Importance Sampling for Chance Constrained Optimization[*]

Vyacheslav Gorchakov [†]

Abstract. Chance constrained problem is stated as minimizing a objective over solutions satisfying, with a given close to one probability, a system of convex constraints. These problems appears in many flavours of engineering. Probabilistic constraints, which appears at chance constrained problems, are not convex in general case. One of the main approach of solving these problems is to approximate the feasible set. Several state-of-the-art convex approximations are known: Markov, Chebyshev, Bernstein and sample average approximation (SAA). Moreover, we have the scenario approximation, which consists in sampling random points according to the nominal noise distribution and solving a deterministic problem afterwards. It is known to be one of the most accurate but time demanding approximation. Our goal is to propose a new way to reduce complexity of scenario approximation method for a linear objective and linear chance constraints with Gaussian noise by utilizing importance sampling.

Key words. chance constrained optimization, importance sampling, scenario approximation

AMS subject classifications. 68Q25, 68R10, 68U05

1. Introduction.

Chance constrained optimization appears in computer vision [7], robotics [11], gas and power systems [15], self-driving cars [14], district heating [16] and transportation networks[3]. The origins of these methods date back to the decision theory in 1950 [1] and, more recently, to robust optimization methods, where we are dealing with probabilistic and non-probabilistic models of robustness [4]. Since a robust solution is usually conservative, it is desired to have a compromise between the robustness and the risk of failure in many practical problems [2]. The robust optimization defines secure as ensuring feasibility for all realizations within a predefined uncertainty set, while chance-constrained optimization seeks to satisfy the constraints with a high probability [6]. Being significantly less conservative then the robust optimization methods, chance-constrained algorithms are widely used in engineering practice.

Being computationally intractable [9], the chance-constrained optimization admits efficient solutions only in a few cases. The case of Gaussian noise at constraints is the simplest approach. To this end, various approximate methods are used in practice. The convexity of constraints simplifies chance constrained problems and there are some state-of-the art approaches as Markov, Chebyshev, Bernstein approximations [12] and sample average approximation for mixed-integer stochastic problems [5]. These approximations find a convex set to approximate the feasible set.

Sample average approximation is a valuable alternative that theoretically can even lead to an optimal solution [10]. Unfortunately, this method requires a number of samples and often expensive for engineering practice. [13] In particular, if a standard deviation of the distribution is much less then a distance to the feasibility boundary the sample average approximation is inefficient as it requires significant efforts to generate even one infeasible point. Importance sampling allows to avoid this problem, by adjusting the distribution to sample from.

[*]Submitted to the editors DATE.

Funding: This work was funded by the Fog Research Institute under contract no. FRI-454.

[†] (ddoe@imag.com, http://www.imag.com/~ddoe/).

This manuscript is for review purposes only.

We investigate importance sampling approach for chance constrained optimization and evaluate its efficiency over simulated and real test cases coming from reliability analysis of power systems.

The paper is organized as follows. Our problem setup is in section 2, experiment design is in section 4, main results are in section 5.

2. Background and Problem Setup.

Let $\{x_i\}_{i=1}^m, x_i \in \mathbb{R}^n$ be a sequence of points. Let $f(x) = c^\top x$ be a linear function. Consider linear probability constraints:

$$\text{Prob}(g(x,\xi) \geq 0) \leq \eta,$$

where ξ is scenario, η is confidence level, $g : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}$ is a function of x, ξ.

The optimization problem is then:

$$\begin{aligned} & c^\top x \to \min_{x \in \mathbb{R}^n} \\ & \text{s.t.: Prob}_\xi(g(x,\xi) \geq 0) \leq 1 - \eta \\ & \quad \xi \sim \mathcal{N}(\mu, \Sigma), \end{aligned}$$

where μ is expected value of gaussian noise and $Sigma$ is its covariance matrix.

Assumption 1. Hinge loss function. Let p, t be parameters. Then function

$$l(x) = \begin{cases} 0, x < p \\ ty + p, x \geq p \end{cases}$$

is called the Hinge loss function.

Firstly, let consider state-of-the-art methods. Our goal is to approximate the feasible set. It is known such approximations of feasible set as Bernshtein approximation, Markov approximation, Tchebyshev approximation and sample average approximation (SAA).

Convex approximations : Markov, Chebyshev, Bernstein. Assume a surrogate function $\psi(u)$ to replace and approximate the 0-1 loss function. $\psi(u) = \max\{1 + u, 0\}$ for Markov approximation, $\psi(u) = (1 + u)^2$ for Chebyshev approximation and $\psi(u) = e^u$ for Bernstein approximation. After convex approximation the original problem becomes problem with constraints as following:

$$\begin{aligned} & c^\top x \to \min_{x \in \mathbb{R}^n}, t \in \mathbb{R} \\ & \text{s.t.} t \cdot \mathbb{E}\left[\psi\left(\frac{g(x,\xi)}{t}\right)\right] - t\eta \geq 0 \\ & \quad \xi \sim \mathcal{N}(\mu, \Sigma) \end{aligned}$$

The constraints becomes a deterministic convex constraint.

State optimization problem:

$$
\begin{aligned}
&-x \to \min_{x \in \mathbb{R}^n} \\
&\text{s.t.: } \operatorname{Prob}_\xi(x\xi \le 1) \le 1 - \eta \\
&\xi \sim \mathcal{N}(\mu, \Sigma),
\end{aligned}
$$

Exact solution:

$$
\operatorname{Prob}_\xi(x\xi \le 0) = \operatorname{Prob}\left(\frac{x\xi}{\sqrt{x\Sigma^2 x}} \le \frac{1}{x\Sigma}\right) = F\left(\frac{1}{x\Sigma}\right)
$$

Then optimization problem is equivalent to:

$$
\begin{aligned}
&-x \to \min_{x \in \mathbb{R}^n} \\
&\text{s.t.: } F^{-1}(1-\eta)x\Sigma \le 1
\end{aligned}
$$

Markov's Approximation.. On the other hand, consider the results of well-known approximations. We start with the Markov's approximation that leads us to the popular CVaR solution:

$$
\begin{aligned}
f^M = x^M = \max_{x,t} x \\
\texttt{s.t.:} \mathbb{E}(x + \xi + t)_+ \le t\eta, \; t \in \mathbb{R}
\end{aligned}
$$

where $(\cdot)_+$ stands for a positive part of the expression in braces. Indeed

$$
\mathbb{P}(x + \xi \ge 0) \le \mathbb{E}\psi((x + \xi)/t)
$$

by the Markov's inequality for any convex ψ and positive t. If $\psi(u) = (u+1)_+$ one has the Markov's bound above.

Computing the expectation, one has

$$
\begin{aligned}
\mathbb{E}(x + \xi + t)_+ &= \int (\xi + x + t)_+ \phi(\xi) d\xi = \int_{\xi \ge -x-t} \xi\phi(\xi)d\xi + (x+t)\int_{\xi \ge -x-t} \phi(\xi)d\xi \\
&= \frac{\sigma^2}{\sqrt{2\pi\sigma^2}} \exp(-(x+t)^2/(2\sigma^2)) + (x+t)\Phi(-(x+t)).
\end{aligned}
$$

Using Feller's bound [**?**, Section 7.1][1] or simply the Taylor series for the tail of the Gaussian c.d.f. anf taking $x + t \to \infty$:

$$
\begin{aligned}
\mathbb{E}(x + \xi + t)_+ &= \frac{\sigma}{\sqrt{2\pi}} \exp(-(x+t)^2/(2\sigma^2)) + \frac{\sigma}{\sqrt{2\pi}} \exp(-(x+t)^2/(2\sigma^2))(1 + O(\sigma^2/(x+t)^2)) \\
&= \frac{2\sigma}{\sqrt{2\pi}} \exp(-(x+t)^2/(2\sigma^2))(1 + O(\sigma^2/(x+t)^2)) \le t\eta.
\end{aligned}
$$

[1] I found this or Pollard's book to be extremely useful.

Minimizing the difference between the left hand side and the right hand side in positive t, one gets $t = -x/2 + \sqrt{t^2 + 4\sigma^2}/2 = \sigma^2/x + O(1/x^2)$. Finally, solving the above in x one has

$$x^M = f^M = -2\sigma\sqrt{\log(1/\eta)} + O(\log\log 1/\eta).$$

A more accurate analysis of residuals shows that $x^M - x^* = \Omega(\log\log n)$ and thus

$$\frac{x^M}{x^*} \to 1, \qquad \frac{\Phi(x^M)}{\Phi(x^*)} \to \log^{\Theta(1)}(1/\eta), \eta \to 1.$$

A numerical analysis of the probability mass ratio shows that $\eta = 0.01$ the ratio $\Phi(x^M)/\Phi(x^*) \approx$ 3.33, and for $\eta = 0.001$ the ratio is $\Phi(x^M)/\Phi(x^*) \approx 3.35$ and independent of σ.

Chebyshev's Approximation. . For this approximation, we consider $\phi(u) = (1+u)_+^2$. In a similar way, one gets the Chebyshev's approximation of the chance constraint:

$$\mathbb{E}(x+\xi) + \sqrt{(1-\eta)\mathbb{E}(x+\xi)^2} = x + \sqrt{(1-\eta)(x^2 + \mathbb{E}\xi^2)} = x + \sqrt{(1-\eta)(x^2+\sigma^2)} \le 0$$

Thus

$$x^2 \ge (1-\eta)x^2 + (1-\eta)\sigma^2 \Leftrightarrow x \le -\sqrt{\frac{1-\eta}{\eta}}\sigma,$$

as x is negative. So that

$$\frac{f^*}{f^C} = \frac{x^*}{x^C} = \sqrt{\frac{1-\eta}{2\eta\log(1/\eta)}} \to \frac{1}{\sqrt{2\eta\log(1/\eta)}}.$$

where f^C and x^C are the the optimal objective and argument of the Chebyshev's approximation. This leads to an exponential mismatch in the tail estimation $\Phi(x^C)/\Phi(x^*) \approx \eta^2\exp(1/(2\eta))$. Numerically for $\eta = 0.01$ the ratio is about 5.

Chernoff's Approximation.. In this case we consider $\phi(u) = \exp(u)$ which yields to

$$\log\mathbb{E}\exp((x+\xi)/t) \le \log\eta.$$

Taking expectation on the left hand side one has[2] for the standard expression for the exponential moment generating function after minimizing in positive t:

$$\min_{t>0}\left\{\frac{x}{t} + \frac{\sigma^2}{2t^2}\right\} = -\frac{x^2}{4\sigma^2} \le \log\eta.$$

Thus $x^B = f^B = -2\sigma\sqrt{\log(1/\eta)}$ and

$$\frac{f^*}{f^B} = \sqrt{\frac{1}{2}}, \qquad \text{and} \qquad \frac{\Phi(x^B)}{\Phi(x^*)} = O(1)\cdot\eta^{3/2} \to +\infty, \eta \to 0.$$

Numerically, the ratio is about 8 for $\eta = 0.01$. Notice, that for a sufficiently large probability $\eta = 0.05$ the Chernoff's approximation is the most accurate, while the Markov's one is the least accurate.

[2] a nice reference paper is here

Sample average approximation. Sample average approximation theoretically solves the problem, especially in cases $\eta \approx 1$, but requires a lot of samples. The main idea of this approach is to replace the probability constraints by sampling.

$$c^\top x \to \min_{x \in \mathbb{R}^n}$$

$$\text{s.t.:} \frac{1}{N} \sum_{i=1}^{N} I_{(0,+\infty)}(g(x, \xi_i)) \le \gamma$$

Notice that case $\gamma > \eta$ corresponds to the solution, which is a lower bound on the original problem. Case $\gamma = 0$ is called the scenario approximation.

3. Algorithm.

Consider a point x and a level set at x of the level $1-\eta$. Notice, that the optimal point x^* is at most at a distance $d = \Phi^{-1}(1-\eta)$ from all the hyperplanes, $g_i^\top x - b_i = 0$. Indeed, in this case the $1-\eta$ level set is completely inside of the feasibility polytope. That leads us to the first step of our algorithm, namely we define the starting point as a solution of

$$\min_y r^\top y$$

$$\texttt{s.t.:} g_i^\top y \le b_i - d/\|g_i\|_2, \ 1 \le i \le m$$

We refer the optimal solution to this problem as y_0. Notice, that the mismatch with the optimal solution is at most $\|r\|_2 d\sqrt{m}$, where m is the number of hyperplanes. The bound in m can be improved to the minimum of the dimension and the number of hyperplanes, which follows directly from the Fritz-Jones ellipsoidal approximation.

On the other hand, the probability of being outside of a hyperplane $g_i^\top x - b_i = 0$ staying at point g^* is $\Phi((b_i - g_i^\top x^*)/(c_i^\top c_i))$. That is all of the hyperplanes can not be too closer to x^*. Then, the following inequalities necessary hold for x^*

$$b_i - g_i^\top x \le g_i^\top g_i \Phi^{-1}(1-\eta).$$

In other words, the following problem is equivalent the chance constrained optimization problem we started with:

$$\min \ r^\top y \tag{3.1a}$$

$$\texttt{s.t.:} \ \mathbb{P}(G(y+\rho) \le b) \ge 1-\eta, \eta > 0 \tag{3.1b}$$

$$\rho \sim \mathcal{N}(0, I_n) \tag{3.1c}$$

$$b_i - g_i^\top x \le g_i^\top g_i \Phi^{-1}(1-\eta). \tag{3.1d}$$

Now let us contribute to the scenario approximation of this problem. Prior to the nest step, notice, that the non-convex part sits in Eq. 3.1c only. Let $p_1, \dots, p_m$ be the distances from the optimal solution y^* of the problem above with Eq. 3.1b and Eq. 3.1d only to the hyperplanes $g_i^\top y - b_i = 0$ respectively.

Our algorithm for scenario approximation consists of the following steps (figures are numbered):

1. Let $\mathcal{P}_0$ be an feasibility polytope when there is no uncertainty, $\mathcal{P}_0 = \{y : \Gamma y \leq \beta\}$ (white area on Fig. (a))
2. Construct inner polytope $\mathcal{P}_1$ (white area on Fig. (b))
$$\mathcal{P}_1 = \{y : -\beta_i + \gamma_i^\top y - \Phi^{-1}(1-\eta) \leq 0,\ 1 \leq i \leq m\}$$
3. Construct inner feasible polytope $\mathcal{P}_2$ (green area on Fig. (c))
$$\mathcal{P}_2 = \{y : -\beta_i + \gamma_i^\top y - \Phi^{-1}(1-\eta) \leq 0,\ 1 \leq i \leq m\}$$
The optimal point is somewhere in the white area.
4. Our "true" feasibility area is in between the above. Marked as green + blue on Fig. (d).
5. Notice, that noise is additive. That is we do not care where is y^* and can use one sample.
6. Let us compute d_i and r_i as shown on Fig. (e)
7. we are interesting only in samples that are outside of the polytope, shown on Fig. (f)
8. Here is our sampling algorithm (following to Owen/Maximov/Chertkov'19 aka the inverse transform sampling):
   (a) Compute $\eta_i = \Phi(d_i)$.
   (b) With probability $\propto \eta_i$ sample $\rho \sim \mathcal{N}(0, I_n)$ s.t. $y^\top \gamma_i \geq b_i$:
      i. Sample $z \sim \mathcal{N}(0, I_n)$
      ii. Sample $u \in U(0,1)$
      iii. Let $w = \Phi^{-1}(u\Phi(-d))$
      iv. Finally, $\rho = -\gamma_i w - (I_n - \gamma_i \gamma_i^\top) z$
9. Use $\rho_1, \ldots, \rho_n$ as a scenario approximation.

The key observation.. Now consider the optimal point y^* and random samples generated by $\rho \sim \mathcal{N}(0, I_n)$. Notice that all samples in the polytope $\mathcal{P} = \{\rho : g_i^\top \rho \leq b_i - p_i\}$ never contribute to the stochastic approximation of the probability (as for the optimal point, they are always inside the feasibility polytope). So we may omit these points from the approximation.

To this end, we only require samples from $\rho \sim \mathcal{N}(0, I_n)$ that are out of the polytope $\mathcal{P}$, $p \notin \mathcal{P}$. As we can not sample directly from it [?], we sample from a density mixture $q_1, \ldots, q_m$ which stand for sampling out of the hyperplane $g_i^\top \rho = b_i - p_i$ each. Let q be a convex combination of the densities, $q(\rho) = \sum_{i=1}^m \alpha_i q_i(\rho)$, $\alpha_i \geq 0$, $\sum_{i=1}^n \alpha_i = 1$. Then we need to consider a weighted sample from q of the form $\{\rho_i \frac{\phi(\rho_i)}{q(\rho_i)}\}$, where $\rho_i \sim q$.

4. Experiment design.

Competitors. We study empirical performance of the following algorithms: Markov, Chebyshev, Chernoff approximations as the state-of-the-art methods and scenarion approximation with Importance sampling.

Place for importance sampling:

- First, one can estimate the "true" probability level by sampling. As all methods guarantees feasibility at least on level $1 - \eta$, one can do conditional sampling beyond the boundary.

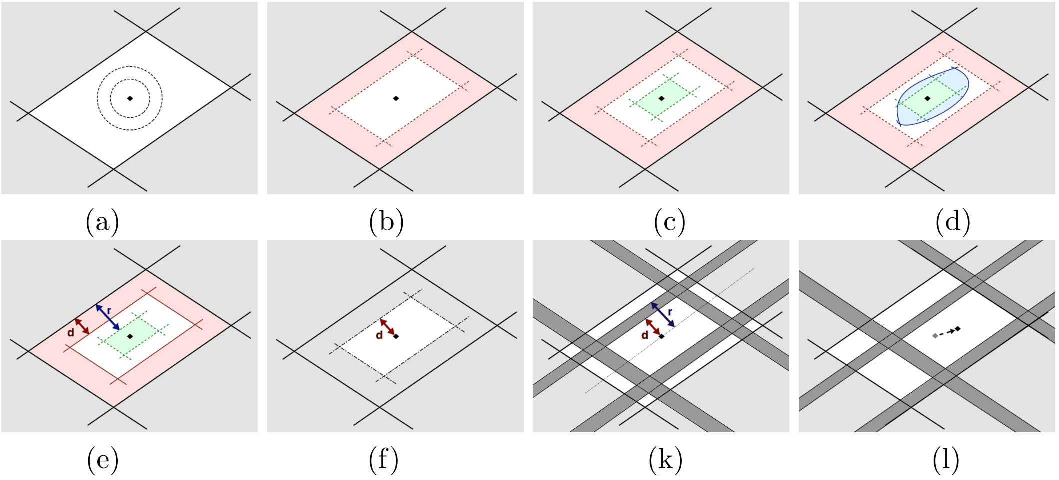


Figure 1. **Top row:** *Left: noiseless feasibility set* $\mathcal{P} = \{x : \; a_i^\top x \le b_i, 1 \le i \le m\}$*, dashed lines – level sets of Gaussian distribution. Black diamond represent an optimization variable* x*. Mid-left: red area does not contain the optimal solution, as the probability* $\max_{i:\, 1 \le i \le m} \mathbb{P}_\xi(a_i^\top(x+\xi) > b_i) > \eta$ *for some* i*,* $1 \le i \le m$*. Mid-right: all points in green area satisfy* $\sum_{i=1}^m \mathbb{P}_\xi(a_i^\top(x+\xi) > b_i) \le \eta$*. Right: feasibility area boundary marked as blue. All points in green and blue satisfy* $\mathbb{P}_\xi\{\bigcap_{i=1}^m a_i^\top(x+\xi) \le b_i\} \ge 1-\eta$*.* **Buttom row***: Left: let* d *and* r *are the distances between the (top-left) polytope boundary and (top-left) boundaries of the green and the red zones. Mid-left: Sampling in white area does not contribute to the error of scenario approximation. Mid-right: More accurately, we are interested in samples only in dark-grey area. Samples that closer diamond are always feasible and samples that are far from diamond are always infeasible. Right: optimization point (black diamond) moved (from the light-grey diamond) and areas of sampling moved accordingly. Operating point can not move further in the same direction.*

- Second, one we have a probability of failure estimate π, and it is less then η at point $\bar{x}$ one can guarantee that the set

$$\{x : \{c^\top x \ge c^\top \bar{x}\} \wedge \{\sqrt{(x-x_0)^\top \Sigma^{-1}(x-x_0)} \le -\log(\Phi^{-1}(1-\pi))\}$$

  So, we can try to use this to make a one-step improvement.
  We will work with confidence bounds on the next step.

Our first experiment is devoted to the simplest case with linear function $g(x,\xi)$:

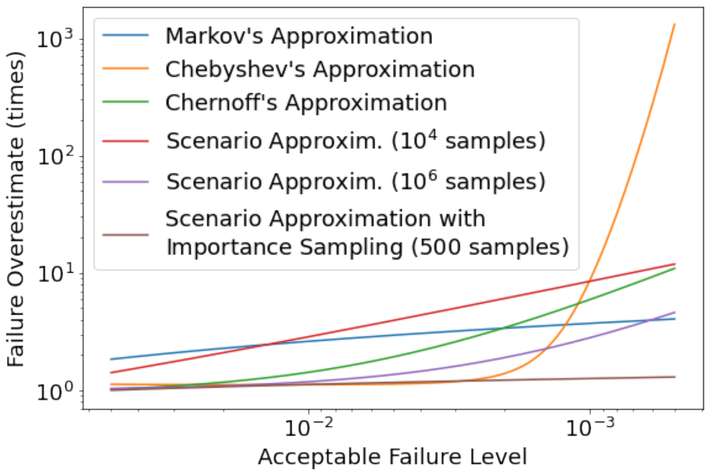
$$\begin{aligned} & c^\top x \to \min \\ & \text{s.t.: } \mathrm{Prob}(a^\top(x+\xi) \le b) \ge 1-\eta, \end{aligned}$$

where $a \in \mathbb{R}^n$ is a fixed vector, x is a an optimization variable, and ξ is Gaussian uncertainty with known mean and covariance $\xi \in \mathcal{N}(\mu, \Sigma)$. In this case the result is known [8], while it is still interesting to understand how various methods perform. Our goal is to figure out which of the algorithms solve the problem exactly and which one solves approximately.

Our second experiment is devoted to portfolio optimization.

5. Main results.

Let plot failure overestimate depending on the confidence level $1-\eta$. Compare the state-of-the-art approaches with mix of scenario approximation and importance



sampling. performances. In particular, compare case $\eta \to 0$, when confidence level is approximately 1. On the figure 2 one can see that scenario approximation performs better than such convex approximations as Markov, Chebyshev and Bernstein, especially in case, when η is approximately 0. On the figure 1 we provide more detailed description of our approach.

REFERENCES

[1] K. J. Arrow, *Decision theory and operations research.*

[2] J. W. Baker, M. Schubert, and M. H. Faber, *On the assessment of robustness*, Structural Safety, 30 (2008), pp. 253–267.

[3] Q. S. Z. Y. D. Bo Zhoua, Guo Chenb, *Robust chance-constrained programming approach for the planning of fastcharging stations in electrified transportation networks.*

[4] G. E. Box, *Robustness in the strategy of scientific model building*, in Robustness in statistics, Elsevier, 1979, pp. 201–236.

[5] M. Branda, *Sample approximation technique for mixed-integer stochastic programming problems with several chance constraints.*

[6] A. Charnes and W. W. Cooper, *Chance-constrained programming*, Management science, 6 (1959), pp. 73–79.

[7] S. M. Gal Dalal, Elad Gilboa and L. Wehenkel, *Chance-constrained outage scheduling using a machine learning proxy.*

[8] N. He, *Chance-constrained programming (ccps).*

[9] L. G. Khachiyan, *On the exact solution of systems of linear inequalities and linear programming problems.*

[10] A. J. Kleywegt, A. Shapiro, and T. Homem-de Mello, *The sample average approximation method for stochastic discrete optimization*, SIAM Journal on Optimization, 12 (2002), pp. 479–502.

[11] Y. K. J. B. Masahiro Ono, Marco Pavone, *Chance-constrained dynamic programming with application to risk-aware robotic space exploration.*

[12] A. Nemirovski and A. Shapiro, *Convex approximations of chance constrained programs.*

[13] B. K. Pagnoncelli, S. Ahmed, and A. Shapiro, *Sample average approximation method for chance constrained programming: theory and applications*, Journal of optimization theory and applications, 142 (2009), pp. 399–416.

[14] A. V. R.Austin Dollar, *Automated vehicles in hazardous merging traffic: A chance-constrained approach.*

[15] L. Roald and G. Andersson, *Chance-constrained ac optimal power flow: Reformulations and efficient*

algorithms.

[16] Z. M. G. S. Zhinong Wei, Juan Sun, *Chance-constrained coordinated optimization for urban electricity and heat networks.*