Epistemic Uncertainty Quantification of Localised Seismic Power Spectral Densities

Marius Bittner[a,b], Marco Behrendt[a], Jasper Behrensdorf[a], and Michael Beer[a,c,d]
[a] Institute for Risk and Reliability, Leibniz Universität Hannover, Germany,
bittner@irz.uni-hannover.de
[b] International Research Training Group 2657 - Computational Mechanics Techniques in High Dimensions, Ecole Normale Supérieure Paris-Saclay, France &
Leibniz University Hannover, Germany
[c] Institute for Risk and Uncertainty, University of Liverpool, United Kingdom
[d] International Joint Research Center for Engineering Reliability and Stochastic Mechanics, Tongji University, Shanghai, China

Abstract: The modelling and quantification of seismic loadings, such as earthquakes, to improve the safe design of structures is a challenging task. In particular, the unpredictable nature of earthquake characteristics like amplitude, dominant frequencies, and duration pose a great risk especially for sensitive structures like power plants, oil rigs, high-rise buildings, or large-span structures. The analysis, understanding and evaluation of those seismic characteristics and their influence on safe structural design is especially important for regions prone to earthquakes. The tectonic mechanisms leading to seismic underground waves are complex but measurements of earthquakes and their mechanical causes on surfaces are available manifold. A new procedure is presented herein for describing uncertainties in the power spectral density (PSD) function of seismic loadings and utilises the novel approach of Sliced-Normal distributions to describe multivariate probability density functions over frequency and amplitude. This representation enables analysts of stochastic dynamic systems the usage of a compact description for PSD functions and to reduce epistemic uncertainties on specific regions prone to earthquake threats. This newly formed PSD function can be used in the simulation of seismic loads via spectral representation or other spectral-based stochastic process generators and is a subsequent development of the already introduced relaxed PSD function.

1. INTRODUCTION

In order to assess the reliability and robustness of buildings and other structures or to design safe structures in the future, it is necessary to carry out extensive simulations. Simulations are often a first step to an abstraction of real structures and to represent them as a simulation model. This is required because a direct application of the safety specifications for sensitive structures in civil engineering is often not possible due to the structural complexity or incomplete information of the system. Such a model can be examined in different scenarios with regard to their excitation. It is thus possible to utilise environmental processes, such as wind or waves, to stress the model or to examine rare extreme events, such as earthquakes. Since these loads, which may also be represented as stochastic processes, can often hardly be described deterministically and are unpredictable, they have a significant influence on systems that exhibit dynamic behaviour [1].

An important tool is the Power Spectral Density (PSD) function, which is based on the Fourier transform and can be used to determine the stochastic processes for their frequency components and amplitude. This is essential to identify whether the natural frequencies of the load coincide with those of the structure, which would result in resonance. The PSD is used in many areas of stochastic dynamics and environmental processes. For example, it can be utilised to determine the response of a structure under dynamic loads, or it can be employed in Monte Carlo (MC) simulations to generate adequate time signals that carry the characteristics of the underlying PSD function [2]. Due to the complexity and the strict mathematical relationship, it is often only possible to estimate the frequency components instead of determining them exactly. An exact determination requires time series of infinite length, do not occur in practice. There are numerous methods for estimating the PSD function from source data, but usually these estimators do not take into account the uncertainties arising from this constraint [3, 4].

The use of real data records supports to generate realistic load models for the simulations. There are a variety of databases that provide different types of data for specific regions and types of environmental processes, such as the Pacific earthquake engineering research centers (PEER) next generation attenuation (NGA) ground-motion databases [5] and others. However, an emerging problem is that real data in general is often subject to uncertainties. Various origins for uncertainties exist, such as simple measurement errors of the sensor, incorrect calibration or total failure. In addition, external influences can affect the measurements, the sensor can be damaged and the placement of the sensor can also influence the recordings. The acquired data are therefore subject to uncertainties that need to be quantified. Particularly in safety-relevant areas, for example when determining whether a building can withstand a certain load, it is important to quantify uncertainties appropriately, as this can have a significant impact on the interpretation of the simulation results. If quantified incorrectly, a result that is actually destructive might be pushed into an acceptable range, and the system can thus be classified as safe, even though it is at high risk of damage or collapse. Instead of discrete simulation results, it is therefore often advantageous to identify a range of possible system response by either determine a potential upper and lower bound with intervals or quantify it probabilistically. In this way, safety margins can be determined instead of discrete values. The consideration of uncertainties in data is therefore indispensable for simulations.

Some approaches have already been introduced for quantifying uncertainties of real data records and the PSD estimation process. In [6, 7], missing data in time series are reconstructed by assuming them to be normally distributed random variables. These are propagated through the Fourier transform and thus an uncertain PSD function is computed. In [8], interval parameters are derived from real seismic records as input parameters for an empirical PSD function. Depending on which bounds are utilised, this results in different representations of the PSD functions and thus a modified excitation. The so-called relaxed PSD, which is a probabilistic representation of an ensemble of data with similar characteristics, was developed in [9]. This approach requires the data to have similar shape, peak frequency and energy in the frequency domain. Due to a high amount of data, reliable statistical information can be extracted that represents each frequency component as a probability density function (PDF) and thus indicates more probable and less probable ranges of the actual underlying PSD. However, the disadvantage of this approach is that correlations and dependencies between frequencies are not taken into account. In this paper, the relaxed PSD is therefore derived using the recently developed Sliced Normal distributions (SN) rather than other feasible methods, such as Copulas.

SN were introduced in [10] and describe a new class of distributions to provide a generalised method for characterising multivariate uncertain quantities. This novel class of distributions is a versatile, generalised model that allows the characterisation of complex (higher dimensional) uncertain parameters with minimal effort. However, the accuracy of characterising the uncertain parameter dependencies is highly dependent on the procedure of designing the SNs, a challenging task seems to be to optimise the estimated SNs towards their ability of enclosing all presented uncertain data and finding a parametric behaviour. A key problem is that in the process of the SNs a mapping from the physical into the so-called feature space is necessary. The tuning of the hyperparameters controlling the SNs and influencing the accuracy of data enclosing however, cannot always ensure optimality and non-convexity in physical and feature space for any arbitrary dimension or data set.

The objective of this work is to develop a more suitable representation of the relaxed PSD using SN. Since the data set is considered as a whole and not as individual frequencies when using SN, correlations and dependencies between the frequencies are also taken into account accordingly. The disadvantage of considering frequencies individually is thus eliminated. The new representation of the relaxed PSD is thus an entire multivariate PDF instead of generating a univariate PDF for each individual frequency.

It is directly applicable for sampling individual PSDs to generate suitable time signals for MC simulation. In this work, SN is considered as a proof-of-concept for a specific optimised feature space. This work is organised as follows: Section 2 provides a brief overview of the theoretical concepts required for this work. The derivation of a relaxed PSD with SN is explained in Section 3. In Section 4, some investigations are carried out with the novel relaxed PSD, such as generating stochastic processes. The work concludes with Section 5.

2. PRELIMINARIES

This section will introduce the basic concepts of the Spectral Representation Method (SRM), ensembles of PSDs and the corresponding relaxed PSD (RPSD), the RPSD constructed from truncated normal distributions (TNRPSD) and the basics of SNs required for the further developments in this work.

2.1. PSD Estimation and Stochastic Process Generation

Stochastic processes are influenced by random occurrences and fluctuations. Such a process cannot be described solely on a deterministic basis. Random variables define the value of the stochastic process at any point in time [11]. Examples of stochastic processes are earthquakes or wind loads subject to high-rise buildings. The estimation of the stationary power spectrum of a given time series x_t can be achieved by the periodogram [4], which is defined by the absolute value of the discrete Fourier transform

$$\hat{S}_X(\omega_k) = \frac{\Delta t^2}{T} \left| \sum_{t=0}^{T-1} x_t e^{-\frac{i2\pi kt}{T}} \right|^2 \tag{1}$$

where Δt describes the time step size, T the total length of the record, t describes the data point index in the record and k is the integer frequency for $\omega_k = 2\pi k/T$.

A stochastic process X_t can be generated utilising an underlying PSD function S_X, either estimated from real data or an analytical model. The autocorrelation function of such a stationary zero-mean process X_t is

$$R_X(\tau) = \sigma^2 \frac{b^4(b^2 - 3\tau^2)}{(b^2 + \tau^2)^3} \qquad -\infty < \tau < \infty \tag{2}$$

and the corresponding analytical expression of the PSD function is given by

$$S_X(\omega) = \tfrac{1}{4}\sigma^2 b^3 \omega^2 e^{-b|\omega|} \qquad -\infty < \omega < \infty. \tag{3}$$

In these equations, σ describes the standard deviation and b is in relation to the correlation of the stochastic process, respectively. Here, $\sigma = 1$ and $b = 1$ is utilised. Both, Eq. (2) and Eq. (3) are used throughout this work when referring to the original ensemble data and to compare the autocorrelation function of such generated processes with their analytical expression [12].

The Spectral Representation Method (SRM) is feasible to adequately generate a stochastic process X_t with the characteristics of the underlying PSD function S_X [12]. SRM reads as follows

$$X_t = \sum_{n=0}^{N_\omega - 1} \sqrt{4 S_X(\omega_n)\Delta\omega}\, (\omega_n t + \varphi_n) \tag{4}$$

where $\omega_n = n\Delta\omega, \quad n = 0,1,2, \dots, N_\omega - 1$, with N_ω as the total number of frequency points, ω_n as the frequency vector, $\Delta\omega$ as frequency step size, t as time vector and φ_n as uniformly distributed random phase angles in the range $[0,2\pi]$. The cut-off frequency is called $\omega_u = (N_\omega - 1)\Delta\omega$, beyond which S_X is assumed to be 0. For a modification of SRM, the concept of Stochastic Harmonic Functions (SHF) was introduced in [13,14] It was shown that for the generation of stochastic processes it is sufficient to capture only partial information of the PSD functions in frequency space. Therefore, it was proposed to divide the frequency space into intervals $[\omega_i, \omega_{i+1}]$. The sampling for the frequency locations was then evenly distributed over these intervals. This results in fewer random variables required overall when considering Eq. (4). However, a new i.i.d. random variable must be introduced for each frequency interval.

2.2. Relaxed Power Spectrum

Epistemic uncertainties are inherent in stochastic processes, especially in real data records. Furthermore, additional uncertainties are introduced by the estimation process, since the estimators, such as the periodogram, do not account for those uncertainties. To capture and quantify these epistemic uncertainties, the relaxed PSD was developed [9]. The relaxed PSD is a probabilistic representation of an ensemble of PSD functions with similar characteristics, such as shape, peak frequency and total power. The ensemble $\{\hat{S}_{X_i}\} \in \mathbb{R}^{N_e \times N_\omega}$ is a set containing N_e PSD functions discretized along ω_n, whereas $s_{i,\omega_n} = \hat{S}_{X_i}(\omega_n), i = 1 \dots N_e$. The relaxed PSD can be used to sample individual PSD functions from the uncertain input space as excitation to approximate the possible response of a system. The strengths of the relaxed PSD can be exploited, especially when a large amount of data is available, as for such an ensemble reliable statistical information can be extracted. For the derivation of the relaxed PSD, it is required to compute the mean value μ_{ω_n} and the standard deviation σ_{ω_n} of the ensemble for each discrete frequency $\omega_n = n\Delta\omega$ with $n = 0,1,2, \dots, N_\omega - 1$. This information can be used to generate a Probability Density Function (PDF) from which random variables are drawn later. In this and previous works, the truncated normal distribution f_{ω_n} is suggested

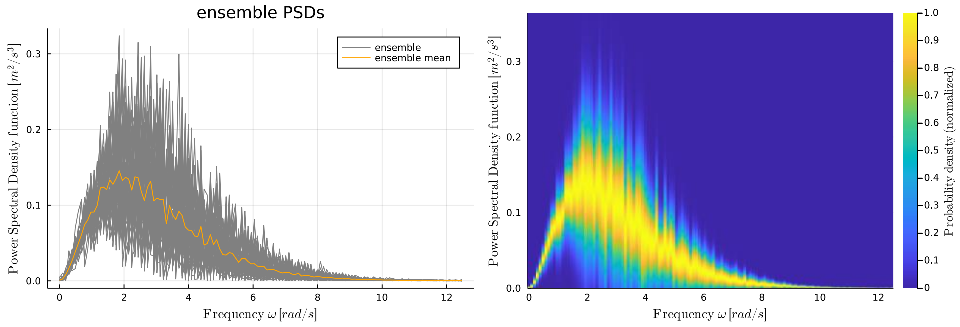
$$f_{\omega_n}(s; \mu, \sigma, a, b) = \frac{1}{\sigma} \frac{\phi\left(\frac{s-\mu}{\sigma}\right)}{\Phi\left(\frac{b-\mu}{\sigma}\right) - \Phi\left(\frac{a-\mu}{\sigma}\right)} \tag{5}$$

where ϕ describes the standard normal distribution and Φ is the corresponding cumulative distribution function, but depending on the appearance of the ensemble, other distribution functions might also be useful. The truncation bounds are denoted by a and b and can be chosen depending on the shape and spectral density values of the ensemble. The only requirement is that the lower bound a must not be smaller than 0 as negative values are physically impossible. Possible ranges are, for instance, $[a = 0, b = \infty]$ or $[a = 0, b = 2\mu_{\omega_n}]$.

An example of an ensemble of PSD functions is given in Fig. 1. The ensemble consists of $N_e = 50$ PSD functions with similar characteristics. The ensemble members are periodograms of Gaussian processes with mean $\mu_{X_t} = 0$ and standard deviation $\sigma_{X_t} = 1$. The relaxed PSD function, estimated from this ensemble, is also depicted. For the estimation the truncation bounds $[a = 0, b = \infty]$ are utilised.

The PDFs for the relaxed PSD must be defined separately for each frequency. Thus, correlations and dependencies between frequencies are not taken into account, which is a disadvantage of this method. To address this issue, the relaxed PSD is derived in this paper using sliced-normal distributions. The generated distribution depicted in Fig.1 on the right side and tuned distributions in Eq. (5) are, for the sake of brevity, referred to as TNRPSD (Truncated Normal Relaxed Power Spectral Density function). Sampling from the TNRPSD is done for each frequency, simple techniques such as inverse sampling can be utilized since the full information for all distributions is available and given with Eq. (5).

Figure 1: Ensemble of PSD functions and derived relaxed PSD function.



2.3. Sliced-normal Distributions

As a proof-of-concept in this work the Sliced Normal distributions (SN) only for a specific optimized feature space is regarded. In the following only, the procedure for this specific design of the SNs is therefore described. Assume data with uncertain and unknown parametric features is available, let this data be denoted by $\delta\colon \mathbb{R}^{n_\delta \times n_o}$, where n_o is the number of available observations and n_δ the dimension of these observations (sometimes referred to as random dimension). One realization of this data retrieved by any stochastic simulation technique (e.g. Monte Carlo simulation) or measurements is called a data sequence and is denoted by $\mathfrak{D}$. From the given data δ a support set $\Delta \in \mathbb{R}^{n_\delta}$ needs to be constructed. Without the presence of outliers, simulation errors or observation errors it can be assumed that $\forall \delta \in \Delta$. In this work we solely assume that the support set is an interval box with the bounds: $\Delta = [min\{\delta\},\ max\{\delta\}] \times [min\{\delta\},\ max\{\delta\}]$. Note that the minimum and maximum bounds can be multidimensional points depending on n_δ.To ensure an optimal enclosing of the data by the SNs a suitable mapping from the physical space (corresponding to the data) into the feature space (which is an artificial augmentation of the physical space) must be formulated. In [10] following mapping function from the physical space δ to a feature space z is suggested

$$z = Z_d(\delta), \tag{6}$$

with $Z_d(\delta)\colon \mathbb{R}^{n_\delta} \to \mathbb{R}^{n_z}$ and the feature space dimension: $n_z = \binom{n_\delta + d}{n_\delta} - 1$. d is an adjustable hyperparameter for the design process of the SNs. d directly corresponds to the dimension of the feature space and the degree of augmentation of the physical space. In [10] $Z_d(\delta)$ is suggested to contain a vector of monomials linked to the data δ and its dimension n_δ. These monomials are sorted in the vector given a specific order, first the monomials are ordered lexicographic and second ordered in ascending degree of the monomials. Note that the lexicographic order is achieved when assuming that the first dimension of the data corresponds to the to the letter a of the alphabet, i.e. $\delta_1 \coloneqq a, \delta_2 \coloneqq b, \ldots$. For example if $n_\delta = 3$ and $d = 2$ the monomials are given as the vectorized functionals $z = [\delta_1, \delta_2, \delta_3, \delta_1^2, \delta_1\delta_2, \delta_2^2, \delta_1\delta_3, \delta_2\delta_3, \delta_3^2]^T$, $z \in \mathbb{R}^{n_\delta \times n_z}$. Given this pre-processing of the data and with the first hyperparameter d we setup the actual construction of the SN.

The Gaussian distribution for $x \in \mathbb{R}^{n_z}$ with mean value $\mu \in \mathbb{R}^{n_z}$ and the inverse of the covariance matrix $P \in \mathbb{R}^{n_z \times n_z}$ is given to be

$$f_{X(x;\,\mu,P)} = \frac{\exp\left(-\frac{(x-\mu)^T P (x-\mu)}{2}\right)}{(2\pi)^{n_\delta/2}\sqrt{(|P^{-1}|)}} \tag{7}$$

Note that μ, P are now additional hyperparameters. P must be symmetric positive definite. Alongside with the formulation of a normalization constant for multivariate Gaussian distributions given as the denominator of Eq. (7) i.e., $c\ (\mu, \mathrm{P}) = (2\pi)^{n_\delta/2}\sqrt{(|P^{-1}|)}$ and Eq. (6) the SN distribution can be given as

$$f(\delta; \mu, P) = \begin{cases} \dfrac{\exp\left(-\dfrac{(Z_d(\delta) - \mu)^T P (Z_d(\delta) - \mu)}{2}\right)}{c(\mu, P)} & \text{for } \delta \in \Delta, \\ 0 & \text{for } \delta \notin \Delta. \end{cases} \tag{8}$$

P can be solely determined by the data sequence $\mathfrak{D}$. However, in [10] and [15] different procedures are introduced for the SNs design and to determine the corresponding hyperparameters. In this work focus only on one procedure presented in [15] under the section and name "Optimized Sliced Normals by Scaling the P Hyperparameter" is laid. Here a convex optimization problem is formulated that maximizes the likelihood in feature space for a given data sequence $\mathfrak{D}$ depending on a scaling value γ for the inverse covariance matrix, i.e., γP. For the sake of brevity, the relation

$$\varphi(\delta, \mu, P) = \exp\left(-\frac{(Z_d(\delta) - \mu)^T P (Z_d(\delta) - \mu)}{2}\right) \tag{9}$$

is introduced. With respect to all previously defined hyperparameters μ, P, d and a suitable construction of the support set Δ, following by γ scalable SN distributions are introduced

$$f(\delta; \mu, \gamma P, \Delta) = \begin{cases} \dfrac{\exp(-\varphi(\delta, \mu, \gamma P))}{c(\mu, \gamma P)} & \text{for } \delta \in \Delta, \\ 0 & \text{for } \delta \notin \Delta. \end{cases} \tag{10}$$

The mean μ is determined by the mean z. $\mu_j = E[z_j] = 1/n_z \sum_{i=1,j}^{n_z} z_{i,j}$. Equivalently P is the covariance matrix corresponding to the data in z. The Gaussian normalization $c(\mu, \gamma P)$ is still related to the rigorous support space Δ, which is not available in a closed expression. Therefore, a numerical estimation of the features of the underlying space must be formulated. In contrast to [10] where a uniform sampling over Δ as hyper ellipsoid is proposed, referenced in [16], in this work the "minimum volume ellipsoid" ϵ is generated using a statistical-geometrical method to find an optimal ellipsoid representing large data sets in high random dimensions, see [17].

Given ϵ which is an approximated representation of Δ, in the feature space, generate n_b samples $u^{n_\delta \times n_b}$ that are in ϵ. With the generated samples u_i , $i = 1, \dots, n_b$, it is possible to estimate the volume V of ϵ. V is an estimation of the volume of the support space Δ. With this relation an approximation of the normalization constant can be given to be

$$c_\Delta(\mu, \gamma P) = \int_\Delta \exp(-\varphi(\delta, \mu, \gamma P)) d\delta \approx \frac{V}{n_b} \sum_{i=1}^{n_b} \exp(-\varphi(u_i, \mu, \gamma P)) \tag{11}$$

The in [15] mentioned convex optimization problem tries to find the optimal scaling parameter γ for a Maximum Likelihood Estimation (MLE) of the SN distribution in Eq. (10) for a specific data sequence $\mathfrak{D}$. The optimization problem which stems from the mentioned MLE is

$$\max_{\gamma \in \mathbb{R}^+} \left\{ \log \prod_{\delta \in \mathfrak{D}} \frac{\exp(-\varphi(\delta, \mu, \gamma P))}{c_\Delta(\mu, \gamma P)} \right\} \tag{12}$$

On the available data $\mathfrak{D}$, with the representation of Δ for the normalization constant Eq. (11) and the corresponding samples u_i following statement is optimized towards γ,

$$\max_{\gamma \in \mathbb{R}^+} \left\{ m \log\left(\frac{1}{c_\Delta(\mu, \gamma P)}\right) - \gamma \mathrm{D} \right\} \tag{13}$$

where $m = n_\delta \cdot n_o$is the total number of available datapoints in δ and scalar $D = \sum_{\delta \in \mathfrak{D}} \varphi(\delta, \mu, P)$ is the unscaled hyperparameter relation in feature space. This procedure to fit the SNs characteristics to the data and ensure an enclosing ability is called "The Covariance Scaling SN Method". In [16] a gradient-based algorithm to localize the global optimum for Eq. (13) is suggested, however in this work the heuristic Nelder-Mead method is applied [18]. For full technical insight on how this implementation is realized, refer to the GitHub repository in [19].

2.3.1 Sampling Procedure

Once the hyperparameters are retrieved from the optimization problem in Eq. (13) a full description of the SN PDF Eq. (10) is possible. To enable a generalized sampling method for the SNs, the Transitional Markov Chain Monte Carlo (TMCMC) scheme as proposed in [20] was implemented. TMCMC can benefit from the a priori definition of the support set Δ, since uniformly distributed prior candidates can be already sampled within Δ. Additionally, since the full PDF can be described a likelihood estimator is readily available. The used implementation of the TMCMC method is available on GitHub [21].

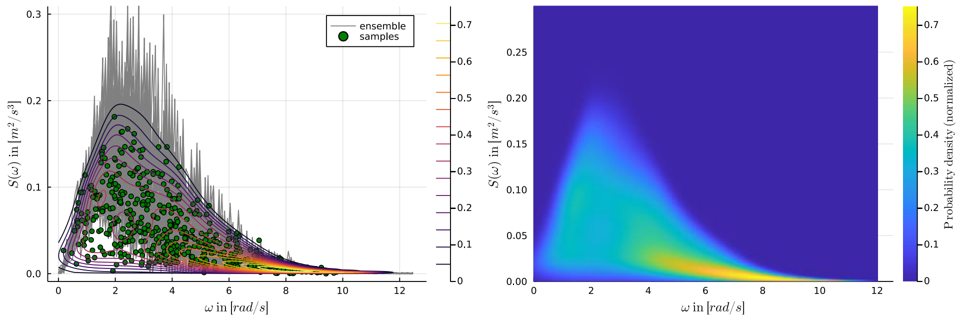
3. RELAXED PSD ESTIMATION USING SLICED-NORMAL DISTRIBUTIONS

The goal of a relaxed PSD estimation is to reduce the epistemic uncertainty stemming from the unprecise or unavailable information of dynamic natural processes. As one illustrative example seismic ground motion is of major interest in this work. The here presented framework utilising novel advances in the design of random variables, namely the tuning and generation of Sliced Normal distributions, shall offer a generalized procedure to estimate future seismic loadings and their magnitudes given past measurements of earthquake accelerograms. The by Sliced Normals designed relaxed Power Spectral Density function shall be denoted as SNRPSD.

3.1 Data Pre-processing

For several measured accelerograms Eq. (1) yields the approximated stationary power spectra, this collection is called an ensemble. In our case only artificial accelerogram time series $\boldsymbol{X_t}$ are regarded as described in Eq. (4). It is important to understand that at this point a major simplification was made because the source spectrum for the ensemble in Fig.1 on the left is a continuous PSD function, as in Eq. (3). The uncertainties in the ensemble in Fig.1 on the left are only stemming from the periodogram estimation in Eq.(1) of the by the continuous PSD function generated Gaussian processes and the fact that contrary to the example presented in [12], Eq. (4) contains $N_\omega = 129$ (In [12] $N_\omega = 128$). The approximated process is Gaussian with $E[X_t] = 0$ and $Var[X_t] = 1$. However, this noisy data is a possibility to represent the actual uncertainty in measured earthquake accelerograms, differences in accelerograms and corresponding PSDs can be found in [1,3,5]. Also it allows for a quick generation of new "observation" data. Additionally, it is possible to analyse the mathematical consistency with the well described example in [12]. Utilizing the procedure described in Chapter 2 to generate a SN distribution, based on the ensemble $\{\hat{S}_{x_i}\}$, with optimality in feature space and scaled hyperparameter P, results in the estimated relaxed PSD in Fig. 2 right. The coverage of the provided data set is not optimal, large values are neglected and values where no density should be estimated, are covered. Mainly the authors assume that this is because an optimality in feature space has been chosen other than the optimality in physical space, as already suggested and visualized in [10] (especially here referring to Fig. 1. and Fig. 2). However, the optimality in physical space can only be achieved by the optimization of a set of parameters. It is not trivial to implement this multi-dimensional optimization problem in a generalized fashion. In this work a simple data processing procedure for functional data of similar shape is proposed. This procedure allows for a simplified handling of data with similar characteristics and enables the usage of the herein presented procedure to design SNs in a generalized fashion.

Figure 2: Ensemble of PSD functions and derived SN relaxed PSD function.



Eventually, the multivariate SNs are serving as multivariate relaxed PSD functions to reduce epistemic uncertainty in data sets containing PSDs of seismic loadings.

In Fig. 1 on the left it is observable that for frequencies larger than 8 (rad/s) the ensemble members converge rapidly to a PSD value of 0, with other words we have a high density of data with similar values. Consequently, in Fig. 2 left and right a large proportion of the by SN sampled points are falling in the tails of the ensemble members and this leads to a high probability density in these areas. For the generation of stochastic processes, a certain number of frequency proportions must be represented by the PSD. Additionally, larger PSD values have a higher impact on the generated process. Therefore, values converging towards zero are less important. A sampling with the standard configuration would result in a poor performance of generating points across the whole frequency range and additionally it might happen, that for a small number of generated samples, information of the PSD is missing. Therefore, a simple data-thinning algorithm is proposed, which is specifically tailored for functional data with similar characteristics. First, the maximum and minimum values in $\{\hat{S}_{x_i}\}$ for each frequency are captured by

$$\begin{aligned} \bar{S}_n &= \arg\max_{\omega_n \in [0,\omega_u]} \hat{S}_{x_i}(\omega_n) \\ \underline{S}_n &= \arg\min_{\omega_n \in [0,\omega_u]} \hat{S}_{x_i}(\omega_n)\,, \end{aligned} \tag{14}$$

next the difference of these vectors $\bar{S}_n$ and $\underline{S}_n$ is calculated to capture the magnitude of deviation, $\grave{S}_n = \bar{S}_n - \underline{S}_n$, from this difference, the maximum and minimum deviation is captured, equivalent to Eq. (13) obtaining, $\bar{s}, \underline{s}$. A simple linear interpolation function is applied

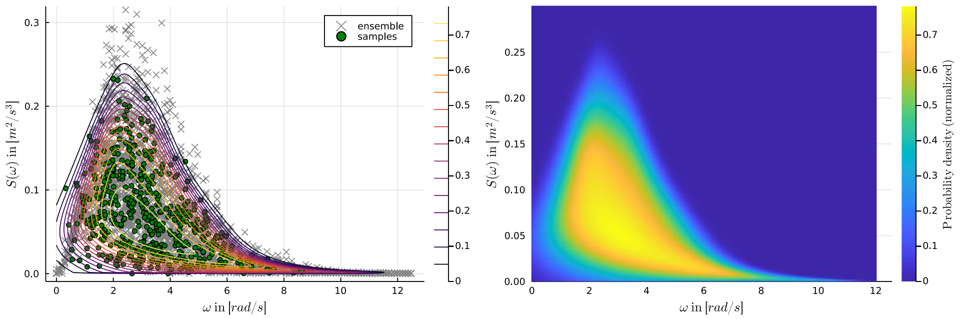
$$y(x) = [y_0(\bar{s} - x) + y_1(x - \underline{s})]/(\bar{s} - \underline{s}) \tag{15}$$

where y_0, y_1 define the rate how many samples shall be kept for the respected differences. To estimate the number of samples that can be deleted from the data for each discretized location following simple relation is given

$$N_{rm,n} = \lfloor N_e(1 - y(\grave{S}_n))\rfloor. \tag{16}$$

for each discretized location in n of $\hat{S}_{x_i}(\omega_n)$ the first $N_{rm,n}$ samples are deleted. With this procedure each discretized value in $\{\hat{S}_{x_i}\}$ is regarded as a single sample. The deletion of samples which could vary in number for different frequencies is an intrusive procedure. However, these samples are only the "training data" for the SN distribution, therefore gaps of information in areas that are not of interest do not directly lead to gaps in the probability density distributed over the parameter space.

Figure 3: Ensemble of thinned-out PSD sample points and derived SN relaxed PSD function.



In Fig. 3 right in comparison to Fig. 2 right a shift of the probability density towards lower frequencies can be regarded, additionally the peak of the ensemble is better approximated, which results in higher possible PSD function values of the PDF. However, problematic are several points. 1. In the first frequency intervals, e.g. $\underline{\overline{\omega_1}} = [0,1]$ the distinct similarity of the data and the steep increase of the PSD cannot be captured well. 2. Some areas are underrepresented and only exhibit very small probability densities, for example the total peaks of the ensemble where $s_p = \hat{S}_{x_i}(\omega_n) > 0.25\frac{m^2}{s^3}$, $f\left(\left\{\omega = 2.5\frac{rad}{s}, s_p\right\}\right) = 4.2412e^{-8}$. 3. The area approximately between $\underline{\overline{\omega}} = [0,2)$ and for $\hat{S}_{x_i}(\omega_n) < 0.05\frac{m^2}{s^3}$ contains a relatively large proportion of probability density values, which should not be the case (refer to Fig.1 left). As already mentioned this problematic coverage of the parameter space results from the optimality in feature space, an optimization approach in physical space could result in a more distinguishable approximation of the probability density in the parameter space. A drawback could then be, that for real measurements with observation errors, the influence of these outliers might be too high, and a special handling of these cases would be necessary. All in all, data-thinning is a low cost and fast way to pre-process functional data to yield a better SN approximation.

3.2 Adaptive Sampling

Since the physical space is now parametrized by the bivariate SN distribution, direct sampling from this distribution type is possible. However, the type of data in the ensemble is functional data and has distinct characteristics. Most importantly is that a larger range of frequency must be covered and filled with information, e.g. $\underline{\overline{\omega}} = [0,8)$. This stands in contrast to SHF in which for each frequency interval a new i.i.d. uniform random variable needed to be introduced.

The whole parametric probability description of the parameter space with a bivariate SN distribution enables the possibility of a one-shot sampling approach (a single initial Monte Carlo simulation). Only a constant number of samples from the whole distribution is sampled and afterwards the samples are sorted according to the predefined frequency intervals to ensure that for each interval at least one sample is present. Additionally, the sorting ensures the functional characteristics of the ensemble data.

Given the SN distribution $f(\{\omega, S\})$ designed with optimality in feature space based on the ensemble data, two dimensional samples $\{\overline{\delta_\iota}\} = \{\delta_\omega, \delta_S\}_l$ can be generated by e.g. TMCMC, with $l = 1, \ldots, N_s$, N_s: Number of desired one-shot samples. The unsorted set elements (each element is in this case a tuple) in $\{\overline{\delta_\iota}\}$ are then sorted according to their ω value in an ascending order. Let the ascendingly sorted sequence be denoted by $_{\triangleleft}\{\overline{\delta_\iota}\}$. To ensure a conform reconstruction of single PSD functions, $N_{\underline{\overline{\omega}}}$ intervals are defined over the whole domain ω. For example, the cut-off frequency $\omega_{\boldsymbol{u}}$ could be set as total upper boundary and $\omega_0 = 0$ the lower boundary. Then a possible interval generation could be:

$$\underline{\overline{\omega}}_i = [\omega_i, \omega_{i+1}), \text{for i} = 0,1, \ldots, N_{\underline{\overline{\omega}}} \tag{17}$$

and $N_{\underline{\overline{\omega}}} = \lceil \omega_u \rceil$. For each interval the first element in the sequence ${}_{\triangleleft}\{\overline{\delta}_\iota\}$. is kept as final PSD function tuple. If no sample is falling in the predefined frequency subinterval, this subinterval is empty. Formally this can be stated as

$$\left\{\overline{S}_i^*\right\} = \begin{cases} {}_{\triangleleft}\{\overline{\delta}_\iota\} & \exists!\, \delta_{\omega,l} \in \underline{\overline{\omega}}_i\,, (l,i) = (1,1),(2,1),\dots,(N_s,1),(1,2),\dots,(N_s,N_{\underline{\overline{\omega}}}) \\ \{\,\} & \text{for } \nexists\, \delta_{\omega,l} \in \underline{\overline{\omega}}_i. \end{cases} \qquad (18)$$

With the recapitulation of the SN theory, the pre-processing of data, i.e. data-thinning and the adaptive sampling a novel technique to describe Relaxed Power Spectral Density Function has been described, the SNRPSD.

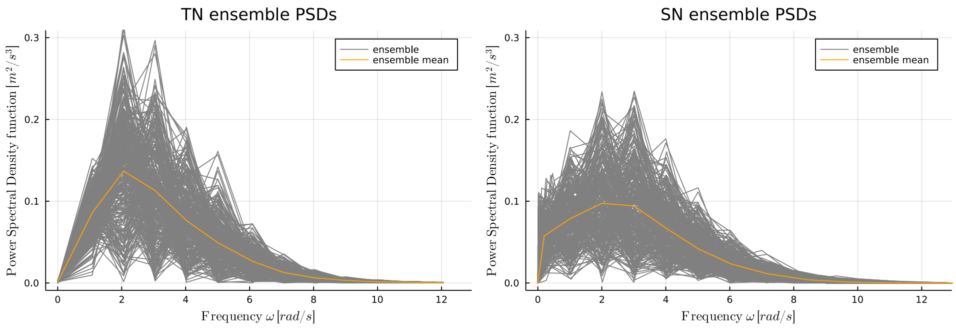
4. STOCHASTIC PROCESS GENERATION

The source ensemble stems from approximated Gaussian processes with mean μ_{X_t}=0 and standard deviation σ_{X_t}=1. Uncertainties of the generated stochastic processes in the sample functions X_t arise from the number of harmonics used. However, the goal is to artificially generate stochastic processes by means of reducing possible epistemic uncertainty present in the source data and at the same time ensure characteristics of the desired artificially simulated process. For this purpose, the global characteristics of the by TNRPSD and SNRPSD generated ensembles and the corresponding stochastic processes are analysed.

4.1 PSD Ensemble Modelling

In Fig. 4 on the left side a sampled ensemble with 300 members utilizing the TNRPSD representation is shown, correspondingly on the right side a sampled ensemble by the SNRPSD is represented.

Figure 4: 300 sampled ensembles by TNRPSD (left) vs SNRPSD (right)

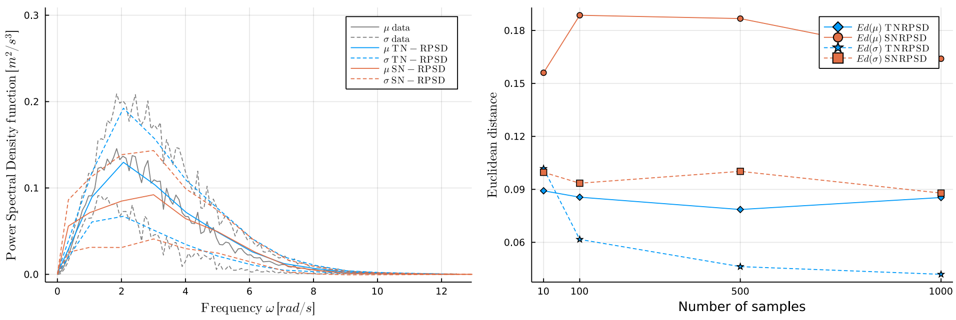


Two major drawbacks of the introduced SNRPSD sampling technique for an artificial ensemble can be seen. First especially the low frequency part of the PSD is not correctly represented by the SN distribution, which can be seen in Fig. 4 where a large proportion of probability mass is in areas where no original data is present. Second the sampled RPSD peak is too small, which is also cause of the probability mass centring around the spots where most data are present. However, despite these drawbacks it is highly important to stress out, that the SNRPSDs are only sampled by a single one-shot sampling of one bivariate random variable, whereas the TNRPSD contains 129 random variables for each frequency discretisation.

4.2. Convergence of Relaxed PSD

In this section the ability of the RPSD methods to reproduce the ensemble data is investigated. In Fig 5 on the left side mean and standard deviation (mean $\mu := E[\hat{S}_{x_i}(\omega_n)]$, $\sigma := \mu \pm \sqrt{Var[\hat{S}_{x_i}(\omega_n)]}$) of two artificial ensembles with 50 members in comparison with the original ensemble are shown. On the right side for 10, 100, 500 and 1000 members respectively the Euclidean distance $Ed(-)$ for mean and standard deviation is calculated. As already recognized earlier it seems, that the SNRPSDs ability to resemble the ensemble is not optimal, however also RPSD exhibits, albeit smaller, deviations.

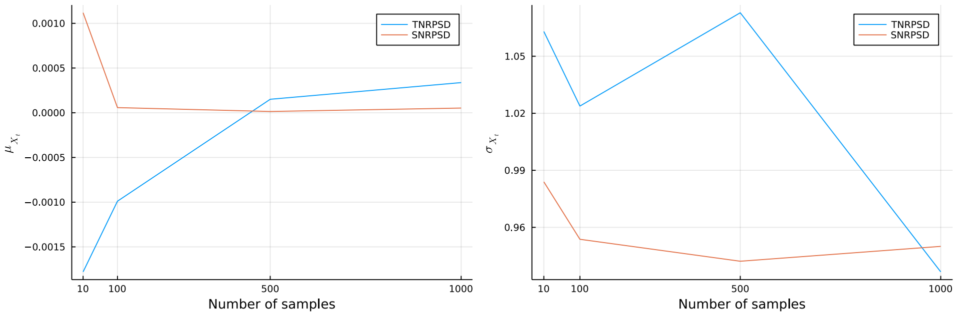
Figure 5 Ensemble data vs TN&SN-RPSD sampled ensembles



4.3. Comparison of Generated Stochastic Processes

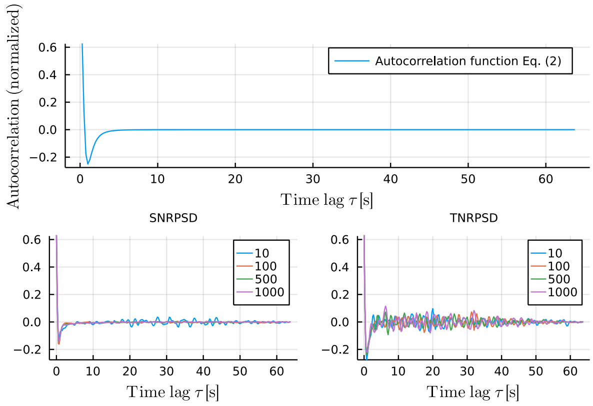
In e.g. [13,14] and [9] some investigations on the number of harmonics N_ω (summation terms in Eq. (4)) needed to approximate Gaussian and non-Gaussian processes sufficiently accurate have been made. For SNRPSD the N_ω can be chosen arbitrarily, similarly to the by Chen et al. introduced SHF [13] only the number of frequency intervals are of importance. For reference see Eq. (17) and $N_{\underline{\overline{\omega}}}$. However, in contrast to SHF and RPSD, the SNRPSD offers a fully continuous description over the frequency space, the intervals are chosen after the SNs are fit to the ensemble data, which offers complete flexibility in choosing the desired number of harmonics in Eq. (4). To compare the following generated stochastic processes, $N_\omega = 129$ was chosen. From the source ensemble which is estimated by Eq. (1) from Gaussian processes with $\mu_{X_t} = 0, \sigma_{X_t} = 1$. Therefore, in average over the generation of numerous samples, this behaviour of the artificially generated stochastic processes by TNRPSD and SNRPSD should yield the same values. A small study has been conducted to analyse the moments of the artificially generated stochastic processes and can be seen in Fig. 6. It seems that SNRPSD exhibits a faster convergence to the desired moments, however a more profound investigation with more samples should be conducted.

Figure 6: Average mean and standard deviation comparison for different numbers of generated stochastic processes



Another key aspect of the generated stochastic processes is the autocorrelation. From [12] the autocorrelation function is known, as given in Eq. (2). In Fig. 7 top this function is depicted. Below in Fig. 7 on the left side the averaged autocorrelation function for the by SNRPSD generated stochastic processes can be seen, equivalently on the right side the autocorrelation of the TNRPSD generated stochastic processes are shown. The numbers in the legend refer to the number of stochastic processes used and averaged. Again, it seems that the SNRPSD exhibits a faster convergence to the Autocorrelation Function Eq. (2).

Figure 7: Averaged autocorrelation comparison for different sample sizes (10, 100, 500, 1000)



5. CONCLUSION

The consideration of uncertainties is of paramount importance in engineering in order to perform simulations correctly and to interpret their results accurately. It is therefore essential to properly quantify these uncertainties in the data sets. In the context of power spectra, the already introduced relaxed PSD offers such a possibility. Its disadvantage, however, is that due to the individual consideration of the frequency components, a high number of PDFs must be generated and thus correlations and dependencies between frequencies are not taken into account. In this work, an enhanced version was therefore presented, which is calculated on the basis of sliced-normal distributions. This guarantees a relatively small modelling effort in contrast to other possibilities, such as copulas. Since all data points in the data set are considered simultaneously, a single multivariate PDF results. Thus, dependencies between frequencies are taken into account. Since high probabilities of SNs are calculated especially in the area of high data density, a data-thinning approach was also employed as part of data pre-processing. This manipulates the data in such a way that the multivariate PDF determined via SN represents the data set more realistically. It was also shown that using the novel relaxed PSD function, it is possible to generate adequate stochastic processes for simulation through an adaptive sampling approach. The SNRPSD offers a single multivariate random variable to capture epistemic uncertainty in PSDs of seismic loads or other time dependant natural processes. This drastic decrease of the needed number of random variables could potentially change sensitivity or reliability analysis for complex systems, additionally meta modelling and model updating procedures could benefit from this generalised low random dimensional approach.

Compared to the regular relaxed PSD, the results still differ slightly, which may be due to the fact that in this work the SNs were calculated via optimality in feature space. An alternative is to calculate the optimality in physical space, in which the SNs become more data-enclosed. The shape of the ensemble of PSDs should then be better captured, especially the peak frequency. Furthermore, areas without data points in the SNs are mostly excluded. The recently introduced sliced-exponential (SE) distributions [23] also offer room for investigation.

Although the use of stationary stochastic processes and the estimation of the resulting stationary PSDs only provides an approximation to the real case, realistic results can be obtained. In a next step, however, evolutionary PSDs should be used, as they represent not only the transformation of the signal in the frequency domain, but the time-frequency domain. Thus, frequency changes over time can also be detected. The SNs and SEs offer both for the optimality in the feature space and in the physical space the possibility to easily extend the dimension to include time.

Acknowledgements

This work is partially funded by the Deutsche Forschungsgemeinschaft (DFG – German Research Foundation) under the research training group GRK 2657 “Computational Mechanics Techniques in High Dimensions”.

References

[1] A. K. Chopra, *Dynamics of Structures: Theory and Applications to Earthquake Engineering*, Prentice Hall, 1995.
[2] G. I. Schuëller, *Efficient Monte Carlo simulation procedures in structural uncertainty and reliability analysis - recent advances*, Structural Engineering and Mechanics 32 (2009) 1–20.
[3] J. Li, J. Chen, *Stochastic Dynamics of Structures*, John Wiley & Sons, 2009.
[4] D. E. Newland, *An Introduction to Random Vibrations, Spectral & Wavelet Analysis*, Courier Corporation, 2012.
[5] T. Kishida, V. Contreras, Y. Bozorgnia, N. Abrahamson, S. Ahdi, T. Ancheta, et al., *NGA-Sub ground motion database*. *UCLA* (2021), Retrieved from https://escholarship.org/uc/item/3bn528xc.
[6] L. Comerford, I. A. Kougioumtzoglou, M. Beer, *On quantifying the uncertainty of stochastic process power spectrum estimates subject to missing data*, International Journal of Sustainable Materials and Structural Systems 2 (2015) 185–206, https://doi.org/10.1504/IJSMSS.2015.078358.
[7] Y. Zhang, L. Comerford, I. A. Kougioumtzoglou, E. Patelli, M. Beer, *Uncertainty Quantification of Power Spectrum and Spectral Moments Estimates Subject to Missing Data*, ASCE-ASME Journal of Risk and Uncertainty in Engineering Systems, Part A: Civil Engineering 3 (2017) 04017020, https://doi.org/10.1061/AJRUA6.0000925.
[8] G. Muscolino, F. Genovese, A. Sofi, *Reliability Bounds for Structural Systems Subjected to a Set of Recorded Accelerograms Leading to Imprecise Seismic Power Spectrum*, ASCE-ASME Journal of Risk and Uncertainty in Engineering Systems, Part A: Civil Engineering 8 (2022) 04022009, https://doi.org/10.1061/AJRUA6.0001215.
[9] M. Behrendt, M. Bittner, L. Comerford, M. Beer, J. Chen, *Relaxed power spectrum estimation from multiple data records utilising subjective probabilities*, Mechanical Systems and Signal Processing 165 (2022) 108346, 10.1016/j.ymssp.2021.108346.
[10] L. G. Crespo, B. K. Colbert, S. P. Kenny, D. P. Giesy, *On the quantification of aleatory and epistemic uncertainty using Sliced-Normal distributions*, Systems & Control Letters 134 (2019) 104560, https://doi.org/10.1016/j.sysconle.2019.104560.
[11] M. Priestley, *Spectral Analysis and Time Series*, *Probability and mathematical statistics: A series of monographs and textbooks*, Academic Press, 1982.
[12] M. Shinozuka, G. Deodatis, *Simulation of Stochastic Processes by Spectral Representation*, Applied Mechanics Reviews 44 (1991) 191–204, https://doi.org/10.1115/1.3119501.
[13] J. Chen, W. Sun, J. Li, J. Xu, *Stochastic Harmonic Function Representation of Stochastic Processes*, Journal of Applied Mechanics 80 (2012), https://doi.org/10.1115/1.4006936.
[14] J. Chen, L. Comerford, Y. Peng, M. Beer, J. Li, *Reduction of random variables in the Stochastic Harmonic Function representation via spectrum-relative dependent random frequencies*, Mechanical Systems and Signal Processing 141 (2020) 106718, https://doi.org/10.1016/j.ymssp.2020.106718
[15] B. K. Colbert, L. G. Crespo, M. M. Peet, *A Convex Optimization Approach to Improving Suboptimal Hyperparameters of Sliced Normal Distributions*, American Control Conference (ACC), 2020, pp. 4478–4483, https://doi.org/10.23919/ACC45564.2020.9147403.
[16] J. Dezert, C. Musso, *An efficient method for generating points uniformly distributed in hyperellipsoids*, Proceedings of the Workshop on Estimation, Tracking and Fusion: A Tribute to Yaakov Bar-Shalom, volume 7, 2001.
[17] M. J. Todd, *Minimum-volume ellipsoids: Theory and algorithms*, Society for Industrial and Applied Mathematics, 2016, https://doi.org/10.1137/1.9781611974386.
[18] J. A. Nelder, R. Mead, *A Simplex Method for Function Minimization*, The Computer Journal 7 (1965) 308–313, https://doi.org/10.1093/comjnl/7.4.308.
[19] J. Behrensdorf, *SlicedNormals.jl*, (no release version) 2022, GitHub repository, https://github.com/FriesischScott/SlicedNormals.jl

[20] J. Ching, Y.-C. Chen, *Transitional Markov chain Monte Carlo method for Bayesian model updating, model class selection, and model averaging*, Journal of Engineering Mechanics 133 (2007) 816–832, https://doi.org/10.1061/(ASCE)0733-9399(2007)133:7(816).
[21] A. Gray, J. Behrensdorf, C. Scherrer, E. Miralles-Dolz, J. Storopoli, *AnderGray/TransitionalMCMC.jl: v0.3.1*, 2021, Zenodo, https://doi.org/10.5281/zenodo.4701274
[22] J. Liang, S. R. Chaudhuri, M. Shinozuka, *Simulation of nonstationary stochastic processes by spectral representation*, Journal of Engineering Mechanics 133 (2007) 616–627, https://doi.org/10.1061/(ASCE)0733-9399(2007)133:6(616).
[23] L. G. Crespo, B. K. Colbert, T. Slager, S. P. Kenny, *Robust Estimation of Sliced-Exponential Distributions*⋆, 60th IEEE Conference on Decision and Control (CDC), 2021, pp. 6742–6748, https://doi.org/10.1109/CDC45484.2021.9683584.