4
Two applications of SAA

In this Chapter we apply SAA to a portfolio problem and to a blending problem. In the first the decision maker must choose the composition of a portfolio of assets such that the expected return is maximized. Due to the chance constraint, the total gain has to be greater than a pre-specified return level v with high probability. When returns follow a multivariate normal distribution, we compute the solution explicitly and compare it with the results of SAA. When returns are lognormally distributed, we have to rely on approximations.

The second problem is a joint version of a two dimensional blending problem. We show that SAA can be readily applied to this situation at no extra cost. Due to the independence assumption, we compute explicit answers for this problem and use them as benchmarks to tune the parameters of SAA.

4.1
Portfolio problem

Consider the following maximization problem subject to a single chance constraint:

$$\begin{array}{ll} \underset{x \in X}{\text{Max}} & \mathbb{E}\left[r^T x\right] \\ \text{s.t.} & \mathsf{Prob}\left\{r^T x \geq v\right\} \geq 1 - \varepsilon. \end{array} \tag{4-1}$$

Here $x \in \mathbb{R}^n$ is vector of decision variables, $r \in \mathbb{R}^n$ is a random vector with known probability distribution, $v \in \mathbb{R}$, $\varepsilon \in (0, 1)$, e is a vector whose components are all equal to 1 and

$$X := \{x \in \mathbb{R}^n \,:\, e^T x = 1,\, x \geq 0\}.$$

Note that, of course, $\mathbb{E}\left[r^T x\right] = \mu^T x$, where $\mu := \mathbb{E}[r]$ is the corresponding mean vector. That is, the objective function of problem (4-1) is linear and deterministic.

The motivation to study (4-1) is the portfolio selection problem going back to Markowitz [MAR]. The vector x represents the percentage of a total wealth of one dollar invested in each of n available assets, r is the vector of random returns of these assets and the decision agent wants to maximize the

mean return subject to having a return greater or equal to a desired level v, with probability at least $1-\varepsilon$. In terms of risk measures, this requirement is equivalent to a Value-at-Risk constraint. We note that problem (4-1) is not realistic because it does not incorporate crucial features of real markets such as cost of transactions, short sales, lower and upper bounds on the holdings, etc. However, it will serve to our purposes as an *example* of an application of the SAA method. For a more realistic model we can refer the reader, e.g., to [WCZ], where the authors include market frictions and discuss the best distribution function for asset returns.

A very similar version of problem (4-1) was analyzed in [ZTSK], where the authors obtained important information about the different policies available in soil management as well as the trade off between net returns and soil loss.

We consider two different situations, namely when the vector of random returns r follows multivariate normal and multivariate lognormal distributions. The two cases are very distinct; on the former one can solve explicitly the chance constraint, while in the latter no explicit formula is known. Under normality, we can compare the quality of the approximations with the true optimal value, while in the lognormal case we have to rely on approximations.

4.1.1
SAA of the portfolio problem

First assume that r follows a multivariate normal distribution with mean vector μ and covariance matrix Σ, written $r \sim \mathcal{N}(\mu, \Sigma)$. In that case $r^T x \sim \mathcal{N}\left(\mu^T x, x^T \Sigma\, x\right)$, and hence (as it is well known) the chance constraint in (4-1) can be written as a convex second order conic constraint (SOCC) as follows.

$$\begin{aligned}
\mathsf{Prob}\left\{r^T x \geq v\right\} &\geq 1-\varepsilon \Leftrightarrow \\
\mathsf{Prob}\left\{\frac{r^T x - \mu^T x}{\sqrt{x^T \Sigma x}} \geq \frac{v - \mu^T x}{\sqrt{x^T \Sigma x}}\right\} &\geq 1-\varepsilon \Leftrightarrow \\
1 - \mathsf{Prob}\left\{\frac{r^T x - \mu^T x}{\sqrt{x^T \Sigma x}} \leq \frac{v - \mu^T x}{\sqrt{x^T \Sigma x}}\right\} &\geq 1-\varepsilon \Leftrightarrow \\
1 - \Phi\left(\frac{v - \mu^T x}{\sqrt{x^T \Sigma x}}\right) &\geq 1-\varepsilon \Leftrightarrow \\
\frac{v - \mu^T x}{\sqrt{x^T \Sigma x}} &\leq z_{1-\varepsilon} \Leftrightarrow \\
v - \mu^T x + z_{1-\varepsilon}\sqrt{x^T \Sigma x} &\leq 0. \qquad (4\text{-}2)
\end{aligned}$$

Using the explicit form (4-2) of the chance constraint, one can efficiently solve the convex problem (4-1) for different values of ε. An *efficient frontier* of

portfolios can be constructed in an objective function value versus confidence level plot, that is, for every confidence level ε we associate the optimal value of problem (4-1). The efficient frontier dates back to Markowitz [MAR]. A discussion of the subject can be found, e.g., in [CM].

If r follows a multivariate lognormal distribution, then no closed form solution for the chance constrained problem (4-1) is available. The related SAA problem (4-1) can be written as

$$\begin{array}{ll} \underset{x \in X}{\text{Max}} & \mu^T x \\ \text{s.t.} & \hat{p}_N(x) \leq \gamma, \end{array} \tag{4-3}$$

where $\hat{p}_N(x) := N^{-1} \sum_{i=1}^{N} \mathbb{1}_{(0,\infty)}(v - r_i^T x)$ and $\gamma \in [0, 1)$. The reason we use γ instead of ε is to suggest that for a fixed ε, a different choice of the parameter γ in (4-3) might be suitable. For instance, if $\gamma = 0$, then the SAA problem (4-3) becomes the linear program

$$\begin{array}{ll} \underset{x \in X}{\text{Max}} & \mu^T x \\ \text{s.t.} & r_i^T x \geq v, \;\; i = 1, \ldots, N. \end{array} \tag{4-4}$$

A recent paper by Campi and Garatti [CG], building on the work of Calafiore and Campi [CC], provides an expression for the probability of an optimal solution $\hat{x}_N$ of the SAA problem (3-3), with $\gamma = 0$, to be infeasible for the true problem (3-2). That is, under the assumptions that the set X and functions $f(\cdot)$ and $G(\cdot, \xi)$, $\xi \in \Xi$, are convex and that w.p.1 the SAA problem attains unique optimal solution, we have that for $N \geq n$,

$$\mathsf{Prob}\left\{p(\hat{x}_N) > \varepsilon\right\} \leq B(n-1; \varepsilon, N), \tag{4-5}$$

and the above bound is tight. We apply this bound to the considered portfolio selection problem to conclude that for a confidence parameter $\beta \in (0, 1)$ and a sample size N^* such that

$$B(n-1; \varepsilon, N^*) \leq \beta, \tag{4-6}$$

the optimal solution of problem (4-4) is feasible for the corresponding true problem (4-1) with probability at least $1 - \beta$.

For $\gamma > 0$, problem (4-3) can be written as the mixed-integer linear program

$$\begin{array}{ll} \underset{x,z}{\text{Max}} & \mu^T x \\ \text{s.t.} & r_i^T x + v z_i \geq v, \\ & \sum_{i=1}^{N} z_i \leq N\gamma, \\ & x \in X, \; z \in \{0,1\}^N, \end{array} \tag{4-7}$$

The equivalence of problems (4-3) and (4-7) was already proved when we

showed the equivalence of problems (2-8) and (2-9) in Chapter 2, Section 2.3.

Given a fixed ε in (4-1), it is not clear what are the best choices of γ and N for approximation (4-7). We believe it is problem dependent and numerical investigations will be performed with different values for both parameters. We know from Proposition 6 that, for $\gamma = \varepsilon$ the larger the N the closer we are to the original problem (4-1). However, the number of samples N must be chosen carefully because problem (4-7) is a binary problem. Even moderate values of N can generate instances that are very hard to solve.

4.1.2 Obtaining candidate solutions

First we perform numerical experiments applying SAA to the portfolio problem (4-1) assuming that $r \sim \mathcal{N}(\mu, \Sigma)$. We considered 10 assets ($n = 10$) and the data for the estimation of the parameters was taken from historical monthly returns adjusted for dividends from 1997 to 2007 of 10 US major companies[1]. The sample was generated by the *Triangular Factorization Method* [BS]. We wrote the codes in GAMS and solved the linear and binary problems using CPLEX 9.0. The computer was a PC with an Intel Core 2 processor and 2GB of RAM.

Let us fix $\varepsilon = 0.10$ and $\beta = 0.01$. For these values, the sample size suggested by (4-6) is $N^* = 183$. We ran 10 independent replications of (4-4) for each of the sample sizes $N = 30, 40, \ldots, 200$ and for $N^* = 183$. We also build an *efficient frontier* plot of optimal portfolios with an objective value versus $\mathsf{Prob}\{r^T x_\varepsilon \geq v\}$ axes, where x_ε is the optimal solution of problem (4-1) for a given ε. We show in the same plot (Figure 4.1) the corresponding objective function values and $\mathsf{Prob}\{r^T \hat{x}_N \geq v\}$ for each optimal solution $\hat{x}_N$ found for the problem (4-4). To identify each point with a sample size, we used a gray scale that attributes light tones of gray to smaller sample sizes and darker ones to larger samples. The efficient frontier curve is calculated for $\varepsilon = 0.8, 0.81, \ldots, 0.99$ and then connected by lines. The vertical and horizontal lines are for reference only: they represent the optimal value for problem (4-1) with $\varepsilon = 0.10$ and the 90% reliability level, respectively.

Figure 4.1 shows interesting features of the SAA problem (4-4). Although larger sample sizes always generate feasible points, the value of the objective function, in general, is quite small if compared with the optimal value 1.004311 of problem (4-1) with $\varepsilon = 0.10$. We also observe the absence of a convergence property: if we increase the sample size, the feasible region of problem (4-4)

[1] JP Morgan, Oracle, Intel, Exxon, Wal-Mart, Apple, Sun Microsystems, Microsoft, Yahoo and Procter & Gamble

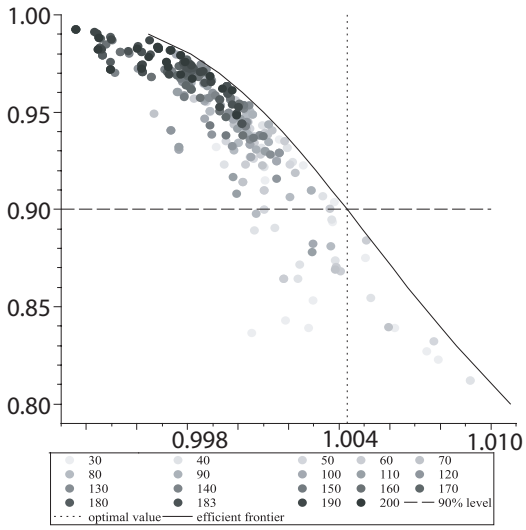

Figure 4.1: Normal returns for $\gamma = 0$.

gets smaller and the approximation becomes more and more conservative and therefore suboptimal. The reason is that for increasingly large samples the condition $r_i^T x \geq v$ for all i approaches the condition $\mathsf{Prob}\{r^T x \geq v\} = 1$.

In order to find better candidate solutions for problem (4-1), we need to solve the SAA problem with $\gamma > 0$, (problem (4-7)), which is a combinatorial problem. Since our portfolio problem is a linear one, we can solve problem (4-3) efficiently for a moderate number (e.g., 200 constraints) of instances. We performed tests for problem (4-3) fixing $\gamma = 0.05$ and 0.10 and changing N as in the sample approximation case. The results are in Figures 4.2 and 4.3.

The best candidate solutions to problem (4-1) were obtained by choosing $\gamma = 0.05$. We considered different sample sizes from 30 to 200. Although several points are infeasible to the original problem (4-1), we observe in Figure 4.2 that whenever a point is feasible it is close to the upper bound. For $\gamma = 0.10$, Figure 4.3 shows us that almost every generated point is infeasible. To further justify this claim, note that among the feasible points in Figure 4.2, more than 70% of them are within 0.2% of the true optimal value 1.004311 of problem (4-1) with $\varepsilon = 0.10$. If we relax the tolerance to 0.3%, then more than 93% of the points are no more than 0.3% away from the optimal value.

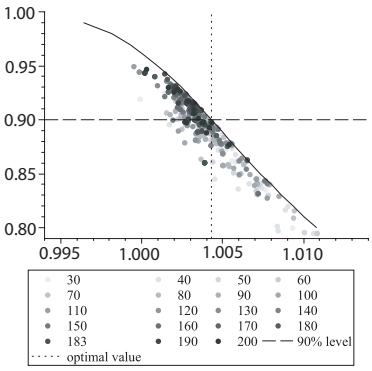

Figure 4.2: Normal with $\gamma = 0.05$.

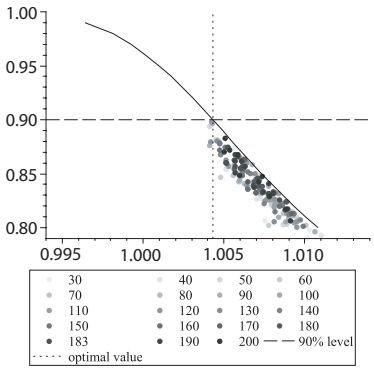

Figure 4.3: Normal with $\gamma = 0.10$.

To investigate the possible choices of γ and N in problem (4-7), we created a three dimensional representation which we call γN-plot. The domain is a discretization of values of γ and N, forming a grid with pairs (γ, N). For each pair we solve an instance of problem (4-7) with these parameters and stored the optimal value and the approximate probability of being feasible to the original problem (4-1). The z-axis represents the optimal value associated to each point in the domain in the grid. Finally, we created a surface of triangles based on this grid as follows. Let i be the index for the values of γ and j for the values of N. If candidate points associated with grid members $(i, j), (i + 1, j)$ and $(i, j + 1)$ or $(i + 1, j + 1), (i + 1, j)$ and $(i, j + 1)$ are feasible to problem (4-1) (with probability greater than or equal to $(1 - \varepsilon)$), then we draw a dark gray triangle connecting the three points in the space. Otherwise, we draw a light gray triangle.

We created a γN-plot for problem (4-1) with normal returns. The result can be seen in Figure 4.4, where we also included the plane corresponding to the optimal solution with $\varepsilon = 0.10$. The values for parameter γ were $0, 0.01, \ldots, 0.10$ and for $N = 30, 40, \ldots, 200$. From Figure 4.4 we see that for any fixed γ small sample sizes tend to generate infeasible solutions and large samples feasible ones. As predicted by the results of Campi and Garatti, when $\gamma = 0$, large sample sizes generate feasible solutions, although they can be seen to be of poor quality judging by the low peaks observed in this region. The concentration of high peaks corresponds to γ values around $\varepsilon/2 = 0.05$ for almost all sample sizes, including small ones (varying from 50 until 120). We generated different instances of Figure 4.4 and the output followed the pattern described here.

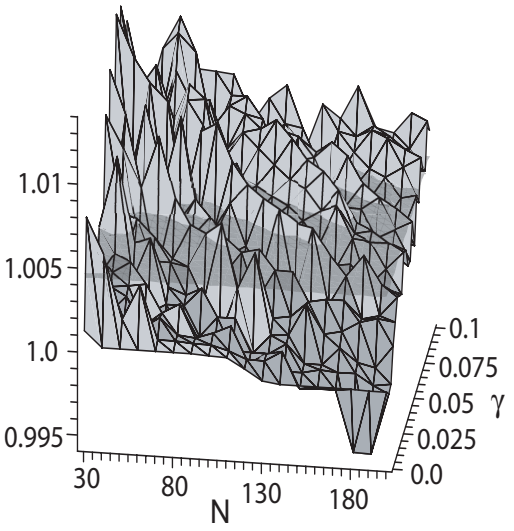

Figure 4.4: γN-plot for the portfolio problem with normal returns.

Even though there are peaks in other regions, Figure 4.4 suggests a strategy to obtain good candidates for chance constrained problems: choose γ close to $\varepsilon/2$, solve instances of SAA problems with small sizes of N (e.g. one third of the Campi-Garatti estimate (4-6)) and keep the best solution. This is fortunate because SAA problems with $\gamma > 0$ are binary problems that can be hard to solve. Our experience with the portfolio problem and with others suggest this strategy works better than trying to solve SAA problems with large sample sizes. The choice $\gamma = \varepsilon/2$ came from our empirical experience. We believe in general the choice of γ is problem dependent.

4.1.3 Upper bounds

A method to compute lower bounds of chance constrained problems of the form (3-1) was suggested in [NS]. We summarized their procedure at the end of Section 3.1, leaving the question of how to choose the constants L, M and N. Given β, M and N, it is straightforward to specify L: it is the largest integer that satisfies (3-12). For a given N, the larger M the better because we are approximating the L-th order statistic of the random variable $\hat{\vartheta}_N$. However, note that M represents the number of problems to be solved and this value is often constrained by computational limitations.

In [NS] an indication of how N should be chosen is not given. It is possible to gain some insight on the magnitude of N by doing some algebra in inequality (3-12). With $\gamma = 0$, the first term ($i = 0$) of the sum (3-12) is

$$\left[1 - (1-\varepsilon)^N\right]^M \approx \left[1 - e^{-N\varepsilon}\right]^M. \tag{4-8}$$

Approximation (4-8) suggests that for small values of ε we should take N of order $O(\varepsilon^{-1})$. If N is much bigger than $1/\varepsilon$ then we would have to choose a very large M in order to honor inequality (3-12). For instance if $\varepsilon = 0.10, \beta = 0.01$ and $N = 100$ instead of $N = 1/\varepsilon = 10$ or $N = 2/\varepsilon = 20$, we need M to be greater than 100 000 in order to satisfy (3-12), which can be computationally intractable for some problems. If $N = 200$ then M has to be grater then 10^9, which is impractical for most applications.

In [LA], the authors applied the same technique to generate bounds on probabilistic versions of the set cover problem and the transportation problem. To construct the bounds they varied N and used $M = 10$ and $L = 1$. For many instances they obtained lower bounds slightly smaller (less than 2%) or even equal to the best optimal values generated by SAA. In the portfolio problem, the choice $L = 1$ generated poor bounds as we will see.

We performed experiments for the portfolio problem with returns now following a lognormal distribution. Figure 4.5 shows the sample points obtained by SAA with $\gamma = 0$ and with the corresponding probability estimated by Monte-Carlo. The reader is referred to [LK] for detailed instructions of how to generate samples from a multivariate lognormal distribution. Since in the lognormal case one cannot compute the efficient frontier, we also included in Figure 4.5 upper bounds[2] for $\varepsilon = 0.02, \ldots, 0.20$, calculated according to (3-12). We fixed $\beta = 0.01$ for all cases and chose three different values for the constants L, M and N.

First we fixed $L = 1$ and $N = \lceil 1/\varepsilon \rceil$ (solid line in Figure 4.5, upper bound A). The constant M was chosen to satisfy the inequality (3-12). The results were not satisfactory, mainly because M ended up being too small. Since the constant M defines the number of samples from $\hat{v}_N$ and since our problem is a linear one, we decided to fix $M = 1\,000$. Then we chose $N = \lceil 1/\varepsilon \rceil$ (dashed line in Figure 4.5, upper bound B) and defined L to be the largest integer such that (3-12) is satisfied. Finally, we generated an upper bound with $M = 1\,000$ and $N = \lceil 2/\varepsilon \rceil$ (dotted line in Figure 4.5, upper bound C).

It is harder to construct upper bounds with $\gamma > 0$. The difficulty lies in solving integer problems and it is hard to find an appropriate choice of the parameters M or N in order to keep the problem size manageable. Based on experiments, a good choice for this problem is $M = 500, N = 50$ and $\gamma = \varepsilon/2 = 0.05$, which originated the dotted upper bound in Figure 4.5. Although it is slightly better than the bounds obtained with $\gamma = 0$, in many situations one often wants an upper bound without much computational effort. If that is the case, it might be appropriate to use equation (3-12) for $\gamma = 0$

[2]The portfolio problem (4-1) is a maximization problem.

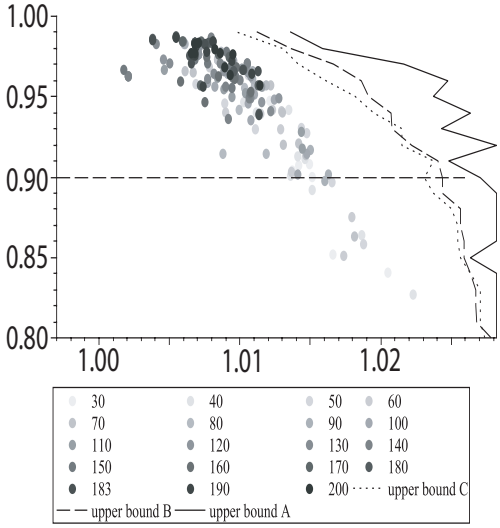

Figure 4.5: Lognormal with $\gamma = 0$.

since the corresponding problems are easier to solve.

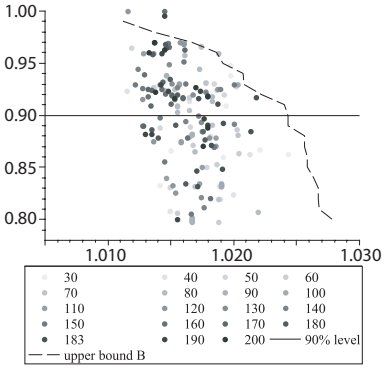

Figure 4.6: Lognormal with $\gamma = 0.05$.

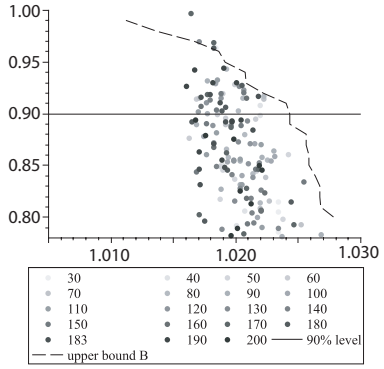

Figure 4.7: Lognormal with $\gamma = 0.10$.

Following the normal case, we performed similar experiments with $\gamma = \varepsilon/2$ and $\gamma = \varepsilon$. The results are in Figures 4.6 and 4.7 respectively, where we only included the dashed upper bound. The experiments for the lognormal case confirmed the tendency observed in the normal case: $\gamma = \varepsilon$ generated infeasible points, $\gamma = 0$ generated feasible points of poor quality if measured by the distance to the upper bound curves and $\gamma = \varepsilon/2$ yielded the best candidate solutions.

4.2 A blending problem

Let us consider a second example of a chance constrained problem. Suppose a farmer has some crop and wants to use fertilizers to increase the

production. He hires an agronomy engineer who recommends $7\,g$ of nutrient A and $4\,g$ of nutrient B. He has two kinds of fertilizers available: the first has $\omega_1\,g$ of nutrient A and $\omega_2\,g$ of nutrient B per kilogram. The second has $1\,g$ of each nutrient per kilogram. The quantities ω_1 and ω_2 are uncertain: we assume they are (independent) continuous uniform random variables with support in the intervals $[1, 4]$ and $[1/3, 1]$ respectively. Furthermore, each fertilizer has a unitary cost per kilogram.

There are several ways to model this blending problem. A detailed discussion can be found in [HV], where the authors use this problem to motivate the field of stochastic programming. We consider a joint chance constrained formulation as follows:

$$\begin{array}{ll} \underset{x_1,x_2}{\text{Min}} & x_1 + x_2 \\ \text{s.t.} & \mathsf{Prob}\{\omega_1 x_1 + x_2 \geq 7, \omega_2 x_1 + x_2 \geq 4\} \geq 1 - \varepsilon, \\ & x_1, x_2 \geq 0, \end{array} \tag{4-9}$$

where x_i represents the quantity of fertilizer i purchased, $i = 1, 2$, and $\varepsilon \in [0, 1]$ is the reliability level. The independence assumption allows us to convert the joint probability in (4-9) into a product of probabilities. After some tedious calculations, one can explicitly solve (4-9) for all values of ε. For $\varepsilon \in [1/2, 1]$ the solution (x_1^*, x_2^*) and the optimal value $v^* = x_1^* + x_2^*$ are

$$x_1^* = \frac{18}{9 + 8(1-\varepsilon)}, \quad x_2^* = \frac{2(9 + 28(1-\varepsilon))}{9 + 8(1-\varepsilon)}, \quad v^* = \frac{4(9 + 14(1-\varepsilon))}{9 + 8(1-\varepsilon)}.$$

For $\varepsilon \in [0, 1/2]$, v^* is

$$v^* = \frac{2(25 - 18(1-\varepsilon))}{11 - 9(1-\varepsilon)}. \tag{4-10}$$

Our goal is to exemplify the use of SAA to joint chance constrained problems. In addition, we use problem (4-9) as a benchmark to gain more understanding of tuning of the underlying parameters of the SAA approach since an explicit solution is available in this case. As mentioned in Section 2.3 of Chapter 2, we can convert a joint chance constrained problem into a problem of the form (3-1) using the min (or max) operators. Problem (4-9) becomes

$$\begin{array}{ll} \underset{x_1,x_2}{\text{Min}} & x_1 + x_2 \\ \text{s.t.} & \mathsf{Prob}\left\{\min\{\omega_1 x_1 + x_2 - 7, \omega_2 x_1 + x_2 - 4\} \geq 0\right\} \geq 1 - \varepsilon, \\ & x_1, x_2 \geq 0. \end{array} \tag{4-11}$$

Introducing one auxiliary variable z_i per scenario, it is possible to formulate the SAA method to problem (4-11) as a mixed integer linear program as follows.

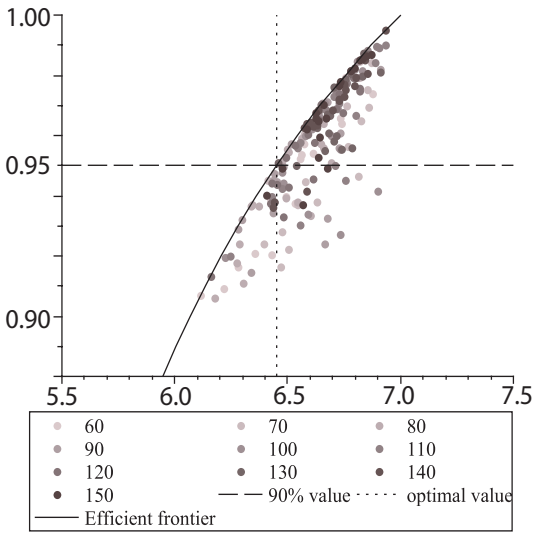

Figure 4.8: SAA for the blending problem with $\gamma = 0.025$.

$$\begin{array}{ll} \underset{x_1,x_2}{\text{Min}} & x_1 + x_2 \\ \text{s.t.} & \omega_1^i x_1 + x_2 - 7 + Kz^i \geq 0, \quad i = 1, \ldots, N, \\ & \omega_2^i x_1 + x_2 - 4 + Kz^i \geq 0, \quad i = 1, \ldots, N, \\ & \sum_{i=1}^{N} z^i \leq N\gamma, \\ & z \in \{0,1\}^N, \\ & x_1, x_2 \geq 0, \end{array} \tag{4-12}$$

where N is the number of samples, ω_1^i and ω_2^i are samples from the random variables ω_1 and ω_2, $\gamma \in (0,1)$ and K is a positive constant greater or equal than 7.

4.2.1 Numerical experiments

We performed experiments similar to the ones for the portfolio problem so we present the results without details. In Figure 4.8 we generated approximations for problem (4-9) with $\varepsilon = 0.05$ using SAA. The sample points were obtained by solving a SAA problem with $\gamma = 0.025$ and sample sizes $N = 60, 70, \ldots, 150$. The Campi-Garatti inequality (4-6) suggested value is $N^* = 130$. In addition, we included the efficient frontier for problem (4-9). We will not show the corresponding Figures for other values of γ, but the pattern observed in the portfolio problem repeated: with $\gamma = 0$ almost every point was feasible but far from the optimal, with $\gamma = \varepsilon = 0.05$ almost every point was infeasible. Again, the parameter choice that generated the best candidate solutions was $\gamma = \varepsilon/2 = 0.025$.

We also show the γN-plot for SAA applied to problem (4-9). We tested γ values in the range $0, 0.005, 0.01, \ldots, 0.05$ and $N = 60, 70, \ldots, 150$. We included

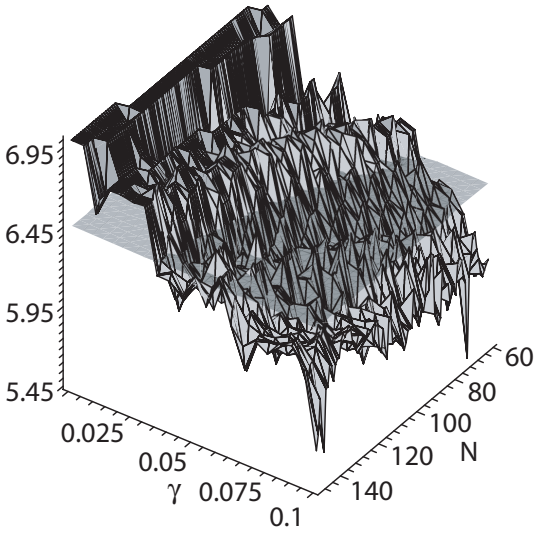

Figure 4.9: γN-plot for the blending problem.

a plane representing the optimal value of problem (4-9) for $\varepsilon = 0.05$, which is readily obtained by applying formula (4-10).

In accordance with Figure 4.4, we note that in Figure 4.9 the best candidate solutions are the ones with γ around 0.025. Even for very small sample sizes we have feasible solutions (dark gray triangles) relatively close to the optimal plane. On the other hand, this experiment gives more evidence that SAA with $\gamma = 0$ is excellent to generate feasible solutions (dark gray triangles) but the quality of the solutions is poor. As shown in Figure 4.9, the high peaks associated with $\gamma = 0$ persist for any sample size, generating points far form the optimal plane. In agreement with Figure 4.4, the candidates obtained for γ close to ε were mostly infeasible.