi_t)$, $t = 1, 2, \ldots$, are i.i.d., zero mean (by (A12)), and satisfy the relation

$$\mathbf{E}\{\mathrm{e}^{\|\chi_t\|_*^2/M_2^2}\} \leq \mathrm{e}$$

by (4(b)); besides this, $h_N(\xi^N) - h = (1/N)\sum_{t=1}^N \chi_t$. Applying Proposition A.1, we get

$$\begin{aligned} &\forall (N \in \mathbf{Z}_+,\ s > 1,\ \lambda \in [0, 2\sqrt{\tau_* N}]) : \\ &\text{Prob}\left\{\|h - h_N(\xi^N)\|_* \geq M_2\left[\frac{\Omega}{2}[1 + s^2] + \lambda\right]N^{-1/2}\right\} \leq \mathrm{e}^{-N(s^2-1)} + \mathrm{e}^{-\lambda^2/\tau_*}. \end{aligned}$$

This combines with (A13), and (A14) to imply (6).

A.3 *Proof of Proposition* 2

Due to similarity reasons, it suffices to prove the proposition for $L = R = 1$. Let $B_{\|\cdot\|_2}$ be the unit Euclidean ball of $\mathbb{R}^n$, and let for a unit $v \in \mathbf{R}^n$ and $0 < \theta \leq \pi/2$, $h_{v,\theta}$ be the spherical cap of $B_{\|\cdot\|_2}$ with 'center' v and angle θ. In other words, if $\delta = 2\sin^2(\theta/2)$ is the 'elevation' of the cap $h_{v,\theta}$ then $h_{v,\theta} = \{x \in B_{\|\cdot\|_2} :\ v^{\mathrm{T}}x \geq 1 - \delta\}$. Observe that for any $\vartheta > 4\theta$ we can straightforwardly build the system D_θ of vectors in the n-dimensional unit sphere S^{n-1} in such a way that the angle between every two distinct vectors of the system is $> 2\theta$, so that the spherical caps $h_{v,\theta}$ with $v \in D_\theta$ are mutually disjoint, while the spherical caps $h_{v,\vartheta}$ cover S_{n-1}. If we denote by $A_{n-1}(\vartheta)$ the area of the spherical cap of angle $\vartheta \leq \pi/2$, then $\text{Card}(D_\theta)A_{n-1}(\vartheta) \geq s_{n-1}(1)$, where $s_{n-1}(r) = 2\pi^{n/2}r^{n-1}/\Gamma(n/2)$ is the area of the n-dimensional sphere of radius r. Note that $A_{n-1}(\vartheta)$ satisfies

$$A_{n-1}(\vartheta) = \int_0^\vartheta s_{n-2}(\sin t)\,\mathrm{d}t = s_{n-2}(1)\int_0^\vartheta \sin^{n-2} t\,\mathrm{d}t \leq s_{n-2}(1)\int_0^\vartheta t^{n-2}\,\mathrm{d}t = s_{n-2}(1)\frac{\vartheta^{n-1}}{n-1}.$$

We conclude that

$$\text{Card}(D_\theta) \geq \frac{s_{n-1}(1)(n-1)}{s_{n-2}(1)\vartheta^{n-1}} \geq 3\vartheta^{1-n}$$

for $n \geq 2$. From now on we fix $\theta = 1/8$ and when choosing ϑ arbitrarily close to $4\theta = \frac{1}{2}$, we conclude that for any $n \geq 2$ one can build D_θ such that $\text{Card}(D_\theta) \geq 2^n$.

Now consider the following construction: for $v \in D_\theta$, let $g_{v,\theta}(\cdot) : B_{\|\cdot\|_2} \to \mathbf{R}$ be defined according to $g_{v,\theta}(x) = [v^{\mathrm{T}}x - (1-\delta)]_+$, where $\delta = 2\sin^2(\theta/2) = 0.0078023\ldots$ is the elevation of $h_{v,\theta}$. Let us put

$$f(x) = \sum_{v \in D_\theta} g_{v,\theta}(x),$$

and consider the optimization problem $\text{Opt} = \min[f(x) :\ x \in B_{\|\cdot\|_2}]$. Since $g_{v,\delta}$ is affine on $h_{v,\delta}$ and vanishes elsewhere on $B_{\|\cdot\|_2}$, and $\|v\|_2 = 1$, we conclude that f is Lipschitz continuous on $B_{\|\cdot\|_2}$ with Lipschitz constant 1. Let now

$$F(x,\xi) = \sum_{v \in D_\theta} 2\xi_v g_{v,\theta}(x),$$

where $\xi_v,\ v \in D_\theta$ are i.i.d. Bernoulli random variables with $\mathrm{Prob}\{\xi_v = 0\} = \mathrm{Prob}\{\xi_v = 1\} = \frac{1}{2}$. Note that $\mathbf{E}_\xi\{F(x,\xi)\} = f(x)\ \forall x \in B_{\|\cdot\|_2}$. Further, for $x \in h_{v,\theta}$, $\mathbf{E}_\xi\{F(x,\xi)^2 - f(x)^2\} = g^2_{v,\theta}(x) \le \delta^2$, and

$$\|F'(x,\xi) - f'(x)\|_2^2 = \|(2\xi_v - 1)g'_{v,\theta}(x)\|_2^2 \le 1.$$

Let us now consider the SAA $f_N(x,\xi^N)$ of f,

$$f_N(x,\xi^N) = \frac{1}{N}\sum_{t=1}^N F(x,\xi_t) = \sum_{v\in D_\theta} \underbrace{\frac{1}{N}\sum_{t=1}^N \xi_{t,v} g_{v,\theta}(x)}_{g^N_{v,\theta}(x)}, \tag{A15}$$

$\xi_t,\ t = 1,\dots,N$, being independent realizations of ξ, and the problem of computing

$$\mathrm{Opt}_N(\xi^N) = \min[f_N(x,\xi^N) :\ x \in B_{\|\cdot\|_2}]. \tag{A16}$$

Note that for a given $v \in D_\theta$, $\mathrm{Prob}\{\sum_{t=1}^N \xi_{t,v} = 0\} = 2^{-N}$. Due to the independence of ξ_v, we have

$$\mathrm{Prob}\left\{\sum_{t=1}^N \xi_{t,v} > 0,\ \forall v \in D_\theta\right\} = (1-2^{-N})^{\mathrm{Card}(D_\theta)} \le (1-2^{-N})^{2^n} \le \mathrm{e}^{-2^n/2^N} \le \exp(-1),$$

for $N \le n$. We conclude that for $N \le n$, with probability $\ge 1 - \mathrm{e}^{-1}$, at least one of the summands in the right-hand side of (A15), let it be $g^N_{v,\theta}(x)$, is identically zero on $B_{\|\cdot\|_2}$. The optimal value $\mathrm{Opt}_N(\xi^N)$ of (A16) being zero, the point $x(\xi^N) = \bar v$ is clearly a minimizer of $f_N(x,\xi^N)$ on $B_{\|\cdot\|_2}$, yet $f(x(\xi^N)) = \delta$, i.e. (7) holds with $c_0 = \delta$.

A.4 *Proof of Proposition* 3

1^0. Let us consider a family of stochastic optimization problems as follows. Let $\|\cdot\| = \|\cdot\|_2$ and let X be the unit $\|\cdot\|_2$-ball in $\mathbf{R}^n$. Given a unit vector h in $\mathbf{R}^n$, positive reals σ, s and δ, d, and setting $\xi = [\eta;\zeta] \sim \mathcal{N}(0, I_2)$, consider two integrands:

$$F_0(x,\xi) = \sigma\eta h^{\mathrm{T}}x + s\zeta, \quad F_1(x,\xi) = (\delta h + \sigma\eta h)^{\mathrm{T}}x + (s\zeta - d),$$

so that

$$f_0(x) := \mathbf{E}_\xi\{F_0(x,\xi)\} = 0, \quad f_1(x) := \mathbf{E}_\xi\{F_1(x,\xi)\} = \delta h^{\mathrm{T}}x - d.$$

Let us now check that F_0 and F_1 verify the assumptions of Proposition 1. In the notation of Proposition 1, we have for F_1

$$L(x,\xi) = \|[\delta h + \sigma\eta h] - \delta h\|_2 = \sigma|\eta|,$$

whence, setting $M_2 = \sigma/\gamma$ with $\gamma^2 = \frac{1}{2}(1-\mathrm{e}^{-2})$,

$$\mathbf{E}_\xi\{\exp\{L(x,\xi)^2/M_2^2\}\} = \exp\{1\}.$$

Similarly, setting $M_1 = \sqrt{\sigma^2+s^2}/\gamma$, we have

$$\mathbf{E}_\xi\{\exp\{(\sigma\eta + s\zeta)^2/M_1^2\}\} = \exp\{1\},$$

so that, for every $z \in [-1,1]$,

$$\mathbf{E}_\xi\{\exp\{(\sigma\eta z + s\zeta)^2/M_1^2\}\} \le \exp\{1\}.$$

When $x \in \mathbf{R}^n$ and $\|x\|_2 \le 1$, we have $F_1(x,\xi) - f_1(x) = \sigma\eta h^{\mathrm{T}}x + s\zeta$, therefore

$$\mathbf{E}_\xi\{\exp\{(F_1(x,\xi) - f_1(x))^2/M_1^2\}\} \le \exp\{1\}.$$

We conclude that $F = F_1$ satisfies the assumptions of Proposition 1 with

$$M_1 = \sqrt{\sigma^2+s^2}/\gamma, \quad M_2 = \sigma/\gamma.$$

It is immediately seen that $F = F_0$ satisfies the assumptions of Proposition 1 with the same $M_1,\ M_2$.

2^0. Now, with $X = \{x \in \mathbf{R}^n : \|x\|_2 \le 1\}$, the optimal values in the problems of minimizing over X the functions f_0 and f_1 are, respectively,

$$\mathrm{Opt}_0 = 0, \quad \mathrm{Opt}_1 = -\delta - d.$$

Suppose that there exists a procedure which, under the assumptions of Proposition 1 with some fixed M_1, M_2, is able, given N observations of $\nabla_x F(\cdot,\xi_t),\ F(\cdot,\xi_t)$, to cover Opt, with confidence $1-\alpha$, by an interval of width W. Note that when $W < |\mathrm{Opt}_1|$, the same procedure can distinguish between the hypotheses stating that the observed first-order information on f comes from F_0 or from F_1, with risk (the maximal probability of rejecting the true hypothesis) α. On the other hand, when $F = F_0$ or $F = F_1$, our observations are deterministic functions of the samples $\omega_1,\dots,\omega_N$ drawn

from the two-dimensional normal distribution $\mathcal{N}\left(\left[\begin{smallmatrix}0\\0\end{smallmatrix}\right], \left[\begin{smallmatrix}\sigma^2 & 0\\ 0 & s^2\end{smallmatrix}\right]\right)$ for $F = F_0$, and $\mathcal{N}\left(\left[\begin{smallmatrix}\delta\\ d\end{smallmatrix}\right], \left[\begin{smallmatrix}\sigma^2 & 0\\ 0 & s^2\end{smallmatrix}\right]\right)$ for $F = F_1$. It is well known that deciding between such hypotheses with risk $\leq \alpha$ is possible only if

$$\sqrt{\frac{\delta^2}{\sigma^2} + \frac{d^2}{s^2}} \geq \frac{2}{\sqrt{N}} q_{\mathcal{N}}(1-\alpha).$$

We arrive at the following lower bound on W, given M_1, M_2, with $M_1 \geq M_2 > 0$:

$$W \geq \max_{\delta\geq 0,\, d\geq 0} \left\{\delta + d : \sqrt{\frac{\delta^2}{\gamma^2 M_2^2} + \frac{d^2}{\gamma^2(M_1^2 - M_2^2)}} \leq \frac{2}{\sqrt{N}} q_{\mathcal{N}}(1-\alpha)\right\} = \frac{2\gamma M_1}{\sqrt{N}} q_{\mathcal{N}}(1-\alpha) = \underline{\mathrm{W}}.$$

A.5 *Proof of Lemma* 1

Without loss of generality we may assume that Opt $= 0$. Let $\bar{x}$ be such that $f_i(\bar{x}) \leq -\varkappa$, $1 \leq i \leq m$. Given $\delta > 0$, there exists $x_\delta \in X$ such that $f_0(x_\delta) + \delta \leq \Phi(-\delta)$ and $f_i(x_\delta) \leq \Phi(-\delta)$, $1 \leq i \leq m$; note that $\Phi(-\delta) > 0$ due to $-\delta < 0 = \mathrm{Opt}$. The point

$$x = \frac{\Phi(-\delta)}{\varkappa + \Phi(-\delta)}\bar{x} + \frac{\varkappa}{\varkappa + \Phi(-\delta)} x_\delta$$

belongs to X and is feasible for (9), since for $i \geq 1$ one has

$$f_i(x) \leq \frac{\Phi(-\delta)}{\varkappa + \Phi(-\delta)} f_i(\bar{x}) + \frac{\varkappa}{\varkappa + \Phi(-\delta)} f_i(x_\delta) \leq -\frac{\Phi(-\delta)\varkappa}{\varkappa + \Phi(-\delta)} + \frac{\varkappa\Phi(-\delta)}{\varkappa + \Phi(-\delta)} = 0.$$

As a result,

$$0 = \mathrm{Opt} \leq f_0(x) \leq \frac{\Phi(-\delta)}{\varkappa + \Phi(-\delta)} f_0(\bar{x}) + \frac{\varkappa}{\varkappa + \Phi(-\delta)} f_0(x_\delta) \leq \frac{\Phi(-\delta)V}{\varkappa + \Phi(-\delta)} + \frac{\varkappa}{\varkappa + \Phi(-\delta)}[\Phi(-\delta) - \delta].$$

The resulting inequality implies $(\Phi(-\delta) - \Phi(0))/\delta = \Phi(-\delta)/\delta \geq \varkappa/(\varkappa + V)$; when passing to the limit as $\delta \to +0$, we get $\vartheta \geq \varkappa/(V + \varkappa)$.

A.6 *Proof of Proposition* 4

Let us fix parameters N, s, λ, μ satisfying the assumptions of the proposition, let ϵ, δ be associated with these parameters according to (11). We denote

$$\begin{aligned}\bar{f}_{0,N}(x, \xi^N) &= f_{0,N}(x, \xi^N) - \mu M_1 N^{-1/2},\\ \bar{f}_{i,N}(x, \xi^N) &= f_{i,N}(x, \xi^N) - \mu M_1 N^{-1/2}, \quad 1 \leq i \leq m,\end{aligned}$$

and set

$$\bar{\Phi}_N(r, \xi^N) = \min_{x\in X} \max[\bar{f}_{0,N}(x, \xi^N) - r, \bar{f}_{1,N}(x, \xi^N), \ldots, \bar{f}_{m,N}(x, \xi^N)].$$

Then $\bar{\Phi}_N(r, \xi^N)$ is a convex nonincreasing function of $r \in \mathbf{R}$ such that

$$\mathrm{Opt}_N(\xi^N) = \min\{r : \bar{\Phi}_N(r, \xi^N) \leq 0\}.$$

Finally, let $\bar{r}$ be the smallest r such that $\Phi(r) \leq \epsilon$. Since (9) is feasible and $\Phi(r) \to \infty$ as $r \to -\infty$, $\bar{r}$ is a well-defined real which is $<$ Opt (since Opt is the smallest root of Φ) and satisfies $\Phi(\bar{r}) = \epsilon$.

Let us set

$$\hat{\Xi} = \underbrace{\{\xi^N : \bar{\Phi}_N(\mathrm{Opt}, \xi^N) \leq 0\}}_{\Xi_1} \cap \underbrace{\{\xi^N : \bar{\Phi}_N(\bar{r}, \xi^N) > 0\}}_{\Xi_2}.$$

Since $\bar{\Phi}_N(r, \xi^N)$ is a nonincreasing function of r and $\mathrm{Opt}_N(\xi^N)$ is the smallest root of $\bar{\Phi}_N(\cdot, \xi^N)$, for $\xi^N \in \hat{\Xi}$ we have $\bar{r} \leq \mathrm{Opt}_N(\xi^N) \leq \mathrm{Opt}$. The left inequality here implies that $\Phi(\mathrm{Opt}_N(\xi^N)) \leq \epsilon$ (recall that Φ is nonincreasing and $\Phi(\bar{r}) = \epsilon$). The bottom line is that when $\xi^N \in \hat{\Xi}$, $\mathrm{Opt}_N(\xi^N)$ ϵ-underestimates Opt. Consequently, all we need to prove is that $\xi^N \notin \hat{\Xi}$ with probability at most δ.

1^0. Let x_* be an optimal solution to (9). Same as in the proof of Proposition 1, for every i, $0 \le i \le m$, we have (see (A10))

$$\mathrm{Prob}\{f_{i,N}(x_*,\xi^N) > f_i(x_*) + \mu M_1 N^{-1/2}\} \le \mathrm{e}^{-\mu^2/4\tau_*},$$

whence for the event

$$\Xi' = \{\xi^N : f_{i,N}(x_*,\xi^N) \le f_i(x_*) + \mu M_1 N^{-1/2},\ 0 \le i \le m\}$$

it holds

$$\mathrm{Prob}\{\xi^N \notin \Xi'\} \le (m+1)\mathrm{e}^{-\mu^2/4\tau_*}. \tag{A17}$$

By the origin of x_* we have $f_0(x_*) \le \mathrm{Opt}$ and $f_i(x_*) \le 0$, $1 \le i \le m$. Therefore, for $\xi^N \in \Xi'$ it holds $\bar{f}_{0,N}(x_*,\xi^N) \le \mathrm{Opt}$ and $\bar{f}_{i,N}(x_*,\xi^N) \le 0$, $1 \le i \le m$, that is,

$$\bar{\Phi}_N(\mathrm{Opt},\xi^N) \le \max[\bar{f}_{0,N}(x_*,\xi^N) - \mathrm{Opt}, \bar{f}_{1,N}(x_*,\xi^N),\ldots,\bar{f}_{m,N}(x_*,\xi^N)] \le 0,$$

implying that $\xi^N \in \Xi_1$. We conclude that $\Xi' \subset \Xi_1$, and, by (A17),

$$\mathrm{Prob}\{\xi^N \notin \Xi_1\} \le (m+1)\mathrm{e}^{-\mu^2/4\tau_*}. \tag{A18}$$

2^0. We have $\epsilon = \Phi(\bar{r}) = \min_{x\in X} \max[f_0(x) - \bar{r}, f_1(x),\ldots,f_m(x)]$, whence by von Neumann's Lemma there exist nonnegative $y_i \ge 0$, $0 \le i \le m$, summing up to 1, such that

$$\begin{aligned}\epsilon &= \min_{x\in X}[\ell(x) := y_0(f_0(x) - \bar{r}) + \sum_{i=1}^m y_i f_i(x)]\\ &= \min_{x\in X}\left[\int_\Xi \underbrace{\left[y_0[F_0(x,\xi) - \bar{r}] + \textstyle\sum_{i=1}^m y_i F_i(x,\xi)\right]}_{\mathcal{L}(x,\xi)} P(\mathrm{d}\xi)\right].\end{aligned}$$

Under the assumptions of the proposition, the integrand F satisfies all assumptions of Proposition 1. Setting

$$\ell_N(x,\xi^N) = \frac{1}{N}\sum_{i=1}^N \mathcal{L}(x,\xi_i)$$

and applying Proposition 1 we get

$$\begin{aligned}&\mathrm{Prob}\left\{\xi^N : \min_{x\in X} \ell_N(x,\xi^N) < \underbrace{\min_{x\in X}\ell(x)}_{=\epsilon} - [\mu M_1 + [\Omega[1+s^2] + 2\lambda]M_2 R]N^{-1/2}\right\}\\ &\quad\le \mathrm{e}^{-N(s^2-1)} + \mathrm{e}^{-\mu^2/4\tau_*} + \mathrm{e}^{-\lambda^2/4\tau_*}.\end{aligned}$$

Now, in view of

$$\ell_N(x,\xi^N) = \underbrace{\lambda_0[\bar{f}_{0,N}(x,\xi^N) - \bar{r}] + \sum_{i=1}^m \lambda_i \bar{f}_{i,N}(x,\xi^N)}_{\bar{\ell}_N(x,\xi^N)} + \mu M_1 N^{-1/2},$$

and due to the evident relation $\min_{x\in X} \bar{\ell}_N(x,\xi^N) \le \bar{\Phi}_N(\bar{r},\xi^N)$, we get

$$\begin{aligned}&\mathrm{Prob}\{\bar{\Phi}_N(\bar{r},\xi^N) < \epsilon - [\mu M_1 + [\Omega[1+s^2]+2\lambda]M_2R]N^{-1/2} - \mu M_1 N^{-1/2}\}\\ &\quad\le \mathrm{Prob}\left\{\min_{x\in X}\ell_N(x,\xi^N) < \epsilon - 2N^{-1/2}\left[\mu M_1 + M_2 R\left[\frac{\Omega}{2}[1+s^2]+\lambda\right]\right]\right\}\\ &\quad\le \mathrm{e}^{-\mu^2/4\tau_*} + \mathrm{e}^{-N(s^2-1)} + \mathrm{e}^{-\lambda^2/4\tau_*}.\end{aligned}$$

By (11), we have

$$\epsilon - 2\left[\mu M_1 + M_2 R\left[\frac{\Omega}{2}[1+s^2]+\lambda\right]\right]N^{-1/2} > 0,$$

and we arrive at

$$\mathrm{Prob}\{\xi^N \notin \Xi_2\} = \mathrm{Prob}\{\bar{\Phi}_N(\bar{r},\xi^N) \le 0\} \le \mathrm{e}^{-\mu^2/4\tau_*} + \mathrm{e}^{-N(s^2-1)} + \mathrm{e}^{-\lambda^2/4\tau_*}.$$

The latter bound combines with (A18) to imply the desired relation

$$\mathrm{Prob}\{\xi^N \notin \Xi\} \le \mathrm{e}^{-\mu^2/4\tau_*} + \mathrm{e}^{-N(s^2-1)} + \mathrm{e}^{-\lambda^2/4\tau_*} + (m+1)\mathrm{e}^{-\mu^2/4\tau_*} = \beta.$$

Appendix 2. Evaluating approximation parameters

For the sake of completeness, we provide here the straightforward derivations of the parameter estimates used to build the bounds in the numerical section.

A.7 *Notation*

Let P be a Borel probability distribution on $\mathbf{R}^k$ and let Ξ be the support of P. Consider the space $\mathcal{C}$ of all Borel functions $g(\cdot):\Xi\to\mathbf{R}$ such that $\mathbf{E}_{\xi\sim P}\{\exp\{g^2(\xi)/M^2\}\}<\infty$ for some $M=M(g)$. For $g\in\mathcal{C}$, we set

$$\pi[g]=\inf\{M\geq 0:\ \mathbf{E}_{\xi\sim P}\{\exp\{g^2(\xi)/M^2\}\}\leq\exp\{1\}\}. \tag{A19}$$

It is well known [29] that $\mathcal{C}$ is a linear subspace in the space of real-valued Borel functions on Ξ and $\pi[\cdot]$ is a semi-norm on this (Orlicz) space. Besides, for a constant $g(\cdot)\equiv a$ we have $\pi[g]=|a|$, and $|g(\cdot)|\leq|h(\cdot)|$ with $h\in\mathcal{C}$ and Borel g implies $g\in\mathcal{C}$ and $\pi[g]\leq\pi[h]$.

Given a convex compact set $X\subset\mathbf{R}^n$, a norm $\|\cdot\|$ on $\mathbf{R}^n$, and a continuously differentiable distance-generating function $\omega(\cdot)$ for the unit ball $B_{\|\cdot\|}$ which is compatible with this norm, let R be the radius of the smallest $\|\cdot\|$-ball containing X. Given a Borel function $F(x,\xi):\mathbf{R}^n\times\Xi\to\mathbf{R}$ which is convex in $x\in\mathbf{R}^n$ and P-summable in ξ for every x, let

$$f(x)=\mathbf{E}\{F(x,\xi)\}:\ X\to\mathbf{R}.$$

We set

$$\begin{aligned}
M_{1,\infty}&=\sup_{x\in X,\xi\in\Xi}|F(x,\xi)-f(x)|,\\
M_{1,\exp}&=\sup_{x\in X}\pi[F(x,\cdot)-f(x)],\\
L(x,\xi)&=\sup_{g\in\partial_x F(x,\xi),h\in\partial f(x)}\|g-h\|_*,\\
M_2&=\sup_{x\in X}\pi[L(x,\cdot)].
\end{aligned}$$

Note that adding to $F(x,\xi)$ a differentiable function g of x: $F(x,\xi)\mapsto F(x,\xi)+g(x)$ does not affect the quantities $M_{1,\infty}$, $M_{1,\exp}$, and M_2.

Our goal is to compute upper bounds on $M_{1,\infty}$, $M_{1,\exp}$, and M_2 in the different settings of Section 3.1.1.

A.8 *Quadratic risk minimization*

In this case

- $X=\{x=[x_1;\dots;x_n]\in\mathbf{R}^n:\ x_1,\dots,x_n\geq 0,\ \sum_{i=1}^n x_i=1\}$,
- Ξ is a part of the unit box $\{\xi=[\xi_1;\dots;\xi_n]\in\mathbf{R}^n:\|\xi\|_\infty\leq 1\}$,
- $F(x,\xi)=\kappa_0\xi^{\mathrm{T}}x+(\kappa_1/2)(\xi^{\mathrm{T}}x)^2$, with $\kappa_1\geq 0$, and $f(x)=\kappa_0\mu^{\mathrm{T}}x+(\kappa_1/2)x^{\mathrm{T}}\mathbf{E}\{\xi\xi^{\mathrm{T}}\}x$, where $\mu=\mathbf{E}\{\xi\}$.

The parameters M_1, M_2, R and Ω of construction can be set according to:

$$M_1\leq 2|\kappa_0|+\frac{\kappa_1}{2},\quad M_2=2|\kappa_0|+\kappa_1,\quad R=1,\quad \Omega=\begin{cases}1, & n=1,\\ \sqrt{2}, & n=2,\\ \ln(n)\sqrt{\dfrac{2\mathrm{e}}{1+\ln(n)}}, & n\geq 3.\end{cases} \tag{A20}$$

Indeed, for $\xi\in\Xi$ and $x\in X$, we get

$$|F(x,\xi)-f(x)|\leq|\kappa_0||(\xi-\mu)^{\mathrm{T}}x|+\frac{\kappa_1}{2}|x^{\mathrm{T}}(V-\xi\xi^{\mathrm{T}})x|\leq|\kappa_0|\|\xi-\mu\|_\infty+\frac{\kappa_1}{2}$$

(indeed, since V is positive semidefinite with $\|V\|_\infty\leq 1$ and $\|\xi\|_\infty\leq 1$, we have $|x^{\mathrm{T}}(V-\xi\xi^{\mathrm{T}})x|\leq 1$ for all x such that $\|x\|_1\leq 1$), and

$$M_{1,\exp}\leq M_{1,\infty}\leq|\kappa_0|(1+\|\mu\|_\infty)+\frac{\kappa_1}{2}\leq 2|\kappa_0|+\frac{\kappa_1}{2}.$$

Further, let us equip $\mathbf{R}^n$ with the norm $\|\cdot\|=\|\cdot\|_1$, so that $\|\cdot\|_*=\|\cdot\|_\infty$, and endow the unit ball of the norm with the distance-generating function[5]

$$\omega(x)=\frac{1}{p\gamma}\sum_{i=1}^n|x_i|^p,\quad p=\begin{cases}2 & \text{for } n\leq 2,\\ 1+1/\ln(n) & \text{for } n\geq 3,\end{cases},\quad \gamma=\begin{cases}1, & n\leq 1\\ \dfrac{1}{2}, & n=2,\\ \dfrac{1}{\mathrm{e}\ln(n)}, & n\geq 3\end{cases} \tag{A21}$$

resulting in $\Omega = \sqrt{2/p\gamma}$ and $R = 1$. Now let $x \in X$ and $\xi \in \Xi$, and let g be a subgradient of $F(x,\xi)$ with respect to x, and h be a subgradient of f at x. We have

$$g = \kappa_0\xi + \kappa_1\xi(\xi^{\mathrm{T}}x), \quad h = \kappa_0\mu + \kappa_1 Vx,$$

thus

$$\|g-h\|_* \le |\kappa_0|\|\xi-\mu\|_\infty + \kappa_1\|V-\xi\xi^\top\|_\infty \le |\kappa_0|(1+\|\mu\|_\infty) + 2\kappa_1.$$

We conclude that

$$M_2 \le |\kappa_0|(1+\|\mu\|_\infty) + 2\kappa_1 \le 2|\kappa_0| + 2\kappa_1.$$

A.9 *Gaussian VaR optimization*

Here the situation is as follows:

- $X = \{x = [x_1;\ldots;x_n] \in \mathbf{R}^n : x_1,\ldots,x_n \ge 0,\ \sum_{i=1}^n x_i = 1\}$,
- $\xi \sim \mathcal{N}(0,\Sigma)$ on $\mathbf{R}^n$, $\Sigma \succ 0$,
- $F(x,\xi) = \kappa_0\xi^{\mathrm{T}}x + \kappa_1|\xi^{\mathrm{T}}x|$, with $\kappa_1 \ge 0$.

We have $f(x) = \sqrt{(2/\pi)}\kappa_1\sigma_x$, with $\sigma_x = \sqrt{x^{\mathrm{T}}\Sigma x}$. In this case one can set Ω and R as in (A20), along with

$$M_1 = \left[\sqrt{\frac{2\mathrm{e}^2}{\mathrm{e}^2-1}}|\kappa_0| + \sqrt{2}\kappa_1\right]\sigma_{\max},$$

$$M_2 = (|\kappa_0|+\kappa_1)\sigma_{\max}\sqrt{2(2+\ln n)} + \kappa_1\sigma_{\max}\sqrt{\frac{2}{\pi}},$$

where $\sigma^2_{\max} = \max_{1\le i\le n}\Sigma_{i,i}$.

Indeed, we have $\xi^{\mathrm{T}}x \sim \mathcal{N}(0,\sigma_x^2)$, we conclude that $f(x) = \kappa_1\sqrt{2/\pi}\sigma_x$, whence

$$|F(x,\xi) - f(x)| \le |\kappa_0||\xi^{\mathrm{T}}x| + \kappa_1||\xi^{\mathrm{T}}x| - \sqrt{2/\pi}\sigma_x| = \sigma_x[|\kappa_0|\,|\eta_x| + \kappa_1||\eta_x| - \sqrt{2/\pi}|]$$

where $\eta_x = \xi^{\mathrm{T}}x/\sigma_x \sim \mathcal{N}(0,1)$. By direct computation we get

$$\pi[|\eta_x|] = \nu := \sqrt{\frac{2\mathrm{e}^2}{\mathrm{e}^2-1}} = 1.52\ldots$$

Next, setting $\vartheta = \sqrt{2/\pi}$ we observe that

$$\begin{aligned}\frac{1}{\sqrt{2\pi}}\int \exp\{||s|-\vartheta|^2/2 - s^2/2\}\,\mathrm{d}s &= \vartheta\int_0^\infty \exp\{[s^2 - 2\vartheta s + \vartheta^2 - s^2]/2\}\,\mathrm{d}s\\ &= \vartheta\int_0^\infty \exp\{\vartheta^2/2 - \vartheta s\}\,\mathrm{d}s = \exp\{\vartheta^2/2\} < \exp\{1\},\end{aligned}$$

implying that

$$\pi[||\eta_x| - \sqrt{2/\pi}|] \le \sqrt{2}.$$

As a result,

$$\pi[F(x,\cdot) - f(x)] \le \sigma_x[|\kappa_0|\pi[|\eta_x|] + \kappa_1\pi[||\eta_x| - \sqrt{2/\pi}|]] \le \sigma_x[\nu|\kappa_0| + \sqrt{2}\kappa_1].$$

Taking into account that for all $x \in X$ $\sigma_x^2 = x^{\mathrm{T}}\Sigma x \le \|\Sigma\|_\infty$, we arrive at

$$M_{1,\exp} \le [\nu|\kappa_0| + \sqrt{2}\kappa_1]\sqrt{\|\Sigma\|_\infty} = [\nu|\kappa_0| + \sqrt{2}\kappa_1]\sigma_{\max}. \tag{A22}$$

Let $x \in X$, and let g be a subgradient with respect to x of $F(x,\xi)$, and h be a subgradient of $f(x)$. We have

$$g = \kappa_0\xi + \kappa_1\xi\chi,$$

with $\chi = \chi(x,\xi) \in [-1,1]$, so that

$$\|g\|_\infty \le [|\kappa_0| + \kappa_1]\|\xi\|_\infty.$$

Note that

$$\partial[\sqrt{x^{\mathrm{T}}\Sigma x}] = \begin{cases}\{(x^{\mathrm{T}}\Sigma x)^{-1/2}\Sigma x\}, & x \ne 0,\\ \{\Sigma^{1/2}u,\ \|u\|_2 \le 1\}, & x = 0.\end{cases}$$

Therefore, for all $h \in \partial f(x)$ one has

$$\|h\|_\infty \le \kappa_1\sqrt{\frac{2}{\pi}}\sup_{x\ne 0}\frac{\|\Sigma x\|_\infty}{\sqrt{x^{\mathrm{T}}\Sigma x}} = \kappa_1\sqrt{\frac{2}{\pi}}\sup_{y\ne 0}\frac{\|\Sigma^{1/2}y\|_\infty}{\|y\|_2} = \kappa_1\sqrt{\frac{2}{\pi}}\max_{1\le i\le n}\|\Sigma_i^{1/2}\|_2 \le \kappa_1\sigma_{\max}\sqrt{\frac{2}{\pi}}$$

(here $\Sigma_i^{1/2}$ stands for the ith row of $\Sigma^{1/2}$), and

$$\|g-h\|_* = \|g-h\|_\infty \le [|\kappa_0|+\kappa_1]\|\xi\|_\infty + \kappa_1\sigma_{\max}\sqrt{\frac{2}{\pi}},$$

that is,

$$L(x,\xi) \le [|\kappa_0|+\kappa_1]\|\xi\|_\infty + \kappa_1\sigma_{\max}\sqrt{\frac{2}{\pi}}.$$

We conclude that

$$\pi[L(x,\cdot)] \le [|\kappa_0|+\kappa_1]\pi[\|\xi\|_\infty] + \kappa_1\sigma_{\max}\sqrt{\frac{2}{\pi}}. \tag{A23}$$

We now use the following simple result.[6]

LEMMA A.2 *Let ξ be a zero-mean Gaussian random vector in $\mathbf{R}^n$, and let $\bar\sigma^2 \ge \max_{1\le i\le n}\mathbf{E}\{\xi_i^2\}$. Then for $M \ge \bar\sigma\sqrt{2(2+\ln n)}$*

$$\mathbf{E}\{\mathrm{e}^{\|\xi\|_\infty^2/M^2}\} \le \mathrm{e}.$$

Proof Let $\eta_n = \max_{1\le i\le n}|\xi_i|$. We have the following well-known fact:

$$\psi_n(r) := \mathrm{Prob}\{\eta_n \ge r\} \le \min\{1, n\mathrm{e}^{-r^2/2\bar\sigma^2}\}.$$

Therefore, for $|t| < (\sqrt{2}\bar\sigma)^{-1}$,

$$\begin{aligned}\mathbf{E}\{\mathrm{e}^{t^2\eta_n^2}\} &= -\int_0^\infty \mathrm{e}^{t^2r^2}\,d\psi_n(r) = 1+\int_0^\infty 2t^2r\mathrm{e}^{t^2r^2}\psi_n(r)\,\mathrm{d}r\\ &\le \mathrm{e}^{2t^2\bar\sigma^2\ln n} + 2nt^2\int_{\bar\sigma\sqrt{2\ln n}}^\infty r\exp\left\{-\frac{(1-2t^2\bar\sigma^2)r^2}{2\bar\sigma^2}\right\}\mathrm{d}r\\ &= \mathrm{e}^{2t^2\bar\sigma^2\ln n} + \frac{2t^2\bar\sigma^2}{1-2t^2\bar\sigma^2}\mathrm{e}^{2t^2\bar\sigma^2\ln n} = \frac{n^{2t^2\bar\sigma^2}}{1-2t^2\bar\sigma^2}.\end{aligned}$$

Note that $\mathrm{e}^{-x} \le 1-x/2$ for $0\le x\le 1$. Thus for all $n\ge 1$ and $t \le (\bar\sigma\sqrt{2(2+\ln n)})^{-1}$,

$$\frac{n^{2t^2\bar\sigma^2}}{1-2t^2\bar\sigma^2} \le \frac{\mathrm{e}^{\ln n/(2+\ln n)}}{1-\frac{1}{2+\ln n}} = \mathrm{e}^{1-2/(2+\ln n)}\frac{2+\ln n}{1+\ln n} \le \mathrm{e}.$$

Finally, using the result of the lemma we conclude from (A23) that one can take for M_2 the expression

$$(|\kappa_0|+\kappa_1)\sigma_{\max}\sqrt{2(2+\ln n)} + \kappa_1\sigma_{\max}\sqrt{\frac{2}{\pi}}. \tag{A24}$$

■

A.10 *CVaR optimization*

Consider the portfolio problem of Section 3.1.3. With some terminology abuse, in what follows, we refer to the special case $n=1$ with $x_1\equiv 1$ as to the case of $n=0$.

- $X = \{x=[x_0;x_1;\ldots;x_n]\in\mathbf{R}^{n+1}:\ |x_0|\le 1,\ x_1,\ldots,x_n\ge 0,\ \sum_{i=1}^n x_i = 1\}$,
- Ξ be a part of the unit box $\{\xi=[\xi_1;\ldots;\xi_n]\in\mathbf{R}^n : \|\xi\|_\infty\le 1\}$,
- $F(x,\xi) = \kappa_0\sum_{i=1}^n\xi_ix_i + \kappa_1[x_0+(1/\epsilon)[\sum_{i=1}^n\xi_ix_i - x_0]_+]$, with $\kappa_0,\kappa_1\in[0,1]$.

The parameters M_1, M_2, R, and Ω of construction can be set according to:

$$M_1 = 2\left(\kappa_0+\frac{\kappa_1}{\epsilon}\right),\quad M_2 = \begin{cases}\dfrac{\kappa_1}{\epsilon}, & n=0,\\ \sqrt{\left(\dfrac{\kappa_1}{\epsilon}\right)^2+4\left(\kappa_0+\dfrac{\kappa_1}{\epsilon}\right)^2}, & n\ge 1,\end{cases}$$

$$R = \begin{cases}1, & n=0\\ \sqrt{2}, & n\ge 1.\end{cases},\quad \Omega = \begin{cases}1, & n=0,\\ \sqrt{2}, & n=1,\\ \sqrt{3}, & n=2,\\ \sqrt{1+\dfrac{2e(\ln(n))^2}{1+\ln(n)}}, & n\ge 3.\end{cases}$$

Indeed, denoting $\xi_x = \xi^{\mathrm{T}} x$ and $\mu_i = \mathbf{E}\{\xi_i\}$, we have

$$f(x) = \kappa_0 \sum_{i=1}^{n} \mu_i x_i + \kappa_1 \left[x_0 + \frac{1}{\epsilon} \mathbf{E}\{[\xi_x - x_0]_+\} \right],$$

whence for $\xi \in \Xi$ and $x \in X$

$$|F(x,\xi) - f(x)| \le \kappa_0 \left| \sum_{i=1}^{n} [\xi_i - \mu_i] x_i \right| + \frac{\kappa_1}{\epsilon} |[\xi_x - x_0]_+ - \mathbf{E}\{[\xi_x - x_0]_+\}|.$$

We have $|\xi_x| \le 1$, whence $0 \le [\xi_x - x_0]_+ \le 1 + [-x_0]_+$ and $0 \le \mathbf{E}\{[\xi_x - x_0]_+\} \le 1 + [-x_0]_+$. Then,

$$-2 \le -1 - [-x_0]_+ \le [\xi_x - x_0]_+ - \mathbf{E}\{[\xi_x - x_0]_+\} \le 1 + [-x_0]_+,$$

so that

$$|[\xi_x - x_0]_+ - \mathbf{E}\{[\xi_x - x_0]_+\}| \le 2.$$

We conclude that

$$M_{1,\exp} \le M_{1,\infty} \le \kappa_0 (1 + \|\mu\|_\infty) + \frac{2\kappa_1}{\epsilon} \le 2 \left(\kappa_0 + \frac{\kappa_1}{\epsilon} \right). \tag{A25}$$

In what follows, for a vector from $\mathbf{R}^{n+1}$, say, $z = [z_0; z_1; \ldots; z_n]$, we set $z' = [z_1; \ldots; z_n]$, so that $z = [z_0; z']$. Let us define norm $\|\cdot\|$ on $\mathbf{R}^{n+1}$ as

$$\|[x_0; x']\| = \sqrt{x_0^2 + \|x'\|_1^2},$$

implying that

$$\|[x_0; x']\|_* = \sqrt{x_0^2 + \|x'\|_\infty^2}.$$

A distance-generating function $\omega([x_0; x'])$ for the unit ball of the norm $\|\cdot\|$ can be taken as

$$\omega([x_0; x']) = \frac{1}{2} x_0^2 + \frac{1}{p\gamma} \sum_{i=1}^{n} |x_i|^p,\ p = \begin{cases} 2, & n \le 2 \\ 1 + 1/\ln(n), & n \ge 3 \end{cases},\quad \gamma = \begin{cases} 1, & n \le 1, \\ \dfrac{1}{2}, & n = 2, \\ \dfrac{1}{\mathrm{e}\ln(n)}, & n \ge 3, \end{cases}$$

resulting in

$$\Omega = \begin{cases} 1, & n = 0 \\ \sqrt{1 + \dfrac{2}{p\gamma}}, & n \ge 1 \end{cases} \quad \text{and} \quad R = \begin{cases} 1, & n = 0, \\ \sqrt{2}, & n \ge 1. \end{cases} \tag{A26}$$

Let $x \in X$ and $\xi \in \Xi$, and let $g = [g_0; g']$ be a subgradient of $F(x,\xi)$ with respect to x, and h be a subgradient of f at x. We clearly have

$$g_0 = \kappa_1 - \frac{\kappa_1}{\epsilon} \chi_0, \quad g' = \kappa_0 \xi + \frac{\kappa_1}{\epsilon} \xi \chi_1, \quad h_0 = \kappa_1 - \frac{\kappa_1}{\epsilon} \chi_2,$$

where $\chi_i \in [0,1]$. Next, for $n \ge 2$,

$$\begin{aligned} |f([x_0; x']) - f([x_0; y'])| &= |\kappa_0 \mu^{\mathrm{T}} (x' - y') + \frac{\kappa_1}{\varepsilon} (\mathbf{E}\{[\xi^{\mathrm{T}} x' - x_0]_+\} - \mathbf{E}\{[\xi^{\mathrm{T}} y' - x_0]_+\})| \\ &\le \kappa_0 \|\mu\|_\infty \|x' - y'\|_1 + \frac{\kappa_1}{\varepsilon} \mathbf{E}\{|\xi^{\mathrm{T}} (x' - y')|\} \\ &\le (\kappa_0 + \frac{\kappa_1}{\varepsilon}) \|x' - y'\|_1. \end{aligned}$$

It follows that $f([x_0; x'])$ is Lipschitz continuous in x' with constant $\kappa_0 + \kappa_1/\varepsilon$ with respect to $\|\cdot\|_1$ and we have $\|h'\|_\infty \le \kappa_0 + \kappa_1/\epsilon$. As a result, we obtain for $n \ge 2$

$$\|g - h\|_* = \sqrt{|g_0 - h_0|^2 + \|g' - h'\|_\infty^2} \le \sqrt{\left(\frac{\kappa_1}{\epsilon}\right)^2 + 4\left(\kappa_0 + \frac{\kappa_1}{\epsilon}\right)^2},$$

while $\|g - h\|_* \le \kappa_1/\epsilon$ for $n = 1$.

We conclude that

$$M_2 = \begin{cases} \dfrac{\kappa_1}{\epsilon}, & n = 0, \\ \sqrt{\left(\dfrac{\kappa_1}{\epsilon}\right)^2 + 4\left(\kappa_0 + \dfrac{\kappa_1}{\epsilon}\right)^2}, & n \ge 1. \end{cases} \tag{A27}$$