Reliability Design and Structural Optimization - Literature Survey

MARTIN TAPANKOV
Reliability Design in Structural Optimization
Literature Survey

Martin Tapankov

June 3, 2013

Contents

Nomenclature 3

1 Introduction 4
   1.1 Survey purpose and goals 4
   1.2 Survey Structure 4

2 Background 5
   2.1 Reliability and safety 5
   2.2 Risk-Based Design 5
   2.3 Limit State Design 6

3 Reliability-Based Design Optimization 8
   3.1 General RBDO formulation 8
   3.2 Early Developments 9
      3.2.1 Hasofer-Lind Reliability index 12
      3.2.2 Finding the Most Probable Point 12
      3.2.3 Transformation in standard normal space 13
   3.3 Double-loop approaches 13
   3.4 Single-loop approaches 14
   3.5 Decoupled approaches 16
   3.6 Comparative studies 18

4 Metamodels 19
   4.1 Design of experiments 20
   4.2 Extensions of RSM 26
      4.2.1 Kriging 27
      4.2.2 Radial basis functions 27
      4.2.3 Sequential RSM 28
      4.2.4 Optimal Regression Models 28

5 Metamodels in RBDO 30

6 Current and future trends 31
Glossary 32
Acronyms 33
References 36
Nomenclature

Reliability and safety

\( \beta, \beta_i \)  
Reliability index (of the \( i \)-th constraint)

\( \text{Cov}[\cdot, \cdot] \)  
Covariance

\( \text{E} [\cdot] \)  
Expected value

\( \mathbf{d} \)  
Deterministic design variables, including probability distribution parameters

\( f() \)  
Optimization objective function

\( g(), g_i() \)  
Optimization constraint function; limit state

\( (\cdot)^U \)  
Upper variable limit

\( (\cdot)^L \)  
Lower variable limit

\( \mu_X = \{\mu_{X_i}\} \)  
Mean value of random variable \( X_i \)

\( P_f \)  
Probability of failure

\( P_S \)  
Probability of safety

\( \text{Pr}[\cdot] \)  
Probability operator

\( \phi(\cdot) \)  
Probability density function of the normal distribution

\( \Phi(\cdot) \)  
Cumulative density function of the normal distribution

\( \sigma_X = \{\sigma_{X_i}\} \)  
Standard deviation of random variable \( X_i \)

\( \text{Var}[\cdot] \)  
Variance

\( \mathbf{u} \)  
Uncorrelated random variables in standard normal space

\( \mathbf{X} \)  
Random variables with arbitrary distributions and correlation

Metamodels

\( \mathbf{\beta} = \{\beta_i\} \)  
Regression coefficients

\( x_i \)  
Predictor variables / regressors

\( \hat{y} \)  
Predicted response

\( y = \{y_i\} \)  
Response variables

\( \mathbf{X} \)  
Experiment matrix, containing the design of experiments

\( \mathbf{M} = \mathbf{X}^T \mathbf{X} \)  
Information matrix

\( \text{NVar}[\hat{g}(\mathbf{x})]/\sigma^2 \)  
Prediction variance

\( \epsilon \)  
Regression error

\( \mathbf{b} = \{b_i\} \)  
Least squares estimators of \( \mathbf{\beta} \)

\( p \)  
Number of model parameters in a regression

\( m \)  
Number of experiments

\( n \)  
Number of design variables

\( k \)  
Number of regression terms
1 Introduction

Reliability-Based Design Optimization (RBDO) has seen rapid development in the recent decades, and a number of studies which document, review and compare the available algorithms have appeared. Most recently, the works of (Aoues and Chateauneuf 2010) and (Valdebenito and Schuëller 2010) provide a solid review of existing RBDO methods as well as an attempt to compare objectively a selection of reviewed algorithms. A large number of methods and approaches have been published in the recent decade, which signifies the level of interest in the field and its relative immaturity. Practical applications of RBDO methods have also been well documented, and some of the methods even gain traction in commercial finite element analysis codes, for example Altair HyperStudy (Altair Engineering Inc. 2012).

None of the available RBDO methods can be considered “best in class” for all possible optimization problems. The numerical difficulties in solving the RBDO problem caused the development of many methods, each with their unique assumptions and features.

1.1 Survey purpose and goals

This survey focuses primarily on the historical development of RBDO and the various classes of methods which emerged through the years. While reviews on contemporary RBDO methods have been done previously, they haven’t incorporated the basis on which such developments were possible in the first place — risk-based design and limit-state design. This study attempts to link the RBDO field with these methodologies for a more complete picture of the field of reliability and safety.

In addition, metamodelling and particularly its use in the context of RBDO is reviewed as well. While a number of research publications use standard metamodels in their exposition, there are many works which attempt to tailor a particular metamodel to the unique properties and features of RBDO.

While this study tries to provide a comprehensive perspective on RBDO methods, their development and related technologies, it does not attempt to compare various RBDO methods. Such studies have already been performed and replicated, see the aforementioned reviews by (Aoues and Chateauneuf 2010) and (Valdebenito and Schuëller 2010).

1.2 Survey Structure

This study is comprised of several distinct sections which treat different aspects of reliability, RBDO and metamodelling. In section 2, some background information is presented on reliability and safety, as well as structural optimization. Section three discusses in detail reliability-based design optimization and the development of various methods and algorithms in the field through the years. Section four gives an overview response surfaces and presents a few other related metamodelling techniques. In the fifth section, the combined use of metamodels and RBDO is explored, particularly metamodels specifically adapted for the needs of RBDO as opposed to general modelling methods presented in the previous section. Finally, the survey concludes with a brief review of current and anticipated future trends in the field.
2 Background

2.1 Reliability and safety

When talking about safety design, the concept of reliability is invariably mentioned. (ISO 2394:1998) defines reliability as

\[ \text{the ability of a structure to comply with given requirements under specified conditions during the intended life.} \]

Structural engineering is one of the first disciplines which recognized the importance of safety and reliability, due to the difference of magnitude of investment, time and cost of failure compared to other engineering fields, for example mechanical engineering. Due to the varying material properties, manufacturing conditions and load uncertainties, the concept of permissible stress design (PSD)\(^1\) was developed. In PSD, the design stresses in a structure are reduced by use of safety factors so that they do not exceed the elastic limit of materials.

The PSD design philosophy is straightforward to understand and is adequate for simple load cases, but its trustworthiness is reduced when more than a few interdependent loads are used. For example, consider a simple rotating shaft supported by two bearings. Some of the factors that can influence the maximum stress in the shaft are: variations in the elastic properties of the material, misalignment of the bearings during assembly, shaft imbalancing during operation, presence of stress concentrators and their severity, etc. Moreover, a combination of these factors can be present simultaneously, and a safety factor should be associated to every possible (or at least most common) combination of these influences. The number of necessary safety factors increases significantly for each new aspect considered, and renders this approach unusable for complicated load cases.

2.2 Risk-Based Design

Risk-based design is an early probabilistic approach, developed in the 1950’s, which tries to follow a probabilistic to the previously used heuristic and experience-based method of determining safety factors. A comprehensive summary of the risk-based design concept can be found in (Freudenthal, Garelt, and Shinozuka 1966). In risk-based design, only two variables are considered — the load of the structure, \( S \), and its resistance or capacity \( R \). Both of these variables are assumed to be random, with given means, variances and probability density functions (PDFs). Failure is defined as the condition in which the resistance of the structure is smaller than the applied load, i.e. the ratio

\[ \nu = \frac{R}{S} \]

is smaller than one. In terms of probability of failure, we can estimate it as follows:

\[ P_f = \Pr(\nu < 1) \]

\(^1\)In USA, PSD is more commonly called “allowable stress design” (ASD).
2.3 Limit State Design

Limit state design (LSD) is a design format that incorporates multiple safety factors and provide a more uniform way of treating multiple load cases. LSD has been introduced by various national industry standard organizations during the 1960’s and the 1970’s, for example under the name ultimate strength design in \( \text{ACI 318-63} \), partial safety factor formats (NKB 1978), load and resistance factor design (LRFD) in USA (Ravindra and Galambos 1978), and as limit state design which is adopted as the design methodology of the Eurocodes \( \text{EN 1990:2002} \).

A central concept in LSD is limit state. \( \text{EN 1990:2002} \) supplies the following definition:

\[ \text{[...]} \text{a condition beyond which a structure (or one or more of its constituent members) is no longer able to satisfy the provisions for which it was designed.} \]

Some examples of limit states from structural mechanics are tensile failure, bending fracture, buckling, wearing, surface scratching etc. Distinction is made between ultimate limit states and serviceability limit states. The first applies to conditions that cause, complete loss of load bearing ability of the structure, including structural collapse, loss of equilibrium, loss of connection between structural members, buckling; as well as structural conditions that present a risk for people’s safety. Serviceability limit states are conditions related to structure’s functionality, durability, usability and appearance that do not compromise the load-bearing capacity of the structure, for example cracks, permanent deformations or vibrations (Madsen, Krenk, and N. C. Lind 1986).

Limit states are represented by mathematical models which are functions of various load, resistance and geometric parameters. The general form of a limit state is as follows:

\[ g(s_{id}, r_{jd}, l_{kd}) > 0 \]  \hspace{1cm} (3)

where \( s_{id}, r_{jd} \) and \( l_{kd} \) are the design values of the load, resistance and geometric parameters, respectively. These design values are determined from a set of characteristic values which are commonly the expected (nominal) values of the corresponding actions, and are augmented by a set of factors which take into account various uncertainty sources and their severity. Load factors typically are greater than one to compensate for above-average loads, impact loads, wear etc., while resistance factors are commonly less than zero to account for varying material properties and manufacturing defects.

For example, a bar loaded in tension, the stress caused by the tensile force is the load parameter, while the allowable yield stress is the resistance parameter. The condition of tensile failure (tensile stress exceeding yield stress) is the limit state. This can be expressed mathematically with the well-known relationship

\[ g = \sigma_{all} - \sigma_{l} \geq 0 \]  \hspace{1cm} (4)

In limit state design, by convention \( g < 0 \) signifies violation of the limit state, i.e. structural failure.

The basic definition can be expanded in several ways: the load and resistance factors can be further decomposed into additional factors that account
for multiple sources of uncertainty; additional load combination and importance factors can be assigned to compensate for positive feedback effects from multiple simultaneous actions; inclusion of accidental or progressive limit states.

Drawbacks of the LSD method

The LSD method has merits for attempting to structure the existing approaches in dealing with uncertainties, however it has some significant drawbacks as well:

Proliferation of factors While the LSD method can be efficient and straightforward for small number of actions, the temptation to add yet another influence and factor to the model can lead to unwieldy tables with a large number of parameters. For example, (EN 1990:2002) distinguishes between three different variable action factors: frequent, quasi-permanent and the combination of the two. For buildings, the standard lists four different types of loads with different categories within each, giving a grand total of more than 30 possible factor values, depending on the load type and supplemental conditions. One can see that, taken to the extreme, this approach ceases to be effective due to the effort required from the designer to navigate the requirements and determine the applicable factors and conditions. Once the factors are determined, however, calculating the design values becomes a simple matter.

Handling of dynamic effects There is no inherent mechanism to account for the dynamics of the structure. Dynamic actions are simply augmented with larger load factors which do not take into account the frequency and amplitude of the dynamic loading. This leads to systemic overdesign of the structure to cope with a worst-case scenario which can occur only in a small fraction of the product life cycle.

Stagnancy Once determined, the factors can continue to be used for extended periods of time without revision while new materials, more precise manufacturing methods and construction techniques are introduced. Due to the purely numerical nature of the factors which don’t have any physical relevance, applying the same factors in a new set of circumstances can be cost-inefficient at best, and life-threatening and worst.

Unquantifiable reliability When using safety factor formats, the reliability of the structure cannot be established accurately, since the factors are chosen in such a way as to impart a certain predefined reliability in advance. There is no way to predict the reliability of a structure with values of its design state. Designs can easily be compared one to another, but the lack of a quantifiable reliability estimation promotes usage of higher than necessary safety factors.
3 Reliability-Based Design Optimization

RBDO is a natural extension of deterministic optimization. In performing the latter, a question that often arise is how sensitive is the optimal solution to changes in design parameters. Estimation of sensitivity is generally not possible without performing a prohibitively costly sensitivity analysis. In structural optimization, this often requires performing a number of additional time-consuming finite element simulations, which is not desireable.

The uncertainties of design parameters are not a new phenomenon, and traditionally designers coped with them by using safety factors which are either prescribed in codified standards or are established internally in each company depending on the availability of manufacturing processes and quality control methods. However, safety factors are rarely more than empirical rules of thumb and it is not obvious how they should be applied to new situations and products without previous experience. Moreover, empirical safety factors tend to become inaccurate in time due to the development of new manufacturing techniques and utilization of new materials.

Both over- and underdesign of machine components carry a cost for a company — overdesign is associated with higher manufacturing and operating costs, while underdesign can cause significant raise in service costs and can have immeasureable operational safety implications.

The reliability-based design optimization methods attempt to include the inherent variability of design parameters into the optimization process itself, thus providing a more

Comprehensive studies on RBDO methods (Aoues and Chateauneuf 2010), (Valdebenito and Schuëller 2010) divide RBDO methods into three categories depending on the methods’ structure: double-loop, single-loop approach and decoupled approaches.

3.1 General RBDO formulation

The basic problem of RBDO is posed as a minimization of an objective function under probabilistic and also possibly deterministic constraints.

\[
\begin{align*}
\min_{X} & \quad \mathbb{E}[f(X)] \\
\text{s.t.} & \quad \Pr[g_i(X) \leq 0] \geq P_{Si}, \quad i = 1, \ldots, m
\end{align*}
\]

where \(X\) is a vector of random variables with known distributions, \(g_i(X)\) denotes a set of constraints, each with a prescribed probability of safety \(P_{Si}\). Deterministic constraints and variables can also be treated without special arrangements, but for simplicity they will not be considered in this text.

The presence of the probabilistic constraint precludes use of traditional gradient-based methods.

The actual probability of failure of the structure is given by integrating over the failure region:

\[
\Pr[g_i(X) > 0] = \int_{g_i(X) > 0} \cdots \int f_X(x_1, x_2, \ldots, x_n) \, dx_1 \, dx_2 \cdots dx_n, \quad i = 1, \ldots, m
\]
Here, $f_X$ is the joint probability density function (JPDF) of all design parameters in $x$. The evaluation of this integral is compounded by two factors: for practical problems, the JPDF is either unknown or very difficult to estimate, especially for interdependent variables; the multi-dimensional nature of the integral makes its numerical computation challenging. To overcome these difficulties, analytical approximations of first and second order have been proposed.

The probability constraint in (5) can also be evaluated using the Monte Carlo method (Metropolis and Ulam 1949). The Monte Carlo method involves generating a large number of sampling data and estimating the probability based on these samples. This is typically considered a heavy-handed, brute-force approach, due to the significant computational effort required. In order to guarantee sufficiently good accuracy and minimize sampling error, a large number of experiments need to be performed. For example, probabilities of the order of 99% i.e. failure rate of 1% would require a few hundred samples to establish reasonable bounds on the sample probability, while for more practical reliability levels with failure rates of 0.01% or lower, tens of thousands of sample points are necessary. Additionally, sufficiently accurate modelling of non-uniform probability distributions of the design variables is required to guarantee that each sample set conforms to the desired distribution parameters, which complicates the random number screening procedure. When finite element analysis is used to provide experimental data, a Monte Carlo simulation becomes extremely time-consuming and practically infeasible.

The reliability estimation is only a part, albeit crucial, of the RBDO problem. Accurate reliability estimation is central to the development of successful reliability-based design optimization algorithms.

### 3.2 Early Developments

One of the first major advances in reliability theory was the idea of (Cornell 1969) to express all uncertainties of the random variables only in terms of their first and second moments (means and covariances). The method introduces a failure function (limit state function) which divides the variable space into failure region and safe region. A safety margin $M$ represents the value of the failure function at each point in the design space:

$$M = g(X_i)$$  \hspace{1cm} (7)

and a reliability index $\beta_C$ is defined as the ratio between the expected value of the safety margin and the distance between the failure function and the expected value of the safety margin:

$$\beta_C = \frac{E[M]}{D[M]}$$  \hspace{1cm} (8)

In the spirit of risk-based design, Cornell proposes the failure function as a difference between the resistance and the load on the system:

$$g(r,s) = r - s$$  \hspace{1cm} (9)

which corresponds to safety margin

$$M = R - S$$  \hspace{1cm} (10)
Using general statistical relations, formulas can be derived for uncorrelated load and resistance variables as functions of their means and standard deviations. The Cornell method, while representing a novel approach to reliability, suffers from two major disadvantages: formulation of a linear safety margin is not possible if the failure function is not a hyperplane; and its arbitrariness precludes using the reliability index as a more universal measure of reliability between different problems, or even two different formulations of the same problem.

A logarithmic scaling of the load and resistance variables has been introduced in (Rosenblueth and Esteva 1972) to restrict the variable domain only to positive values, which allows for a more natural treatment of physical variables without introducing additional constraints. The safety margin becomes a nonlinear function in this case, which requires the use of a linearization procedure. This formulation circumvents some of the drawbacks of the reliability index defined by Cornell, but the question of arbitrariness of the failure function still remains, further compounded by the arbitrariness in selecting the linearization point.

For a general limit state as a function of many random variables, a first-order approximation around the mean point can be written as

\[ g(X) \approx \tilde{g}(X) = g(\mu_X) + \sum_{i=1}^{n} \frac{\partial g}{\partial X_i} (X_i - \mu_X) + \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \frac{\partial^2 g}{\partial X_i \partial X_j} (X_i - \mu_X) (X_j - \mu_X) + \cdots \]  

(11)

If we collect only the linear terms, we can estimate the first-order approximate mean and variance of the limit state function as

\[ \mu_{\tilde{g}} \approx \tilde{g}(\mu_{X_1}, \mu_{X_2}, \ldots, \mu_{X_n}) \]

\[ \sigma^2_{\tilde{g}} \approx \sum_{i=1}^{n} \sum_{j=1}^{n} \frac{\partial g}{\partial X_i} \frac{\partial g}{\partial X_j} \text{Cov}(X_i, X_j) \]  

(12)

This is possible due to the properties of the normal distribution — linear combinations of normally distributed variables are also normally distributed, and their means and variances are calculated as shown above. For uncorrelated variables, the standard deviation is simply

\[ \sigma_{\tilde{g}} \approx \sqrt{\sum_{i=1}^{n} \left( \frac{\partial g}{\partial X_i} \right)^2 \sigma^2_{X_i}} \]  

(13)

where \( \beta \) is the reliability index of the design and is calculated as the ratio between the mean and the standard deviation of the limit state:

\[ \beta = \frac{\mu_{\tilde{g}}}{\sigma_{\tilde{g}}} \]  

(14)

The corresponding probability constraint is then estimated as follows:

\[ \text{Pr}[g(X \leq 0)] \approx \Phi \left( \frac{0 - \mu_{\tilde{g}}}{\sigma_{\tilde{g}}} \right) = \Phi(\beta) \]  

(15)
This approximation of the probability of failure using only first-order information as well as the first and second moments of the probability distributions of the design variables is known as first-order reliability method (FORM). If the Taylor expansion is performed around the mean point of the limit state, the method bears the name mean value first-order second-moment method (MVFOSM). A significant drawback of MVFOSM is that no information on the actual distribution of the random variables is used. For linear limit states, the estimated reliability is accurate, but for non-linear failure functions the omission of high-order terms can introduce a significant error.

The first-order approximations of the probability of failure in (15) is only appropriate when the limit state is linear or close to linear in the region of the current design point: general non-linear limit states cannot be modelled appropriately using only linear terms. Approximations which incorporate second-order information, called second-order reliability method (SORM), were first investigated by (Fiessler, Neumann, and R. Rackwitz 1979). A closed-form solution was later provided by (Breitung 1984), as follows:

\[
\Pr[g(X \leq 0)] \approx \Phi(\beta) \prod_{i=1}^{n-1} \frac{1}{\sqrt{1 + 2\beta \kappa_i}} \tag{16}
\]

where \( \kappa_i \) are the principal curvatures of the limit state at the MPP, and \( \beta \) is the first-order reliability index obtained using FORM. The principal curvatures are obtained by first performing an affine rotation around the origin of the design space, so that it coincides with the gradient vector \( \alpha \) of the limit state at the MPP. The rotation matrix \( R \) is calculated by orthonormalising of the matrix \( R_0 \), defined as:

\[
R_0 = \begin{bmatrix}
1 & 0 & \cdots & 0 \\
0 & 1 & \cdots & 0 \\
\vdots & \vdots & \ddots & \vdots \\
\alpha_1 & \alpha_2 & \cdots & \alpha_n
\end{bmatrix} \tag{17}
\]

The principal curvatures are the eigenvalues of the matrix \( A \), whose elements \( a_{ij} \) are computed as

\[
a_{ij} = \frac{(RGR^T)_{ij}}{||\nabla^2 g(z^*)||}, \quad i = 1, \ldots, n - 1
\]

\[
G = \nabla^2 g(z^*)
\]

The Breitung approximation includes only parabolic terms — mixed second-order terms and their derivatives are ignored. SORM estimates using this formula are generally only accurate for high target reliability. An alternative formulation has been proposed by (Hohenbichler et al. 1987), as follows:

\[
\Pr[g(X \leq 0)] \approx \Phi(\beta) \prod_{i=1}^{n-1} \frac{1}{\sqrt{1 + 2\Phi(\beta) \kappa_i}} \tag{19}
\]

Another method is proposed by (Tvrdt 1990) which uses both parabolic and a second-order approximation of the limit state, but without using asymptotic approximations. This method circumvents the low accuracy of the Breitung formula at low reliability levels, while still remaining accurate at high target reliabilities.
3.2.1 Hasofer-Lind Reliability index

Hasofer and Lind (1974) introduced a method that builds on the ideas of Cornell, particularly the use of the reliability index and the first and second moments only to estimate the reliability. An important feature of the proposed reliability criterion is the invariance of the reliability index under equivalent transformations of the design variables. Hasofer and Lind introduce a nonhomogeneous linear mapping between the original set of random variables $X_i$ onto a new set $Z_i$ of normalized and uncorrelated variables. Formally, this means that the expected values of the new variables are zero, and the covariance matrix is the unit matrix:

$$E[Z] = 0$$
$$\text{Cov}(Z, Z^T) = I$$

In the newly-constructed design space of variables $Z_i$, the geometrical distance between the origin and the transformed failure surface $\tilde{g}(Z)$ corresponds to the distance (in standard deviation units) from the mean point in $x$-space to $g(X)$. In other words,

$$\beta_{HL}(z) = \sqrt{z^T z}$$

(20)

For normal variables, this is readily achieved by transforming the variables into standard normal form with zero mean and unity standard deviation:

$$Z_i = \frac{X_i - \mu_{X_i}}{\sigma_{X_i}}, \quad i = 1, \ldots, n$$

(21)

For other distributions, however, such transformations are not possible, which puts a severe restriction on the method’s applicability. Still, near-equivalent transformations to normally distributed uncorrelated variables are possible, for example the two-parameter equivalent normal transformation due to Rackwitz and Fiessler (1976). With this transformation, the cumulative distribution function (CDF) and the PDF are made equal at the design point on the failure surface. While appropriate for symmetric distributions, the Rackwitz-Fiessler method is increasingly inaccurate for highly-skewed distributions such as extreme value distributions like Gumbel, Fréchet or Weibull.

The point on the failure surface which corresponds to the reliability index is called design point or most probable point (MPP). When the failure surface is a hyperplane, the reliability indices of Cornell and Hasofer-Lind coincide. The MPP can be selected as a linearization point, and this approach is commonly known as advanced first-order second-moment method (AFOSM) in literature.

3.2.2 Finding the Most Probable Point

Using equation (20), the most probable point is the point on the failure surface closest to the origin in the reduced coordinates. Formally, this corresponds to solving the following optimization problem:

$$\min_z \sqrt{z^T z}$$
$$\text{s.t. } \tilde{g}(z) = 0$$

(22)
Initially, an iterative algorithm was devised for solving such problems, due to Rackwitz and Fiessler (1978), later expanded by Ditlevsen (1981) to cover dependent variables. The algorithm is simple to implement and understand, and it uses first-order derivative information to calculate the point of next iteration. However, its convergence is not guaranteed, and standard constrained optimization algorithms can be successfully used instead.

### 3.2.3 Transformation in standard normal space

The Hasofer-Lind reliability index is defined only for normally distributed design variables; the transformation to standard normal space is not valid for other probability distributions. The Rosenblatt transformation (Rosenblatt 1952) can be used to obtain a set of uncorrelated variables with standard normal distribution, provided that the joint CDF of all original design variables is available.

Alternative methods, such as Rackwitz-Fiessler method (Rüdiger Rackwitz and Fiessler 1976), can be used to provide a two-parameter equivalent transformation into standard normal variables. With this algorithm, the CDF and the PDF at the current design point $x^* = \{x_i^*\}$ should be equivalent in the actual and the normalized variables:

$$
\Phi \left( \frac{x_i^* - \mu_{X_i}^N}{\sigma_{X_i}^N} \right) = F_{X_i}(x_i^*), \quad i = 1, \ldots, n
$$

which results in the following expressions for the mean and standard deviation of the normalized variables:

$$
\begin{align*}
\mu_{X_i}^N &= x_i^* - \Phi^{-1} (F_{X_i}(x_i^*)) \sigma_{X_i}^N \\
\sigma_{X_i}^N &= \frac{\phi (\Phi^{-1} (F_{X_i}(x_i^*)))}{f_{X_i}(x_i^*)}
\end{align*}
$$

This transformation is fairly accurate for close-to-symmetric probability distributions, but the estimates of the mean and standard deviation are often not satisfactory for highly skewed or naturally bounded distributions, such as Fréchet or beta distributions. Extensions of the Rackwitz-Fiessler algorithm have also been published, for example the three-parameter transformation due to Chen-Lind (X. Chen and N. Lind 1982), which in addition to the mean and standard deviation also uses a scale factor that equates the slopes of the original and the transformed distributions at the design point.

### 3.3 Double-loop approaches

Nikolaidis and Burdisso (1988) utilized the idea of Hasofer and Lind and proposed an iterative nested optimization procedure to solve the RBDO problem as stated in (5). The inner optimization loop estimates the reliability at the current solution, while the outer loop optimizes the overall cost function using the reliability calculated in the inner loop. The authors used both MVFOSM and AFOSTM for reliability assessment in the inner loop and ultimately concluded that the latter is more accurate and robust. This approach is commonly referred in the literature as reliability index approach (RIA).
The reliability index method, while capable of solving the RBDO problem, was later found to be slowly convergent, numerically unstable and computationally expensive, especially at high reliability levels. Tu and Choi (1999) introduced the performance measure approach (PMA) which performs inverse reliability analysis, under the assumption that it is simpler to optimize a complex function with simple constraints than vice versa.

The method is shown to be more robust and efficient in a comparative study performed by (J.-O. Lee, Y.-S. Yang, and Ruy 2002). The main difference between RIA and PMA is the formulation of the MPP problem. While in RIA the MPP is found by minimizing the distance to the failure surface, in PMA the objective and the constraint have been interchanged:

\[
\min_{z} g(z) \\
\text{s.t. } ||z|| = \beta_a
\]  
(25)

The constraint always has the shape of a hypersphere, which simplifies significantly the optimization problem as specialized algorithms for solving them are devised.

### 3.4 Single-loop approaches

Double-loop approaches require significant computational effort due to the nested optimization and probabilistic loops, and efforts to increase the performance of RBDO algorithms lead to the work of (Xiaoguang Chen, Hasselman, and Neill 1997), which replaces the probabilistic constraints with approximate deterministic ones. Thus, the inner probabilistic loop is eliminated, resulting in a more efficient algorithm. This method is known in the literature as single-loop, single-vector approach (SLSV). A similar work has been published previously by (Thanedar and Kodiyalam 1992). However, the method was shown to be inadequate for non-linear problems due to the linear assumption of FORM (R. Yang and Gu 2004).

**Single-loop approaches based on Karush-Kuhn Tucker conditions (KKT)**

Parallel to the development of SLSV, (Madsen and Hansen 1992), and later improvements by (Kuschel and Rüdiger Rackwitz 1997), approached the problem from a different perspective — the reliability analysis is replaced by the corresponding KKT optimality conditions, thus allowing the use of standard optimization algorithms. Two formulations of RBDO problems are suggested, which differ somewhat from the problems typically studied: total cost minimization (CRP) and reliability maximization (RCP). The CRP has a general formulation as follows:

\[
\min_{u} \ C_t(p) = C_i(p) + C_f(p)\Phi(- ||u||) \\
\text{s.t. } \begin{cases} 
  g(u, p) = 0 \\
  u^T \nabla g(u, p) + \|u\| \cdot \|\nabla g(u, p)\| = 0
\end{cases}
\]  
(26)

where \(C_t, C_i, C_f\) are the total, initial and failure cost respectively; \(u\) is a transformation of the original vector \(X\) of random variables into uncorrelated variables with standard normal distribution, and \(p\) is a vector of deterministic variables.
and distribution parameters. Additional constraints on the design parameters can also be specified as necessary.

While the method does away with the inner optimization loop, it requires calculation of second-order derivatives of the constraints, which can be computationally costly and inaccurate. Only FORM assumption of the probability of failure is made, with authors claiming that expansion to second-order reliability method can be problematic. The probabilistic transformation from $X$ to $u$ needs to be specified explicitly, with its associated approximation errors.

However, the probabilistic transformation need to be specified explicitly, and while only first-order optimality conditions are used, second-order derivatives of the constraints are required as well. The authors also observe sensitivity to initial values, and the necessity of monotonic transformation of the objective and the constraints to improve convergence.

This work is later extended in (Kuschel and Rüdiger Rackwitz 2000) to solve optimization problems with time-variant constraints, although under the assumption of Poisson failure process and limited to first-order reliability estimation.

In a related paper, (Agarwal et al. 2007) applied the same first-order KKT necessary conditions, but this time replacing the RIA-formulated reliability loop used in (Kuschel and Rüdiger Rackwitz 2000) with PMA. The authors observe improved computational efficiency and robustness compared to double-loop approaches.

Improvements to the SLSV were suggested in (Liang, Mourelatos, and Nikolaidis 2007), which relates the approximations to the KKT conditions of the inverse FORM. The suggested optimization problem to solve looks like

$$\begin{aligned}
&\min_{\mu_X} \quad f(\mu_X) \\
&\text{s.t.} \quad g_i(x^k) \leq 0, \quad i = 1, \ldots, m \\
&\quad x^k = \mu^k_X - a^k_i \sigma_X \beta^T \\
&\quad a^k_i = \left\{ \sigma_X \nabla g_i(x^{k-1}) \right\} \\
&\quad \sigma_X = \{\sigma_X\}
\end{aligned}$$

This approach is commonly referred to as single-loop approach (SLA). No MPPs are calculated in this formulation — by solving the KKT conditions, an approximation of the MPPs of the active constraints is used. The KKT conditions involve the normalized sensitivities $a_i$ which are approximate by the MPP estimate from the previous iteration. One of the major differences with the approach of (Xiaoguang Chen, Hasselman, and Neill 1997) is that the normalized sensitivities are calculated exactly instead of being approximated or assumed.

The authors recognize a drawback of SLA, particularly problems with converging to the global optimum, if the limit states are very non-linear. Since this approach uses approximations of MPPs instead of actually calculating them, it is more sensitive to non-linearities than double-loop methods.

Another method that can be classified as single-loop approach is the procedure outlined in (Shan and G. G. Wang 2008). In contrast to other methods, this algorithm first determines a reliable design space in which the reliability
constraints are enforced throughout, and subsequently performs a determin-
istic optimization within this constrained space. The gradients at the current
solution are used as approximation for the gradients at the MPP, thus eliminat-
ing the reliability analysis altogether. This has the advantage of not needing a
specialized optimization algorithm which embeds reliability analysis, and stan-
dard deterministic optimization methods can be used instead. The approach is
shown to work for a set of known analytical benchmarks, however its accuracy
can be suboptimal for very non-linear problems.

3.5 Decoupled approaches

Reliability analysis has proven to be the bottleneck of most RBDO methods, and
which spurred the development of a class of methods that avoid performing
reliability analysis at each step of the optimization. Those methods are com-
monly referred to as decoupling approaches. A key feature of all decoupling
approaches is the deterministic approximation of the reliability index.

Sequential optimization and reliability assessment (SORA), introduced in
(Xiaoping Du and Wei Chen 2004), performs reliability analysis and determinis-
tic optimization in two alternating cycles in each optimization step. In the first
cycle, the deterministic solution of the optimization problem is first obtained,
using standard optimization procedures. Since the reliability requirements of
most constraints will be likely violated, in the second cycle they are shifted
into the probabilistic feasible region by means of a shifting vector

$$ s = \mu_X - x_{MPP} $$ (28)

The inverse MPPs of the constraints are calculated after the second cycle using
the method of (Tu, Kyung K. Choi, and Park 1999), and since the feasible space
has been reduced, the reliabilities of the violated constraints are expected to
improve. The method formulation can be more formally expressed as follows:

$$ \min_{\mu_X} f(d, \mu_X) $$

$$ \text{s.t. } g(d, u_X - s) \leq 0 $$ (29)

The two cycles are repeated until the objective converges and the shifting
distances become zero, i.e. the reliability requirements are satisfied. In this
formulation, the design variables are assumed to be normally distributed and
uncorrelated, however arbitrary distributions can be used provided that a
transformation to standard normal random variables is available.

An extension of SORA has been suggested by (Cho and B. Lee 2010), using
method of moving asymptotes (MMA) and convex linearization (CONLIN)
approximations instead of linear Taylor expansion. CONLIN, developed by
(Fleury and Braibant 1986), and MMA, first presented by (Svanberg 1987) and
further extended in (Svanberg 2002), are adaptive approximation schemes
which can utilize the specific structure of the problem and linearize the ob-
jectives and constraints with respect to the reciprocals of the design variables.
Cho et al. tested the method on various types of problem and its accuracy and
numerical performance match or exceed those of SORA.
Approximating reliability index  A class of decoupled approaches forego the reliability analysis altogether by utilizing an approximation of the reliability index. An early attempt in this direction can be found in (Royset, Kiureghian, and Polak 2001), where the reliability index is replaced by its approximation using deterministic functions.

This class of methods utilize a direct constraint on the reliability index rather than a probabilistic constraint:

\[
\begin{align*}
\min_{d} f(d) \\
\text{s.t.} \quad \{ \beta_t \leq \beta(d) \} \\
\end{align*}
\]

(Chandu and Grandhi 1995) proposed a first-order approximation of the reliability index based on linear Taylor expansion:

\[
\beta_k(d) \approx \beta(d^{k-1}) + (\nabla \beta(d^{k-1}))^T (d - d^{k-1})
\]

where the sensitivities of the reliability index are calculated according to (L. Wang, Grandhi, and Hopkins 1995). However, this method of approximating the reliability index is not very efficient and holds little advantage over traditional double-loop methods. (Cheng, Xu, and Jiang 2006) proposed a sequential approximate programming (SAP) strategy which provides an approximation of the reliability index and its gradient derived from the necessary KKT conditions:

\[
\begin{align*}
\beta_k(d) & \approx \hat{\beta}(d^{k-1}) + (\nabla \hat{\beta}(d^{k-1}))^T (d - d^{k-1}) \\
\lambda^{k-1} & = \frac{1}{\| \nabla g(d^{k-1}, u^{k-1}) \|} \\
\hat{\beta}(d^{k-1}) & = \lambda^{k-1} \left( g(d^{k-1}, u^{k-1}) - (u^{k-1})^T \nabla g(d^{k-1}, u^{k-1}) \right) \\
u^k & = -\hat{\beta}(d^{k-1}) \lambda^{k-1} \nabla g(d^{k-1}, u^{k-1})
\end{align*}
\]

A similar approach using an active set strategy has been proposed by (Zou and Mahadevan 2006).

A sequential linear programming approach has been suggested by (Chan, Skerlos, and Papalambros 2007). At each iteration, a new step \( s \) is calculated within a trust region, which is used to obtain the new design point

\[
\mu_\chi^{k+1} = \mu_\chi + s
\]

The constraints can be approximated using either FORM or SORM assumptions; in case of FORM, the LP subproblem has the form

\[
\begin{align*}
\min_{s^k} f^T(\mu_\chi^k) \cdot s^k \\
\text{s.t.} \quad \{ g^k(x_{MPP}^k) + (\nabla g^k(x_{MPP}^k))^T \cdot s^k \leq 0 \} \\
\end{align*}
\]

where

\[
x_{MPP}^k = \mu_\chi^k + \sigma_\chi \beta_i \frac{\nabla g_i}{\| \nabla g_i \|}
\]
The MPPs are derived using KKT conditions in a similar fashion as in (Liang, Mourelatos, and Nikolaidis 2007). SORM-based SLP uses the principal curvatures $\kappa$ to estimate the reliability index.

### 3.6 Comparative studies

There is no shortage of RBDO methods developed in the last decade, and while there have been attempts to categorize and compare different approaches to find the “best” one, there is currently no method that is recognized as superior in all situations. The main reason for this is that RBDO methods are often developed to solve a particular type of problems, factoring number and distributions of design variables, linearity of the problem, system vs. component reliability, static vs. stochastic loading, large or small number of limit states, etc. Direct, objective comparison between methods is thus problematic. Still, a number of studies attempted independent tests of various RBDO approaches. (R. Yang and Gu 2004) provided an early study comparing SORA and SLSV to double-loop approaches for a variety of analytical and structural optimization problems. In this study, SORA and SLSV consistently showed superior performance and accuracy compared to traditional double-loop methods.

In recent years, (Aoues and Chateauneuf 2010) provided a comparison between the most well-known methods currently in use, including RIA, PMA, KKT-based methods, SLA, SORA and SAP. A variety of problems with distinct characteristics were shown. SLA, SORA and SAP show overall good performance and accuracy for problems with nonlinear limit states, large number of design variables and non-normally distributed random parameters. Other methods were shown to perform more favourably in certain situations — KKT converges quickly for problems with small number of variables and constraints, but can be numerically unstable and inefficient for problems with non-normally distributed variables; PMA and RIA have adequate efficiency when limit state sensitivities are given analytically, and work well with variables with non-Gaussian distribution, but are generally the slowes, sometimes by a factor of three or more.

Another comprehensive study comparing various methods that appeared recently is due to (Valdebenito and Schuëller 2010). The authors assert that objective testing of RBDO algorithms is problematic and instead discuss various aspects which differentiate the methods, such as dimensionality of the design variables and uncertain parameters, application of meta-modelling, component vs. system reliability.
4 Metamodels

Computer models of various processes have been used extensively in the last few decades — for example, finite element simulations, climate models, molecular models, etc. Such abstractions have proven their accuracy in prediction of the properties of the observed process, however obtaining results from utilizing these models can be time-consuming and computationally intensive. For instance, a typical finite element simulation in industrial setting can run from an hour to a week — clearly undesirable situation if a large number of simulations need to be carried out. For this reason, metamodels (i.e. “models of models”) have been developed. Metamodels (also referred sometimes as “surrogate models”) are not phenomenological, but purely mechanistic — instead of simulating the actual physical process, they mimic its behaviour in a limited domain, for which they have been deemed sufficiently accurate. Metamodels are typically very fast to evaluate, and are a valuable tool when a large number of results are desired. Process optimization also benefits greatly from using metamodels.

Response surface method (RSM) is a popular metamodeling technique. According to (Roux, Stander, and Haftka 1998), response surface methodology is defined as

\[ y = \beta_0 + \sum_{j=1}^{k} \beta_j x_j + \epsilon \] (35)

Here, \( \epsilon \) is the regression error term with assumed \( \text{E}(\epsilon) = 0 \) and \( \text{Var}(\epsilon) = \sigma^2 \), and uncorrelated components.

Given \( m \) observations, we can write an expression for each experiment using (35):

\[ y_i = \beta_0 + \sum_{j=1}^{k} \beta_j x_{ij} + \epsilon_i, \quad i = 1, \ldots, m \] (36)
We can write (36) in matrix form as follows:

\[
y = X\beta + \epsilon
\]

\[
y = \begin{bmatrix}
y_1 \\
y_2 \\
\vdots \\
y_m
\end{bmatrix}, \quad X = \begin{bmatrix}
1 & x_{11} & x_{12} & \cdots & x_{1k} \\
1 & x_{21} & x_{22} & \cdots & x_{2k} \\
\vdots & \vdots & \vdots & \ddots & \vdots \\
1 & x_{m1} & x_{m2} & \cdots & x_{mk}
\end{bmatrix}
\]

\[
\beta = \begin{bmatrix}
\beta_1 \\
\beta_2 \\
\vdots \\
\beta_k
\end{bmatrix}, \quad \epsilon = \begin{bmatrix}
\epsilon_1 \\
\epsilon_2 \\
\vdots \\
\epsilon_m
\end{bmatrix}
\]

Minimizing the deviation of the regression model from the actual response at the design points is equivalent to minimizing \(\epsilon_i\) in the least squares sense. The least squares estimator of \(\beta\) and the fitted regression model can then be expressed by

\[
b = (X^TX)^{-1}X^Ty \quad \hat{y} = Xb
\]

The least squares estimator is unbiased — its expected value is equal to the regression coefficients \(\beta\). The covariance of \(b\) can be expressed as

\[
\text{Cov}(b) = \sigma^2(X^TX)^{-1}
\]

where an unbiased estimator of \(\sigma^2\) is given by

\[
\hat{\sigma}^2 = \frac{SS_E}{m-k} = \frac{\sum_{i=1}^{m}(y_i - \hat{y}_i)^2}{m-k}
\]

Quadratic and higher-order models are also used, however the increased number of interactions between variables requires more regression coefficients to be calculated and more experiments to create an adequate model.

A comprehensive historical perspective of the development of RSM can be found in (Myers, Khuri, and Carter Jr. 1989). An excellent introduction to RSM and DoE can be found in (Myers and Montgomery 2002).

4.1 Design of experiments

One of the most important elements that influence the accuracy and efficiency of the regression model is how the design points are selected in the feasible design space. Important considerations when building response surface models are coverage and size, as the number of experiments one is allowed to perform is typically constrained by cost and/or available time. From time and cost perspective, it is advantageous to minimize the number of experiments — physical tests require significant effort to set up and execute, and a single numerical simulation of a large model can easily take from one hour to one week or more wall clock time. However, large samples allow building more complex and accurate models, and have generally better predictive capabilities than smaller samples.
Efficient designs of experiments have a number of desirable properties that affect their accuracy and prediction power:

- **Prediction variance.** The variance of a predicted response is given by

\[
\text{Var}[\hat{y}(x)] = \xi(x)^T(X^T X)^{-1}\xi(x) \cdot \sigma^2 \tag{41}
\]

where \(\xi(x)\) is the selected basis of the model as a function of the design points. This makes the prediction variance an explicit function of the selected experimental design.

The scaled prediction variance is useful when comparing various sampling schemes, as it is dimensionless:

\[
v(x) = \frac{N\text{Var}[\hat{y}(x)]}{\sigma^2} \tag{42}
\]

A stable prediction variance (rather, its scaled version) is preferable feature of experimental designs, as it provides assurance that predicted responses would be similarly accurate regardless of the position of the measured point in the design space.

- **Orthogonality.** For a design to be orthogonal, the scalar product of any two column vectors in the design must evaluate to zero; in other words, the information matrix \(X^T X\) becomes diagonal. This minimizes prediction variance over the regression coefficients when using linear regressions.

- **Rotatability.** The concept of rotatability is introduced in (Box and Hunter 1957) and refer to a design for which the prediction variance is constant at any point situated at the same distance from the design center. Thus, spherical designs are inherently rotatable and thus preferable.

**Rule-based designs**  Factorial designs are simple and straight-forward to implement — a few (two or three are most common) levels of each variables are selected, and a DoE capturing all possible combinations between all levels of design variables are created. This type of designs is suitable for screening experiments to identify the most important variables in the modelled process. \(2^k\) designs are suitable for fitting a first-order model, and are often the basis for more complex experiment setups.

While adequate for small number of variables, the number of experiments grow exponentially with number of variables: the number of necessary design points \(m\) can be calculated using the expression

\[
m = \prod_{i=1}^{n} l_i \tag{43}
\]

where \(l_i\) is the number of levels for each design variable and \(n\) is the number of design variables. For example, a \(2^k\) design of ten variables, each with two levels, would require 1024 experiments, which often is prohibitively expensive and time-consuming for practical problems. For this reason, more compact designs of experiments have been devised through the years that scale better with increasing the number of variables.
One instance where factorial designs of low order are particularly valuable is screening experiments — eliminating factors with negligible effect on the response. This allows exclusion of the associated design variables from further investigations, which yields simpler and faster to construct and evaluate metamodels.

Central composite designs (CCDs), introduced in (Box and Wilson 1951), is a sequential scheme which combines a full two-level factorial designs with axial and central runs to create a spherical coverage of the design domain. According to (Myers and Montgomery 2002, p. 321), this design is one of the most used second-order schemes. The rationale for this kind of design is to provide a stable prediction variance through the use of factorial design, and at the same time give information about coupling effects between variables. An example CCD matrix for three design variables is shown below:

\[
D = \begin{bmatrix}
-1 & -1 & -1 \\
-1 & 1 & 1 \\
1 & -1 & -1 \\
1 & -1 & 1 \\
1 & 1 & -1 \\
1 & 1 & 1 \\
-\alpha & 0 & 0 \\
\alpha & 0 & 0 \\
0 & -\alpha & 0 \\
0 & \alpha & 0 \\
0 & 0 & -\alpha \\
0 & 0 & \alpha \\
0 & 0 & 0
\end{bmatrix}
\] (44)

The model has an adjustable parameter, \( \alpha \), which controls the extents of the axial runs. \( \alpha \) is commonly set to achieve spherical, thus rotatable design.

Face-centered designs (FCDs) are similar to CCD, with the design spanning a cube rather than a sphere, i.e. the parameter \( \alpha \) is set to unity. These are applicable when the design variables have strict upper and lower limits.

Box-Behnken designs (BBDs), due to (Box and Behnken 1960), feature three levels of the design variables and are suitable for fitting second-order response surfaces. Box-Behnken designs explore all possible combinations of the extremes of the design variables, taken two at a time, while the remaining variables are set at their mid-level. A Box-Behnken design of three normalized
variables is given below:

\[
D = \begin{bmatrix}
-1 & -1 & 0 \\
-1 & 1 & 0 \\
1 & -1 & 0 \\
1 & 1 & 0 \\
-1 & 0 & -1 \\
-1 & 0 & 1 \\
1 & 0 & -1 \\
1 & 0 & 1 \\
0 & -1 & -1 \\
0 & -1 & 1 \\
0 & 1 & -1 \\
0 & 1 & 1 \\
0 & 0 & 0
\end{bmatrix}
\]  

(45)

BBDs occupy spherical region of interest, and as such, the scaled prediction variance \( N \text{Var}[\hat{y}(x)]/\sigma^2 \) is constant. This is a desirable feature of spherical designs, as it provides some assurance of how good \( \hat{y}(x) \) is as a response prediction. However, if the corners of the design space are of interest, BBD is not an appropriate choice, as there are no runs that cover all variables at their extremes.

**Compact designs** A class of compact designs are also available, for situations where every experiment counts — these designs feature adequate coverage of the design domain, but with reduced density and prediction accuracy. 

**Small composite designs (SCDs)** are a variation of CCD, but without a full factorial design. Augmented and central runs are used in the same fashion as in CCD. A symmetry in selecting which \( 2^k \) runs to include in the design is desirable to stabilize the prediction variance. A compelling feature of SCD is the possibility for sequential experimentation — after the factorial part of the design is executed, the remaining runs can be foregone if a linear model is a sufficiently good fit.

Koshal designs, presented in (Koshal 1933), examine the response of the design variables taken only one at a time. This leads to a compact design, but with limited possibilities of estimating interaction terms.

First-order Koshal designs follow the structure below:

\[
D = \begin{bmatrix}
0 & 0 & \cdots & 0 \\
1 & 0 & \cdots & 0 \\
0 & 1 & \cdots & 0 \\
\vdots & \vdots & \ddots & \vdots \\
0 & 0 & \cdots & 1
\end{bmatrix}
\]  

(46)

Hybrid designs (Roquemore 1976) combine central composite designs of reduced order with levels of the remaining variable selected in a way that produces near-rotatable and near-orthogonal design. The number of experiments is close to the minimum required to establish the desired model. While efficient, these designs are nearly saturated, which leaves little room for compensating for lack of fit of the chosen regression model. Another drawback
is the seemingly random levels at which the design variables are measured. A design with three variables and ten experiments (dubbed hybrid 310 by Roquemore) is given below:

\[
D = \begin{bmatrix}
0 & 0 & 1.2906 \\
0 & 0 & -0.1360 \\
-1 & -1 & 0.6386 \\
1 & -1 & 0.6386 \\
-1 & 1 & 0.6386 \\
1 & 1 & 0.6386 \\
1.736 & 0 & -0.9273 \\
-1.736 & 0 & -0.9273 \\
0 & -1.736 & -0.9273 \\
0 & 1.736 & -0.9273
\end{bmatrix}
\] (47)

**Optimality criteria designs**  Rule-based designs are generally efficient for regular design spaces where all combinations of variable values between their minimum and maximum are accepted. However, these schemes do not work for irregular regions of interest, or in situations where there is a hard limit on the number of experiments and efficient use of the available tests is desirable. A set of optimality criteria, sometimes referred to as alphabetic optimality criteria, are used. These criteria aim at reducing the variance of the regression parameters or the minimization of the covariance matrix \( \text{Cov}(\beta) \), given by

\[
\text{Cov}(\beta) = \sigma^2 (X^T X)^{-1}
\] (48)

In all optimality criteria that follow, the design space is assumed to be normalized to the unit cube.

**D-Optimality**  This criterion minimizes the generalized variance of the model coefficient estimates, so that they Span the largest volume possible in the design space. An equivalent description is minimizing the volume of the confidence regions to obtain the actual design parameters. This is achieved by maximizing the determinant of the information matrix \( M = X^T X \):

\[
\max_x \det M(x)
\] (49)

Related to D-optimality is the D-efficiency, which measures the relative volume span of the DoE to the total design space:

\[
D_{eff} = \frac{\det (M)^{1/p}}{k}
\] (50)

where \( p \) is the rank of \( M \). Normalized design variables yield maximum D-efficiency of unity. This measure is useful to compare the quality of various DoE schemes. D-optimality is one of the best known and often used criteria, and has been thoroughly studied, see for example (St. John and Draper 1975) for an early review on D-optimal designs and related practical considerations, as well as (Myers, Khuri, and Carter Jr. 1989).
D-optimal designs inherently depend on the type of regression model used, and are prone to biasing if inadequate regression models are used; additionally, have the tendency to have repeated experiments at exactly the same design points, especially for non-regular design domains. The latter issue is particularly problematic — while repeated trials are acceptable for physical testing situations, such experiments are wasteful in cases of fully deterministic numerical simulations. A Bayesian modification has been proposed by (Du-Mouchel and B. Jones 1994) and recently investigated by (Magnus Hofwing and Niclas Strömberg 2009), which adds higher-order terms to the regression model to reduce occurrence of duplicates. As D-optimality is a highly non-linear problem, algorithms to construct D-optimal designs do not converge to the global optimums. Search strategies based on simulated annealing and genetic algorithms have been successfully applied, see for example (Broudiscou, Leardi, and Phan-Tan-Luu 1996).

A-Optimality  Minimize the average variance of the model coefficient estimates without taking into consideration covariances among regression coefficients — that is, only the trace of the information matrix $M$ is considered. Formally, A-optimality is defined as

$$\min_x \text{tr}[M(x)]^{-1}$$ (51)

S-Optimality  S-optimal designs maximize the geometric mean of the distances between nearest neighbours of the design points:

$$\max_x \frac{\sum_{i=1}^{n} \min |x_i - x_j|}{n}$$ (52)

Unlike other optimality criteria, S-optimal designs are independent on the assumed regression model. S-optimal designs inherently do not produce duplicates, however finding the global optimum or even a satisfactory solution can be difficult due to the presence of multiple local optima. (Magnus Hofwing and Niclas Strömberg 2009) applied a hybrid genetic algorithm together with sequential linear programming (SLP) to find S-optimal designs in irregular design domains.

I-Optimality  Minimize the average of the expected variance (taken as an integral) over the region of prediction:

$$\min_x \frac{1}{K} \int_{\Omega} f^T(x) M^{-1} f(x) w(x) \, dx, \quad K = \int_{\Omega} \, dx$$ (53)

where $f(x)$ is a row from the design matrix $X$ and $w(x)$ is a weighting function. This formulation is equivalent to the following optimization problem:

$$\min_x \text{tr}[M^{-1} B], \quad B = \frac{1}{K} \int_{\Omega} f(x)f^T(x)w(x) \, dx$$ (54)
Given the trace operator and the integral in the definition above, direct application of standard optimization packages is difficult. Zeroth order optimization algorithms, for example simulated annealing in (Haines 1987), are commonly applied.

**Prediction variance criteria** A class of optimality criteria focus on measures directly related to controlling the prediction variance by limiting its maximum or average over the whole design domain, a part of it, or only using a subset of the available design points.

**G-Optimality** G-optimal designs minimize the maximum prediction variance is minimized:

\[
\min_{\xi} \left[ \max_{v \in R} v(x) \right]
\]  

(55)

**V-Optimality** In V-optimality, the goal is to minimize the average prediction variance over a set of points of interest, which may not necessarily be the design points used to construct the response surface.

**Q-Optimality** In similar vein to V-optimality, Q-optimal designs minimize the average prediction variance, but instead of performing this on the whole design space, a region of interest \( R \) is used instead. The Q-optimal design is given by

\[
\min_{x} \frac{1}{K} \int_{R} v(x) \, dx, \quad K = \int_{R} dx
\]  

(56)

**Space-filling designs** As an alternative to alphabetical optimality criteria, other heuristic sampling-based methods can be used. Latin hypercube sampling (LHS) is a popular choice as it could be both efficient and easy to construct, as evidenced by a comparison study between random, stratified and Latin hypercube sampling found in (McKay, Beckman, and Conover 1979) which first introduced the concept. LHS designs show less estimate variance compared to simple random sampling, as detailed in (Stein 1987). Multiple LHS-conformant designs can exist for the majority of problems, and choosing an “optimal” among those has been the goal of multiple studies: (Palmer and Tsui 2001) presented a LHS scheme which minimizes the integrated square basis over the region of interest; (Morris and Mitchell 1995) describe a maximin-optimized LHS which maximizes the minimum distance between neighbour points in the DoE; (Tang 1993) utilize orthogonal arrays to control the distribution of points in the design domain.

### 4.2 Extensions of RSM

The basic response surface method is a powerful modelling tool which has proven its usefulness in a number of practical applications. However, it is not always accurate or reliable enough for very non-linear problems. Several methods have emerged as extensions to RSM, notably Kriging method, radial basis functions (RBFs) and successive response surface method (SRSM).
4.2.1 Kriging

The method was suggested originally by Danie Krige in his Master thesis (Krige 1951) and further developed by (Matheron 1963). While these works provide rigorous derivation of the method, a more gentle introduction to Kriging can be found in e.g. (D. R. Jones 2001). The Kriging predictor is unbiased, based on minimum expected squared prediction error and is a linear combination of the measured responses.

The regression is modelled as a realization of a stochastic process, by estimating the uncertainty of the value of the function at an arbitrary point. A mean \( \mu \) and variance \( \sigma^2 \), can be ascribed to this uncertainty, as well as a correlation function between any two sample points in the design space. The correlation function is modelled as exponentially decaying function with parameters \( \theta \) and \( p \), with the former controlling the rate of decay of correlation as a function of distance between the points, and the latter determining the smoothness. The parameters \( \mu, \sigma^2, \theta \) and \( p \) are estimated in a maximum likelihood sense with respect to the observed data, thus obtaining a model which is most consistent with the observations.

Kriging approximations have been used heavily in recent years, due to their superior prediction capabilities and adaptability, since no basis model needs to be assumed in advance. A key difference between Kriging and ordinary response surfaces is the presence of parameters in the basis functions, which allow ‘fine tuning’ of the model to the particular observation data. Additionally, no model scaling and normalization is necessary to achieve accurate response, which is often necessary in response surface modelling.

4.2.2 Radial basis functions

Radial basis functions are commonly used for function approximations instead of polynomial regressions or RSM in problems with large number of dimensions. The approximating function is represented by a finite sum of radial basis functions with certain centers and weight coefficients. Formally, this approximation can be written as

\[
y(\mathbf{x}) = \sum_{i=1}^{n} \lambda_i \phi(||\mathbf{x} - \zeta_i||) \tag{57}
\]

where \( \phi(\cdot) \) is a radial function, \( \lambda = \{\lambda_i\} \) is a vector of associated weights, and \( \zeta_i \) is the center point of each radial function. For a given set of function values \( \mathbf{f} = \{f_i\} \) corresponding to \( \mathbf{x} \) which is to be interpolated, the weight coefficients can be obtained by solving

\[
\begin{pmatrix}
 f_1 \\
 \vdots \\
 f_n 
\end{pmatrix} =
\begin{pmatrix}
 A_{11} & \cdots & A_{1n} \\
 \vdots & \ddots & \vdots \\
 A_{n1} & \cdots & A_{nn} 
\end{pmatrix}
\begin{pmatrix}
 \lambda_1 \\
 \vdots \\
 \lambda_n 
\end{pmatrix} \tag{58}
\]

\[
A_{ij} = \phi(||\mathbf{x}_i - \zeta_j||), \quad i,j = 1,2,\ldots,n
\]

Provided that the matrix \( \mathbf{A} \) is invertible, the weight coefficients can be uniquely determined through \( \lambda = \mathbf{A}^{-1}\mathbf{f} \). If a function fitting instead of interpolation
is desired, linear least squares method can be used to minimize the sum of squares of $|| \mathbf{A} \lambda - \mathbf{f} ||^2$.

An excellent introduction to RBF, as well as efficient methods to obtain the approximant can be found in (Broomhead and Lowe 1988), (Buhmann 2000) and references therein. (Jin, X. Du, and W. Chen 2003) shows in a comparative study between RBF, kriging, polynomial regression and multivariate adaptive regression splines that radial basis functions shows good accuracy and performance over a set of 14 test cases with varying degrees of nonlinearity, noise and number of variables.

4.2.3 Sequential RSM

In structural optimization, the number of experiments that can be performed is often limited, as each numerical simulation can be computationally expensive; thus the quality of fit and the prediction accuracy of the regression model can be largely unknown for complex non-linear simulations. In the setting of structural optimization. Moreover, local non-linearities and optima can go completely undetected, and a potentially better solution missed altogether.

To overcome these difficulties, (Stander and Craig 2002) suggests a method called SRSM. In essence, after each regression model is created and an optimum solution based on it is found, a new, reduced region of interest as well as a corresponding DoE is created around the current optimum, with the expectation that the smaller the design domain, the more accurate the regression model is. The authors suggest a simple panning and zooming technique to determine the region of interest after each iteration, based on how closely the previous and the current optimal design are — zooming is more rapid if the solutions are close to each other, suggesting that the prediction is close to an actual optimum of the original model. While Stander and Craig originally applied the model on standard regression models, other models have also been suggested, for example neural networks in (Gustafsson and Niclas Strömberg 2007).

The advantage of this method is its simplicity and adaptability, as it can be viewed as an extension of RSM, not as a completely different type of regression model. Stander and Craig argue that SRSM can be used instead of numerical gradient-based methods such as SQP, particularly for noisy and discontinuous functions. As a disadvantage, the method require increased number of experiments, which can be alleviated somewhat by the fact that only a few successive DoE evaluations are typically required.

4.2.4 Optimal Regression Models

The general form of the regression equation in (38) does not stipulate the use of a particular basis of the regression — while linear and quadratic bases are well-known and common, other bases can be used as well. In (M. Hofwing, N. Strömberg, and Tapankov 2011), an optimal regression model is suggested, which utilizes a parametric polynomial basis with variables the exponents of the design variables, with a generic form

$$y_i = \beta_1 x_{i1}^{\gamma_{11}} x_{i2}^{\gamma_{12}} \ldots x_{ik}^{\gamma_{1k}} + \beta_2 x_{i1}^{\gamma_{21}} x_{i2}^{\gamma_{22}} \ldots x_{ik}^{\gamma_{2k}} + \ldots + \beta_m x_{i1}^{\gamma_{m1}} x_{i2}^{\gamma_{m2}} \ldots x_{ik}^{\gamma_{mk}}$$

(59)
A genetic algorithm is used in conjunction with the normal equation in (38) — a number of sets of exponents are generated, and for each, the regression coefficients are calculated. The sum of square errors is used to rank the candidate designs.

This method has proven to be easy to implement and superior to linear and quadratic regression models for the same number of design points, especially for highly non-linear responses. Computationally, it is significantly more expensive, however for practical applications this is of little consequence — a single numerical simulation can take hours or days to complete, and the vast majority of time in building a regression model would be spent performing the simulations themselves rather than constructing the regression model.

A limitation of this regression model is the practical number of design variables that can be used — the design space for the genetic algorithm grows exponentially with introducing of new variables, which slows down significantly the discovery of sufficiently good regression models.

A similar technique has also been suggested in (Zhao, Kyung K Choi, et al. 2010), which utilizes genetic algorithm to find optimal polynomial basis for Kriging approximations.
Due to the increased computational effort of RBDO problems, metamodels are a natural complement to them. The works combining metamodels and RBDO can be separated in two broad groups.

In some cases, metamodels are used without modifications specific to RBDO, for example (Jin, X. Du, and W. Chen 2003) compares RBF, kriging and polynomial regressions for optimization under uncertainty. (Zhao, Kyung K. Choi, et al. 2009) utilize a dynamic Kriging method for sampling-based RBDO for a small set of analytical and numerical problems, with results showing the increased accuracy and efficiency of this method for low-dimensional problems, although at the expense of higher computational effort.

Modifications specifically tailored to the RBDO have also been regularly proposed — for example, an early work by (Qu et al. 2001) integrate RIA with a RSM, although with low efficiency, mostly due to the limitations of RIA; later, (Youn and Kyung K. Choi 2004) embed a response surface within a hybrid mean value (HMV) method for RBDO which shows greater performance and accuracy; for example, (T. Lee and Jung 2008) suggest a constraint boundary sampling technique which emphasises the importance of accurate modelling of limit states, and show that this method performs better than comparable sampling methods for use with space-filling designs.

(Papadrakakis and Lagaros 2002) utilizes a combination of neural networks, evolution strategies and Monte Carlo simulations. In the training phase, a neural network is constructed using sampling data obtained with Monte Carlo simulations, and is validated using the deterministic and probabilistic constraints of the problem. In the optimization phase, an evolution strategy is applied to generate large number of possible designs, and the trained network is used as a predictor of response and validates both the deterministic and the probabilistic constraints. This method is completely gradient-free and potentially very efficient due to the possibility to apply all algorithms on many designs simultaneously by applying parallel or distributed computing.

In (Kang, Koh, and Choo 2010), a moving least squares method is used instead of the least-squares approximation widely adopted in RSM. Points close to the MPP are weighted more heavily, which increases the accuracy of representation of the design space in the region around it.

This technique has been further investigated in (Song and J. Lee 2011), which proposes a constraint-feasible version of the moving least-squares method. The constructed regression model is guaranteed to be always feasible, i.e. the approximation errors shift the response only in the feasible region, which ensures feasible optimal solution.

(Kim and Kyung K. Choi 2008) adopts a standard RSM which is refined by additional sampling at the MPPs for the active constraints as part of the RBDO algorithm. A prediction interval on the reliability is used as a metric for accuracy of the RBDO, and a lower and upper bounds are placed to avoid too conservative solutions.
6 Current and future trends

With the advancement of RBDO methods and the falling price of computing resources, RBDO has started to see traction in the field of topology optimization as well. (Maute and Frangopol 2003) published an early work in the field, applying PMA and topology optimization on compliant mechanisms. (Kharmanda et al. 2004) provides a more comprehensive study of how reliability affects the topology, and gives visual comparison for a few basic topology optimization problems. Both these works follow a more or less standard approach to topology optimization (compliance minimization, SIMP, filtering, optimality criteria method) without modifications related to reliability. (Kim, S. Wang, et al. 2006) proposes a volume-minimization topology optimization formulation with reliability constraints, with reliability computed using PMA and RIA. The work of (Eom et al. 2011) uses a quadratic response surface to approximate the response of 3D structures instead of direct evaluation. Three-dimensional standard problems are studied with volume minimization under probabilistic constraint on the maximum displacement. A FORM approximation is used for reliability assessment. In (Silva et al. 2010), the reliability assessment is performed using a single-loop method — a marked difference from previous works which only used double-loop approaches — coupled with standard topology optimization method.

Very few works are published in the field of reliability-based topology optimization, but there’s a clear desire to combine the more advanced methods in RBDO with topology optimization for imparting reliability into structural design as early as possible in the design process.
Glossary

characteristic value
A value of a parameter chosen to satisfy a given reliability level, commonly taken at . Often augmented with additional factors to compensate for load uncertainties, dynamic effects, variations in material properties etc. 6, 32

design of experiments
A structured sampling plan, as opposed to random sampling, which aims at maximizing metamodel, accuracy and prediction power by efficient choice of sample points. 19

design value
A value of a parameter that is used in limit state relationships, and is chosen to satisfy a given reliability level. It is a function of the corresponding parameter’s characteristic value and its load or resistance factor. 6, 32

factor
A dimensionless coefficient which modifies a characteristic value to obtain the design value used in a limit state expression. Load factors are usually greater than one, while resistance factors are less than unity. 6, 32

failure function
See limit state. 9

limit state
A condition beyond which a structure (or one or more of its constituent members) is no longer able to satisfy the provisions for which it was designed (EN 1990:2002). Limit states are usually represented as functions of various load, resistance and geometric parameters. 6, 32

reliability index
A non-dimensional measure of the relative reliability of a design. By convention, positive values indicate safe design, while negative values correspond to failure. See (Cornell 1969) and (Hasofer and Niels C. Lind 1974). 9

run
A combination of values of the design variables used to conduct a single experiment. 22

serviceability limit state
Conditions which imply deformations but without exceeding the load-carrying capacity of the structure (Madsen, Krenk, and N. C. Lind 1986). Examples include cracks, vibrations, surface damage etc. 6
single-loop approach
A reliability-based design optimization method in which the reliability assessment and the cost function optimization are performed in a single optimization cycle. 8

ultimate limit state
Corresponds to limit of load-carrying capacity of a member or the structure as a whole (Madsen, Krenk, and N. C. Lind 1986). Examples include plastic yielding, brittle or fatigue fracture, buckling. 6

Acronyms

AFOSM
Advanced first-order second-moment method. A reliability estimation method based on the reliability index concept introduced by Hasofer and Lind (1974). 12

BBD
Box-Behnken design. A three-level design featuring combinations of all design variables, taken two at a time, at their extremes. First suggested by (Box and Behnken 1960). 22

CCD
Central composite design. A spherical design of experiments featuring a two-level factorial design augmented with central and axial runs. 22, 33

CDF
Cumulative distribution function. The probability that a random variable with a given probability distribution will be found at or less than a certain given value. 12, 13

CONLIN
Convex linearization. An approximation method in structural optimization that utilizes linearization of the reciprocal of the design variables to reduce approximation error. First presented in (Fleury and Braibant 1986). 16

FCD
Face-centered design. A rectangular design, variation of CCD with axial runs on the faces of the design domain hypercube. 22

FORM
First-order reliability method. A common designation for all reliability assessment methods which utilize first-order linearization of the failure function. 11
JPDF

Joint probability density function. A multivariate version of the probability density function. 9

KKT

Karush-Kuhn-Tucker conditions are the first-order necessary optimality conditions for a given solution to be optimal. 14

LHS

A sampling technique in which the domain of each design variables is divided into equally spaced levels, and a set of design points is picked to cover each level of each design variables. 26

LRFD

Load and Resistance Factor Design. 6, see also LSD

LSD

Limit state design. A design method used in structural engineering that incorporates multiple safety factors. 6

MMA

Method of moving asymptotes. An efficient optimization technique first proposed in (Svanberg 1987). 16

MPP

Most probable point. The point on the failure surface at which failure is most likely to occur. 12

MVFOSM

Mean value first-order second-moment method. A reliability estimation method based on the reliability index concept of Cornell (1969). 11

PDF

Probability density function. The relative likelihood that a random variable would take a certain value. 5

PMA


RBDO

Reliability-based design optimization is an optimization methodology which takes into consideration the uncertainties of the design variables during the optimization process, with target reliability prescribed to each design constraint. 4, 8, 35
RBF
A class of real-valued functions whose value depend only on the distance from the origin or a designated center point. 26

RIA

SAP
Sequential approximate programming. 17

SCD
Small composite designs is a compact DoE that feature an incomplete factorial design augmented with central runs. 23, see

SLA
A single-loop RBDO method, developed by (Liang, Mourelatos, and Nikolaidis 2007). 15

SLP
Sequential linear programming is a gradient-based optimization method, in which successive linear approximations of the problem are created around the current iterate, which are solved using linear programming. 25

SLSV
Single-loop, single-vector approach — one of the first single-loop methods, suggested by (Xiaoguang Chen, Hasselman, and Neill 1997). 14

SORA
Sequential Optimization and Reliability Assessment method. A popular decoupled RBDO method, due to (Xiaoping Du and Wei Chen 2004). 16

SORM
Second-order reliability method. Second-order approximation of the probability of failure, which provides more accurate estimation of the reliability index for very non-linear limit states. Most commonly used SORM formulations can be found in (Breitung 1984), (Hohenbichler et al. 1987) and (Tvedt 1990). 11

SRSM
Successive response surface method. A sequential approximation of RSM with progressively smaller regions of interest around the optimal solution obtained in the previous optimization step. First proposed by (Stander and Craig 2002). 26, 28
References


Chen, Xiaoguang, Timothy Hasselman, and Douglas Neill (Apr. 1997). “Reliability based structural design optimization for practical applications”.


Gustafsson, Erik and Niclas Strömberg (May 2007). “Successive Response Surface Methodology by using Neural Networks”. In: 7th World Congress on Structural and Multidisciplinary Optimization, pp. 820–829.


Shan, Songqing and G. Gary Wang (2008). “Reliable design space and complete single-loop reliability-based design optimization”. In: Reliability Engineering


