Efficient Solution of Chance-Constrained nonlinear Optimization Problems with non-Gaussian Uncertainties

Dr. Abebe Geletu
Joint work with M. Klöppel, A. Hoffmann, P. Li
(Supported by DFG, Grant Nr. LI806/8-2)
Ilmenau University of Technology

EurOpt2010, September 5 - 9, 2010, Lisbon, Portugal

September 8, 2010
Topics

1. Motivation
2. A nonlinear MPC with chance-constraints
   2.1. A chance constrained nonlinear optimization
3. Strategies for the computation of chance constraints
4. Efficient integration techniques for probability integrals
5. Cost-optimal wind power generation - a preliminary investigation
6. Conclusions
1. Motivation

Control, design, planning and investment in the presence of uncertainties demand for not only optimality, but also call for reliability and risk aversion. Chance constrained optimization models can be applied in:

- financial risk management (using risk metrics like VaR and cVaR);
- planning and budget management;
- process control and optimization under uncertainties;
- water reservoir management;
- reliability based design optimization (RBDO);
- Control and optimization based on prediction (say using weather forecast data etc.).
Motivation...

In general,

- process input variables can be Gaussian or non-Gaussian;
- the process can be dynamic and non-linear;
⇒ process outputs are time-dependent and uncertain;
⇒ deterministic constraints on (future) outputs or design specifications of the process can be violated;

Many practical applications are tolerant to a degree of constraint violations.
⇒ such constraints can be required to be satisfied with a specified degree of reliability
⇒ chance constraints.
2. A nonlinear MPC with chance constraints

Consider

\[
(\text{CCNMPc}) \quad \min_u \{ E[f_1(X, U, \xi)] + \gamma \text{Var}[f_2(X, U, \xi)] \} \tag{1}
\]

\[s.t.\]

\[x_{k+1} = g(x_k, u_k, \xi_k), \text{ with given } x_0; \tag{2}\]

\[\Pr \{ x_{k,i}^{\text{min}} \leq x_{k,i} \leq x_{k,i}^{\text{max}} \} \geq \alpha_i, i \in I; \tag{3}\]

\[u_k \in U, \xi_k \in \mathcal{X}; \tag{4}\]

\[k = 0, 1, \ldots, N - 1. \tag{5}\]

- \(x_k \in \mathbb{R}^n, u_k \in \mathbb{R}^m\) - state and control variables, resp.;
- \(\xi_k \in \mathbb{R}^p\) - random disturbances with known probability distribution;
- \(E(\cdot), \text{Var}(\cdot)\) - expectation and variance operators;
- \( Pr(\cdot) \) probability measure;
- \( \alpha_i \) - given reliability levels; usually, \( 1/2 \leq \alpha_i \leq 1 \);
- \( \gamma > 0 \) - a weighting factor.
- \( \gamma > 0 \) - a weighting factor.
- \( N \) = length of the prediction horizon;
- commonly, \( \xi_k \) are uncertain input variables;
- \( \phi(\xi_k) \) probability density for uncertain variables \( \xi_k, k = 0, 1, \ldots, N - 1 \), over the prediction horizon \( \Rightarrow \) assumed known based on predicted, experimental or historical data of the process under consideration;
- \( x_k \) output variables - uncertain due to the relation \( x_{k+1} = g(x_k, u_k, \xi_k) \);
- the constraint \( Pr \{ x_{k,i}^{\min} \leq x_{k,i} \leq x_{k,i}^{\max} \} \geq \alpha_i \) demands that the constraint \( x_{k,i}^{\min} \leq x_{k,i} \leq x_{k,i}^{\max} \) to be satisfied with reliability level \( \alpha_i \);
- the functions \( f_1, f_2, g \) are twice differentiable.
2.1. A chance constrained NLOP

On each prediction horizon to solve an optimization problem of the form:

\begin{equation}
\text{(CCOPT)} \quad \min_{u} \{ E[f_1(x, u, \xi)] + \gamma \text{Var}[f_2(x, u, \xi)] \}
\end{equation}

s.t.

\begin{align*}
G(x, u, \xi) &= 0 \\
Pr\{x_i^{\min} \leq x_i \leq x_i^{\max}\} &\geq \alpha_i, i \in I \subset \{1, \ldots, n\} \\
u \in \mathcal{U}, \xi \in \mathcal{X}.
\end{align*}

**Assumption:** for a given \(u\) and realizations of \(\xi\), the system \(G(x, u, \xi) = 0\) there is a representation \(x(u, \xi)\). Hence:

- \(x_i(u, \xi)'s\) are random;
- the distribution of \(x_i(u, \xi)'s\) can be difficult to determine.
3. Computation of chance constraints

Except for some "special cases", the computation of the values of and gradients of the functions

\[ h_i(u) := \Pr\{x_i^{\min} \leq x_i(u, \xi) \leq x_i^{\max}\} \geq \alpha_i, i \in I; \]
\[ F_1(u) := E[f_1(x(u, \xi), u, \xi)] \quad \text{and} \quad F_2(u) := \text{Var}\left[f_2(x(u, \xi, u, \xi)\right] \]

pose enormous difficulties.

**Idea**: transform a chance constraint
\[ \Pr\{h^{\min} \leq h(u, \xi) \leq h^{\max}\} \geq \alpha \] to a tractable one.

Three major approaches:

1. (3.1.) Back-projection
2. (3.2.) Sample average approximation (SAA)
3. (3.3.) Analytic Approximation
3.1. Back-projection

Transform the chance constraint
\[ Pr\{h^{\text{min}} \leq h(u, \xi) \leq h^{\text{max}}\} \geq \alpha \]
to the space of the random variables \(\xi\) with a known distribution.

Find a \(\xi_j\) such that
- either \(\xi_j\) increases \(\Rightarrow h(u, \cdot)\) increases; denoted \(\xi_j \uparrow h(u, \cdot)\)
- or \(\xi_j\) increases \(\Rightarrow h(u, \cdot)\) decreases; denoted \(\xi_j \downarrow h(u, \cdot)\)

Hence, for instance, \(\xi_j \uparrow h(u, \cdot)\) \(\Rightarrow\)

\[
h(u) = Pr\{h^{\text{min}} \leq h(u, \xi) \leq h^{\text{max}}\} = \\
= Pr\{\xi_j^{\text{min}}(u, h^{\text{min}}, \tilde{\xi}) \leq \xi_j \leq \xi_j^{\text{max}}(u, h^{\text{max}}, \tilde{\xi})\} \\
= \int_{-\infty}^{+\infty} \ldots \int_{-\infty}^{+\infty} \int_{\xi_j^{\text{min}}(u, h^{\text{min}}, \tilde{\xi})}^{\xi_j^{\text{max}}(u, h^{\text{max}}, \tilde{\xi})} \phi(\xi) d\xi \]
where $\tilde{\xi} = (\xi_1, \ldots, \xi_{j-1}, \xi_{j+1}, \ldots, \xi_p)$. 

- $\xi^{min}_j(u, h^{min}, \tilde{\xi})$ and $\xi^{max}_j(u, h^{max}, \tilde{\xi})$ are determined from $G(x, u, \xi) = 0$ for given $u$ and realization of $\xi$. 

Back-projection contd...

Monotony relations

- can be determined based on mathematical analysis of the model equations; or using simulation techniques;
- commonly applicable to small-scale models of process with input/output relations;

Advantages:

- Can provide an exact representation of chance constraints.

Disadvantages,

- monotony relations may be not exist; can be also difficult to identify.
3.2. Sample average approximation (SAA)

Replace \( Pr\{h_{min} \leq h(u, \xi) \leq h_{max}\} \geq \alpha \) with its equivalent
\[
Pr\{h(u, \xi) < h_{min}\} + Pr\{h(u, \xi) > h_{max}\} \leq 1 - \alpha. \quad (\star)
\]

Set
\[
H^{(1)}(u, \xi) := -h(u, \xi) + h_{min}, \quad \text{and} \quad H^{(2)}(u, \xi) := h(u, \xi) - h_{max}
\]
and
\[
h^{(1)}(u) = E \left[ \mathbb{I}_{(0, +\infty)}(H^{(1)}(u, \xi)) \right] \quad \text{and} \quad h^{(2)}(u) = E \left[ \mathbb{I}_{(0, +\infty)}(H^{(2)}(u, \xi)) \right]
\]
where
\[
\mathbb{I}_{(0, +\infty)}(z) = \begin{cases} 
1, & \text{if } z > 0 \\
0, & \text{else.}
\end{cases}
\]

Hence, \((\star)\) is equivalent to
\[
h^{(1)}(u) + h^{(2)}(u) \leq 1 - \alpha
\]
Instead of \( h^{(j)}(u) \) use \( h_N^{(j)} = \frac{1}{N} \sum_{k=1}^{N} \mathbb{I}_{(0, +\infty)} (H^{(j)}(u, \xi_k)) \), \( j = 1, 2; \)

where \( \xi_1, \xi_2, \ldots, \xi_N \) are random samples (or scenarios).

**Advantages:**

- (SAA) can be used wether \( \xi \) is Gaussian or non-Gaussian;
- can lead to a convex problem if \( H^{(j)}(\cdot, \xi) \) is convex for \( j = 1, 2; \)

**Disadvantages:**

- requires very large sample size \( \implies \) large number of nonlinear equality constraints for the approximation of (CCOPT) \( \implies \) expensive computation;
- resulting problem is non-smooth;
- resulting solution can be infeasible for the (CCOPT).
3.3. Analytic approximation

Uses ideas from approximation of probability inequalities.

- **Take a convex function**
  \[ \varphi : \mathbb{R} \rightarrow \mathbb{R}_+ : \varphi(z) > \varphi(0) = 1, \forall z > 0. \]
  **Example**: \( \varphi(z) = \exp(z) \).

- Define
  \[ \psi(u, t) := tE[\varphi(t^{-1}H(u, \xi))] \]
  corresponding to \( \Pr\{H(u, \xi) > 0\} \leq \alpha \).

**Theorem** (Nemirovski and Shapiro 2006)

Suppose there \( t : \psi(u, t) \leq t\alpha \).

\[ \inf_{t > 0} \{\psi(u, t) - t\alpha\} \leq 0 \quad \iff \quad \Pr\{H(u, \xi) > 0\} \leq \alpha. \]
Corollary: Let $\alpha^{(1)} + \alpha^{(2)} = 1 - \alpha$ and
\[
\psi^{(1)}(u, t) := tE \left[ \varphi \left( t^{-1} \left\{ -h(u, \xi) + h^{\text{min}} \right\} \right) \right],
\]
\[
\psi^{(2)}(u, t) := tE \left[ \varphi \left( t^{-1} \left\{ h(u, \xi) - h^{\text{max}} \right\} \right) \right].
\]

If
\[
\inf_{t > 0} \left\{ \psi^{(j)}(u, t) - t\alpha^{(j)} \right\} \leq 0, j = 1, 2;
\]
then $Pr\{h(u, \xi) < h^{\text{min}}\} + Pr\{h(u, \xi) > h^{\text{max}}\} \leq 1 - \alpha$
\[
\Leftrightarrow Pr \left\{ h^{\text{min}} \leq h(u, \xi) \leq h^{\text{min}} \right\} \geq \alpha.
\]
Advantages:

- results in a smooth approximation of chance constraints;
- can be used irrespective of the distribution of the random variable.

Disadvantage:

- there is a danger of over or under estimation of chance constraints
- very little work in this direction

Note that: Approaches (I) & (III) require the computation of multi-dimensional integrals:
\[ \int_{\Omega} f(u, \xi) \phi(\xi) \, d\xi \] for each fixed \( u \).

- Fast and efficient computation of integrals is crucial!
4. Computation of probability integrals

Given a function $f : \mathbb{R}^p \to \mathbb{R}$ and a weight (probability density) function $\phi(\xi)$, how to efficiently compute the integral:

$$E[f] = \int_{\Omega} f(\xi)\phi(\xi) d\xi$$

with $\Omega \subset \mathbb{R}^p$ is the support of $\phi$?

**Efficient methods for the numerical computation of $E[f]$:**

Deterministic cubature rules

- Full-grid integration (cubature) rules (Cools 2002)
- Sparse-grid integration rules (Smolyak 1963, Bungarthy and Griebel 2004)

**The state-of-the-art:** Fewer integration nodes & higher polynomial exactness $\implies$ Sparse-grid techniques.

**Note:**

- Cubature techniques are constructed based one-dimensional quadrature rules.

- One-dimensional **interpolatory Gauss quadrature rules** (and their extensions) are found to be efficient, due to their higher **degree of accuracy**.
Let $f : \mathbb{R} \to \mathbb{R}$ and $\Omega_1 \subset \mathbb{R}$, then $I[f] = \int_{\Omega_1} f(\xi_1)\phi_1(\xi_1)d\xi_1$

can be approximated by $Q^{(1)}[f] = \sum_{k=1}^{N_1} w^{(k)}_1 f\left(\xi_1^{(k)}\right)$,

where the integration nodes $\mathcal{X}_1 = \{\xi_1^{(1)}, \ldots, \xi_1^{(N)}\}$ and weights $w^{(1)}_1, \ldots, w^{(N)}_1$ are constructed based orthogonal polynomials associated with $\phi_1$ and $\Omega_1$. 
Basic assumptions

Assumptions:

(A1): \( \phi(\xi) = \phi(\xi_1, \xi_2, \ldots, \xi_p) \) can be written as

\[
\phi(\xi) = \prod_{j=1}^{p} \phi_j(\xi_j) \leftarrow \text{product weight function;}
\]

where \( \phi_j : \mathbb{R} \rightarrow \mathbb{R}_+ \) and \( \phi_1 \equiv \phi_2 \equiv \ldots \equiv \phi_p \). This holds true when \( \xi_1, \xi_2, \ldots, \xi_p \) are independent random variables.

(A2): the domain of integration

\[
\Omega = \Omega_1 \times \Omega_2 \times \ldots \Omega_p;
\]

where \( \Omega_j \subset \mathbb{R} \) and \( \Omega_1 = \ldots = \Omega_p \).

Example: \( \phi(\xi) = \prod_{j=1}^{p} m_{\alpha_j, \beta_j} \xi_j^{\alpha_j} (1 - \xi_j)^{\beta_j} \) with

\[
\phi_j(\xi_j) = m_{\alpha_j, \beta_j} \xi_j^{\alpha_j} (1 - \xi_j)^{\beta_j} \quad \text{and} \quad \Omega_j = [0, 1], j = 1, \ldots, p.\]
Full-grid integration techniques

Suppose assumptions (A1) and (A2) hold true. Let, for each $j = 1, \ldots, p$,

- $\mathcal{X}_j = \{x_j^{(1)}, x_j^{(2)}, \ldots, x_j^{(N_j)}\} \subset \Omega_j$ are Gauss quadrature nodes;
- $w_j^{(1)}, w_j^{(2)}, \ldots, w_j^{(N_j)}$ corresponding weights for the one-dimensional integral on $\Omega_j$ with weight function $\phi_j$.

Full-grid integration rule:

$$Q[f] = \left( Q_1^{(1)} \otimes Q_2^{(1)} \otimes \ldots \otimes Q_p^{(1)} \right) [f] =$$

$$\sum_{k_1=1}^{N_1} \sum_{k_2=1}^{N_2} \ldots \sum_{k_p=1}^{N_p} \left( w_1^{(k_1)} \cdot w_2^{(k_2)} \cdot \ldots \cdot w_p^{(k_p)} \right) f \left( \xi_1^{(k_1)} \cdot x_2^{(k_2)} \cdot \ldots \cdot \xi_p^{(k_p)} \right).$$
How good are full-grid integration rules?

- **Number of grid-points (integration-nodes) in** $Q[·]$

  $$\#X = N_1 \times N_2 \times \ldots \times N_p.$$  

  If $N_1 = N_2 = \ldots = N_p$, then  

  $$\#X = N^p.$$ 

  ⇒ the number of grid-points in $Q[·]$ grows exponentially w.r.t. with the dimension of the integral 

  ⇒ known as the **curse-of-dimensions**.

- if each quadrature rule $Q^{(1)}[·]$ has a degree of **polynomial exactness** equal to $2N_j - 1$, then the degree of exactness of $Q[·]$ is equal to 

  $$\max_{1 \leq i \leq n} \{2N_i - 1\}.$$
If $N_1 = N_2 = \ldots = N_p = N$, then the degree of exactness of $Q[\cdot]$ is $2N - 1$; But $N^p$ integration nodes are too many ⇒ redundancy in the full-grid method (Mysovskikh 1968).

**Mysovskikh 1968**: to attain a degree of polynomial exactness equal to $d$, the **required number of integration nodes** in $Q[\cdot]$ has an upper bound given by

$$N_{\text{opt}} \leq \binom{p + d}{d} = N_{\text{max}}.$$  

How to construct such a cubature rule? **Use Smolyak’s technique**! (Smolyak 1963).
4.2. Sparse-grid integration rules

Since $\Omega_1 = \Omega_2 = \ldots = \Omega_p$ and $\phi_1 = \phi_2 = \ldots = \phi_p \implies$ the same quadrature rule on each $\Omega_i$ using $\phi_i$, so drop the index $i$.

**Assumption (A3):** For each one-dimensional cubature rule on $\Omega \subset \mathbb{R}$, there is a sequence of sets of nodes $\mathcal{X}^{(i)}, \mathcal{X}^{(i+1)}, \ldots$ such that

$$\# \mathcal{X}^{(i)} \leq \# \mathcal{X}^{(i+1)}, \ldots, i = 1, 2, \ldots.$$

For $f_1 : \mathbb{R} \to \mathbb{R}$, the quadrature rule $Q^{(i)}_1$ with nodes in $\mathcal{X}^{(i)} = \{\xi_1^{(i)}, \xi_2^{(i)}, \ldots, \xi_{N_i}^{(i)}\}$ is

$$Q^{(1)}_i[f_1] = \sum_{k_i=1}^{N_i} w^{(i)}_{k_i} f_1(\xi_{k_i}^{(i)}).$$

Hence, $Q^{(1)}_i[f_1] \leq Q^{(1)}_{i+1}[f_1], i = 1, 2, \ldots$. 
Sparse-grid integration rules ...

**Definition (Smolyak’s tensor product cubature rule)**

Suppose \( Q^{(1)}_1, Q^{(1)}_2, \ldots \) is a sequence of quadrature rules such that \( Q^{(1)}_i \) is a quadrature with \( N_i \) nodes and \( Q^{(1)}_0[f_1] = 0 \) (i.e. \( N_0 = 0 \)). The Smolyak’s cubature rule with accuracy \( l \in \mathbb{N} \) for \( p \)-dimensional integration is defined as

\[
S_{l,p} [f] = \sum_{p \leq \|i\| \leq p+l-1} (\Delta_{i_1} \otimes \Delta_{i_2} \otimes \ldots \otimes \Delta_{i_p}) [f]
\]

where \( \Delta_i = Q^{(1)}_{i+1} - Q^{(1)}_i, i = 0, 1, 2, \ldots \); and multi-index \( i = (i_1, i_2, \ldots, i_p) \in \mathbb{N}^p \) with \( \|i\| := i_1 + i_2 + \ldots + i_p \).

\Rightarrow \text{a tensor product of difference of quadrature rules.}
Theorem (Wasilkowski & Wozniakowski 1995)

The sparse-grid integration technique for a given integer \( l \) (with \( l \geq p \)) is defined as

\[
S_{l,p}[f] = \sum_{l-p \leq \|i\| \leq l-1} (-1)^{l-1-\|i\|} \left( \begin{array}{c} p-1 \\ l-1-\|i\| \end{array} \right) \times
\left( Q_{1}^{(i_1)} \otimes Q_{1}^{(i_2)} \otimes \ldots \otimes Q_{1}^{(i_p)} \right) [f];
\]

where

\[
\left( Q_{1}^{(i_1)} \otimes Q_{1}^{(i_2)} \otimes \ldots \otimes Q_{1}^{(i_p)} \right) [f] = \sum_{k_{i_1}} N_{i_1} \sum_{k_{i_2}} N_{i_2} \ldots \sum_{k_{i_p}} N_{i_p} \left( w_{k_{i_1}} \cdot w_{k_{i_2}} \cdot \ldots \cdot w_{k_{i_p}} \right) f \left( \xi_{k_{i_1}} \cdot \xi_{k_{i_2}} \cdot \ldots \cdot \xi_{k_{i_p}} \right).
\]
Properties of sparse-grid techniques

- The set of all nodes in the sparse grid rule $S_{l,p}[\cdot]$ is
  \[ \mathcal{X}_{l,p} = \bigcup_{l-p \leq \|i\| \leq l-1} (\mathcal{X}^{(i_1)} \times \mathcal{X}^{(i_1)} \times \ldots \times \mathcal{X}^{(i_p)}) . \]

- The number of nodes in the sparse grid rule $S_{l,p}[\cdot]$ is
  \[ \#\mathcal{X}_{l,p} \approx \frac{2^l}{l!} p^l. \]

$\Rightarrow$ $\#\mathcal{X}_{l,p}$ has a polynomial dependence w.r.t. the dimension $p$, for a fixed $l$.

$\Rightarrow$ sparse-grid techniques need few integration nodes as compared to full-grid rules

- the sparse-grid integration rule $S_{l,p}[\cdot]$, can attain a degree of polynomial exactness at least $2l - 1$. 

Properties of sparse-grids techniques...

**Advantages**

- a few integration nodes are enough to yield a good approximation of integrals; thus, saving CPU time;
- the construction of the integration nodes & weights is done using the weight function \( \phi(\cdot) \) and the integration domain \( \Omega \), **independent of the function to be integrated**. Thus, grid-points and weights can be computed only once and used repeatedly;
- functions which are polynomial with respect to the uncertain variable can be integrated exactly.

**Disadvantages of Sparse-Grid**

- pure sparse-grid integration **may not be good for non-smooth functions** (i.e. for functions with singularities)
Objective: Cost-optimal and reliable operational strategies for wind power generation plants based on nonlinear model predictive control under chance constraints.

- Wind power generation has a nonlinear dynamics and depends on wind-speed prediction;
- Uncertainty in wind speed $\Rightarrow$ uncertain power output $\Rightarrow$ deterministic constraints may be violated;
- Uncertain (random) variables are known to have non-Gaussian distributions; like Weibull- or Beta-distributions, etc.
- Expected operational costs can be higher
Cost optimal wind power generation ...

Difficulties:
- for each wind-speed prediction; the parameters of the distribution function needs to be recomputed \(\Rightarrow\) re-computation of integration grid points on each prediction horizon;
- analytic approximation chance constraints could not provided good results;
6. Conclusions

- Use chance constrained optimization for **optimal performance with higher reliability**.
- There should be a compromise between the optimal cost vs. the required level of reliability.
- In the CCNMPC, a longer prediction horizon:
  - may affect the reliability level for holding chance constraints;
  - incurs large number of uncertain variables $\Rightarrow$ increased computation time.
- Construction of sparse-grid integration methods based on quadrature rules using non-Gaussian weight functions is not a trivial task.
- There should be more work on analytic approximation of chance constraints.
- There needs to be a feasibility study for the approximate solutions to the original chance
Thank You!!