Basis concepts of stochastic analysis

Grundlagen zur Robustheitsbewertung
virtueller Designentwürfe

Christian Bucher

Vienna University of Technology
& Dynardo Austria GmbH, Vienna

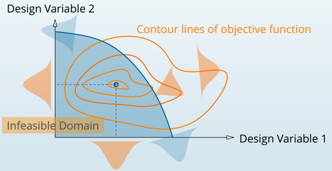
12. Weimarer Optimierungs- und Stochastiktage, 5./6. November 2015

Overview

- Introduction
- Probability theory
- Estimation
- Regression
- Failure probability
- Sampling methods
- Examples

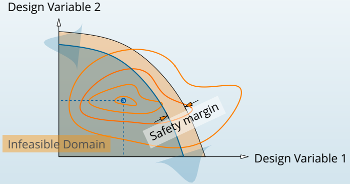
Uncertainties in optimization

- Design variables (e.g. manufacturing tolerances)
- Objective function (e.g. tolerances, external factors)
- Constraints (e.g. tolerances, external factors)

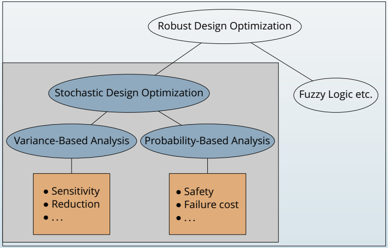


Traditional design approach

- Introduce "safety factors" into the constraints
- Leads to results satisfying safety requirement, but not necessarily optimal designs

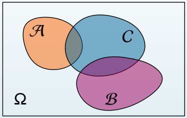


Tools for optimal robust design



Probability

- Events

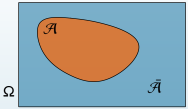


- Axioms(Kolmogorov)

$$\text{I}: \quad 0 \le \text{P}[\mathcal{A}] \le 1$$
$$\text{II}: \quad \mathbf{P}[\Omega] = 1$$
$$\text{III}: \quad \mathbf{P}[\mathcal{A} \cup \mathcal{B}] = \mathbf{P}[\mathcal{A}] + \mathbf{P}[\mathcal{B}]$$

$$\mathbf{P}[\mathcal{A} \cup \mathcal{C}] = \mathbf{P}[\mathcal{A}] + \mathbf{P}[\mathcal{C}] - \mathbf{P}[\mathcal{A} \cap \mathcal{C}]$$

Complementary Event



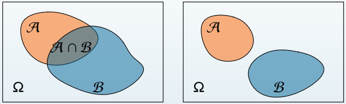
- An event can either happen or not happen

$$\mathbf{P}[\mathcal{A}] + \mathbf{P}[\bar{\mathcal{A}}] = \mathbf{P}[\Omega] = 1$$

- An event cannot happen and not happen at the same time

$$\mathbf{P}[\mathcal{A} \cap \bar{\mathcal{A}}] = \mathbf{P}[\emptyset] = 0$$

Conditional Probability



- Definition

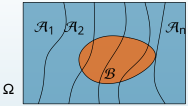
$$\mathbf{P}[\mathcal{A}|\mathcal{B}] = \frac{\mathbf{P}[\mathcal{A} \cap \mathcal{B}]}{\mathbf{P}[\mathcal{B}]}$$

- Independence

$$\mathbf{P}[\mathcal{A}|\mathcal{B}] = \mathbf{P}[\mathcal{A}]$$

$$\rightarrow \mathbf{P}[\mathcal{A} \cap \mathcal{B}] = \mathbf{P}[\mathcal{A}]\mathbf{P}[\mathcal{B}]$$

Decomposition of event space



- Total probability

$$\mathbf{P}[\mathcal{B}] = \mathbf{P}[\mathcal{B}|\mathcal{A}_1]\,\mathrm{P}[\mathcal{A}_1] + \ldots + \mathbf{P}[\mathcal{B}|\mathcal{A}_n]\mathbf{P}[\mathcal{A}_n]$$

- Bayes' theorem

$$\mathbf{P}[\mathcal{A}_i|\mathcal{B}] = \frac{\mathbf{P}[\mathcal{B}|\mathcal{A}_i]\mathbf{P}[\mathcal{A}_i]}{\mathbf{P}[\mathcal{B}|\mathcal{A}_1]\mathbf{P}[\mathcal{A}_1] + \ldots + \mathbf{P}[\mathcal{B}|\mathcal{A}_n]\mathbf{P}[\mathcal{A}_n]}$$

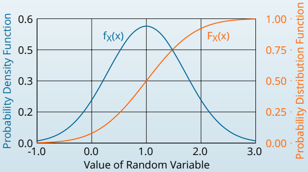
Random Variables

- Distribution function

$$F_X(x) = \mathbf{P}[X < x]; \lim_{x\to-\infty} F_X(x) = 0; \lim_{x\to+-\infty} F_X(x) = 1$$

- Probability density function

$$f_X(x) = \frac{d}{dx} F_X(x)$$



Expected values

- Definition

$$\mathbf{E}[g(X)] = \int_{-\infty}^{\infty} g(x) f_X(x) dx$$

- Mean value

$$\bar{X} = \mathbf{E}[X] = \int_{-\infty}^{\infty} x f_X(x) dx$$

- Variance (square of standard deviation)

$$\sigma_X^2 = \mathbf{E}[(X - \bar{X})^2] = \int_{-\infty}^{\infty} (x - \bar{X})^2 f_X(x) dx$$

- Coefficient of variation (dimensionless)

$$V_X = \frac{\sigma_X}{\bar{X}}; \quad \bar{X} \neq 0$$

- Expectation is a linear operator

$$\mathbf{E}[g + h] = \mathbf{E}[g] + \mathbf{E}[h]; \quad \mathbf{E}[\lambda g] = \lambda \mathbf{E}[g]$$

Standardization

- Definition of standardized variables

$$Y = \frac{X - \bar{X}}{\sigma_X}$$

- Mean value

$$\mathbf{E}[Y] = \frac{1}{\sigma_X}(\mathbf{E}[X] - \mathbf{E}[\bar{X}]) = 0$$

- Variance

$$\mathbf{E}[Y^2] = \frac{1}{\sigma_X^2}\mathbf{E}[(X - \bar{X})^2] = \frac{\sigma_X^2}{\sigma_X^2} = 1$$

Estimation

- Estimator Γ for an unknown parameter γ (e.g. mean value) from independent observations $X_k; k = 1 \dots n$
- Consistency

$$\Gamma : \Gamma_n = \Gamma(X_1, \dots X_n)$$

$$\forall \epsilon > 0 : \quad \lim_{n\to\infty} P[|\Gamma_n - \gamma| < \epsilon] = 1$$

- Unbiasedness

$$\mathbf{E}[\Gamma_n] = \gamma$$

- Asymptotic unbiasedness

$$\lim_{n\to\infty} \mathbf{E}[\Gamma_n] = \gamma$$

- Any estimate based on finite sample size contains some uncertainty which should be made sufficiently small (usually by adjusting the sample size)

Estimators for mean and variance

- Arithmetic mean is a consistent and unbiased estimator for the mean value $\bar{X}$

$$M_n = \frac{1}{n}\sum_{i=1}^{n} X_i$$

- The variance estimator

$$S_n'^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - M_n)^2$$

is asymptotically unbiased

$$\mathbf{E}[S_n'^2] = \frac{n-1}{n}\sigma_X^2$$

- Therefore

$$S_n^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - M_n)^2$$

is an unbiased estimator for the variance σ_X^2

Estimation error

- Limited number of samples leads to random deviation of the estimate from the true expected value
- Example: estimator for the mean value

$$m_X = \frac{1}{n}\sum_{i=1}^{n} X_i$$

- Variance of the estimated value

$$\sigma_m^2 = \mathbf{E}[(m - \bar{X})^2]$$

- Estimator for the variance of the mean value estimator

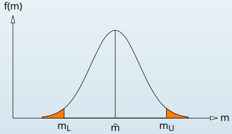
$$S_m^2 = \frac{1}{n(n-1)}\sum_{i=1}^{n}(m - X_i)^2 = \frac{1}{n}S_X^2$$

Confidence interval

- Statistical error (standard deviation) of the estimator

$$S_m = \frac{S_X}{\sqrt{n}}$$

- Assume normally distributed error → Compute confidence interval for estimated value



$$\mathbf{P}[m_L \le m \le m_U] = 1 - \alpha$$

Two distribution functions

- Normal (Gaussian) distribution

$$f_X(x) = \frac{1}{\sqrt{2\pi}\sigma_X} \exp\left[-\frac{(x-\bar{X})^2}{2\sigma_X^2}\right]; \quad -\infty < x < \infty$$

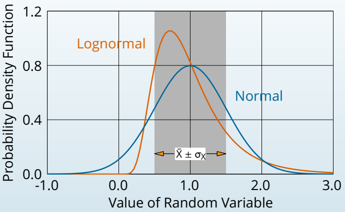
$$F_X(x) = \Phi\left(\frac{x-\bar{X}}{\sigma_X}\right); \quad \Phi(z) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{z} \exp\left(-\frac{u^2}{2}\right) du$$

- Log-normal distribution

$$f_X(x) = \frac{1}{x\sqrt{2\pi}s} \exp\left[-\frac{(\log\frac{x}{\mu})^2)}{2s^2}\right]; \quad 0 \le x < \infty$$

$$F_X(x) = \Phi\left(\frac{\log\frac{x}{\mu}}{s}\right); \quad \mu = \bar{X}\exp\left(-\frac{s^2}{2}\right); \quad s = \sqrt{\ln\left(\frac{\sigma_X^2}{\bar{X}^2}+1\right)}$$

Normal and log-normal density functions



Random Vectors

- Collect all random variables into a random vector

$$\mathbf{X} = [X_1, X_2, \dots X_n]^T$$

- Mean value by applying expectation operator to all components

$$\bar{\mathbf{X}} = \mathbf{E}[\mathbf{X}] = [\bar{X}_1, \bar{X}_2, \dots \bar{X}_n]^T$$

- Covariance matrix

$$\mathbf{C}_{\mathbf{XX}} = \mathbf{E}[(\mathbf{X} - \bar{\mathbf{X}})(\mathbf{X} - \bar{\mathbf{X}})^T]$$

- Coefficient of correlation

$$\rho_{ik} = \frac{\mathbf{E}[(X_i - \bar{X}_i)(X_k - \bar{X}_k)]}{\sigma_{X_i}\sigma_{X_k}}$$

Standardization

- Cholesky decomposition of covariance matrix

$$\mathbf{C}_{\mathbf{XX}} = \mathbf{L}\mathbf{L}^\mathbf{T}$$

- Linear transformation

$$\mathbf{Y} = \mathbf{L}^{-1}(\mathbf{X} - \bar{\mathbf{X}}); \quad \mathbf{X} = \mathbf{L}\mathbf{Y} + \bar{\mathbf{X}}$$

- Transformed vector has zero mean and unit covariance matrix

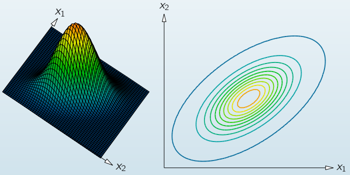
$$\begin{aligned}
\mathbf{E}[\mathbf{Y}] &= \mathbf{E}[\mathbf{L}^{-1}(\mathbf{X} - \bar{\mathbf{X}})] = \mathbf{L}^{-1}\mathbf{E}[\mathbf{X} - \bar{\mathbf{X}}] = \mathbf{0} \\
\mathbf{C}_{\mathbf{YY}} &= \mathbf{E}[\mathbf{Y}\mathbf{Y}^T] = \mathbf{E}[\mathbf{L}^{-1}(\mathbf{X} - \bar{\mathbf{X}})(\mathbf{X} - \bar{\mathbf{X}})^T\mathbf{L}^{-1T}] = \\
&= \mathbf{L}^{-1}\mathbf{L}\mathbf{L}^T\mathbf{L}^{-1T} = \mathbf{I} \\
\mathbf{E}[Y_i^2] &= 1 \quad \forall i; \quad \mathbf{E}[Y_iY_k] = 0 \quad \forall i \neq k
\end{aligned}$$

Joint probability density function

- Multi-dimensional normal distribution

$$f_{\mathbf{X}}(\mathbf{x}) = \frac{1}{(2\pi)^{\frac{n}{2}}\sqrt{\det \mathbf{C_{XX}}}} \exp\left[-\frac{1}{2}(\mathbf{x}-\bar{\mathbf{X}})^T \mathbf{C}_{\mathbf{XX}}^{-1}(\mathbf{x}-\bar{\mathbf{X}})\right];\ \mathbf{x} \in \mathbb{R}^n$$

- Two-dimensional case



Nataf model

- Transformation of correlated non-Gaussian random variables (ρ_{ik}) to correlated standard Gaussian variables (ρ'_{ik})

$$\{X_i; f_{X_i}(x_i)\} \leftrightarrow \{V_i; \varphi(v_i)\}$$

- Mapping

$$V_i = \Phi^{-1}[F_{X_i}(X_i)]$$

- Properties

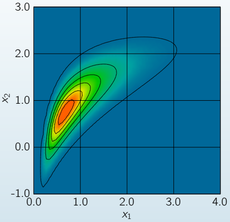
$$\mathbf{E}[V_i] = 0; \quad \mathbf{E}[V_i^2] = 1; \quad \mathbf{E}[V_i V_k] = \rho'_{ik}$$

- Assumption of a multi-dimensional Gaussian distribution

$$f_{\mathbf{V}}(\mathbf{v}) = \frac{1}{(2\pi)^{\frac{n}{2}}\sqrt{\det \mathbf{R_{VV}}}} \exp\left(-\frac{1}{2}\mathbf{v}^T \mathbf{R}_{\mathbf{VV}}^{-1}\mathbf{v}\right)$$

Joint probability density function

- Example: two correlated random variables
- X_1 . . . Lognormally distributed
- X_2 . . . Normally distributes
- Both variables have mean values 1, standard deviations 0.5, correlation ρ_{12} = 0.5

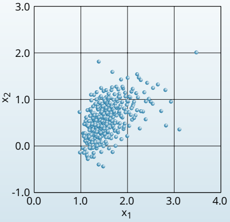


Simulation of correlated variables

- Generate uncorrelated variables with zero mean and unit standard deviation
- Apply linear transformation shown before to obtain correlated Guassian variables
- In this simple form suitable for Gaussian variables
- For non-Gaussian variables additional nonlinear Rosenblatt-Transformation is required
- Special case: Assumption of Gaussian copula (leads to Nataf-model)

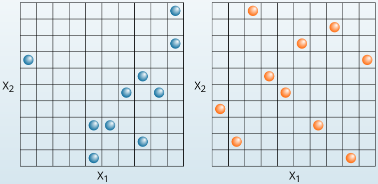
Simulation of samples

- Example: two correlated random variables as before
- Correlation $\rho_{12} = 0.5$
- 300 samples



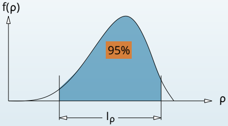
Plain Monte Carlo vs. Latin Hypercube

- 10 samples of uniformly distributed independent random variables



Estimation of correlations

- Repeated simulations lead to different results → estimator for ρ is randomly distributed

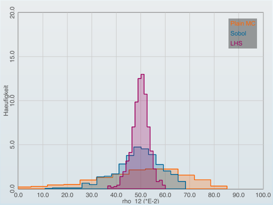


$$I_\rho = [\tanh(z_{ij} - \frac{z_c}{\sqrt{N-3}}),\ \tanh(z_{ij} + \frac{z_c}{\sqrt{N-3}})]$$

$$z_{ij} = \frac{1}{2} \log \frac{1+\rho_{ij}}{1-\rho_{ij}}; \quad z_c = \Phi^{-1}(1-\alpha'/2)$$

Example

- Repeated simulation of two correlated Gaussian variables
- Estimate coefficient of correlation from samples
- Perform statistics on the estimators



Regression

- Adjust a model to experiments

$$Y = f(X, \mathbf{p})$$

- Set of parameters

$$\mathbf{p} = [p_1, p_2, \dots, p_n]^T$$

- Experimental values for input X and output Y

$$(X^{(k)}, Y^{(k)}), k = 1 \dots m$$

- Search for best model by minimizing the residual

$$S(\mathbf{p}) = \sum_{k=1}^{m} \left[Y^{(k)} - f(X^{(k)}, \mathbf{p})\right]^2; \quad \mathbf{p}^* = \text{argmin}\, S(\mathbf{p})$$

Linear regression

- Linear dependence on parameters (not on variables!)

$$f(X, \mathbf{p}) = \sum_{i=1}^{n} p_i g_i(X)$$

- Necessary condition for a minimum

$$\frac{\partial S}{\partial p_j} = 0; \quad j = 1 \dots n$$

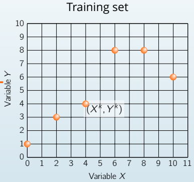
- Solution

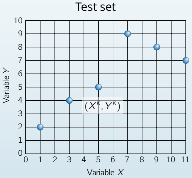
$$\sum_{k=1}^{m} \left\{ g_j(X^k)[Y^k - \sum_{i=1}^{n} p_i g_i(X^k)] \right\} = 0; \quad j = 1 \dots n$$

$$\mathbf{Qp} = \mathbf{q}$$

Example

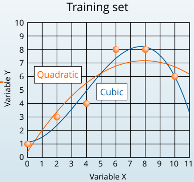
- Adapt polynomial function to 12 data points

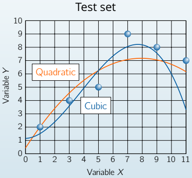




Regression result

- Adjust model to training data, cross-check with test data





Quality of regression

- Coeffient of determination (CoD): correlation between experimental data and model predictions

$$R^2 = \left(\frac{\mathbf{E}[Y \cdot Z]}{\sigma_Y \sigma_Z}\right)^2 ; Z = \sum_{i=1}^{n} p_i g_i(X)$$

- Adjusted (reduced) CoD for small sample sizes

$$R^2_{adj} = R^2 - \frac{n-1}{m-n}\left(1 - R^2\right)$$

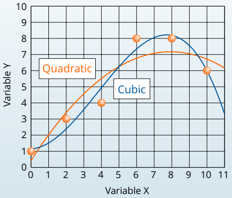
- If an additional test data set T is available: coefficient of prognosis (CoP)

$$CoP = \left(\frac{\mathbf{E}[T \cdot Z_T]}{\sigma_Y \sigma_Z}\right)^2 ;\ Z_T = \sum_{i=1}^{n} p_i g_i(X_T);\ 0 \le CoP \le 1$$

Quality depending on model order

- Repeat regression and test for different polynomial orders

n	R^2	CoP	R^2_{adj}
1	0.83	0.83	0.79
2	0.93	0.92	0.88
3	0.98	0.81	0.94
4	0.98	0.73	0.90
5	1.00	0.70	-



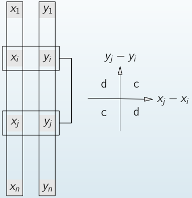
Outliers and correlation

- Only physical effects considered here (no numerical “glitches”)
- Events with small probability destabilize the statistics
- Need stable correlation measures for identification of important variables
- Solutions:
  - Remove outliers. Difficult to define criteria!
  - Use rank-based statistics: Spearman rank order correlation, Kendall tau statistic. Much more stable.

Kendall tau

- Compute differences of elements within vectors x and y
- Compare signs of differences
- Same sign → concordant, different sign → discordant
- Count number of concordant and discordant pairs
- Normalize by possible number of pairs

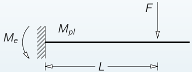
$$\tau = \frac{2}{n(n-1)}\left(\sum_{i,j} c_{ij} - \sum_{i,j} d_{ij}\right)$$



$$\begin{cases} c_{ij} = 1 & (x_j - x_i)(y_j - y_i) > 0 \\ d_{ij} = 1 & (x_j - x_i)(y_j - y_i) < 0 \end{cases}$$

Reliability analysis

- Mechanical system



- Failure condition

$$\mathcal{F} = \{(F, L, M_{pl}) : FL \geq M_{pl}\} = \{(F, L, M_{pl}) : 1 - \frac{FL}{M_{pl}} \leq 0\}$$

- Failure probability

$$\mathbf{P}(\mathcal{F}) = \mathbf{P}[\{\mathbf{X} : g(\mathbf{X}) \leq 0\}$$

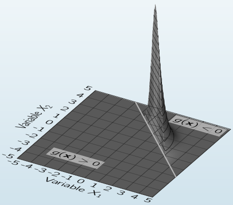
$$\mathbf{P}(\mathcal{F}) = \int_{-\infty}^{\infty} \dots \int_{-\infty}^{\infty} I_g(x) f_{X_1 \dots X_n} dx_1 \dots dx_n$$

$I_g(x_1 \dots x_n) = 1$ if $g(x_1 \dots x_n) \leq 0$ and $I_g(.) = 0$ else

Computational Challenge

- Integrand is non-zero only in a small region
- Difficult to find appropriate integration points
- Example in standard Gaussian space

$$g(x_1, x_2) = 3 - x_1 - x_2$$



First order reliability method (FORM)

- Transformation to standard Gaussian space (here: Rosenblatt transform for Nataf-model)

$$Y_i = \Phi^{-1}[F_{X_i}(X_i)]; \quad i = 1 \dots n$$

$$\mathbf{U} = \mathbf{L}^{-1}\mathbf{Y}; \quad \mathbf{C_{YY}} = \mathbf{LL}^T$$

- Inverse transformation

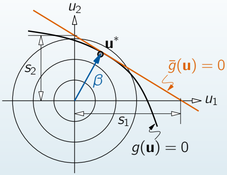
$$X_i = F_{X_i}^{-1}\left[\Phi\left(\sum_{k=1}^{n} L_{ik} U_k\right)\right]$$

- Computation of "design point"

$$\mathbf{u}^* : \mathbf{u}^T\mathbf{u} \rightarrow \text{Min.}; \quad \text{subjectto} : g[\mathbf{x}(\mathbf{u})] = 0$$

- Linearize the limit state function at the design point in standard Gaussian space

FORM Procedure



$$\bar{g} : -\sum_{i=1}^{n} \frac{u_i}{s_i} + 1 = 0; \quad \sum_{i=1}^{n} \frac{1}{s_i^2} = \frac{1}{\beta^2}$$

$$\mathbf{P}(\mathcal{F}) = \Phi(-\beta)$$

Monte Carlo estimation

- Write failure probability as expectation

$$\mathbf{P}(\mathcal{F}) = p_f = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\dots\int_{-\infty}^{\infty} I_g(x) f_{X_1\dots X_n} dx_1 \dots dx_n$$

- Indicator function

$$I_g(x_1 \dots x_n) = \begin{cases} 1 \text{ for } g(x_1 \dots x_n) \le 0 \\ 0 \text{ else} \end{cases}$$

- Consistent and unbiased estimator (arithmetic mean)

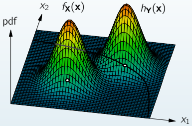
$$\bar{p}_f = \frac{1}{m}\sum_{k=1}^{m} I_g(\mathbf{x}^{(k)})$$

- Variance of estimator

$$\sigma^2_{\bar{p}_f} = \frac{p_f}{m} - \frac{p_f^2}{m} \approx \frac{p_f}{m} \to \sigma_{\bar{p}_f} = \sqrt{\frac{p_f}{m}}$$

Importance sampling

- Simulation density



- Estimator of failure probability

$$\bar{P}(\mathcal{F}) = \frac{1}{m}\sum_{k=1}^{m} \frac{f_{\mathbf{X}}(\mathbf{x})}{h_{\mathbf{Y}}(\mathbf{x})} I_g(\mathbf{x}) = \mathbf{E}\left[\frac{f_{\mathbf{X}}(\mathbf{x})}{h_{\mathbf{Y}}(\mathbf{x})} I_g(\mathbf{x})\right]$$

- Variance of estimator

$$\sigma^2_{\bar{P}(\mathcal{F})} = \frac{1}{m}\mathbf{E}\left[\frac{f_{\mathbf{X}}(\mathbf{x})^2}{h_{\mathbf{Y}}(\mathbf{x})^2} I_g(\mathbf{x})\right]$$

Choice of sampling density 1

- Consider a one-dimensional problem

$$f_X(x) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2}\right); \quad g(x) = \beta - x$$

- Gaussian sampling density with the same variance

$$h_Y(x) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{(x-\bar{Y})^2}{2}\right)$$

- Variance of estimator

$$\begin{aligned}
\sigma^2_{\bar{p}_f} &= \frac{1}{m} \int_\beta^\infty \frac{f_X(x)^2}{h_Y(x)^2} h_Y(x) dx \\
&= \frac{1}{m} \int_\beta^\infty \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{2x^2}{2} + \frac{(x-\bar{Y})^2}{2}\right) dx \\
&= \frac{1}{m} \exp(\bar{Y}^2) \Phi[-(\beta+\bar{Y})]
\end{aligned}$$

Choice of sampling density 2

- Minimize variance of estimator with respect to mean value of sampling density

$$\frac{\partial}{\partial \bar{Y}}(\sigma^2_{\bar{p}_f}) = 0$$

$$\rightarrow 2\bar{Y}\Phi[-(\beta+\bar{Y})] - \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{(\beta+\bar{Y})^2}{2}\right) = 0$$

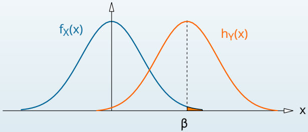
- Asymptotic relation (Mill's ratio)

$$\Phi(-z) \approx \frac{1}{z\sqrt{2\pi}} \exp\left(-\frac{z^2}{2}\right)$$

- Optimal mean value of sampling density

$$\frac{2\bar{Y}}{\beta+\bar{Y}} - 1 = 0 \rightarrow \bar{Y} = \beta$$

Choice of sampling density 3



Importance sampling at the design point

- Determine design point $\mathbf{u}^*$ in standard Gaussian space (e.g. using FORM)
- Construct a multi-dimensional standard Gaussian sampling density centered at the design point with unit covariance matrix (identical to that of the actual random variables in standard Gaussian space)

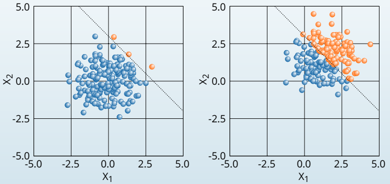
$$h_{\mathbf{Y}}(\mathbf{u}) = \frac{1}{(2\pi)^{\frac{n}{2}}} \exp\left[-\frac{1}{2}(\mathbf{u}-\mathbf{u}^*)^T(\mathbf{u}-\mathbf{u}^*)\right]$$

- Carry out random sampling and estimation of the failure probability

Example

- Two standard Gaussian random variables

$$g(X_1, X_2) = 3 - X_1 - X_2; \quad \mathbf{x}^* = [1.5, 1.5]^T$$



Monte Carlo

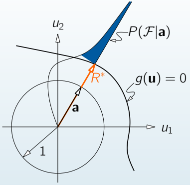
Importance sampling

Directional sampling

- Transformation into standard Gaussian space
- Generate random unit direction vector
- Compute the distance from the origin to the failure domain in this direction (typically using bisection)
- Compute conditional failure probability for this direction (Chi-Square distribution)
- Statistical analysis (estimation of mean and variance)

Procedure

- Conditional failure probability $P(\mathcal{F}|\mathbf{a})$ for one direction $\mathbf{a}$



$$P(\mathcal{F}|\mathbf{a}) = \int_{R^*(\mathbf{a})}^{\infty} f_{R|\mathbf{A}}(r|\mathbf{a})dr =$$

$$= S_n r^{n-1} \frac{1}{\pi^{\frac{n}{2}}} \exp\left(-\frac{r^2}{2}\right) dr = 1 - \chi^2_n[R^*(\mathbf{a})^2]$$

Further reading

- C. Bucher: Computational Analysis of Randomness in Structural Mechanics, Taylor & Francis, 2009.