Manuscript, October 2015

A discretization procedure for rare events in Bayesian networks

Kilian Zwirglmaier & Daniel Straub

Engineering Risk Analysis Group, Technische Universität München

(kilian.zwirglmaier@tum.de, straub@tum.de, www.era.bgu.tum.de)

Abstract

Discrete Bayesian networks (BNs) can be effective for risk- and reliability assessments, in which probability estimates of (rare) failure events should be continuously updated with new information. To solve such reliability problems accurately in BNs, the discretization of continuous random variables must be performed carefully. To this end, we develop an efficient discretization scheme, which is based on finding an optimal discretization for the linear approximation of the reliability problem obtained from the First-Order Reliability Method (FORM). Because the probability estimate should be accurate under all possible future information scenarios, the discretization scheme is optimized with respected to the expected posterior error. To simplify application of the method, we establish parametric formulations for efficient discretization of random variables in BNs for reliability problems based on numerical investigations. The procedure is implemented into a software prototype. Finally, it is applied to a verification example and an application example, the prediction of runway overrun probabilities of a landing aircraft.

Keywords

Bayesian networks; discretization; near-real-time; structural reliability; updating

1 Introduction

For operational risk and reliability management, it is often desirable to compute the probability of a rare event F under potentially evolving information. Examples include warning systems for natural and technical hazards, or the planning of inspection and intervention actions in infrastructure systems. Ideally, this is achieved through Bayesian updating of $\Pr(F)$ with the new information Z to the posterior probability $\Pr(F|Z)$. When physically-based or empirical models for predicting the rare event exist, such updating is possible with structural reliability methods (SRM) (Sindel and Rackwitz, 1998, Straub, 2011). However, it is often difficult to perform the required computations in near-real-time, due to a lack of efficiency or robustness. A modeling and computational framework that does facilitate efficient Bayesian updating is the discrete Bayesian network (BN). Hence it was proposed to combine SRMs with discrete Bayesian networks for near-real-time computations (Friis-Hansen, 2000, Straub and Der Kiureghian, 2010a, Straub and Der Kiureghian, 2010b).

BNs are based on directed acyclic graphs (DAGs), to efficiently define a joint probability distribution $p(\mathbf{Y})$ over a random vector $\mathbf{Y}$ (Jensen and Nielsen, 2007, Kjaerulff and Madsen, 2013). The DAG of a BN, which is often referred to as the qualitative part of a BN, consists of a node for each variable in $\mathbf{Y}$ and a set of directed links among nodes representing dependence among the variables. In the case of discrete BNs, conditional probability tables (CPTs) quantitatively define the type and strength of the dependence among the variables. The entries of the CPT of a variable Y_i are the probabilities for each state of Y_i conditional on all possible combinations of states of its parents.

For hybrid BNs, which include both discrete and continuous variables, exact inference is available only for two special cases, which are BNs with Gaussian nodes, whose means are linear functions of their parents, and BNs, whose nodes are defined as a mixture of truncated basic functions (MoTBFs) (Langseth et al., 2009, Langseth et al., 2012). Otherwise, approximate inference algorithms are available for hybrid BNs based on sampling techniques, e.g. (Lerner, 2002, Hanea et al., 2006). However, these are computationally demanding and not suitable for most near-real-time decision support (Hanea et al., 2015). As an alternative, the continuous random variables can be discretized, which enables the use of exact inference algorithms that exist for general discrete BNs. These include the variable elimination algorithm (Zhang and Poole, 1994) and the junction tree algorithm (Lauritzen and Spiegelhalter, 1988, Jensen et al., 1990).

The size of discrete BNs, and the associated computational effort, increases approximately exponentially with the number of discrete states of its nodes, which motivates the development of efficient discretization algorithms. While efficient discretization in the context of machine learning and BNs in general has been investigated by multiple researchers (Dougherty et al., 1995, Kotsiantis and Kanellopoulos, 2006), research on efficient discretization in the context of engineering risk analysis or structural reliability has been

limited. In general, it is to be distinguished between static and dynamic discretization. While the former discretizes the BN a-priori before entering evidence (offline), the latter is based on an iterative scheme that updates the discretization scheme in function of the evidence (online).

Dynamic discretization for risk analysis applications has been developed mainly by (Neil et al., 2008), based on the work by (Kozlov and Koller, 1997). The procedure starts with an initial discretization of a hybrid BN, for which an approximate entropy error is calculated. If the error complies with a convergence criterion, the current discretization is accepted. Otherwise the discretization is iteratively altered, by splitting the intervals with the highest entropy error, until the convergence criterion is fulfilled. The approach is implemented in the software AgenaRisk (Agena, 2005). Other dynamic discretization algorithms for reliability analysis have been proposed, e.g. in (Zhu and Collette, 2015) for dynamic BNs. The advantage of dynamic discretization is its flexibility when evidence is entered in the BN, i.e. when the model is updated with new observation.

Static discretization has the advantage of being computationally more efficient and simple to implement. Some considerations for static discretization of BNs in reliability applications have been presented in (Friis-Hansen, 2000, Straub, 2009, Straub and Der Kiureghian, 2010a). As pointed out by (Friis-Hansen, 2000), for applications in which extreme events are important, discretization of the distribution tails should be performed with care. Static discretization facilitates a careful representation of these tails. However, the accuracy of the static discretization varies with the available evidence. The difficulty is thus to find a discretization scheme that is optimal under a wide variety of posterior distributions.

In this paper we derive a procedure for efficiently performing static discretization of continuous reliability problems. An optimal discretization scheme is sought, which minimizes the expected approximation error with respect to possible future observations (evidence). To solve this optimization problem, we propose to approximate the reliability problem by the First-Order Reliability Method (FORM). Section 2 of the paper describes the methodology applied. Section 3 presents numerical parameter studies, and simple parametric relations for defining an efficient discretization scheme are derived. In Section 4, the procedure is applied to a set of verification examples and to the computation of the probability of runway overrun of a landing aircraft.

2 Methodology

2.1 Structural reliability

Since the 1970s structural reliability methods have been developed and applied in the engineering community to estimate failure probabilities $\Pr(F)$ of components or systems, based on physical or empirical models. The performance of engineering components is described by a limit state function (LSF) $g(\mathbf{x})$, where $\mathbf{X} = [X_1; \dots; X_n]$ is a vector of basic random variables influencing the performance of the component. By definition, failure corresponds to $g(\mathbf{x})$ taking non-positive values, i.e. the failure event is $F = \{g(\mathbf{X}) \leq 0\}$. $g(\mathbf{x})$ includes the physical or engineering model, which is often computationally demanding. The probability of failure is calculated by integrating the probability density function (PDF) of $\mathbf{X}$, $f_{\mathbf{X}}(\mathbf{x})$, over the failure domain:

$$\Pr(F) = \int_{g(\mathbf{x})\leq 0} f_{\mathbf{X}}(\mathbf{x}) d\mathbf{x} \tag{2}$$

The formulation can be extended to the reliability of general systems by defining the failure domain as a combination of series and parallel systems (Ditlevsen and Madsen, 2007). In the general case, there is no analytical solution to Eq. 2 and the integral is potentially high-dimensional. For this reason, structural reliability methods (SRMs) are applied to approximate it. These include the first- and the second order reliability method (FORM and SORM) as well as a large variety of sampling methods, including importance sampling methods such as directional importance sampling, and sequential sampling methods such as subset simulation. These methods are well-documented in the literature (Au and Beck, 2001, Rackwitz, 2001, Der Kiureghian, 2005, Ditlevsen and Madsen, 2007).

2.2 First order reliability method (FORM)

To obtain an approximation of the probability of failure through FORM, the LSF $g(\mathbf{X})$ is transformed to an equivalent LSF $G(\mathbf{U})$ in the space of uncorrelated standard normal random variables $\mathbf{U} = [U_1; \dots; U_n]$ (Fig. 1). The transformation is probability conserving, so that $\Pr[g(\mathbf{X}) \leq 0] = \Pr[G(\mathbf{U}) \leq 0] = \Pr(F)$. A suitable transformation for this purpose, which is consistent with the BN, is the Rosenblatt transformation (Hohenbichler and Rackwitz, 1981). In case all basic random variables are independent, this transformation reduces to the marginal transformations: $U_i = \Phi^{-1}\left[F_{X_i}(X_i)\right]$, with Φ^{-1} being the inverse standard normal CDF.

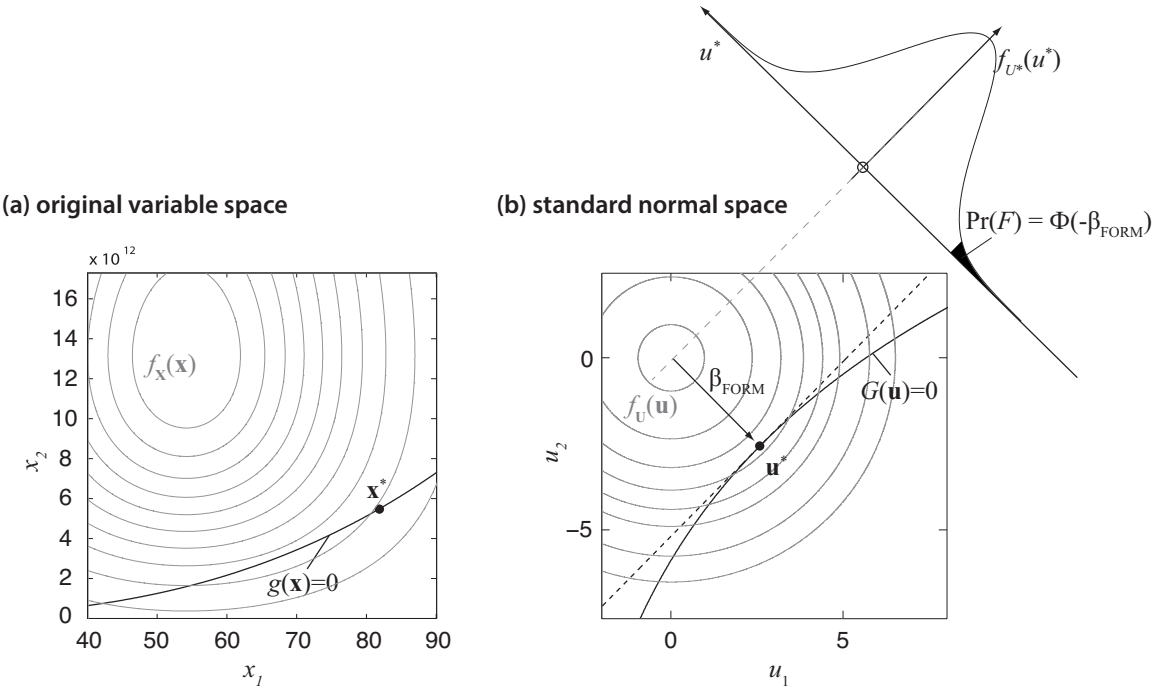

Figure 1. Design point and linear approximation of the limit state surface. Left side: original random variable space; right side: standard normal space (from (Straub, 2014a)).

The FORM approximation of $\Pr(F)$ is obtained by substituting the LSF in U-space $G(\mathbf{U})$ by a linear function $G_L(\mathbf{U})$, i.e. a first-order Taylor expansion of $G(\mathbf{U})$. The key idea of FORM is to choose as the expansion point the so-called design point $\mathbf{u}^*$, which is the point that minimizes $\|\mathbf{u}^*\|$ subject to $G_L(\mathbf{U}) \leq 0$. It is also known as the most likely failure point, as it is the point in the failure domain with the highest probability density. Since all marginal distributions of the standard uncorrelated multinormal distribution are standard normal, it can be shown that the FORM probability of failure $\Pr[G_L(\mathbf{U}) \leq 0]$ is:

$$\Pr[G_L(\mathbf{U}) \leq 0] = \Phi(-\beta_{FORM}) \tag{3}$$

where Φ is the standard normal CDF and β_{FORM} is the distance from the origin to the design point, i.e. $\beta_{FORM} = \|\mathbf{u}^*\|$. The problem thus reduces to finding the design point $\mathbf{u}^*$. If $G(\mathbf{U})$ is linear, the FORM solution of the probability of failure is exact, otherwise it is an approximation, which however is sufficiently accurate in most practical applications with limited numbers of random variables (Rackwitz, 2001).

The linearized LSF $G_L(\mathbf{U})$ can be written as:

$$G_L(\mathbf{U}) = \beta_{FORM} - \alpha^{\mathrm{T}}\mathbf{U} \tag{4}$$

where $\boldsymbol{\alpha} = [\alpha_1, \ldots, \alpha_n]$ is the vector of FORM importance measures. These importance measures are defined as:

$$\alpha_i = \frac{u_i^*}{\beta_{FORM}} \tag{5}$$

where u_i^* is the i-th component of the design point coordinates. The α_i's take values between -1 and 1, and it is $\|\boldsymbol{\alpha}\| = 1$. α_i is 0, if the uncertainty on U_i has no influence on $\Pr(G_L(\mathbf{U}) \leq 0)$, and it is 1 or -1, if U_i is the only random variable affecting $\Pr(g_L(\mathbf{U}) \leq 0)$. When the original random variables X_i are mutually independent, the α_i's are readily applicable also in the original space, otherwise the α_i's can be transformed as described in (Der Kiureghian, 2005).

2.3 Treatment of a reliability problem in a BN

We combine discrete BNs and structural reliability concepts to facilitate updating of rare event (failure) probabilities under new observations. The general problem setting is illustrated in the BN of Fig. 2. We here limit ourselves to component reliability problems; system problems are considered later. The binary random variable 'Component performance' is described by the LSF $g(\mathbf{X})$.

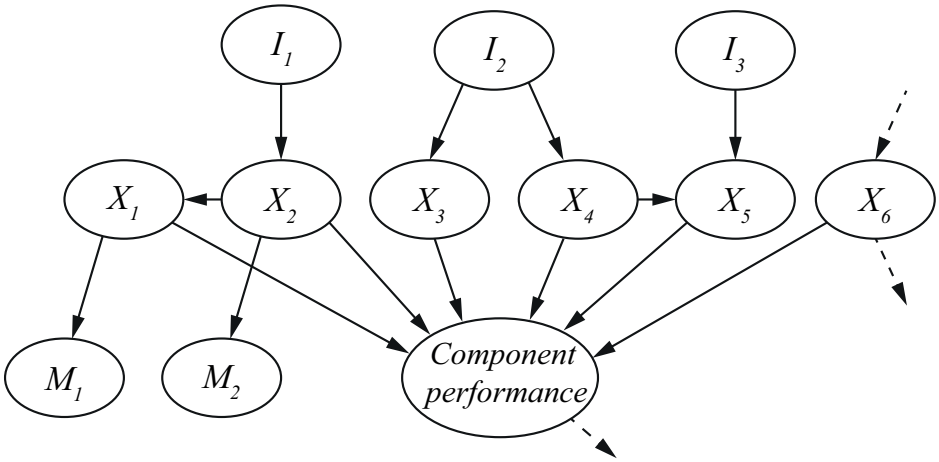

Figure 2. A general BN including a component reliability problem.

The basic random variables **X** of the model are included in the BN as parents of 'Component performance'. The nodes M_i represent measurements of individual random variables X_i, and nodes I_j represent factors influencing the basic random variables. Dependence between the variables in **X** is modeled either directly by links among them (here $X_2 \rightarrow X_1$ and $X_4 \rightarrow X_5$) or through common influencing factors (here $I_2 \rightarrow X_3$ and $I_2 \rightarrow X_4$) the component performance node can have (multiple) child nodes. However that does not have an impact on the discretization of the reliability problem.

Ultimately, the goal is to predict the component performance, i.e. $\Pr(F)$, conditional on observations of other variables, typically of the measurement variables M_i, but possibly of any other random variable in the BN, such as the influencing variables I_j. Whenever new

evidence on these variables is available, the BN should be evaluated (in near-real) time utilizing exact BN inference algorithms.

To enable exact inference algorithms, all continuous random variables are discretized. These include the $\mathbf{X}$, and possibily the M_i and I_j. In the general case, the computational effort for solving the BN is a direct function of the CPT size of 'Component performance'. The size of this CPT is $2\prod_{i=1}^{n} n_i$, where n is the number of random variables in $\mathbf{X}$, and n_i is the number of states used for discretizing X_i. In this paper we do not describe the discretization of random variables M_i and I_j, since it is typically straightforward and does not contribute significantly to computational performance. The key parameter for computational efficiency and accuracy is the discretization scheme for $\mathbf{X}$, which is described in the next section.

2.4 Simplification of BNs through node removal

Removing random variables from a BN is one possibility to reduce the computational effort associated with a model. A formal approach for removing nodes from a BN is described in (Straub and Der Kiureghian, 2010b). In order to decide which nodes to remove from the BN the following questions should be considered:

- Which random variables are relevant for prediction? (This includes 'Component performance'.)
- Which random variables can potentially be observed? (This includes the measurement variables.)
- Which random variables simplify the modeling of dependencies? (These are e.g. common influencing factors such as I_2 in Fig. 2.)
- For which random variables is it desirable to explicitly show their influence on component performance?

If a random variable does not belong to any of these categories, the corresponding node in the BN can be removed. Since the computational efficiency of the model is governed by the size of the CPT of the 'Component state' node, the primary interest is in removing basic random variables from the network. As a measure for the relevance of a basic random variable, importance measures α_i from a FORM analysis may be used. To better understand the relation between α_i and X_i's relevance for prediction consider a linearized LSF $G_L(\mathbf{U})$. Following (Der Kiureghian, 2005) the variance of $G_L(\mathbf{U})$ can be decomposed as:

$$\sigma_{G_L}^2 = \|\nabla G\|^2(\alpha_1^2 + \alpha_2^2 + \cdots + \alpha_n^2) \tag{6}$$

where ∇G denotes the gradient vector of the non-linearized LSF $G(\mathbf{U})$. From Eq. 6 it is seen that a random variable X_i with corresponding α_i accounts for $\alpha_i^2 \cdot 100\%$ of the variance $\sigma_{G_L}^2$. Therefore, observing a random variable X_i with $\alpha_i = 0.1$ will reduce the variance $\sigma_{G_L}^2$ by 1%, whereas the fixing of X_j with $\alpha_j = 0.5$ will reduce $\sigma_{G_L}^2$ by 25%.

2.5 Discretization of basic random variables

For ease of presentation, we first consider discretization of statistically independent basic random variables **X**, i.e. the special case of the BN in Fig. 2 in which the X_i's have no parents. This is extended to the general case thereafter.

2.5.1 Independent basic random variables

The situation is illustrated in Fig. 3. The performance of the component depends on n statistically independent random variables and is described by a LSF $g(\mathbf{X}) = g(X_1, \dots, X_n)$. For all basic random variables X_i, corresponding measurements M_i can be performed. To obtain an equivalent discrete BN, the continuous X_i are replaced by the discrete random variables Y_i, and the LSF is replaced by the CPT of component performance conditional on $\mathbf{Y} = [Y_1; \dots; Y_n]$. For each discrete random variable Y_i with n_i states $1,2, \dots, n_i$, we define a discretization scheme $D_i = [d_0, d_1, \dots, d_{n_i-1}, d_{n_i}]$ consisting of $n_i + 1$ interval boundaries. The first and the last interval boundaries are given by the boundaries of $X_i's$ outcome space.

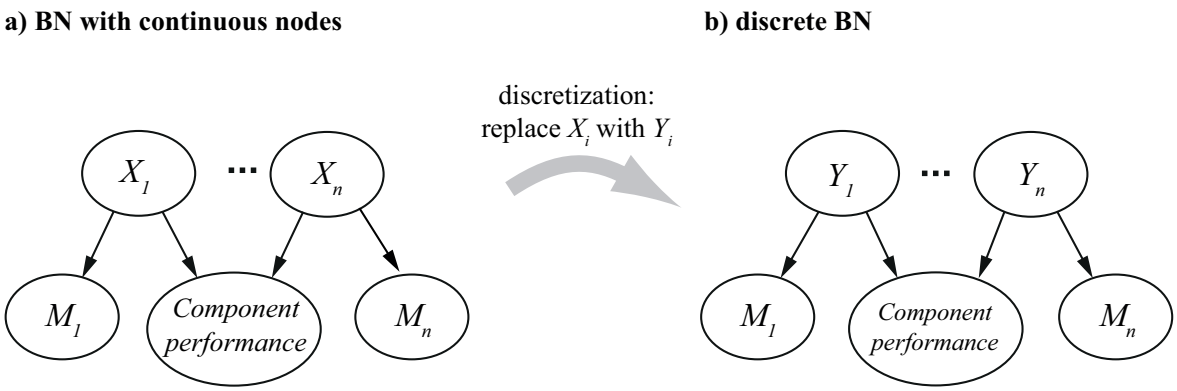

Figure 3. Representation of a basic reliability problem with n independent basic random variables in a BN. Left: original problem with continuous basic random variables X_i, right: discrete BN, in which X_i are substituted with discrete nodes Y_i.

Since here the X_i, and thus the Y_i, have no parents, the PMF of Y_i is defined as:

$$p_{Y_i}(j) = F_{X_i}(d_j) - F_{X_i}(d_{j-1}) \tag{7}$$

where F_{X_i} denotes the cumulative distribution function (CDF) of X_i. The probability of failure corresponding to the discrete BN in Fig. 3b can be calculated as:

$$\Pr(F) = \sum_{y_1=1}^{n_1} \dots \sum_{y_n=1}^{n_n} p_{Y_1}(y_1) \cdot \dots \cdot p_{Y_n}(y_n) \cdot \Pr(F|Y_1 = y_1 \cap \dots \cap Y_n = y_n) \tag{8}$$

Once measurements from the nodes $\mathbf{M} = [M_1; \dots; M_n]$ are available, the conditional failure probability can be calculated as:

$$\Pr(F|\mathbf{M} = \mathbf{m}) \approx \frac{1}{p_{\mathbf{M}}(\mathbf{m})} \sum_{y_1=1}^{n_1} \dots \sum_{y_n=1}^{n_n} p_{Y_1}(y_1) \cdot p_{M_1|Y_1}(m_1|y_1) \cdot \dots \cdot p_{Y_n}(y_n) \cdot p_{M_n|Y_n}(m_n|y_n) \cdot \Pr(F|Y_1 = y_1 \cap \dots \cap Y_n = y_n) \tag{9}$$

where $\Pr(F|Y_1 = y_1, \dots, Y_n = y_n)$ is the conditional probability of component failure given $y_1, \dots, y_n$. If no measurements are available for some of the basic random variables, the corresponding likelihood terms $p_{M_j|Y_j}(m_j|y_j)$ are simply omitted in Eq. 9.

While the computation of the unconditional failure probability following Eq. 8 is exact, the computation of the conditional failure probability through Eq. 9 is only an approximation. The reason is that the dependence between the measurement variable M_i and the 'Component performance' variable is not fully captured in the discrete BN (see also Straub and Der Kiureghian, 2010b). In Fig. 4, this is illustrated for a reliability problem with one basic random variable X_i. Both the continuous distribution (Fig. 4a) and the corresponding discretized distribution (Fig. 4b) are updated correctly after observing M_1. However, for Eq. 9 to be exact, also the conditional failure probabilities $\Pr(F|Y_1 = y_1)$ would need to be updated. This can be observed in Fig. 4a: in interval $Y_1 = 3$, which is the one cut by the limit state surface, the ratio of the probability mass in the failure domain to that in the safe domain changes from the prior to the posterior case. The fact that $\Pr(F|Y_1 = 3) \neq \Pr(F|Y_1 = 3, M_1 = m_1)$ shows that the independence assumption underlying Eq. 9, namely $\Pr(F|Y_i = y_i) = \Pr(F|Y_i = 3, M_i = m_i)$ is only an approximation. The error occurs only in the intervals that are cut by the limit state surface. In the simple one-dimensional case of Fig. 4, an optimal discretization approach would be to discretize the whole outcome space in two intervals, one capturing the survival and one the failure domain. This discretization would have zero approximation error. However, already in a two-dimensional case, such a solution is not possible. This is illustrated in Fig. 5, where the cells cut by the limit state surface are indicated in grey. The failure probability conditional on measurements calculated according to Eq. 9 will necessarily be an approximation. The approximation error will be small, if the contribution of the cells cut by the limit state surface (the grey cells in Fig. 5) to the total failure probability is small. An efficient discretization will thus limit this contribution with as few intervals as possible.

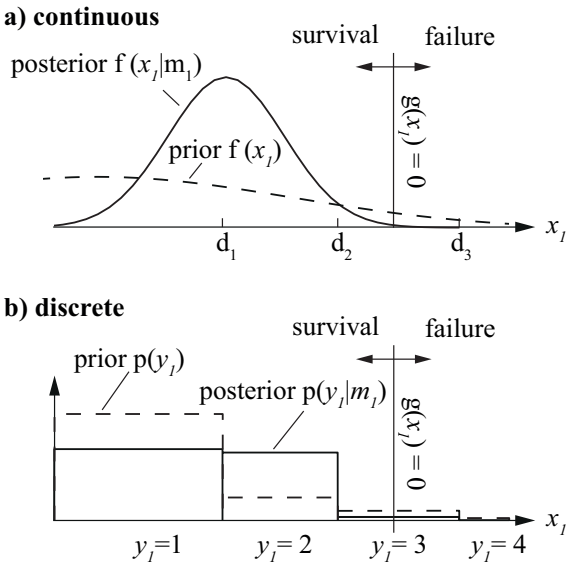

Figure 4. Discretization error in 1D.

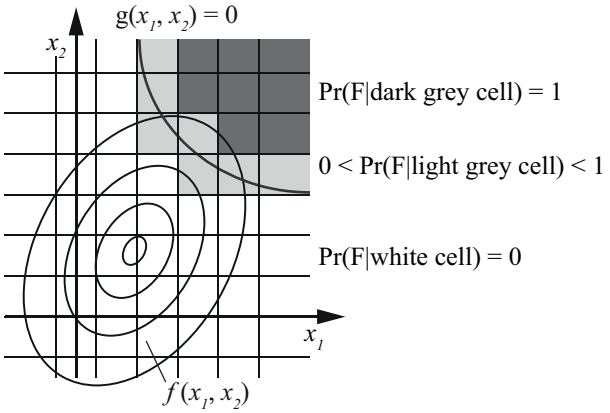

Figure 5. Discretization error in 2D.

2.5.2 Dependent basic random variables

Eqs. (7-9) must be adjusted when dependence among the X_i's is present, in accordance with the case-specific BN structure. However, the principles outlined above for independent $X_1, \dots, X_n$ hold equally for dependent basic random variables: The discretization error is a function of the cells cut by the limit state function.

When determining an optimal discretization, we propose in the following to find the FORM approximation of the reliability problem, which can readily account for the dependence among the random variables. Hence, there is no need to distinguish between the cases with independent or dependent random variables.

2.6 Efficient discretization

2.6.1 Optimal discretization of linear problems in standard normal space

To find an efficient discretization of $\mathbf{X}$, we consider the FORM solution to the reliability problem. Evaluating the linearized FORM LSF $G_L(\mathbf{U})$ is computationally inexpensive once the design point $\mathbf{u}^*$ is known. Therefore, it is feasible to find a discretization of $\mathbf{U}$ that is optimal for the event $\{G_L(\mathbf{U}) \leq 0\}$ through optimization. If $G(\mathbf{U})$ is not strongly non-linear, this solution will be an efficient discretization for $\{G(\mathbf{U}) \leq 0\}$ and, after a transformation to the original space, also for $\{g(\mathbf{X}) \leq 0\}$.

As discussed in Section 2.5.1, the approximation error of the discretization is associated with the change from the prior to the posterior distribution of the basic random variables. A measure of optimality must thus consider possible measurements of $\mathbf{X}$ or $\mathbf{U}$. We consider hypothetical measurements $\widetilde{M}_i$ of all U_i with additive measurement error ε_i:

$$\widetilde{M}_i = U_i + \varepsilon_i \tag{10}$$

ε_i is modeled as a normal distribution with zero mean and standard deviation σ_ε. The conditional distribution of U_i given a measurement outcome $\widetilde{M}_i = \tilde{m}_i$ is the normal distribution with mean $\frac{1}{1+\varepsilon_i^2}\tilde{m}_i$ and standard deviation $\sqrt{\left(1-\frac{1}{1+\varepsilon_i^2}\right)}$.

We define an error measure based on comparing the true posterior probability of failure $\mathrm{P}_{\mathrm{F}|\widetilde{\boldsymbol{M}}}(\tilde{\mathbf{m}})$ with the posterior probability of failure calculated from the discretized $\mathbf{U}$, denoted by $\widehat{\mathrm{P}}_{\mathrm{F}|\widetilde{\boldsymbol{M}}}(\mathbf{d};\tilde{\mathbf{m}})$. Here, $\mathbf{d}$ are the parameters defining the discretization. The proposed error measure is:

$$e(\mathbf{d},\tilde{\mathbf{m}}) = \left|\frac{\log_{10}\widehat{\mathrm{P}}_{\mathrm{F}|\widetilde{\boldsymbol{M}}}(\mathbf{d};\tilde{\mathbf{m}}) - \log_{10}\mathrm{P}_{\mathrm{F}|\widetilde{\boldsymbol{M}}}(\tilde{\mathbf{m}})}{\log_{10}\mathrm{P}_{\mathrm{F}|\widetilde{\boldsymbol{M}}}(\tilde{\mathbf{m}})}\right| \tag{11}$$

$e(\mathbf{d},\tilde{\mathbf{m}})$ is a relative measure of the posterior error, weighted by the magnitude of the probability of failure.

A-priori, the measurement outcomes are not known. Hence we define the optimal discretization as the one that minimizes the expected preposterior error $\mathrm{E}_{\widetilde{\boldsymbol{M}}}\left[e\left(\mathbf{d},\widetilde{\boldsymbol{M}}\right)\right]$:

$$\mathbf{d}^{opt} = \arg\min_{\mathbf{d}} \mathrm{E}_{\widetilde{\boldsymbol{M}}}[e(\mathbf{d},\widetilde{\boldsymbol{M}})] = \arg\min_{\mathbf{d}} \int e(\mathbf{d},\tilde{\mathbf{m}})\, \mathrm{f}_{\widetilde{\boldsymbol{M}}}(\tilde{\mathbf{m}})\mathrm{d}\tilde{\mathbf{m}} \tag{12}$$

The optimization is thus based on the computation of an expected value with respect to the possible measurements outcomes $\widetilde{\boldsymbol{M}}$ before having taken any measurements. This is analogous to a preposterior analysis (Raiffa and Schlaifer, 1961, Straub, 2014b). However, unlike in traditional preposterior analysis, the objective is not to identify an optimal action under future available information, but to find the optimal discretization parameters $\mathbf{d}^{opt}$. The integral in Eq. 12 is evaluated through a simple Monte Carlo approach. All $\widetilde{M}_i$ have the normal distribution with zero mean and variance $1+\sigma_\epsilon^2$.

The parameters in $\mathbf{d}$ describing the discretization scheme are:

- n_i: number of intervals used to discretize each random variable U_i,
- w_i: width of the discretization frame in the dimension of U_i, and
- v_i: position of the midpoint of the discretization frame relative to the design point

These parameters are illustrated in Fig. 6. For a problem with n basic random variables, the full set of optimization parameters is $\mathbf{d} = [w_1,\dots,w_n,n_1,\dots,n_{n-1},v_1,\dots,v_n]$.

Clearly, the discretization error reduces with increasing n_i. Because the computational efficiency of the final BN is a direct function of the size of the CPT associated with component performance, which is $\prod_{i=1}^{n} n_i$, we constrain its size. To this end, we define c_{up} as the maximum number of parameters of the CPT of the component state node. This puts a constraint on the optimization of Eq. 12:

$$c_{up} \leq \prod_{i=1}^{n} n_i \tag{13}$$

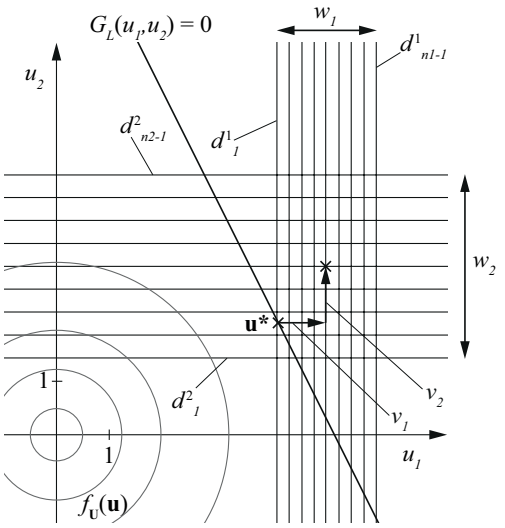

Figure 6. Schematic representation of a discretization of a linear 2D reliability problem. w_i is the distance between interval boundaries d_1^i and $d_{n_i-1}^i$. All intervals between these boundaries are equi-spaced. v_i is the position of the midpoint of the discretization frame relative to the design point $\mathbf{u}^*$ in dimension i.

The optimization is implemented through a two-level approach. The optimization of the continuous parameters width w_i and position of the discretization frame v_i for all $i = 1, \dots, n$ is carried out using unconstrained nonlinear optimization for fixed values of n_i. The optimization of the discrete n_i is performed through a local search algorithm.

2.6.2 Efficient discretization of the original random variables **X**

Since the nodes in the BN represent random variables **X** in their original outcome space, the discretization schemes, which are derived for the corresponding standard normal random variables **U**, need to be transformed to the X-space. In the case of mutually independent random variables X_i, any point on the i-th interval boundary in U-space – if transformed – will result in the same corresponding i-th interval boundary in X-space. This is not the case for dependent random variables X_i, where a mapping of the interval boundaries in U-space to X-space will not lead to an orthogonal discretization scheme in X-space, even if it is orthogonal in U-space. To preserve orthogonality throughout the transformation, we propose to represent each interval boundary through a characteristic point and determine the boundary in X-space through a transformation of this point. For transforming the interval boundary of X_i, the characteristic point is selected as the design point $\mathbf{u}^*$, where the i-th element is substituted by the coordinate of the interval boundary. In Fig. 7 this is shown for an example with $n = 2$ random variables.

a) Discretization in U-space

b) Transformed discretization

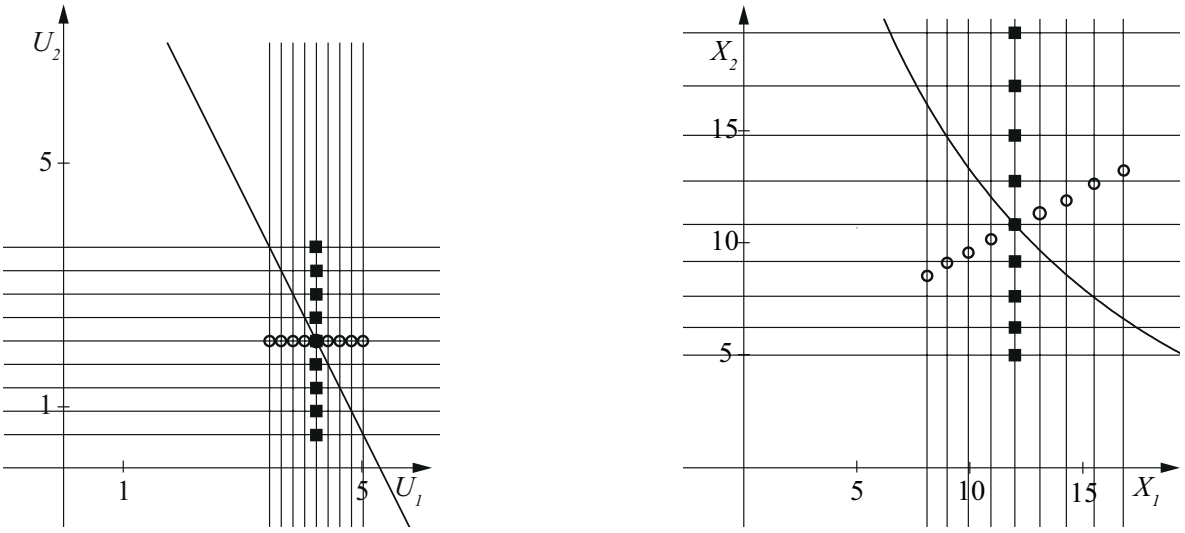

Figure 7. Transformation of a discretization scheme from U-space to X-space. To preserve orthogonality each interval boundary in U-space is represented by a characteristic point. The random variables X_1 and X_2 are Weibull distributed with scale and shape parameter 1 and their correlation is 0.5.

3 Development of an efficient discretization procedure

3.1 Optimization of the FORM approximation

We present the optimal discretization for the FORM approximation $G_L(\mathbf{U})$ for $n = 2$ and $n = 3$ dimensions. Extension to higher numbers of random variables is discussed. Because the linear LSF employed in FORM is described only by the reliability index β_{FORM} and the vector $\boldsymbol{\alpha}$ of FORM sensitives, following Eq. 4, it facilitates parametric studies.

Initially, we consider a reliability index $\beta_{FORM} = 4.26$, corresponding to a probability of failure of 10^{-5}. The standard deviation of the additive measurement error is set to either $\sigma_\varepsilon = 0.5$ or $\sigma_\varepsilon = 1.0$. Different combinations of FORM sensitivity values α_i are selected, to investigate their effect on the optimal discretization. In all investigated cases, we find that the position of the midpoint of the optimal discretization frame coincides with the design point, i.e. $v_i^{opt} = 0$. Furthermore, the optimal number of intervals n_i^{opt} is approximately the same for all random variables in all investigated cases, independent of the α_i values. We therefore fix these two optimization parameters at $v_i = 0$ and $n_i = \left\lfloor c_{up}^{1/n} \right\rfloor$. The optimal discretization widths w_i^{opt}, however, vary significantly with the importance measures α_i.

3.1.1 Dependence of the optimal discretization width on α_i

At first sight, the dependence of w_i^{opt} on α_i is not obvious, but a clear trend can be observed by plotting the probability mass enclosed by w_i^{opt} against α_i, as shown in Fig. 8. The width w_i describes the domain in which a fine discretization mesh is applied (Fig. 6). The results of Fig. 8 indicate that the probability mass contained within this interval should be a direct function of the random variable's importance, as expressed through α_i. The more important

the variable, the finer the discretization around the design point should become. The observed relationship between this probability mass and α_i follows a clear trend, and a function can be fitted (Fig. 8). Neither the dimensionality of the problem nor the standard deviation of the measurement error appear to have an influence on this relation. However, as shown in the following section, it is found that the relation does depend on the prior failure probability of the problem (i.e. on β_{FORM}) and on the number of intervals n_i used to discretize the domain.

To facilitate the application in practice and extending the results to larger numbers of random variables, in section 3.2 parametric functions are fitted to the optimization results to capture the dependency between the optimal discretization width w_i^{opt} and the FORM importance measures α_i.

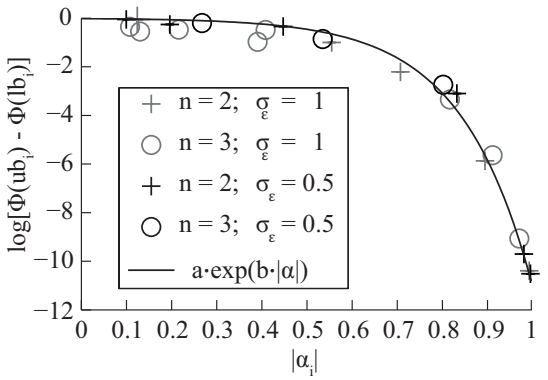

Figure 8. Logarithm of the probability mass enclosed by the discretization frame plotted against α_i. Φ denotes the standard normal CDF and ub_i respectively lb_i the last (upper) and the first (lower) interval bound in dimension i.

3.1.2 Dependence of the optimal discretization on the reliability index β and the number of discretization cells c_{up}

The influence of the prior failure probability and the maximum size of the CPT, c_{up}, on the optimal discretization is investigated through 10 problems with $n = 2$ random variables in standard normal space. The FORM importance measures of the random variables are selected between 0.1 to 0.995 and the standard deviation of the measurement error is fixed to $\sigma_\varepsilon = 1.0$. We find that the optimal discretization frame is again centered at the design point, i.e. $v_i^{opt} = 0$, and that the intervals are distributed uniformly among the dimensions.

Firstly, we vary the maximum CPT size c_{up}, i.e. the total number of discretization cells. The reliability index is $\beta_{FORM} = 5.2$. Fig. 9 shows the influence of c_{up} on the resulting width of the discretization frame w_i. Three cases are considered: $c_{up} = 25$, $c_{up} = 100$ and $c_{up} = 400$. These choices correspond to 5, 10 and 20 intervals for each random variable. The left side of Fig. 9 shows the relation between the optimal w_i and $|\alpha_i|$. The right side of Fig. 9 shows the same relation, where the w_i's are scaled as in Fig. 8, i.e. the logarithm of the probability mass enclosed by the outer interval boundaries is depicted. As in Fig. 8, there is a clear dependence between the scaled w_i values and the $|\alpha_i|$'s. The interval frames increase with increasing number of random variables.

Secondly, we vary the prior failure probability from 10^{-3} ($\beta = 3.1$) to 10^{-7} ($\beta = 5.2$). The results are shown in Fig. 10. Again, a distinct dependence between the scaled w_i values and the $|\alpha_i|$'s is found. The interval frames decrease with increasing reliability index (with decreasing failure probability).

3.2 Parametric function of optimal discretion frame

As evident from Fig. 9 and Fig. 10, and discussed above, there is a clear dependence of the probability mass enclosed by the optimal discretization frame (with width w_i) on the FORM sensitivity values $|\alpha_i|$. The following parameteric function captures this dependence:

$$\log\big(\Phi(ub_i) - \Phi(lb_i)\big) = a \cdot \exp\,(b \cdot |\alpha_i|) \tag{14}$$

ub_i is the upper and lb_i the lower interval boundary in dimension i, such that $w_i = ub_i - lb_i$. a and b are the parameters of the exponential function. This function is depicted in Figs. 9 and 10. Tab. 1 shows the parameter values a and b for the different combinations of the prior reliability index β and number of intervals per dimension

Table 1. Parameters a and b of Eq. 14 for $\beta = 3.1$, $\beta = 4.3$ and $\beta = 5.2$ as well as 5, 10 and 20 intervals per dimension.

$\begin{bmatrix} a, \\ b \end{bmatrix}$	$n_i = 5$	$n_i = 10$	$n_i = 20$
$\beta = 3.1$	$\begin{bmatrix} -0.28, \\ 2.9 \end{bmatrix}$	$\begin{bmatrix} -1.6 \cdot 10^{-2}, \\ 5.8 \end{bmatrix}$	$\begin{bmatrix} -9.8 \cdot 10^{-4}, \\ 8.7 \end{bmatrix}$
$\beta = 4.3$	$\begin{bmatrix} -0.15, \\ 4.3 \end{bmatrix}$	$\begin{bmatrix} -2.4 \cdot 10^{-2}, \\ 6.1 \end{bmatrix}$	$\begin{bmatrix} -2.1 \cdot 10^{-2}, \\ 6.2 \end{bmatrix}$
$\beta = 5.2$	$\begin{bmatrix} -0.36, \\ 3.7 \end{bmatrix}$	$\begin{bmatrix} -0.11, \\ 5.0 \end{bmatrix}$	$\begin{bmatrix} -3.7 \cdot 10^{-2}, \\ 6.0 \end{bmatrix}$

From the left sides of Fig. 9 and Fig. 10, it can be observed that the relation between $|\alpha_i|$ and w_i is fairly diffuse for random variables with $|\alpha_i| < 0.6$. Here, the parametric relationship of Eq. 14 is less accurate. However, these random variables by definition have lower importance on the reliability estimate. Hence, the inaccuracy of Eq. 14 for random variables with $|\alpha_i| < 0.6$ is not critical, as is confirmed by the numerical investigations performed in the remainder of the paper.

The parameter values of Tab. 1 are derived from two-dimensional problems. In Fig. 8 it is shown that there are no notable differences between two and three dimensions. On this basis, it is hypothesized that the heuristics are applicable also to problems with higher dimensions. This assumption is furthermore supported by the verification examples presented in chapter 4, where the heuristics are applied also to four-dimensional problems without any notable deterioration in the results.

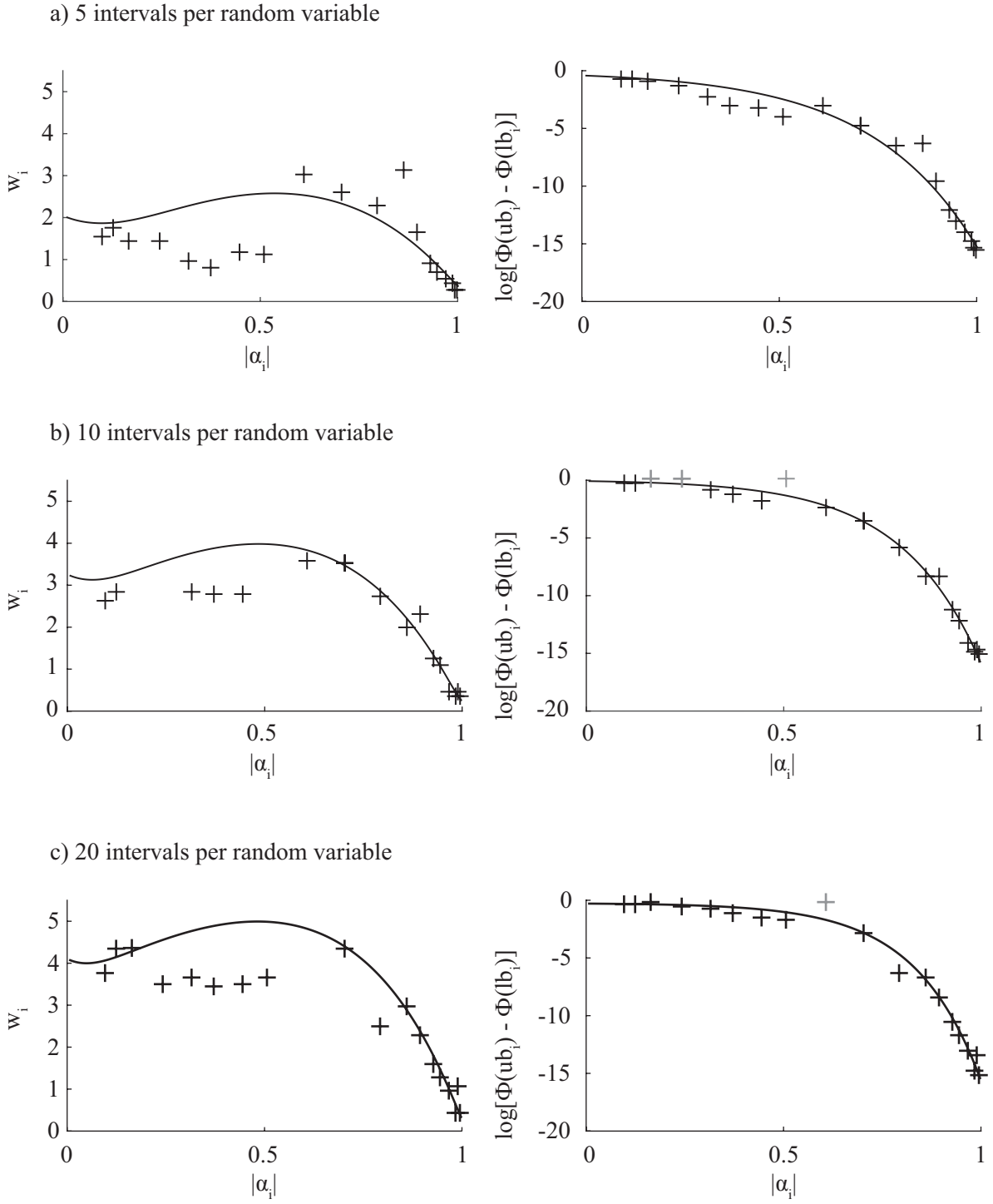

Figure 9. Optimization results for 10 two-dimensional, linear problems in standard normal space, which are discretized with 5, 10 and 20 intervals per dimension. In all cases the prior failure probability is 10^{-7} ($\beta = 5.2$). The crosses represent the optimization results. The solid lines are the fitted parametric function (Eq. 14). The left-hand side shows the relation between the width of a discretization frame w_i and $|\alpha_i|$ and the right-hand side shows the relation between the probability mass enclosed by the discretization frame with width w_i and $|\alpha_i|$. The grey crosses on the right side represent outliers (i.e. results were the optimization was not successful) and are not shown in the figures on the left.

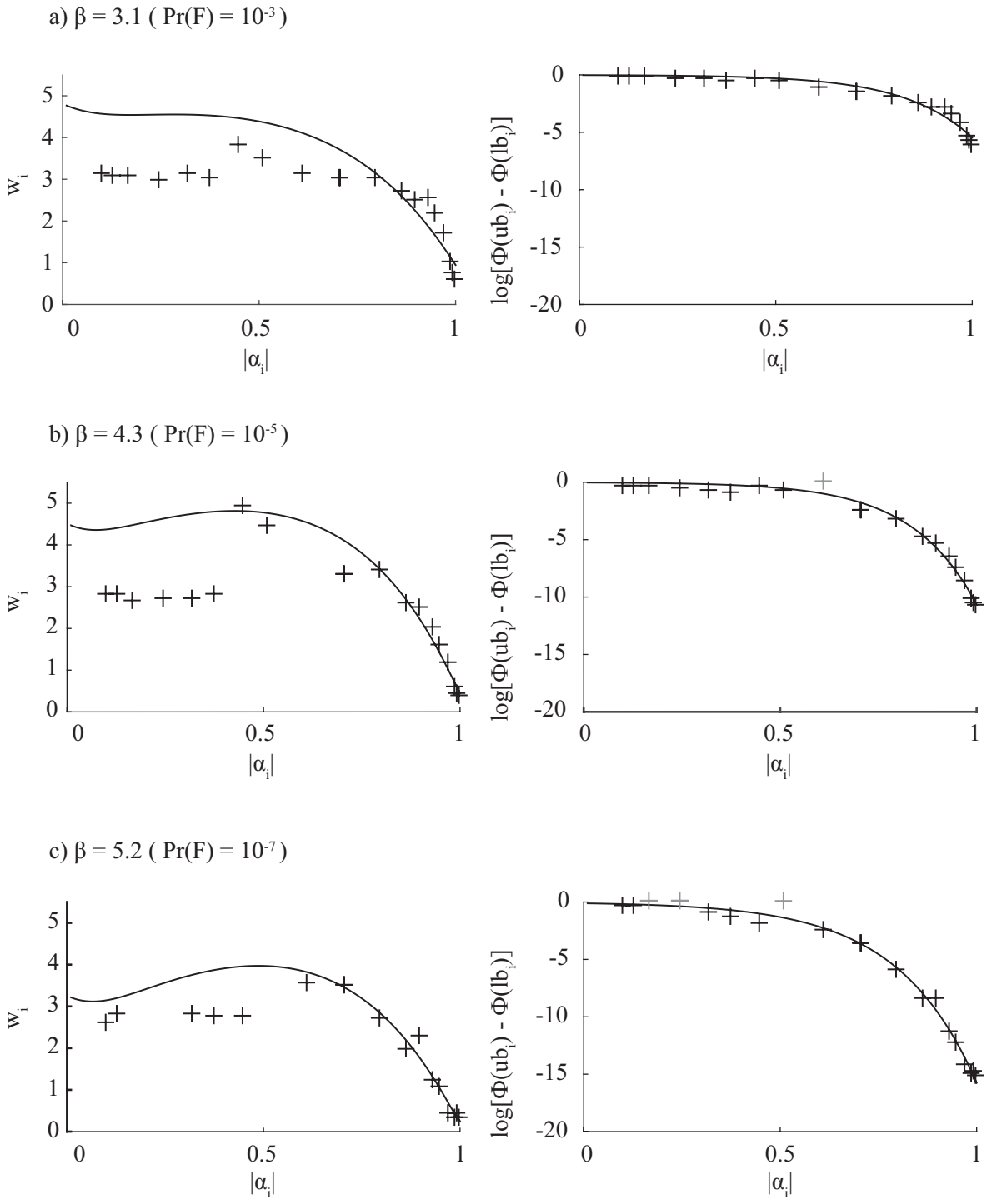

Figure 10. Optimization results for 10 two-dimensional, linear problems in standard normal space, which are discretized with 10 intervals per dimension. The prior failure probabilities are $10^{-3}(\beta = 3.1)$, $10^{-5}(\beta = 4.3)$ and $10^{-7}(\beta = 5.2)$. The crosses represent the optimization results. The solid lines are the fitted parametric function (Eq. 14). The left-hand side shows the relation between the width of a discretization frame w_i and $|\alpha_i|$ and the right-hand side shows the relation between the probability mass enclosed by the discretization frame with width w_i and $|\alpha_i|$. The grey crosses on the right side represent outliers (i.e. results were the optimization was not successful) and are not shown in the figures on the left.

3.3 Summary of the proposed procedure

The steps of the proposed procedure are:

1. Formulate the reliability problem
2. Set up the corresponding BN
3. Perform a FORM analysis for the reliability problem
4. Simplify the BN by removing nodes based on:
    a. Their importance for prediction
    b. Their observability
    c. Whether or not a node simplifies modeling of dependencies
    d. Whether or not it is desired to explicitly show a node in the BN for communication purposes
5. Find the discretization scheme in U-space based on the proposed heuristics i.e.:
    a. The discretization scheme is centered at the design point
    b. The same number of intervals is used for each random variable
    c. The width of the discretization frame follows Eq. 14
6. Transform the discretization scheme to X-space
7. Compute the CPTs of the component state node and the basic random variables

A MATLAB based software tool performing these steps (except step 1) is available for download under www.era.bgu.tum.de/software.

4 Applications

4.1 Verification example I

For verification purposes, we apply the proposed methodology to the discretization of a general limit state with non-normal dependent random variables. The approximation error made by this discretization is investigated for different measurement outcomes.

Failure is defined through the linear LSF $g(\mathbf{x})$:

$$g(\mathbf{x}) = a - \prod_{i=1}^{n} X_i \tag{15}$$

i.e., failure corresponds to the event $\{\prod_{i=1}^{n} X_i \geq a\}$.

The basic random variables are distributed as $X_1 \sim LN(0,0.5)$ and $X_2, \dots, X_n \sim LN(1,0.3)$ (values in parenthesis are the parameters of the lognormal distribution). The statistical dependence among the X_i is described through a Gaussian copula model, with pairwise correlation coefficients ρ_{ij}. The parameters a and ρ_{ij} determine the prior failure probability P_F. Measurements $M_\mathrm{i} = m_\mathrm{i}$ are available for all basic random variables; they are associated with multiplicative measurement errors $\varepsilon_\mathrm{i} \sim \mathrm{LN}(0,0.71)$. In Tabs. 2 and 3, different cases

with 3 and 4 random variables are shown. These cases differ with respect to the prior failure probability P_F, the correlation between the random variables ρ_{ij} and the observed measurements **m**. For each case, a reference solutions $P_{F|\mathbf{M}}$ is calculated analytically.

Table 2. Evaluation of the discretization error for different measurements **m**, for problems with $n = 3$ random variables. a is the constant of the LSF, Eq. 15; ρ_{ij} is the correlation coefficient between X_i and X_j for all $i \neq j$; P_F and $P_{F|\boldsymbol{M}}$ denote the analytically calculated prior and posterior failure probabilities, respectively; $\widehat{P}_{F|\boldsymbol{M}}$ is the conditional failure probability calculated with the discrete BN.

a	c_{up}	ρ_{ij}	P_F	**m**	$P_{F\|\boldsymbol{M}}$	$\widehat{P}_{F\|\boldsymbol{M}}$	Absolute error	Relative error
100	10^3	0	$3.6E-5$	[3.0,2.9,2,9]	$4.3E-5$	$4.5E-5$	$3E-6$	6
100	10^3	0	$3.6E-5$	[2.3,1.1,2,1]	$4.6E-6$	$5.3E-6$	$7E-7$	14
100	10^3	0	$3.6E-5$	[0.9,2.4,0.9]	$2.8E-7$	$3.5E-7$	$7E-8$	25
200	15^3	0.5	$1.6E-4$	[1.6,2.0,1.2]	$1.4E-6$	$1.4E-6$	$1E-7$	4
400	8^3	0.5	$6.4E-6$	[2.6,3.0,3.2]	$8.2E-7$	$8.9E-7$	$7E-8$	9
400	12^3	0.5	$6.4E-6$	[3.6,3.3,4,3]	$4.9E-6$	$5.0E-6$	$1E-9$	3

Table 3. Evaluation of the discretization error for different measurements **m**. The number of random variables $n = 4$; a is the constant of the LSF, Eq. 15; ρ_{ij} is the correlation coefficient between X_i and X_j for all $i \neq j$; P_F and $P_{F|\boldsymbol{M}}$ denote the analytically calculated prior and posterior failure probabilities, respectively; $\widehat{P}_{F|\boldsymbol{M}}$ is the conditional failure probability calculated with the discrete BN.

a	c_{up}	ρ_{ij}	P_F	**m**	$P_{F\|\boldsymbol{M}}$	$\widehat{P}_{F\|\boldsymbol{M}}$	Absolute error	Relative error [%]
400	10^4	0	$1.7E-5$	[2.2,3.2,2.4,3.4]	$9.5E-6$	$1.0E-5$	$9E-7$	9
400	10^4	0	$1.7E-5$	[1.6,1.6,1.6,2.0]	$6.5E-7$	$7.9E-7$	$1E-7$	21
400	10^4	0	$1.7E-5$	[1.1,2.3,1.9,1.2]	$2.4E-7$	$3.0E-7$	$6E-8$	26
600	10^4	0.5	$1.3E-3$	[3.3,1.7,2.8,2.6]	$4.2E-4$	$4.3E-4$	$2E-5$	4
800	8^4	0.5	$5.3E-4$	[1.9,2.0,1.9,2.4]	$1.8E-5$	$1.9E-5$	$1E-6$	8

The results in Tables 2 and 3 show that the proposed methodology for discretization leads to generally acceptable errors in the posterior probability estimate. (It is reminded that the prior error is zero.) As expected, the relative error is larger when the posterior probability is low, and the absolute error is larger when the posterior probability is high. This follows from the

error measure defined in Eq. 11, which balances the relative with the absolute error. In addition, the results do not display any apparent effect of correlation on the accuracy.

To assess the effect of the choice of the number of discretization intervals, the failure probability $\widehat{\mathrm{P}}_{\mathrm{F}|\boldsymbol{M}}$ was calculated for a discretization scheme with up to 20 intervals per RV for the fourth measurement case in Tab. 2. The estimated failure probabilities $\widehat{\mathrm{P}}_{\mathrm{F}|\boldsymbol{M}}$ are plotted together with the exact solution in Fig. 11.

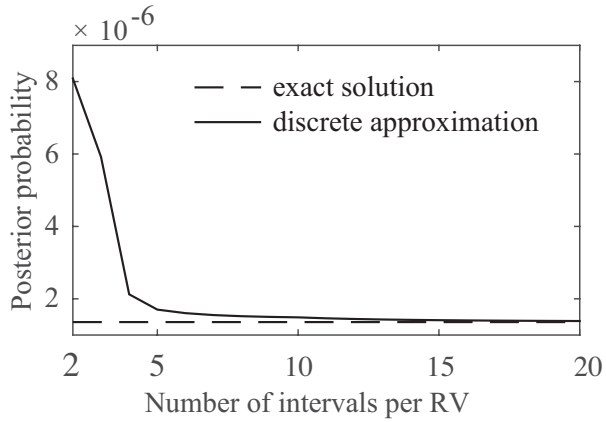

Figure 11. Posterior probability $\widehat{\mathrm{P}}_{\mathrm{F}|\boldsymbol{M}}$ as a function of the number of intervals per random variable together with the exact (analytical) solution $\mathrm{P}_{\mathrm{F}|\boldsymbol{M}}$ for the fourth measurement case in Tab 2.

4.2 Verification example II

The failure criterion applied in verification example I (Eq. 15) leads to a linear LSF in U-space. To verify the accuracy of the proposed method for problems with non-linear LSFs in U-space, we additionally investigate the following LSF:

$$g(\mathbf{x}) = a - \sum_{i=1}^{n} X_i \tag{16}$$

Again the basic random variables X_1 to X_n are distributed as $X_1 \sim LN(0,0.5)$ and $X_2, \dots, X_n \sim LN(1,0.3)$. Different cases with $n = 2$, 3 and 4 random variables are investigated. Measurements $M_\mathrm{i} = m_\mathrm{i}$ are available for all basic random variables; associated to these measurement are multiplicative measurement errors $\varepsilon_\mathrm{i} \sim \mathrm{LN}(0,0.71)$. For independent random variables X_i it is possible to determine posterior distributions $f_{X_i|M_i}(x_i|m_i)$ analytically. The posterior failure probabilities $P_{F|\boldsymbol{M}}$, which are used as reference solutions, are calculated through importance sampling with 10^7 samples. The results are presented in Tab. 4.

Table 4. Evaluation of the discretization error for different measurements **m**. The problems have $n = 2, 3$ or 4 random variables; a is the constant of the LSF, Eq. 16; ρ_{ij} is the correlation coefficient between X_i and X_j for all $i \neq j$; P_F and $P_{F|\boldsymbol{M}}$ denote the prior respectively posterior failure probabilities, which are calculated through importance sampling with 10^7 samples; $\widehat{P}_{F|\boldsymbol{M}}$ is the conditional failure probability calculated with the discrete BN.

a	c_{up}	ρ_i	P_F	**m**	$P_{F\|\boldsymbol{M}}$	$\widehat{P}_{F\|\boldsymbol{M}}$	Absolute error	Relative error [%]
$\underline{n = 2:}$								
12	10^2	0	$1.3E-5$	[2.8,4.5]	$1.4E-5$	$1.2E-5$	$2E-6$	15
12	10^2	0	$1.3E-5$	[2.3,2.4]	$3.3E-6$	$3.5E-6$	$2E-7$	6
10	12^2	0	$1.7E-4$	[4.0,3.2]	$4.0E-4$	$3.7E-4$	$3E-5$	7
$\underline{n = 3:}$								
15	10^3	0	$3.7E-5$	[2.1,5.6,5.0]	$4.8E-5$	$4.5E-5$	$4E-6$	7
15	10^3	0	$3.7E-5$	[1.1,3.7,3.4]	$1.5E-5$	$1.8E-5$	$3E-6$	20
13	12^3	0	$5.0E-4$	[3.0,3.0,3.0]	$5.4E-4$	$5.4E-5$	$3E-6$	1
$\underline{n = 4:}$								
20	8^4	0	$7.4E-6$	[2.0,4.0,3.4,3.0]	$4.9E-6$	$5.6E-6$	$6E-7$	13
17	8^4	0	$3.1E-4$	[1.0,1.4,1.2,2.0]	$4.5E-5$	$5.2E-5$	$7E-6$	15
17	12^4	0	$3.1E-4$	[3.1,2.0,3.3,2.4]	$2.5E-4$	$2.5E-4$	$7E-6$	3

The results in Tab. 4 do not differ substantially from Tabs. 2 and 3. This indicates that the (weak) non-linearity of the LSF function describing failure does not affect the accuracy significantly.

4.3 Runway overrun

Runway overrun (RWO) of a landing aircraft is one of the most critical accidents types in civil aviation. A RWO warning system is developed with the proposed discretization procedure. It provides RWO probabilities conditional on observations of the landing-weight, the headwind and the approach speed for different aircraft types and different airports. For a detailed description of how this problem can be treated in BN framework we refer to (Zwirglmaier and Straub, 2015).

Formally, RWO can be expressed as the event of the operational landing distance exceeding the available runway length (Fig. 12). Correspondingly, a LSF for runway overrun can be defined as:

$$g(\mathbf{X}) = \text{Runway length} - \text{Operational landing distance}(\mathbf{X}) \quad (17)$$

with **X** representing the basic random variables of the problem.

(Drees and Holzapfel, 2012) proposed a model for the operational landing distance required by a landing aircraft, which is applied here. The model, as well as the basic random variables **X**, are presented in (Zwirglmaier et al., 2014), which also includes a detailed description of the reliability and sensitivity analysis.

We consider two different airports (AP I and AP II) and two different aircraft types (AC A and AC B). While the aircraft type affects the landing-weight, the airport affects both the headwind and the approach speed. The distribution models for landing-weight, headwind and approach speed deviation at the different airports and with the different aircraft types are given in Tabs. 5–7. All other basic random variables of the problem are not affected by the airport and aircraft type and are as in (Zwirglmaier et al., 2014).

Tab. 8 summarizes the FORM importance measures of all random variables **X** computed for the four combinations of aircrafts and airports.

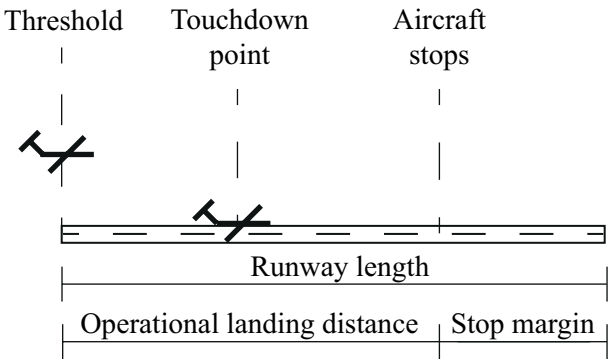

Figure 12. Runway definitions.

Table 5. Distribution models for landing weight conditional on the aircraft.

	Landing weight [t]		
	Distribution	Mean	Std. deviation
AC A	Weibull (min)	59.25	1.69
AC B	Weibull (min)	64.25	1.69

Table 6. Distribution models for head wind conditional on the airport.

	Head wind [kts]		
	Distribution	Mean	Std. deviation
AP I	Normal	5.42	5.75
AP II	Normal	6.51	5.75

Table 7. Distribution models for approach speed deviation conditional on the airport.

	Approach speed deviation [kts]		
	Distribution	Mean	Std. deviation
AP I	Gumbel (max)	4.69	4.21
AP II	Gumbel (max)	5.63	4.21

Table 8. FORM importance measures α_i for each aircraft-airport combination and every basic random variable in the RWO application.

	α_i				
Random variable	(I/A)	(I/B)	(II/A)	(II/B)	Annotation
Landing weight [t]	0.09	0.10	0.11	0.09	Modeled
Headwind [kts]	−0.65	−0.61	−0.67	−0.60	Modeled
Temperature [°C]	0.03	−0.00	−0.03	−0.03	Not important
Air pressure [hPa]	0.01	−0.01	−0.01	−0.00	Not important
Touchdown point [m]	0.20	0.16	0.18	0.20	Modeled
Approach speed deviation [kts]	0.20	0.21	0.20	0.24	Not observable
Time of spoiler deployment [s]	−0.00	−0.00	0.01	0.01	Not important
Time of breaking initiation [s]	0.70	0.74	0.68	0.73	Not observable
Time of reverser deployment [s]	0.03	0.04	0.06	0.05	Not important
Time of breaking end [s]	−0.02	−0.01	0.01	0.02	Not important

4.3.1 Selection of relevant random variables

The applied RWO model includes 10 basic random variables. However, it is sufficient to include only a selection of these explicitly in the BN. Random variables that are not relevant for the prediction of RWO in the considered scenarios can be excluded. This is the case for

552 random variables with a low FORM importance, whose value does not depend significantly
553 on airport and aircraft type. Here, all random variables, whose absolute value of the FORM
554 importance measure $|\alpha_i|'s$ is smaller than 0.1, are excluded (see Tab. 8). The one exception is
555 landing weight, since this variable is substantially influenced by the aircraft type.

556 Furthermore, one can exclude random variables that cannot be measured before the decision
557 on whether to land or not is made. This holds for Touchdown point and the time at which the
558 pilot initiates breaking. Since these basic random variables are also not needed to simplify the
559 modeling of dependencies, it is not necessary to explicitly model them in the BN, as indicated
560 in Tab. 8.

561 ### 4.3.2 BN model

562 The resulting BN of the RWO warning system is shown in Fig. 13. During the aircraft
563 approach, measurements can be obtained for the three basic random variables included in the
564 BN.

565 The random variables were discretized separately for each aircraft-airport combination (joint
566 states of discrete parents) with 8 intervals each, following the proposed discretization
567 procedure. In a second step, the discretization schemes are merged, i.e. the regions of the
568 outcome space, which are discretized with fine intervals for at least one of the aircraft-airport
569 combinations, are discretized with the respective fine intervals also in the merged
570 discretization scheme. In the end 15 (landing-weight), 10 (headwind) and 9 (approach speed
571 deviation) intervals are used to discretize the three basic random variables.

572 For all observable quantities, the measurements m_i are modeled with an additive observation
573 error:

$$m_i = x_i + \varepsilon_i \tag{18}$$

574 ε_i is modeled by a normal distribution with zero mean and standard deviation σ_{ε_i}.

575 For the random variable landing weight (at landing time) we model the standard deviation of
576 the measurement error as $\sigma_{\varepsilon_{LW}} = 0.34\,t$. Due to turbulences governing wind speeds, the
577 measurement of the head wind speed at the time of the measurement is a less reliable
578 indicator for the head wind speed at landing time; we model the measurement error with a
579 standard deviation $\sigma_{\varepsilon_{HW}} = 2.88\,kts$. A high uncertainty is also assumed for the approach
580 speed deviation at landing $\sigma_{\varepsilon_{ASD}} = 4.21\,kts$.

581 49 (Measurement LW), 57 (Measurement HW) and 57 (Measurement ASD) intervals are used
582 to discretize the measurement nodes.

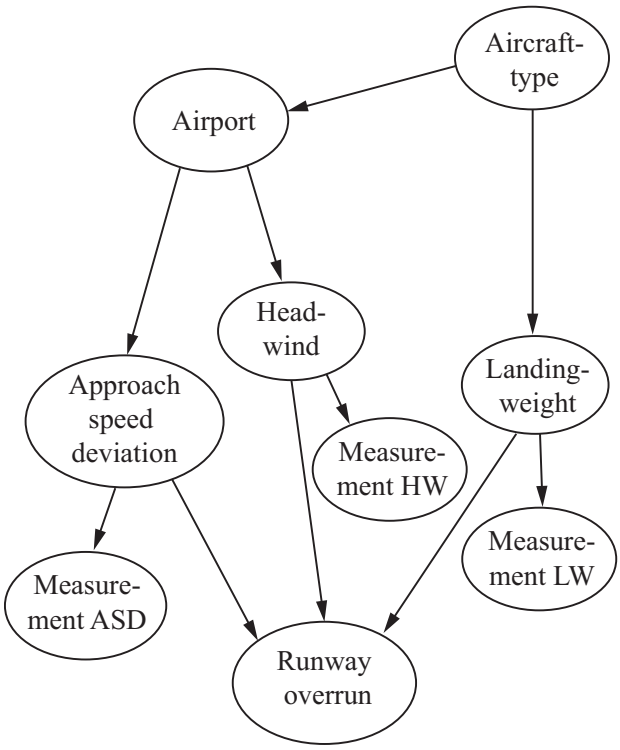

Figure 13. BN structure for a RWO warning system.

4.3.3 Results

In Tab. 9, RWO probabilities for the different airports and aircrafts obtained with the discrete BN are compared to reference solutions, which were calculated by importance sampling around the design point.

Table 9. RWO probabilities for the different airports and aircrafts calculated with the discrete BN p_{BN}, together with reference solutions calculated by importance sampling around the design point p_{DS}. The reference solution has a sampling error with coefficient of variation in the order of 10%.

AP/AC	p_{BN}	p_{DS}
I/A	$2.0e-7$	$1.9e-7$
I/B	$1.0e-6$	$9.2e-7$
II/A	$1.3e-7$	$1.3e-7$
II/B	$6.9e-7$	$6.5e-7$

In Tab. 10, results obtained with the BN for different hypothetical cases of aircrafts approaching an airport are presented. In each of these cases, measurements associated with landing weight, headwind and the approach speed deviation are made. A threshold is used to decide, whether or not the pilot should continue landing or fly to the alternate airport

600 respectively try a second approach. Here we assume that up to a probability of runway
601 overrun of 10^{-6} the pilot should continue landing.

602

603 Table 10. Probabilities of RWO and corresponding decision on landing, computed with the BN for different
604 sets of observations.

Cas	Airpo	Aircra	Meas. LW	Meas. HW	Meas. ASD [kts]	Pr (RWO)	Landing
a)	AP I	AC B	63	0	10.5	$2.5e-8$	Yes
b)	AP I	AC A	61	-10	5	$4.8e-6$	No
c)	AP II	AC B	67	3	0	$6.5e-10$	Yes
d)	AP II	AC A	57.5	-12	3	$1.3e-6$	No

605 ## 5 Discussion

606 When modeling with BNs, it is often necessary or beneficial to discretize continuous random
607 variables. When the BN includes rare events that are a function of such random variables, the
608 choice of the discretization scheme is non-trivial. In this contribution, we investigate this
609 discretization based on FORM concepts, and propose a heuristic procedure for an efficient
610 discretization in these cases. This is based on importance measures α_i obtained through a
611 FORM analysis, which represent the influence of the uncertainty associated with a random
612 variable X_i.

613 The most important finding is that discretization should focus on the area around the most
614 likely failure point (design point), identified by a FORM analysis. Furthermore, we find that
615 optimally all random variables should be discretized with approximately equal numbers of
616 intervals, independent of their importance, as long as $|\alpha_i|$ is not close to zero. The widths of
617 the intervals should be selected based on the FORM importance α_i of the random variables.
618 With increasing importance, the interval width should be reduced, leading to finer
619 discretization for larger $|\alpha_i|$. This relation is particularly evident for $|\alpha_i| \geq 0.8$. We show that
620 it is possible to fit a parametric function to approximate the relation between $|\alpha_i|$ and the
621 optimal width of the region on which the discretization should focus.

622 This parametric function is used to derive a heuristic procedure for finding an efficient
623 discretization. This allows the extrapolation of the optimization results to problems with more
624 random variables. As demonstrated by the verification examples, the heuristic procedure leads
625 to accurate results.

626 In this paper, we restrict ourselves to static discretization. Application of the proposed
627 procedure within dynamic discretization (e.g. (Neil et al., 2008)) should be investigated. The
628 results of the procedure can serve as an initial discretization scheme, which is iteratively

adjusted within dynamic discretization. This might strongly enhance the convergence performance of these algorithms.

Here, we consider only component reliability problems, which are characterized by a single design point. Nevertheless the heuristics derived can also be applied to system reliability problems. System reliability problems can in general be treated as combinations of component reliability problems. Parallel and serial systems are to be distinguished. For parallel systems discretization should be performed based on the joint design point of the problem. For serial systems, following the same line of thought as in the runway overrun example, discretization can be performed separately for each component problem (corresponds to the discrete cases i.e. airport- aircraft combinations in the RWO example). In a second step the discretization schemes can be merged. In the same way it is possible to apply the heuristic to multi state components. One can treat each limit state surface (LSF) defining the boundary between two states separately and merge the discretization schemes afterwards.

The number of basic random variables in a single LSF that can be modeled explicitly in a BN is limited to around 5 to 8. This is due to the exponential growth of the target nodes CPT with increasing number of parents and is independent of the discretization method. Despite this limitation, BNs are applicable to many practical problems – particularly if one considers that usually not all basic random variables need to be modeled explicitly as nodes, as demonstrated in the presented example.

While in this paper the focus was on the discretization of the basic random variables, it is straightforward to incorporate the BNs discussed into larger models.

6 Conclusion

We investigated discretization of continuous reliability problems such that they can be treated in a discrete Bayesian network framework. Reliability problems with linear LSF in standard normal space were considered. These can be seen as FORM approximations of reliability problems. For these linear LSFs optimal discretization schemes were found, which are optimal with respect to an error measure calculated through a preposterior analysis. Since FORM is known to give good approximations also for most non-linear reliability problem, the resulting discretization schemes are efficient also for non-linear LSFs. The main findings presented in this paper are:

- An optimal discretization scheme should discretize finely the area around the FORM design point.
- The size of the sub-region of the outcome space of a random variable X_i can be reduced significantly for random variables whose corresponding uncertainty is dominating the reliability problem

- The number of intervals used for discretization should be approximately equal for all basic random variables

On this basis, we proposed a heuristic that can be used to find an efficient discretization scheme. In verification examples, this heuristic is found to give good accuracy

7 References

AGENA. 2005. Available: http://www.agenarisk.com.

AU, S.-K. & BECK, J. L. 2001. Estimation of small failure probabilities in high dimensions by subset simulation. *Probabilistic Engineering Mechanics,* 16**,** 263-277.

DER KIUREGHIAN, A. 2005. First- and Second-Order Reliability Methods. *In:* NIKOLAIDIS, E., GHIOCEL, D. M. & SINGHAL, S. (eds.) *Engineering Design Reliability: Handbook.* CRC PressINC.

DITLEVSEN, O. & MADSEN, H. O. 2007. Structural Reliability Methods. John Wiley & Sons.

DOUGHERTY, J., KOHAVI, R. & MEHRAN, S. Supervised and Unsupervised Discretization of Continuous Features. *In:* PRIEDITIS, A. & RUSSEL, S., eds. Machine Learning: Proceedings of the Twelfth International Conference, 1995 San Francisco.

DREES, L. & HOLZAPFEL, F. 2012. Determining and Quantifying Hazard Chains and their Contribution to Incident Probabilities in Flight Operation. *AIAA Modeling and Simulation Technologies Conference.* American Institute of Aeronautics and Astronautics.

FRIIS-HANSEN, A. 2000. *Bayesian Networks as a Decision Support Tool in Marine Applications.* PhD, Technical university of denmark.

HANEA, A., NAPOLES, O. M. & ABABEI, D. 2015. Non-parametric Bayesian networks: Improving theory and reviewing applications. *Reliability Engineering & System Safety,* 144**,** 265-284.

HANEA, A. M., KUROWICKA, D. & COOKE, R. M. 2006. Hybrid method for quantifying and analyzing Bayesian belief nets. *Quality and Reliability Engineering International,* 22**,** 709-729.

HOHENBICHLER, M. & RACKWITZ, R. 1981. Nonnormal dependent vectors in structural safety. *Journal of Engineering Mechanics,* 107**,** 1227-1238.

JENSEN, F. V., LAURITZEN, S. L. & OLESEN, K. G. 1990. Bayesian updating in causal probabilistic networks by local computations. *Computational statistics quarterly,* 4**,** 269-282.

JENSEN, F. V. & NIELSEN, T. D. 2007. *Bayesian networks and decision graphs,* New York [u.a.], Springer.

KJAERULFF, U. B. & MADSEN, A. L. 2013. *Bayesian Networks and Influence Diagrams: A Guide to Construction and Analysis*, Springer Publishing Company, Incorporated.

KOTSIANTIS, S. & KANELLOPOULOS, D. 2006. Discretization techniques: A recent survey. *GESTS International Transactions on Computer Science and Engineering,* 32**,** 47-58.

KOZLOV, A. V. & KOLLER, D. 1997. Nonuniform dynamic discretization in hybrid networks. *13th Annual Conference on Uncertainty in AI (UAI).* Providence, Rhode Island.

LANGSETH, H., NIELSEN, T. D., RUMI, R. & SALMERÓN, A. 2012. Mixtures of truncated basis functions. *International Journal of Approximate Reasoning,* 53**,** 212-227.

LANGSETH, H., NIELSEN, T. D., RUMÍ, R. & SALMERÓN, A. 2009. Inference in hybrid Bayesian networks. *Reliability Engineering & System Safety,* 94**,** 1499-1509.

LAURITZEN, S. L. & SPIEGELHALTER, D. J. 1988. Local computations with probabilities on graphical structures and their application to expert systems. *Journal of the Royal Statistical Society. Series B (Methodological)*, 157-224.

LERNER, U. N. 2002. *Hybrid Bayesian networks for reasoning about complex systems.* PhD, Stanford University.

NEIL, M., TAILOR, M., MARQUEZ, D., FENTON, N. & HEARTY, P. 2008. Modelling dependable systems using hybrid Bayesian networks. *Reliability Engineering & System Safety,* 93, 933-939.

RACKWITZ, R. 2001. Reliability analysis—a review and some perspectives. *Structural Safety,* 23, 365-395.

RAIFFA, H. & SCHLAIFER, R. 1961. *Applied statistical decision theory,* Boston, Division of Research, Graduate School of Business Adminitration, Harvard University.

SINDEL, R. & RACKWITZ, R. 1998. Problems and Solution Strategies in Reliability Updating. *Journal of Offshore Mechanics and Arctic Engineering,* 120, 109-114.

STRAUB, D. 2009. Stochastic Modeling of Deterioration Processes through Dynamic Bayesian Networks. *Journal of Engineering Mechanics,* 135, 1089-1099.

STRAUB, D. 2011. Reliability updating with equality information. *Probabilistic Engineering Mechanics,* 26, 254-258.

STRAUB, D. 2014a. Engineering Risk Assessment. *Risk – A Multidisciplinary Introduction.* Springer.

STRAUB, D. 2014b. Value of information analysis with structural reliability methods. *Structural Safety,* 49, 75-85.

STRAUB, D. & DER KIUREGHIAN, A. 2010a. Bayesian Network Enhanced with Structural Reliability Methods: Application. *Journal of Engineering Mechanics,* 136, 1259-1270.

STRAUB, D. & DER KIUREGHIAN, A. 2010b. Bayesian Network Enhanced with Structural Reliability Methods: Methodology. *Journal of Engineering Mechanics,* 136, 1248-1258.

ZHANG, N. L. & POOLE, D. A simple approach to Bayesian network computations. Proc. of the Tenth Canadian Conference on Artificial Intelligence, 1994.

ZHU, J. & COLLETTE, M. 2015. A dynamic discretization method for reliability inference in Dynamic Bayesian Networks. *Reliability Engineering & System Safety,* 138, 242-252.

ZWIRGLMAIER, K., DREES, L., HOLZAPFEL, F. & STRAUB, D. 2014. Reliability analysis for runway overrun using subset simulation. *ESREL 2014.* Wroclaw, Poland.

ZWIRGLMAIER, K. & STRAUB, D. 2015. Discretization of Structural Reliability Problems: An Application to Runway Overrun. *12th International Conference on Applications of Statistics and Probability in Civil Engineering, ICASP12.* Vancouver, Canada.