**A Computational Framework for Quantifying Extreme-Event Statistics
in Nonlinear Systems with Stochastic Input**

by

Xianliang Gong

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Naval Architecture and Marine Engineering and Scientific Computing)
in the University of Michigan
2024

Doctoral Committee:

      Assistant Professor Yulin Pan, Chair
      Assistant Professor Xun Huan
      Professor Kevin John Maki
      Professor Joaquim R R A Martins

Xianliang Gong

xlgong@umich.edu

ORCID iD: 0000-0003-2817-2704

ACKNOWLEDGEMENTS

I would like to express my most sincere gratitude to many individuals whose support has been invaluable during my Ph.D. journey:

My advisor, Prof. Yulin Pan, for his continuous guidance, feedback, and support during the past five years. I have learned a tremendous amount of knowledge and life lessons from him and I cannot thank him enough for everything he did for me. I am also grateful to my committee advisors, Prof. Xun Huan, Prof. Kevin John Maki, and Prof. Joaquim R R A Martins. Their expertise has significantly enriched and deepened my contributions to this dissertation.

My colleagues at the FPEL lab, Zhou Zhang, Alexander Hrabski, Guangyao Wang, Cynthia Wu, Jeongbin Park, Ashleigh Simonis, Zih-En Tseng, Shayesteh Hafezi, and former members Yukun Sun, Yuminghao Xiao, and Katerina Siavelis. Thank you a lot for the wonderful time shared together.

My running mates in Ann Arbor Next Level running group, my language mates in Allison's Sunday conversation group, and my friends in the Department of Naval Architecture & Marine Engineering, Joint Institute, and JP Morgan. Your friendship and support have been a source of joy and motivation.

My parents Li Sun and Fuzheng Gong, for their unconditional love, sacrifices, and support since the beginning, and my extended family for their constant encouragement. I also want to thank my girlfriend Yiqiu Sun for being my partner in running, climbing, and coding over the past four years. Your companionship has been my pillar of strength. Finally, I want to express my deepest love for my cat, Gomez Gong.

TABLE OF CONTENTS

CHAPTER

LIST OF FIGURES

FIGURE

LIST OF TABLES

TABLE

LIST OF APPENDICES

ABSTRACT

Extreme events happen in many stochastic natural and engineering systems. Although these events occur with a low probability, they are often associated with catastrophic consequences, making the quantification of their statistics vitally important for a better understanding or design of the system. In this dissertation, we aim to build efficient and accurate methods for the resolution of extreme-event statistics, using a master example of extreme ship motions in random seas.

A direct computation of the extreme ship motion statistics requires running numerical ship simulations to a long wave signal covering all wave conditions. The computational cost of this approach, however, is prohibitively high considering the high complexity of the random wave field, the rareness of the extreme motions, and the expensiveness of the numerical simulation. One critical effort to reduce the computational cost is reducing the complexity of the random sources, i.e., parameterizing the wave field to a low-dimensional feature space. The original problem then becomes a standard uncertainty quantification task to quantify the extreme response statistics given an input-to-response (ItR) function (that needs to be learned) with known input (feature) probability. Many methods have been proposed to address such problems, with one method we are particularly interested in—surrogate modeling trained with active learning. In detail, one can train a surrogate to approximate the ItR function. The training samples are sequentially selected by optimizing an acquisition function based on the existing samples, which facilitates the convergence of the extreme-event statistics.

In this dissertation, we design a set of efficient computational methods to resolve

extreme-event statistics measured in various forms. Regarding the problem of ship motion in random waves, we first introduce a basic framework following existing methods in wave group parameterization and sequential sampling that targets the tail part of the ship motion PDF. In addition to some algorithmic improvements on these two components, we also enrich the framework by considering complete system dynamics through nonlinear wave simulation and ship-wave interaction CFD simulation. In this basic framework, the ship response statistics are defined in terms of the maximum motion in each wave group, which is easy to implement but not straightforward to interpret in practice. We next adapt the framework to quantify a more robust measure, the temporal exceeding probability as the fraction of time that responses exceed a given threshold, with novel developments introduced for both the surrogate model and acquisition function in sequential sampling. While group parameterization significantly speeds up the computation in the above two works, the uncertainties introduced by reduced complexities have not been quantified. To incorporate the lost information, we consider systems characterized by a stochastic ItR with heteroscedastic randomness. These systems are common in physics and engineering with ItR randomness either intrinsic or additional due to dimension reduction. In addressing this problem, we leverage the variational heteroscedastic Gaussian process regression as the surrogate, along with a new acquisition function considering randomness from both input and ItR.

In addition, we develop a multi-fidelity method to further reduce the computational cost of the framework. The key idea here is to leverage low-fidelity models, e.g., a low-resolution ship response CFD whose cost is only a certain fraction of its high-resolution counterpart. In particular, we employ the multi-fidelity Gaussian process regression as a surrogate model and design a new acquisition function that allows the selection of the next-best sample in terms of both the location and fidelity level. An inexpensive analytical evaluation of the acquisition with its derivative is also devel-

oped, enabling gradient-based sampling selection in relatively high-dimensional space. We further adapt the multi-fidelity framework to quantify exceeding/failure probability over a threshold in the context of reliability analysis of connected and autonomous vehicles (CAV). Our acquisition is formulated through information-theoretic consideration which is not only desired to reduce the cost of CAV evaluation but also valuable to the general field of reliability analysis.

We next improve a likelihood-weighted acquisition (algorithm) that was initially designed for rare-event statistics and later extended to many other applications. The improvement comes from the generalized form with two additional parameters addressing two weaknesses of the original likelihood weight: (1) that the input space associated with rare-event responses is not sufficiently stressed; (2) that the surrogate model trained from data may have a significant deviation from the true ItR function. A critical procedure in Monte-Carlo discrete optimization of the acquisition function is also developed that achieves orders of magnitude acceleration compared to existing approaches. In the final part of this dissertation, we present an ongoing work on batch sampling and conclude with a discussion of limitations and future research directions.

CHAPTER 1

Introduction

1.1 Background

Extreme events are generally abnormal system responses that can occur in many stochastic natural and engineering systems. Typical examples include tsunamis, extreme precipitations, ship capsizing, and pandemic spikes. Although these events occur with a low probability, they can potentially result in catastrophic consequences to the environment, industry, and society [33, 40, 111, 109, 42]. Therefore, the accurate quantification of extreme-event statistics is crucial for the assessment and improvement of the system reliability.

Resolving extreme-event statistics, however, is a non-trivial task. For instance, consider a ship at sea with motions excited by stochastic waves. With the increase in computational power, it is now possible to simulate an individual extreme event with high-fidelity simulations, e.g., computational fluid dynamics (CFD). A direct computation of the extreme motion statistics, requiring running CFD to cover ensembles of all wave field realizations or an infinitely long wave signal, is still computationally prohibitive due to the high complexity of the random wave field, the rareness of the extreme motions, and the expensiveness of the numerical simulation. One critical effort to reduce the computational cost is to reduce the dimension of the random source. In ocean engineering, many methods have been proposed to convert the
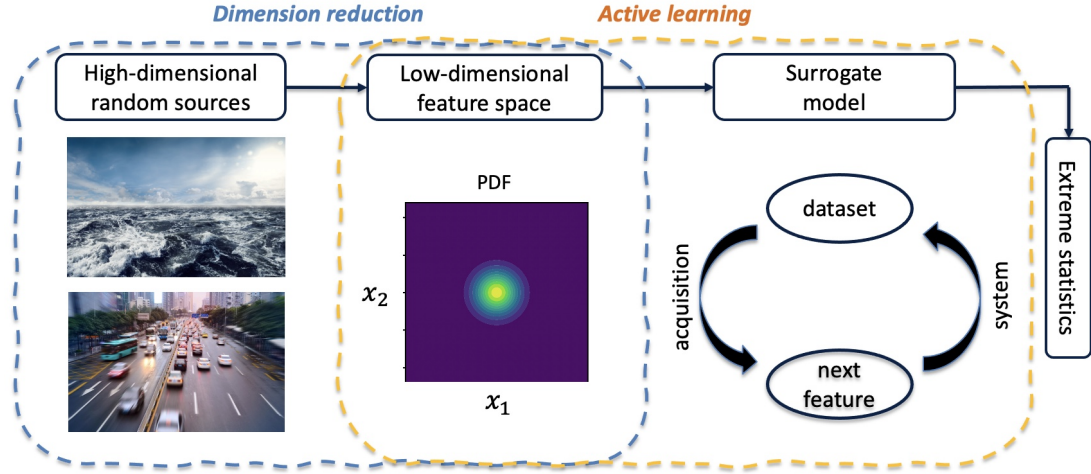
Figure 1.1: Schematic of the computational framework. Random sources with high dimensionality/complexity, e.g., irregular wave fields in the ship example and traffic conditions in the CAV example, are first parameterized to low-dimensional feature (input) space. A surrogate of the input-to-response (ItR) function is then trained by (sequential) active learning. Finally, we compute the extreme statistics from the cheap surrogate.

high-dimensional wave fields into a low-dimensional parameter space of wave groups consisting of a small number of consecutive single waves. For example, wave groups can be parameterized by the peak amplitude, the number of waves overrunning a threshold, and the average period, with their occurrence probability evaluated via a Markov-chain property of the wave process [77, 70, 3, 128]. For a narrow-band wave field, the wave groups can be approximated by Gaussian forms and more conveniently parameterized by the height and length of the group [26, 94]. Similarly, for reliability analysis of connected and autonomous vehicles (CAV), converged statistics of CAV accident rate require hundreds of millions of miles for each configuration of CAV [66]. To reduce the testing cost, scenario-based approaches have been developed [114, 97, 141, 37], where the scenarios (and their distribution) describing certain traffic environments are parameterized from the Naturalistic Driving Data.

The responses of the ship/CAV can then be evaluated for a given group/scenario as input, and their statistics are approximated by collecting responses considering the

probability distribution of the low-dimensional input. In this sense, the original problem becomes a standard uncertainty quantification task: given an input-to-response (ItR) function (that needs to be learned) with known input probability characterizing the system of interests, our objective is to quantify the extreme response statistics (measured in various forms). The ItR is usually expensive to evaluate through numerical simulations or physical experiments, restricting the number of samples that can be placed for system evaluations even in the low-dimensional space. In order to reduce the required number of samples, methods using importance sampling [161], control variate [156], first/second-order reliability methods [55], and large deviation theory [140, 139, 27] have been developed and extensively studied. However, these methods usually target only a single metric, lacking a general view of the extreme-event statistics. In addition, they often make strong assumptions and deal with cases where extreme events of interest occur in an isolated region of the input space, which is not necessarily true for a complex nonlinear response function.

Another type of approach relies on a surrogate model of the ItR function that, in principle, overcomes the issues mentioned above. The learning of the surrogate can be achieved by a standard machine learning algorithm but needs to be conducted with limited data (i.e., samples with system evaluations). A typical method involved here is active learning (or Bayesian experimental design for probabilistic surrogates) which sequentially selects the next sample most informative to extreme response statistics through optimizing an acquisition function. Numerous methods have been proposed with varying surrogates, acquisitions, measures of extreme-event statistics, and applications. Among them, one popular area is structural reliability analysis aiming for the failure probability. However, many existing acquisitions are rather empirical [31, 15, 149], leaving much room for a rigorous development. In ocean engineering, the first application (as far as we know) is [95] to estimate the whole probability density function (PDF) of the marine structure response in random waves

with emphasis on the tail part.

1.2 Roadmap

In this dissertation, we design an efficient computational framework (schematized in figure 1.1) including novel developments in both dimension reduction and active learning, with a master example of quantifying extreme ship responses in random seas. In the following, we will walk readers through each chapter.

In §2, we introduce the basic computational framework in the context of quantifying ship response PDF focusing on the tail part, including wave group parameterization following [26], sequential sampling following [95] with Gaussian process regression as the surrogate. In addition to some algorithmic improvements on these two components, we also consider complete system dynamics through nonlinear wave simulation and ship-wave interaction CFD. We test the validity of the framework for the cases of ship response calculated by a nonlinear roll equation and show the importance of wave nonlinearity to extreme response statistics. Finally, the framework is coupled with the CFD model to demonstrate its applicability to more realistic and general ship motion problems. The subsequent chapters can be seen as improvements of §2 in terms of specific problems raised in ocean engineering or more broadly the active learning algorithms.

The ship response statistics in both §2 and [95] are defined in terms of the maximum motion in each (isolated) wave group, which is not straightforward to interpret in practice (e.g., its value depends on the subjective definition of the wave group). In §3, we further adapt the computational framework to quantify a more robust measure, the temporal exceeding probability as the fraction of time that responses exceed a given threshold. For this goal, novel developments need to be introduced for both the surrogate model and acquisition function. We validate our framework with a

nonlinear roll equation, in terms of the agreement of our results to the true solution and the independence of our results to the criterion of defining groups. Finally, we demonstrate the coupling of the framework with CFD simulations.

While group parameterization in the above two chapters significantly speeds up the computation, the uncertainties introduced in complexities reduction have not been quantified. These uncertainties result in stochastic ItR, often with heteroscedastic randomness for different inputs, in contrast to deterministic ItR in §2 and §3. To incorporate the lost information, in §4, we consider systems characterized by two random sources: (1) parameterized input with a known probability distribution as before, and (2) stochastic ItR function with heteroscedastic randomness. The problem setup often arises in physics and engineering problems, with randomness in ItR coming from either intrinsic uncertainties or additional uncertainties as a result of dimension reduction to the original high-dimensional space, e.g., wave group parameterization. Our method leverages the variational heteroscedastic Gaussian process regression to account for the stochastic ItR, along with a new acquisition function considering randomness from both input and ItR. The validity of the method is first tested in two synthetic problems with the stochastic ItR functions defined artificially. Finally, we demonstrate the application to estimate the extreme ship motion probability, where the uncertainty in ItR naturally originates from wave group parameterization reducing the original high-dimensional wave field into a low-dimensional group space.

In §5, we develop a multi-fidelity sequential sampling method to further reduce the computational cost of the framework in §2. The key idea here is to leverage low-fidelity samples whose responses can be computed with a cost of a certain fraction of that for high-fidelity samples, in an optimized configuration. In particular, we employ a multi-fidelity Gaussian process regression as the surrogate and develop a multi-fidelity acquisition to select the next-best sample in terms of both its location and the fidelity level. In addition, we develop an inexpensive analytical evaluation of

the acquisition and its derivative, avoiding numerical integrations that are prohibitive for high-dimensional problems. The new method is mainly tested in a bi-fidelity context for a series of synthetic problems with varying dimensions, low-fidelity model accuracy, and computational costs. Compared with the single-fidelity method and the bi-fidelity method with a pre-defined fidelity hierarchy, our method consistently shows the best (or among the best) performance for all the test cases. Finally, we demonstrate the superiority of our method in quantifying extreme ship responses, using CFD configured in §2 with two different grid resolutions as the high and low-fidelity models.

In §6, we adapt the multi-fidelity method in §5 to quantify exceeding/failure probability over a threshold in the context of reliability analysis of connected and autonomous vehicles. Our acquisition is formulated through information-theoretic consideration which is not only desired to reduce the cost of CAV evaluation but also valuable to the general field of reliability analysis as an improvement to many empirical acquisitions. The developed method is tested in a widely considered two-dimensional cut-in problem for CAVs, where Intelligent Driving Model with different time resolutions are used to construct the high and low-fidelity models. We show that our single-fidelity method outperforms the existing approach for the same problem, and the bi-fidelity method can further save half of the computational cost to reach a similar accuracy in estimating the accident rate.

In §7, we generalize the likelihood-weighted (LW) acquisition (as the basis to build multi-fidelity acquisition in §5), a popular one in recent years that was initially designed for rare-event statistics and later extended to many other applications. The improvement in our acquisition comes from the generalized form with two additional parameters, by varying which one can target and address two weaknesses of the original LW acquisition: (1) that the input space associated with rare-event responses is not sufficiently stressed; (2) that the surrogate model trained from data may have a

significant deviation from the true ItR function. In addition, we develop a critical procedure in Monte-Carlo discrete optimization of the acquisition function, which achieves orders of magnitude acceleration compared to existing approaches for such type of problems. The superior performance of our new acquisition to the original LW acquisition is demonstrated in a number of test cases, including some cases that were designed to show the effectiveness of the original LW acquisition.

Finally, we present an ongoing work about batch sampling in §8. In batch sampling, a certain number of samples are collectively selected and parallelly evaluated to reduce the wall computational time and, at the same time, improve (or keep) the sampling efficiency. We conclude this dissertation in §9. The limitations and future work will also be discussed in §9 in terms of (i) extension of the framework to broadband wave fields, (ii) choices of other surrogates, (iii) pre-trained sampling policy via reinforcement learning, and (iv) inclusion of derivative observations.

CHAPTER 2

Basic Framework for Group-Based Extreme-Event Statistics of Ship Motions

2.1 Introduction

In this chapter, we present a computational framework for quantifying extreme ship motions in stochastic oceanic waves. While extreme waves are generally recognized as an important factor, extreme motions can also be induced in moderate wave conditions through nonlinear wave-body interaction mechanisms. These wave-body interactions may lead to parametric roll resonance, surf-riding, broaching, and other dynamical phenomena. A complete dynamical model for the extreme ship motion needs to account for the nonlinearity in both wave field and wave-ship interactions.

As we discussed in §1, with the increase of computational power, it is now possible to simulate an individual extreme event with high fidelity, while the resolution of the extreme motion statistics is computationally intractable. To obtain a practical solution of extreme ship motion statistics, earlier method [129] relies on linear extrapolation between $\ln(T)$ and $1/H_s^2$, with T the failure return period and H_s the significant wave height. However, this is later shown to result in under-estimation of T, i.e., conservative results [125]. More recent approaches mainly rely on the principle of separation [10], which splits the computation into rare (R) and non-rare (NR) sub-problems. Within the category of R-NR approaches, the (Envelope) Peaks-Over-

Threshold (POT or EPOT) method [90, 22, 150] identifies the upcrossing rate over a moderate threshold S_m in the NR problem, and extrapolates the distribution to a larger threshold based on the asymptotic extreme value theory in the R problem; the split-time method [9, 8] identifies the upcrossing rate and distribution of derivative process (say roll rate) at S_m in the NR problem, and computes the probability of failure conditioning on the derivative process at S_m in the R problem; The critical wave group approach [137, 3, 1, 2] uses the NR problem to evaluate the distribution of ship initial conditions encountering a wave group, and relates the extreme statistics to the probability of critical wave groups based on the Markov process of wave crest in the R problem. All three methods achieve certain levels of success. In particular, [8] addresses the validity of EPOT and split-time methods; Positive results are reported for the critical wave group approach [81, 126], with its experimental implementation discussed in [5, 6, 7, 3].

In spite of the success of previous methods, an efficient method that incorporates the nonlinearity from both wave field and wave-body interactions is still lacking. In this chapter, following recent methods of wave group parameterization [26] and sequential sampling enabled by Bayesian experimental design [95, 59, 31], we implement a framework which enables an efficient resolution of the statistics of extreme ship responses in irregular nonlinear wave fields. In addition to some algorithmic improvement on the existing methods, our framework allows the wave nonlinearity and wave-ship interaction captured respectively through simulations by high-order spectral method [29, 151] and CFD. We restrict the application of the framework to two-dimensional narrow-band wave fields and define the ship motion statistics in terms of the maximum motion response in each (isolated) wave group (hereafter group-maximum motion statistics). The developed framework is benchmarked for its effectiveness in accurately resolving the group-maximum statistics in a problem where a roll equation is used to compute the group-maximum motion response. The effects
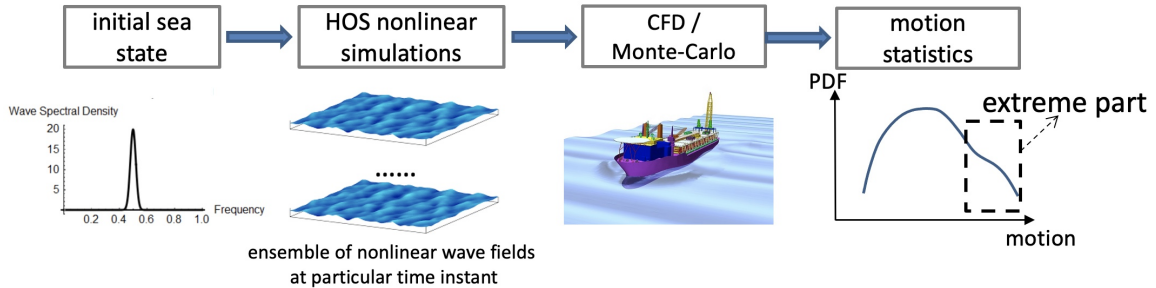
Figure 2.1: Outline of the brute-force way to compute the extreme PDF of ship motion

of wave nonlinearity on extreme response statistics are illustrated in an example of an evolving nonlinear wave field. We finally demonstrate the coupling of the framework with CFD simulations to study the realistic roll-motion statistics of a two-dimensional square-shaped hull.

This chapter is adapted from [50]. The Python code for the proposed algorithm, named wavefinder, is available on Github[1].

2.2 Computational framework

Given a narrow-band wave spectrum (but otherwise arbitrary) and a certain ship geometry, the purpose of our computation is to resolve the probability distribution function (PDF) of the group-maximum ship motion with high precision on its tail part. An outline of the full (brute-force) computation is illustrated in figure 2.1. In this process, the procedure of computing the ensemble of nonlinear wave fields from a given spectrum ((a)-(b)) can be accomplished by an ensemble run of the high-order spectral (HOS) method [29, 151] starting from different initial random phase distributions. This is in essence a full Monte-Carlo method, which is computationally viable due to the highly efficient spectral algorithm in HOS. The difficulty in the computation of figure 2.1 lies in the process from (b) to (d), where CFD simulations are used to

---

[1]https://github.com/umbrellagong/wavefinder

10

sample the ensemble of nonlinear wave fields to resolve the motion PDF. Due to the high dimensionality of the wave fields and the rarity of the extreme motion response, a large number of CFD simulations are required to obtain converged statistics for the tail of the motion PDF. This can become computationally prohibitive for complex problems where each CFD simulation is already expensive.

To enable the computation in figure 2.1, efficient algorithms have to be developed to realize the process from (b) to (d). We next describe two key methods for this purpose. In particular, we use a wave group parameterization technique to reduce the dimensionality of a narrow-band nonlinear wave field, and a sequential sampling enabled by Bayesian experimental design to reduce the number of samples (in the low-dimensional space) to resolve the extreme motion statistics.

2.2.1 Wave group parameterization

Wave groups are structures with successive large waves embedded in random wave fields. To parameterize the wave groups, we first convert the wave elevation field $\eta(x)$ into an envelope process through the Hilbert transform [127] (see figure 2.2(a) for a typical case):

$$\eta(x) = Re\{\rho(x)e^{ik_0x+i\phi(x)}\}, \tag{2.1}$$

where $\rho(x)$ is the envelope process, k_0 the carrier wavenumber, and $\phi(x)$ the phase modulation. Here, we restrict our method only to sufficiently narrow-band wave field where the low-dimensional structure in figure 2.2 (in particular the long group in $\rho(x)$ and nearly constant $\phi(x)$ in each group) can be clearly identified. Under this situation, each wave group of the envelope can be approximated by a localized Gaussian function:

$$\rho_c(x) \sim A \exp \frac{-(x - x_c)^2}{2L^2}, \tag{2.2}$$
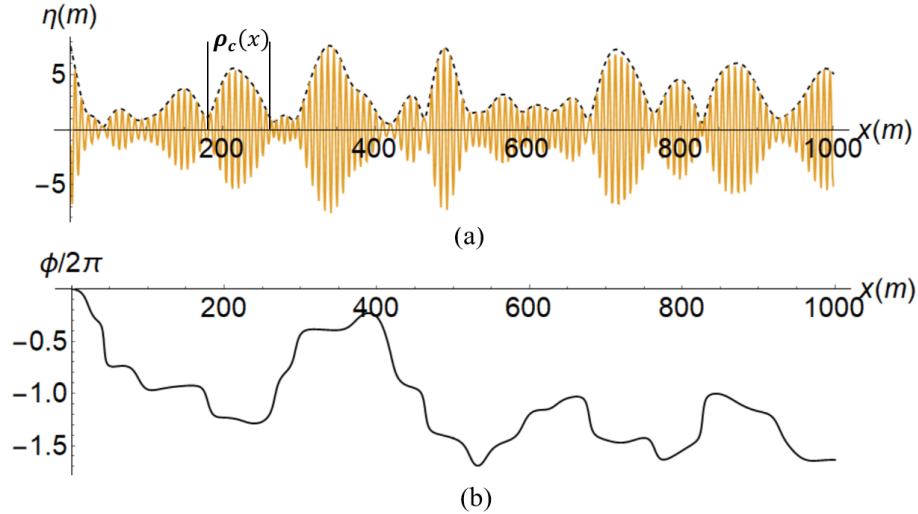
Figure 2.2: (a) envelope process $\rho(x)$ (- - -) and (b) phase modulation $\phi(x)$ computed from the surface elevation $\eta(x)$ (—— in (a)) using the Hilbert transform. A localized wave group $\rho_c(x)$ is indicated in (a).

where A is the group amplitude, L the length scale and x_c the location of the group, $\rho_c(x)$ the segment of the envelope $\rho(x)$ in the vicinity of x_c, corresponding to a wave group.

Our purpose here is to detect L and A for each wave group in the envelope $\rho(x)$, so that a low-dimensional description of the (ensemble) wave field can be established in terms of the joint PDF of L and A (x_c is not important since the ship-motion response does not depend on x_c). Our method builds on [26, 25], originated from a scale-selection method in the computer vision field for feature detection in images [75]. Specifically, the scale-space representation function $S(x, l)$ of $\rho(x)$ is defined by

$$S(x, l) = G(x, l) * \rho(x), \tag{2.3}$$

where $G(x, l) = (1/\sqrt{2\pi}l) \exp^{-x^2/(2l^2)}$ is a Gaussian kernel with scale l at location x, and $*$ denotes convolution. The detection algorithm finds the local minimum of the

normalized second-order derivative of S and set

$$(x_c, L_0) = \text{argmin}_{l,x}\ l^2\frac{\partial^2}{\partial x^2}S(x, l), \tag{2.4}$$

$$L = \frac{1}{\sqrt{2}}L_0. \tag{2.5}$$

The second equation of (2.5) is obtained from theoretical consideration when $\rho(x)$ is in an exact Gaussian shape [75, 26]. Given x_c, A for the same wave group is determined correspondingly as $A = \rho(x_c)$. In [26], it is demonstrated that the direct detection by (2.5) results in a number of fake groups with low similarity with the $\rho(x)$. To remove these wave groups, a discrepancy index is defined as

$$D(L, A, x_c) = \frac{\|\rho_c(x) - Ae^{-\frac{(x-x_c)^2}{2L^2}}\|_2}{\|Ae^{-\frac{(x-x_c)^2}{2L^2}}\|_2}, \tag{2.6}$$

The groups with a discrepancy index $D < 0.25$ are kept.

While the detection algorithm is simple to implement, we find that the scale L is in general under-predicted. Figure 2.3(a) shows a typical result where this point is elucidated. This is due to the finite length of the actual group $\rho_c(x)$, in contrast to an ideal Gaussian group with infinite length. As a result, the adjacent signal $\rho(x)$ around $\rho_c(x)$ affects the scale resolution. To address this issue, we consider the value of L calculated by (2.5) as an initial guess, and conduct another optimization to directly minimize the discrepancy index (2.6), and set

$$L = \text{argmin}_L\ D(L, A, x_c). \tag{2.7}$$

We use Newton's iteration method to solve (2.7). The result after applying (2.7) is shown in figure 2.3(b), where clear improvement (in terms of the closeness between the detected groups and original signal) can be visualized compared to figure 2.3(a). For this wave field, the average value of D decreases from 0.22 to 0.09 after applying

13
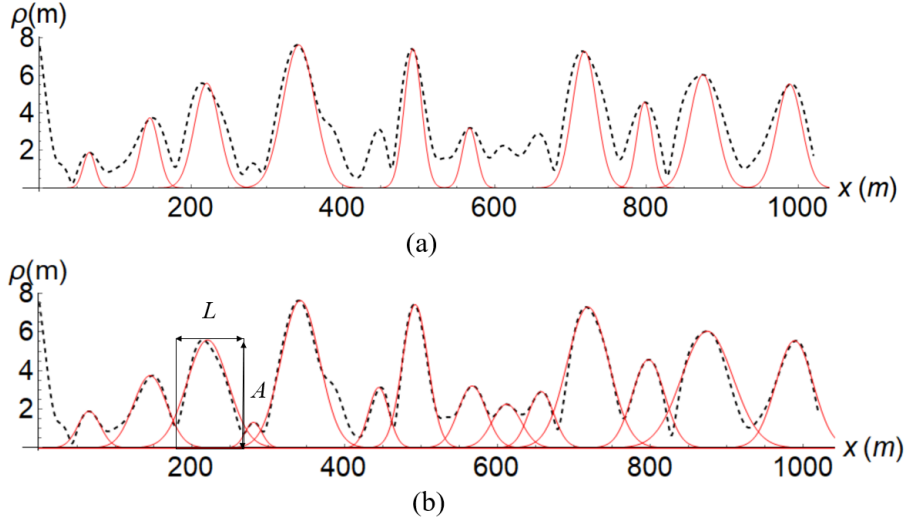
Figure 2.3: The localized Gaussian wave groups (—) calculated using (a) existing method with (2.5) and (b) improved method with (2.5) and (2.7), for an envelope process (---).

(2.7).

The group detection algorithm can be applied to the ensemble of nonlinear wave fields to resolve a collection of pairs L and A, which are then used to calculate $p_{LA}(L, A)$, the joint PDF of L and A (see figure 2.4 for an example). This joint PDF provides a low-dimensional probabilistic description of the nonlinear wave fields, and can be sampled as input to the CFD simulations.

2.2.2 Sequential sampling

Given $p_{\mathbf{x}}(\mathbf{x} \equiv A, L)$, the PDF of a ship motion can be computed from sampling the space of $\mathbf{x}$. Mathematically, we consider a map $r(\mathbf{x}) : \mathbb{R}^2 \to \mathbb{R}$ which maps a wave group parameter $\mathbf{x}$ to the group-maximum ship motion. Our objective is to resolve the PDF of the ship motion $p_r(r)$, with high precision on its tail part. In practice, the computation of map r requires expensive CFD computations. Hence only a limited number of computations can be conducted. Even though the dimension of sample space is low, the required number of CFD simulations is still too large due to the low

14
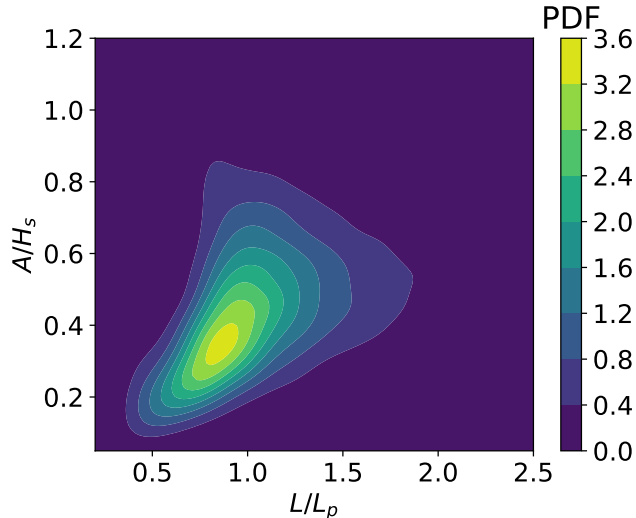
Figure 2.4: An example of the joint PDF $p_{LA}(L, A)$ of L and A, normalized respectively by spectral peak wavelength L_p and significant wave height H_s.

probability of extreme motions.

We next describe a sequential sampling method which significantly reduces the number of samples (thus computational cost). Compared with the full Monte Carlo (or importance sampling) which generates all samples based only on $p_{\mathbf{x}}(\mathbf{x})$, our sequential sampling generates samples making use of the previous samples and the corresponding values of r from simulations to stress the tail part. The implementation requires (i) a surrogate model to approximate the map r, and (ii) solution of an optimization problem to obtain the next-best sample which provides the fastest convergence rate for the tail part of $p_r(r)$. The two components are next described in details.

For the surrogate model, we use the Gaussian process regression (GPR, a.k.a. Kriging), which is a well-developed method in machine learning [113] and geostatistics [65]. Given a number of available parameter-to-response pairs $\mathcal{D} = \{\mathbf{x}^i, r(\mathbf{x}^i)\}_{i=1}^n$, the GPR predicts the function $r(\mathbf{x})$ as a posterior Gaussian process: $r(\mathbf{x})|\mathcal{D} \sim \mathcal{GP}\big(\mathbb{E}(r(\mathbf{x})|\mathcal{D}), \mathrm{cov}(r(\mathbf{x}), r(\mathbf{x}')|\mathcal{D})\big)$ with detailed formulae of mean and covariance
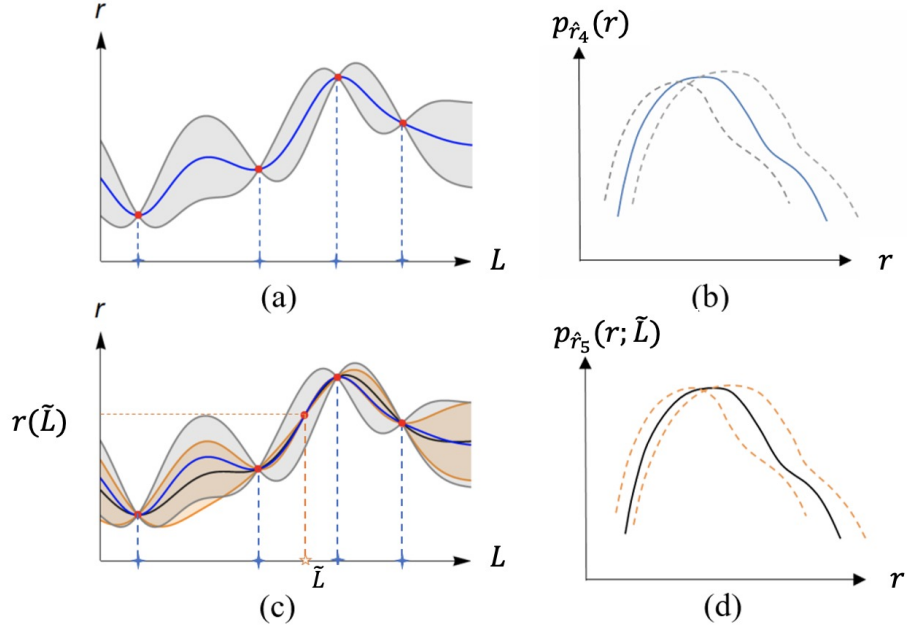
Figure 2.5: A schematic plot for the Gaussian process regression (GPR) and the resulted response PDF, by assuming only the group length L as parameter and $n = 4$. (a) predictive mean $\hat{r}_4(L)$ (——) and uncertainty bounds $\hat{r}_4(L) \pm \sigma_4(L)$ (——) obtained from four parameter-to-response pairs. (b) $p_{\hat{r}_4}(r)$ (——), $p_{r_4^+}(r)$ and $p_{r_4^-}(r)$ (- - -) calculated from $\hat{r}_4$, $\hat{r}_4 + \sigma_4$ and $\hat{r}_4 - \sigma_4$. (c) predictive mean $\hat{r}_5(L; \tilde{L})$ (——) and uncertainty bounds $\hat{r}_5(L; \tilde{L}) \pm \sigma_5(L; \tilde{L})$ (——) obtained from existing four parameter-to-response pairs and $(\tilde{L}, \hat{r}_4(\tilde{L}))$. (d) $p_{\hat{r}_5}(r)$ (——), $p_{r_5^+}(r)$ and $p_{r_5^-}(r)$ (- - -) calculated from $\hat{r}_5$, $\hat{r}_5 + \sigma_5$ and $\hat{r}_5 - \sigma_5$.

summarized in Appendix A. For convenience, we hereafter denote the n-pair predictive mean and standard deviation as $\hat{r}_n(\mathbf{x})$ and $\sigma_n(\mathbf{x})$.

To visualize the concept, we show a schematic plot in figure 2.5(a), where we use $n = 4$ and group length L as the parameter for simplicity. Taking the predictive mean $\hat{r}_n$ as a (cheap) surrogate model, we are able to calculate the PDF of the response $p_{\hat{r}_n}(r)$ using a large number of samples. In addition, the upper and lower uncertainty bounds of the PDF, $p_{r_n^+}(r)$ and $p_{r_n^-}(r)$, can be calculated from, say, $\hat{r}_n + \sigma_n$ and $\hat{r}_n - \sigma_n$ (figure 2.5(b)). Suppose again we have a dataset $\mathcal{D} = \{\mathbf{x}^i, r(\mathbf{x}^i)\}_{i=1}^n$ where $r(\mathbf{x}^i)$ is computed by CFD simulations. We aim to find the next-best sample which leads to the fastest convergence of the tail part of $p_r(r)$. Given an arbitrary next sample $\tilde{\mathbf{x}}$, the surrogate model $\hat{r}_n$ can be used to predict its response $\hat{r}_n(\tilde{\mathbf{x}})$. This provides

us with $n+1$ parameter-to-response pairs $\mathcal{D} \cup (\tilde{\mathbf{x}}, \hat{r}_n(\tilde{\mathbf{x}}))$, from which a new GRP can be performed to develop an updated surrogate model $\hat{r}_{n+1}(\mathbf{x}; \tilde{\mathbf{x}})$, as well as the resulted PDF $p_{\hat{r}_{n+1}}(r; \tilde{\mathbf{x}})$. Due to the additional information provided by $(\tilde{\mathbf{x}}, \hat{r}_n(\tilde{\mathbf{x}}))$, the uncertainties of both $\hat{r}_{n+1}(\mathbf{x}; \tilde{\mathbf{x}})$ and $p_{\hat{r}_{n+1}}(r; \tilde{\mathbf{x}})$ are reduced, compared with $\hat{r}_n(\mathbf{x})$ and $p_{\hat{r}_n}(r)$. The schematic plot of this procedure is shown in figure 2.5 (c) and (d).

We note that both $\hat{r}_{n+1}(\mathbf{x}; \tilde{\mathbf{x}})$ and $p_{\hat{r}_{n+1}}(r; \tilde{\mathbf{x}})$ depend on the $(n+1)_{\text{th}}$ sample $\tilde{\mathbf{x}}$. To find the optimal $\tilde{\mathbf{x}}$ for the resolution of the tail of $p_r(r)$, we construct an optimization problem with an acquisition function

$$U(\tilde{\mathbf{x}}) = \int |p_{r_{n+1}^+}(r; \tilde{\mathbf{x}}) - p_{r_{n+1}^-}(r; \tilde{\mathbf{x}})| r^m \mathrm{d}r, \qquad (2.8)$$

where $p_{r_{n+1}^{\pm}}(r; \tilde{\mathbf{x}})$ are uncertainty bounds of $p_{\hat{r}_{n+1}}(r; \tilde{\mathbf{x}})$ calculated from $\hat{r}_{n+1} + \hat{\sigma}_{n+1}$ and $\hat{r}_{n+1} - \hat{\sigma}_{n+1}$. r^m is a weighting factor which gives more weight for a larger value of r (i.e., tail of $p_{\hat{r}_{n+1}}(r; \tilde{\mathbf{x}})$) with $m \gg 1$. We use $m = 6$ in our current work.

The acquisition function (2.8) is different from the one used in [95], which stresses the low probability part of $p_{\hat{r}_{n+1}}(r; \tilde{\mathbf{x}})$ by defining U based on the difference of log-PDF without the weighting factor. However, a low probability does not necessarily mean a large response (it may also correspond to an extremely small response). In contrast, the function (2.8) provides a direct measure of the uncertainty of $p_{\hat{r}_{n+1}}(r; \tilde{\mathbf{x}})$ focusing on the part of extreme (large) responses, with the level of 'extreme' tunable by the value of m. We also remark that the sequential sampling based on (2.8) can explore all extremes in a multi-modal response function given sufficient samples.

The next-best sample $\mathbf{x}^*$ is chosen from the sample space by solving the optimization problem

$$\mathbf{x}^* = \operatorname{argmin}_{\tilde{\mathbf{x}}} U(\tilde{\mathbf{x}}). \qquad (2.9)$$

In our current work, (2.9) is numerically solved using a particle swarm method [110]. The sequential sampling process is repeated for the next-best sampling until satisfac-

tory convergence is achieved for the tail part of $p_r(r)$. We summarize the algorithm of the whole process in the Algorithm **1**.

Algorithm 1 Bayesian experimental design for extreme ship response statistics

---
Require: Number of initial samples n_{init} and sequential samples n_{seq}
Input: Initial dataset $\mathcal{D} = \{\mathbf{x}^i, r(\mathbf{x}^i)\}_{i=1}^{n_{init}}$
   Initialization $n = n_{init}$
   while $n < n_{init} + n_{seq}$ **do**
      Train the surrogate model with $\mathcal{D}$
      Solve (2.9) to find the next best sample $\mathbf{x}^*$
      Implement numerical simulation to get $r(\mathbf{x}^*)$
      Update the dataset $\mathcal{D} = \mathcal{D} \cup \{\mathbf{x}^*, r(\mathbf{x}^*)\}$
      $n = n + 1$
   end while
Output: Compute the ship motion PDF based on the surrogate model

---

2.3 Validation of the framework

For validation, it is desirable to compare PDF from our reduced-order approach with the exact motion response PDF $p_r(r)$. This requires a cheap parameter-to-response map r such that $p_r(r)$ can be efficiently calculated. For this purpose, we use a nonlinear roll equation [142] to calculate the time series of ship roll $\xi(t)$ from a given group signal $\eta(t; \mathbf{x})$:

$$\ddot{\xi} + \alpha_1 \dot{\xi} + \alpha_2 \dot{\xi}^3 + (\beta_1 + \epsilon_1 \cos(\chi)\eta(t; \mathbf{x}))\xi + \beta_2 \xi^3 = \epsilon_2 \sin(\chi)\eta(t; \mathbf{x}), \quad (2.10)$$

which phenomenologically models the ship roll motion due to nonlinear resonance and parametric roll in oblique irregular waves. We use empirical coefficients $\alpha_1 = 0.25$, $\alpha_2 = 0.1$, $\beta_1 = 0.04$, $\beta_2 = -0.01$, $\epsilon_1 = 0.006$, $\epsilon_2 = 0.008$ and $\chi = \pi/6$. (2.10) is numerically integrated with a 4th-order Runge-Kutta method from zero initial condition to obtain the group-maximum response $r(\mathbf{x}) = \max(\xi(t; \mathbf{x}))$. The values of parameters are tuned such that the r is not sensitive to the initial conditions for

$\eta(t; \mathbf{x})$ described by (2.1) and (2.2).

To obtain the time series $\eta(t; \mathbf{x})$ for a given $\mathbf{x}$, we first construct the localized Gaussian group from (2.2) and then the spatial surface elevation $\eta(x; \mathbf{x})$ from (2.1). The effect of $\phi(x)$ is neglected in (2.1), since the phase modulation $\phi(x)$ can be considered almost a constant within a wave group (see figure 2.2(b)) for a sufficiently narrow-band wave spectrum. The relatively large variation of $\phi(x)$ occurs near the end of the wave group, which is expected to have a much smaller impact on the extreme motion response than $\mathbf{x} \equiv (L, A)$. Given $\eta(x; \mathbf{x})$, $\eta(t; \mathbf{x})$ is then constructed by considering the propagation of each wave mode by linear dispersion velocity.

To calculate the exact motion PDF $p_r(r)$, we generate $\rho_c(t)$ from $\rho_c(x)$ (figure 2.2(a)) in a similar manner, which is then substituted to (2.10) in replacement of $\eta(t; \mathbf{x})$. This is conducted for all $\rho_c(x)$ in the ensemble of nonlinear wave fields to obtain a collection of $r(\mathbf{x})$, which is then used to calculate $p_r(r)$.

We next describe the implementation of our framework on this simplified ship roll problem. We consider an initial narrow-band wave spectrum in a Gaussian form:

$$F(k) \sim \exp \frac{-(k - k_0)^2}{2\mathcal{K}^2}, \tag{2.11}$$

with significant wave height $H_s = 12m$, peak (carrier) wavenumber $k_0 = 0.018m^{-1}$ (corresponding to peak period $T_p = 15s$), and $\mathcal{K} = 0.05k_0$. Four hundred HOS simulations are run, where each of them has a domain length of 64 times the peak wavelength. The ensemble of nonlinear wave fields are collected at $t = 50T_p$, for which the wave parameterization method discussed before is used to generate the joint PDF of L and A.

The sampling process in the space (L, A) is performed as follows. We first generate six random samples (following $p_{LA}(L, A)$) and calculate their response r from (2.10) as the initial dataset. Then we conduct twelve sequential samplings to compute $p_{\hat{r}_{18}}^{seq}(r)$.
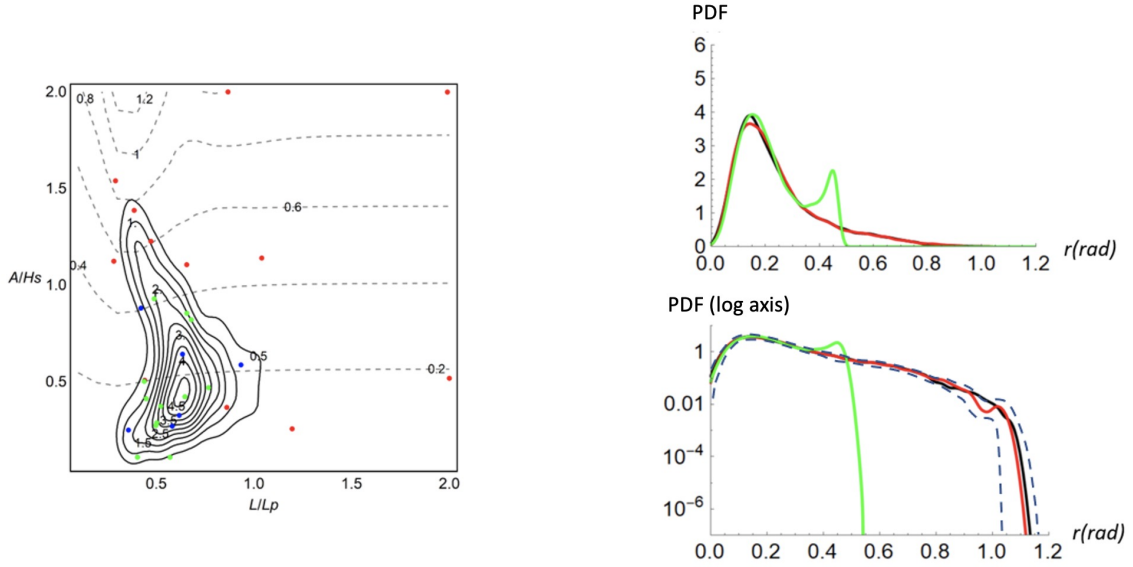
Figure 2.6: Samples and PDFs from sequential and random samplings. Left: Initial 6 random samples (•), subsequently 12 sequential samples (•) and 12 random samples (•) in contour plot of $p_{LA}(L, A)$ (——) and $r(L, A)$ (- - -). Right: The exact roll PDF $p_r(r)$ (——), sequential-sampling PDF $p^{seq}_{\hat{r}_{18}}(r)$ (——) and random-sampling PDF $p^{ran}_{\hat{r}_{18}}(r)$ (——) plotted on both linear and log axes. The 95% confidence interval for $p^{seq}_{\hat{r}_{18}}(r)$ is included (- - -) in the log-axis plot.
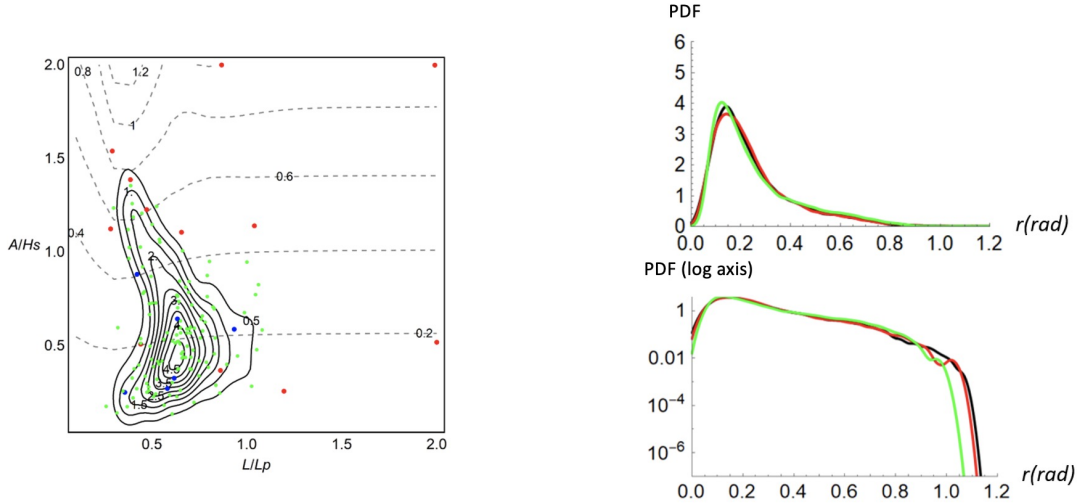
Figure 2.7: Samples and PDFs from sequential and random samplings. Left: initial 6 random samples (•), subsequently 12 sequential samples (•) and 120 random samples (•) in contour plot of $p_{LA}(L, A)$ (——) and $r(L, A)$ (- - -). Right: The exact roll PDF $p_r(r)$ (——), sequential-sampling PDF $p^{seq}_{\hat{r}_{18}}(r)$ (——) and random-sampling PDF $p^{ran}_{\hat{r}_{126}}(r)$ (——) plotted on both linear and log axes.

20

For comparison, a random sampling approach with an equal number of samples is also conducted to compute $p^{ran}_{\hat{r}_{18}}(r)$ (based on the GPR surrogate model constructed from 18 random samples).

The PDFs $p^{seq}_{\hat{r}_{18}}(r)$, $p^{ran}_{\hat{r}_{18}}(r)$ and $p_r(r)$, as well as the 95% confidence interval for $p^{seq}_{\hat{r}_{18}}(r)$, are plotted in figure 2.6, on both linear and log scales. It can be seen that the result from the sequential sampling is much closer to the exact PDF compared to the result from the random sampling. With the left sub-figure showing the locations of the samples, we see that the random samples are concentrated in the high probability region of $p_{LA}(L, A)$, while the sequential samples explore the region with large response (combined with nontrivial probability). In order to obtain comparable roll PDF using random sampling, at least one order of magnitude more samples are needed. Figure 2.7 shows the result $p^{ran}_{\hat{r}_{126}}(r)$ obtained from 120 random samples (after the 6 initial ones), which is still less accurate than $p^{seq}_{\hat{r}_{18}}(r)$ in terms of the tail part of the PDF.

2.4   Effects of wave nonlinearity

Our framework allows the direct resolution of wave nonlinearity effect on the ship response statistics. It has been experimentally and numerically demonstrated that the modulational instability of nonlinear waves result in the non-Gaussian statistics (in terms of a heavy-tail PDF) of the surface elevations [103, 155]. Based on this, we expect the extreme ship response statistics to be enhanced when nonlinearity effect is included for the wave field.

To illustrate the effect of wave nonlinearity, we generate an initial linear wave field from the spectrum (2.11) using independent and random phase distributions, i.e., with surface elevation following Gaussian statistics. An ensemble of four hundred HOS simulations (with domain length of 64 times of the peak wavelength) are used
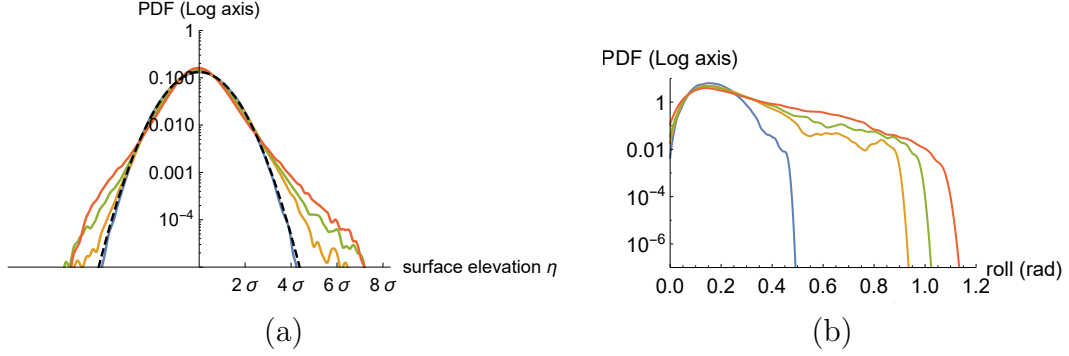
Figure 2.8: PDFs of (a) surface elevation and (b) roll response at $t = 0$ (———), $20T_p$ (———), $30T_p$ (———) and $50T_p$ (———). The Gaussian PDF is plotted in (a) with (- - -).

to compute the evolution of wave statistics with time. Figure 2.8(a) shows the PDFs of surface elevations at $t = 0$, $20T_p$, $30T_p$ and $50T_p$ respectively. It is clear that the PDF at $t = 0$ follows closely a Gaussian tail. With the increase of time, a heavier tail develops due to the nonlinear wave effect, indicating a higher probability of extreme waves in the field. We compute the response PDF using roll equation and sequential sampling with wave group statistics calculated from the wave fields in each instant of figure 2.8(a), with the results shown in figure 2.8(b). We can clearly observe the development of enhanced extreme ship motion statistics with the increase of time. Since the evolution of wave spectrum in $O(10 - 50)T_p$ is relatively mild, it is not unreasonable to associate the enhanced motion statistics to non-Gaussian wave statistics developed due to the wave nonlinearity. While this case qualitatively demonstrates the influence of wave nonlinearity on extreme ship motion statistics, more studies are warrantied in future work to quantify this effect.

2.5 Coupling with CFD

In this section, we use CFD simulations to compute the ship motion response in a given incident wave group described by parameters L and A. For each sample in the wave parameter space, we define the initial condition of CFD simulation using

the profile of a propagating wave group with parameters L and A, and compute the maximum response as the output of the simulation. For simplicity, we consider the motion of a two-dimensional (2D), square-shaped hull geometry with $40m \times 40m$ cross section and density $\rho_h = 0.5\rho_w$ with ρ_w being the water density. The turbulence effects are neglected in the simulations. Our interest is to resolve the extreme roll statistics using the framework described above. In spite of the simplification (in terms of the 2D geometry), this computation is sufficient to demonstrate the effectiveness of the new framework when coupled to CFD, which enables more realistic problems to be studied.

2.5.1 CFD model

The CFD simulations in this work are performed using the open-source code Open-FOAM [63]. We next describe the details of our model.

2.5.1.1 Mathematical formulation

The air-water interface in CFD is modeled by the volume fraction γ ($\gamma = 0$ for air and $\gamma = 1$ for water). To capture the evolution of the γ field, we use the interFoam solver, which is based on an algebraic volume of fluid (AVOF) approach. In AVOF, the flux of volume fraction γ is computed algebraically without a geometric reconstruction of the interface [93]. An interfacial compression term is used to mitigate the effects of numerical smearing of the interface [28]. The governing equations of the problem include the continuity equation, the momentum equation and the volume fraction equation:

$$\nabla \cdot \mathbf{u} = 0, \tag{2.12}$$

$$\frac{\partial(\rho\mathbf{u})}{\partial t} + \nabla \cdot (\rho\mathbf{u}\mathbf{u}) = -\nabla p_d + \nabla \cdot (\mu\nabla\mathbf{u})$$
$$+ \nabla\mathbf{u} \cdot \nabla\mu - \mathbf{g} \cdot \mathbf{x}\nabla\rho, \tag{2.13}$$

$$\frac{\partial \gamma}{\partial t} + \nabla \cdot (\mathbf{u}\gamma) + \nabla \cdot (\mathbf{u}_r \gamma(1-\gamma)) = 0, \tag{2.14}$$

where $\mathbf{u}$ is the velocity field, $p_d = p - \rho\mathbf{g} \cdot \mathbf{x}$ is the modified pressure, with p being the pressure, and $\mathbf{g}$ the gravitational acceleration vector, and $\mathbf{x}$ the position vector. The fluid properties, including the density ρ and the dynamic viscosity μ are calculated as weighted averages based on γ,

$$\rho = \gamma \rho_w + (1-\gamma)\rho_a, \tag{2.15}$$

$$\mu = \gamma \mu_w + (1-\gamma)\mu_a, \tag{2.16}$$

where the subscripts w and a correspond to water and air respectively. In (2.14), the last term (with tuned parameter $\mathbf{u}_r$) represents an artificial compression of the interface to mitigate the numerical diffusion [12].

The motion of the hull is calculated based on the force exerted by flow pressure and shear stress. The hull is considered as a rigid body, whose motion is solved by numerical integration implemented by the Newmark method [99]:

$$\dot{X}_{n+1} = \dot{X}_n + (1-\alpha)\Delta t \ddot{X}_n + \alpha \Delta t \ddot{X}_{n+1}, \tag{2.17}$$

$$X_{n+1} = X_n + \Delta t \dot{X}_n + \frac{1-2\beta}{2}\Delta t^2 \ddot{X}_n + \beta \Delta t^2 \ddot{X}_{n+1}, \tag{2.18}$$

where X represents the linear displacement of the heave and sway motion, and the angular displacement of the roll motion, Δt is the time step, α and β are user-defined coefficients. Typically, $\alpha = 0.5, \beta = 0.25$ are the most common choices and are used in our study.

2.5.1.2 Computational grids

The 2D computational domain is discretized by a Cartesian mesh with refined grids near the free surface (see figure 2.9(a)). The dynamic mesh is used around the hull,
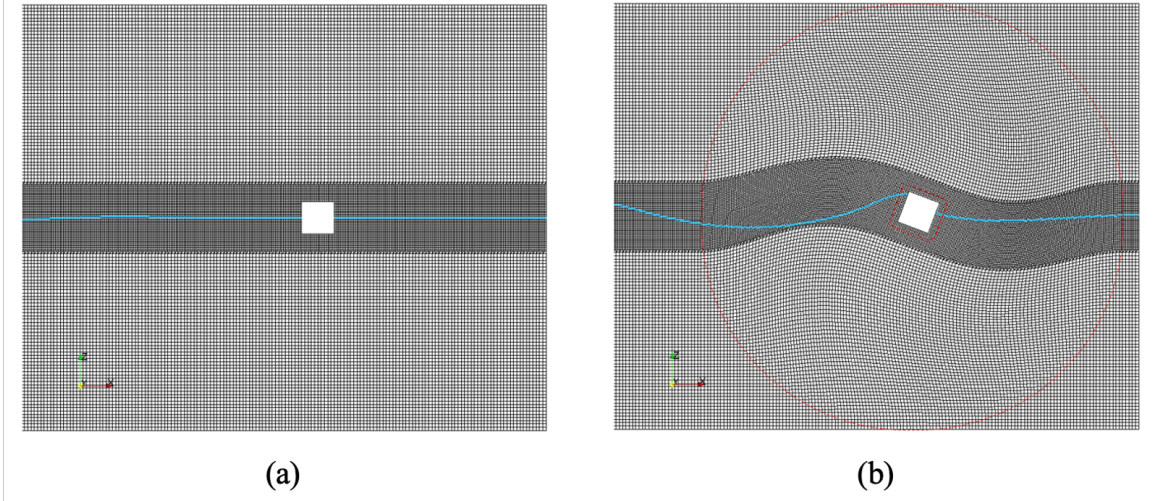
Figure 2.9: (a) The mesh around the hull with a refined region near the free surface; (b) The deformed mesh around the moving hull. The interface is marked by a blue solid line. The boundaries of inner distance and outer distance are marked by red dashed lines.

allowing the mesh to deform when the hull moves. The region of mesh deformation is controlled both inner and outer distance: the grids within the inner distance from the body surface move with the hull as a rigid body; the grids between the inner distance and the outer distance can be morphed; and the grids outside the outer distance do not move (see figure 2.9(b)).

2.5.1.3 Initial condition and boundary conditions

The initial condition of the simulation includes the initial fields $\gamma_0(x, z) \equiv \gamma(x, z, t = 0)$ and $\mathbf{u}_0(x, z) \equiv \mathbf{u}(x, z, t = 0)$, where z is the coordinate in the vertical direction and t is the time. These initial fields are defined separately for the left and right parts of the computation domain (see figure 2.10). For the right part, we consider the situation of a stationary hull floating on still water, with a corresponding volume fraction $\gamma_0(x, z)$, and velocity $\mathbf{u}_0(x, z) = \mathbf{0}$. For the left part, the initial condition represents the profile of a propagating wave group with parameters L and A. In particular, the $\gamma_0(x, z)$ field is specified from the free surface position $\eta_0(x) = \eta(x; A, L)$ (see (2.1)

25
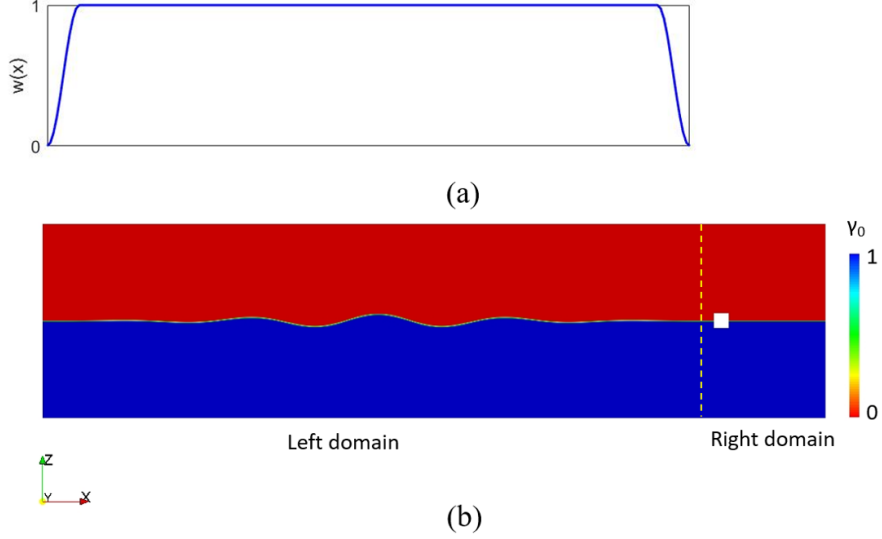
Figure 2.10: (a) the window function $w(x)$; (b) a typical initial field $\gamma_0(x, z)$. The border of the left and right domains is indicated by a dashed line.

and (2.2)) by

$$\gamma_0(x, z) = \begin{cases} 0, & z > \eta_0(x) \\ 1, & z < \eta_0(x) \end{cases}. \tag{2.19}$$

The $\mathbf{u}_0(x, z)$ field is calculated correspondingly assuming a linear dispersion relation for each wave mode of $\eta_0(x)$, namely with the horizontal and vertical velocity components

$$u_0(x, z) = \sum_{j=1}^{N} a_j \omega_j \frac{\cosh[k_j(z + h)]}{\sinh(k_j h)} \cos(k_j x + \psi_j), \tag{2.20}$$

$$v_0(x, z) = \sum_{j=1}^{N} a_j \omega_j \frac{\sinh[k_j(z + h)]}{\sinh(k_j h)} \sin(k_j x + \psi_j), \tag{2.21}$$

where $a_j \cos(k_j x + \psi_j)$ is the mode j of $\eta_0(x)$, with a_j the amplitude, k_j the modal wavenumber, and ψ_j the modal phase. $\omega_j = \sqrt{gk_j \tanh(k_j h)}$ is the angular frequency of mode j, h is the water depth (300m in our case).

To obtain a smooth field as the initial condition, a window function $w(x)$ is multiplied to $\gamma_0(x, z)$ and $\mathbf{u}_0(x, z)$ to remove the discontinuity between the wave field and the still-water field at both ends of the wave group. This is illustrated in figure 2.10

26

along with the resulted initial $\gamma_0(x, z)$ field.

Periodic boundary conditions are used at the left and right ends of the domain, allowing the wave to keep propagating without reflection after interacting with the hull. We have tested that the domain is sufficiently large such that the periodic boundary does not result in spurious waves interacting with the body (in the time period of the simulation). At the bottom of the domain, a slip-wall boundary condition is used. At the top, the pressureInletOutletlVelocity boundary condition is used, which is a modified zero-gradient velocity condition. For the boundaries on the floating hull, the no-slip boundary condition is used, which sets the fluid velocity to the same value as the velocity of the moving rigid body.

2.5.1.4   A typical case

We show the result of a typical case with $A = 8.1m$ and $L = 332.7m$ in figure 2.11. A snapshot of the $\gamma(x, z)$ field is shown in figure 2.11(a) in the process of a wave group interacting with the hull. The time series of roll angle $\xi(t)$ is shown in figure 2.11(b). The group-maximum response is then taken as the maximum of $\xi(t)$, which is used in the sequential sampling.

2.5.2   Results

The wave parameter space $p_{\mathbf{x}}(\mathbf{x})$ is generated using the initial condition (2.11) with data collected at $t = 20T_p$. The sequential sampling is coupled with CFD simulations to compute $p_r(r)$. We again use 6 random samples (as well as $r(\mathbf{x})$ obtained from CFD) as the initial dataset. Then 14 sequential samples are performed to compute $p_{\hat{r}_{20}}(r)$. Since the exact $p_r(r)$ is not available for this problem, we plot $p_{\hat{r}_n}(r)$ for different n in figure 2.12. We see that there is a clear trend of convergence for the tail of $p_{\hat{r}_n}(r)$ as n increases. For the last few samples ($n = 16 \sim 20$), the tail of the PDF oscillates in small regions, indicating the convergence of the extreme statistics.
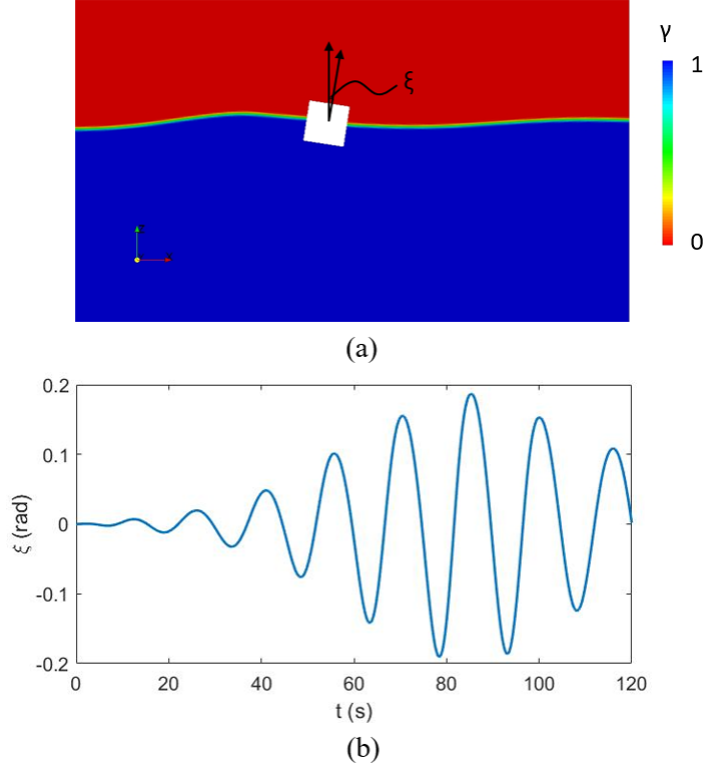
Figure 2.11: The result from a typical case with $A = 8.1m$ and $L = 332.7m$. (a) a snapshot of the $\gamma(x, z)$ field in the process of a wave group interacting with the hull; (b) time series of $\xi(t)$.

We also plot $U(\mathbf{x}^n)$ as a function of the sample number n in figure 2.13(a). We see that at the end of 20 samples, the value of $U(\mathbf{x}^n)$ almost converges to a constant level. Nevertheless, this constant level is higher than U at previous samples, e.g., $U(\mathbf{x}^{20}) > U(\mathbf{x}^{10})$. This is due to the shift of the tail of $p_{\hat{r}_n}(r)$ to the right (or larger values of r), which results in an increase of U through the factor r^m (see (2.8)). An alternative metric to quantify the uncertainty level of the PDF tail can be defined as $U' = U(\mathbf{x}^n)/\max(\hat{r}_n(\mathbf{x}))^7$, where $\max(\hat{r}_n(\mathbf{x}))$ provides the maximum value of response from all parameters $\mathbf{x}$. Compared to U, U' removes the effect of the right-shifting of the PDF tail, and focuses only on the difference between the upper and lower bounds of the tail. The metric $U'(\mathbf{x}^n)$ as a function of n is plotted in figure 2.13(b), showing a satisfactory convergence over all samples.
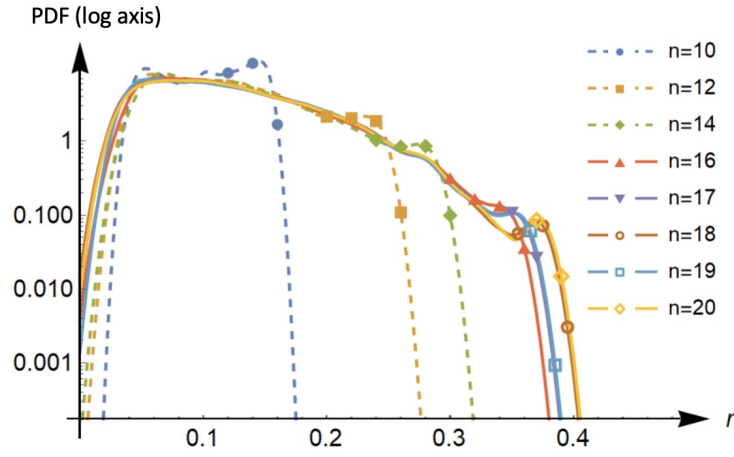
Figure 2.12: The PDFs $p_{\hat{r}_n}(r)$ for different n.
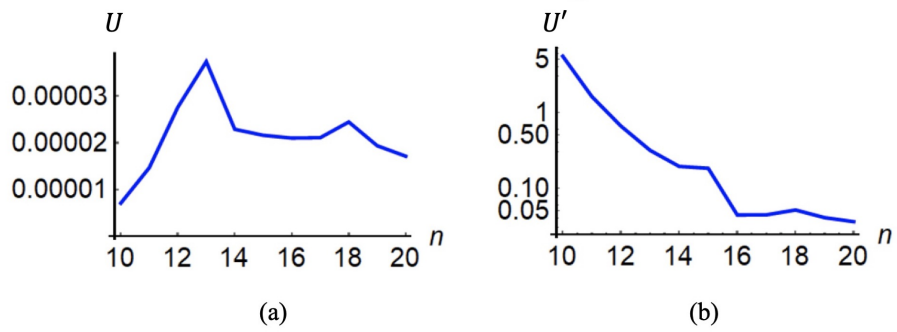
Figure 2.13: (a) $U(\mathbf{x}^n)$ as a function of the sample number n; (b) $U'(\mathbf{x}^n)$ as a function of the sample number n

2.6 Summary

Building on existing methods, we implement a computational framework which allows an efficient resolution of extreme ship motion statistics in narrow-band nonlinear wave fields. Three key components are included in the framework: (1) generation of an ensemble of nonlinear wave fields using the high-order spectral method; (2) wave group parameterization to reduce the high-dimensional wave field to a low-dimensional space of (L, A); and (3) sequential sampling to obtain the motion response PDF $p_r(r)$ with fastest convergence rate of its tail, i.e., extreme motion part. In addition to some improvements to the existing methods, our framework allows the effect of wave nonlinearity to be incorporated in the computation of ship response statistics. The framework is validated through a simplified problem of roll motion predicted by a nonlinear roll equation, where the sequential sampling are shown to be effective in obtaining accurate $p_r(r)$ with significantly reduced computational cost. The capability of the framework to include nonlinearity of the waves have also been demonstrated in an evolving nonlinear wave field. We finally demonstrate the coupling of the framework with CFD to resolve the extreme roll statistics of a two-dimensional, square-shaped hull.

We note that three approximations in our method require further considerations: (1) The Gaussian wave group representation neglects the deviation of realistic wave groups from Gaussian functions; (2) The application of a constant phase modulation ignores its variation within a wave group; (3) The use of zero initial conditions neglects the high variability of initial conditions in a wave field with multiple groups. The impact of factors (1) and (2) can be considered insignificant in the limit of narrow-band wave field, with their effects increasing with the increase of spectral bandwidth. The impact of (3) is negligible only for dynamical systems where the maximum responses are not sensitive to initial conditions (such as the roll equation considered in this chapter). For general cases, the uncertainties associated with (1)-(3) have to

be considered, which will be discussed in §4. We also remark that the computation of the acquisition (2.8) involves resolving the upper and lower bound PDFs, which is relatively expensive. Different methods to reduce this cost will be introduced later in §5 and §7.

CHAPTER 3

Efficient Computation of Temporal Exceeding Probability of Ship Motions

3.1 Introduction

In the previous chapter, we have built a computational framework targeting group-maximum statistics defined in terms of the maximum motion response in each (isolated) wave group. While the group-maximum statistics is indeed a natural metric to compute under the wave group parameterization, this concept is not sufficient under certain circumstances. First, the group-maximum statistics, say, exceedance probability of the group-maximum response, depends on additional parameters to define wave groups, e.g., the overrun threshold in the Markov-chain method or the similarity score in the Gaussian group method. Depending on values of these additional parameters, different numbers of groups can be defined in a time series, which affects the denominator in calculating the group-based probability. In most cases, in order to make physical sense of this probability, one needs to combine it with the encounter rate of groups in a wave field, which requires an additional step in the computation. Second, if both the group-maximum exceedance probability and group encounter rate are available, one can further compute valuable derivatives such as the probability of the number of exceedances in a given exposure time period. However, this probability quantity is still not informative enough in two aspects: (1) the group-maximum

approach ignores the situation when multiple exceedances occur in one wave group, which would be counted as only one exceedance since only the group-maximum response is computed. Therefore, the probability on the number of exceedance is only strictly valid if the exceeding threshold is sufficiently high (so that in each group there is at most one exceedance). (2) Knowing the number of exceedances does not provide information about the "severity" of the exceedance, e.g., how long the exceedance lasts.

In contrast, the temporal exceeding probability (i.e., percentage of the exposure time that ship responses are greater than a given large threshold) provides an important complement to the group-maximum exceeding probability, with the following advantages: (a) the calculation of temporal exceeding probability, as will be discussed in detail, naturally incorporates the information of wave group encounter rate so that no additional step is needed. (b) It safely accounts for multiple exceedances in one group and quantifies the 'severity' of the exceedance by the exceeding time. In computing the temporal exceeding probability, the group parameterization remains as an effective way for dimension reduction of the wave field (although the final result does not depend on the defined groups as we will demonstrate later in the chapter). One may expect to further consider the motion exceeding time as the response function of group parameters and apply the sequential sampling or Bayesian experimental design (BED). However, this operation introduces additional difficulties that need to be resolved by substantial developments in both components involved in sequential sampling — the surrogate model and acquisition function. While a standard Gaussian process regression (GPR) is sufficient as a surrogate model to compute the group-maximum response (as a smoothly varying response function), great challenges, i.e., prediction errors, arise if the group exceeding time is used as a response function. This is because the group exceeding time as a response function is characterized by two drastically different scales of variation in the regions of zero values (for the major-

ity of the inputs) and positive values (for 'dangerous' critical wave groups) separated by the limiting state. In addition, the existing acquisition functions for exceeding probability [136, 31, 15, 149, 59, 74, 135, 163] mainly focus on sampling at the limiting state, which becomes insufficient for temporal exceeding probability for which the wave groups leading to longer exceeding time matter more than those at the limiting state.

In this chapter, we address the above two problems in the sequential BED procedure, enabling an efficient computation of temporal exceeding probability of ship responses in a random wave field. Specifically, we construct a uniformly-varying (i.e., varying with the same or comparable length scale) response function derived from the group exceeding time which eliminates the prediction errors in GPR while not affecting the final solution of exceeding probability. We then formulate a new acquisition function focusing on sampling wave groups associated with significant exceeding time rather than at the limiting state. We validate our developed computational framework in a case of ship response calculated by a nonlinear roll equation, in terms of the agreement of our obtained result with the true solution, and the independence of our result to different criteria to define wave groups. Finally, we demonstrate the coupling of our framework to CFD models to enable the computation for more realistic and general ship response problems.

This chapter is adapted from [49]. The Python code for the proposed algorithm, named gpship, is available on Github[1].

---

[1]https://github.com/umbrellagong/gpship

3.2 Method

3.2.1 Problem setup

We start from a narrow-band wave field with a sufficiently long time series of wave elevation $\eta(t), t \in [0, T_{end}]$. There is no restriction on how $\eta(t)$ should be obtained. It can be extracted from real oceanic data, nonlinear wave simulations [29, 151, 148], or linear wave construction given a spectrum [102] (which we use in the current work). Our objective is to compute the temporal exceeding probability (i.e., percentage of time that the ship response is larger than a threshold r_s) when a ship goes through the wave field described by time series $\eta(t)$, defined as

$$P_{temp} = \frac{\int_0^{T_{end}} \mathbf{1}\big(|\xi(t)| - r_s\big)\mathrm{d}t}{T_{end}}, \tag{3.1}$$

where $\xi(t)$ is the time series of ship response caused by waves $\eta(t)$, r_s the response threshold, and $\mathbf{1}(\cdot)$ is a Heaviside function:

$$\mathbf{1}(x) = \begin{cases} 1, & if\ x > 0 \\ 0, & if\ x \leq 0 \end{cases}. \tag{3.2}$$

The computation of P_{temp}, as defined in (3.1), involves a simulation of the ship response from $\eta(t)$, which needs to be extremely long to cover (many times) all wave conditions associated with a given wave spectrum. If high-fidelity models (e.g., CFD) are used for this computation, the computational cost can become prohibitively high.

In order to reduce the overall computational cost, we can parameterize the time series $\eta(t)$ into wave groups. Specifically, we compute the envelope process $\rho(t)$ from $\eta(t)$ through the Hilbert transform [127] (figure 3.1(a)), and then construct Gaussian-
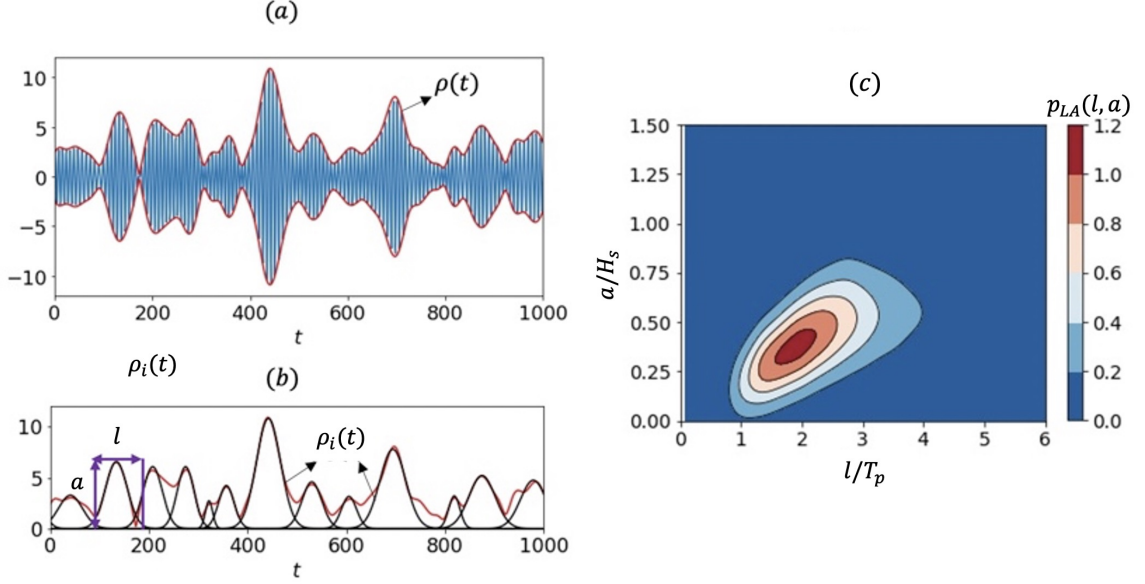
Figure 3.1: (a) surface elevation $\eta(t)$ (——) and the corresponding envelope process $\rho(t)$ (——) in a random wave field. (b) $\rho(t)$ (——) fitted by an ensemble of Gaussian wave groups $\rho_i(t)$(——) with parameters l and a. (c) probability distribution $p_{LA}(l, a)$ obtained from the whole wave field.

like wave groups $\rho_i(t)$ which best fit $\rho(t)$ locally:

$$\rho_i(t) = a_i \exp \frac{-(t - t_i)^2}{2l_i^2}, \tag{3.3}$$

where t_i, a_i, and l_i are respectively the temporal location, amplitude and length of group i that are selected to fit $\rho(t)$ (see figure 3.1(b)). This construction relies on the group detection algorithm detailed in §2.2.1. The final identification of a wave group is based on the tolerance of a discrepancy index

$$D_i(l_i, a_i, t_i) = \frac{\int_{t_i-2l_i}^{t_i+2l_i} \left(\rho(t) - \rho_i(t)\right)^2 \mathrm{d}t}{\int_{t_i-2l_i}^{t_i+2l_i} \rho_i(t)^2 \mathrm{d}t}. \tag{3.4}$$

Only groups with $D_i < D^{thr}$ are qualified as wave groups in the time series $\eta(t)$, with D^{thr} serving as a user-defined (subjective) criterion to define wave groups.

Based on the detected groups, we can construct an approximation of P_{temp} as

$$P_{temp}^a = \frac{\sum_{i=1}^m S(l_i, a_i)}{T_{end}} \equiv \frac{m}{T_{end}} \int S(l, a) p_{LA}(l, a) \, \mathrm{d}l \mathrm{d}a, \tag{3.5}$$

where the input-to-response (ItR) function $S(l_i, a_i) = \int \mathbf{1}\big(|r(t; l_i, a_i)| - r_s\big) \mathrm{d}t$ is the time of responses $r(t; l_i, a_i)$ exceeding r_s in group (l_i, a_i), m is the total number of groups in $\eta(t)$. In (3.5), P_{temp}^a is expressed in two equivalent ways, the first through the summation of exceeding times over all groups in $\eta(t)$, and the second through sampling in the parameter space (L, A) with known probability distribution $p_{LA}(l, a)$ obtained from $\eta(t)$ (see figure 3.1(c)). The equivalence of these two expressions can be demonstrated by computing $\int S(l, a) p_{LA}(l, a) \, \mathrm{d}l \mathrm{d}a$ with sufficiently large m samples in a Monte Carlo way, i.e., $\int S(l, a) p_{LA}(l, a) \, \mathrm{d}l \mathrm{d}a = \sum_{i=1}^m S(l_i, a_i)/m$. We note that P_{temp}^a is an approximation to P_{temp} ((3.5) relative to (3.1)) since certain information is lost when the group parameterization is conducted. For a narrow-band wave field, the most influential information lost is the initial condition of the ship encountering a wave group that can cause deviation of P_{temp}^a from P_{temp}, as carefully analyzed in [44]. In this work, we focus on the efficient sampling method to compute P_{temp}^a, and discuss how the proposed method can be further developed for the computation of P_{temp} in §3.3.

We also remark that although the definition in (3.5) involves group numbers m (that depends on the group definition criterion through D^{thr}), P_{temp}^a can be considered to be independent of D^{thr} in the practical calculation. This can be easily understood from (3.5) especially through the definition with summation of S. As long as the large groups leading to positive $S(l_i, a_i)$ are correctly identified, P_{temp}^a remains constant even though the number of small wave groups varies. This is generally true with our group detection algorithm where the threshold D^{thr} only affects the identification of small wave groups (that are ambiguous in nature). One may alternatively argue that

another measure could be defined as the number of groups leading to the exceedance within $[0, T_{end}]$, which is also invariant due to the above argument and involves only the group maximum in its computation. However, such a measure provides less information since $S(l_i, a_i)$ additionally captures the 'severity' of a given wave group as discussed in §3.1.

Our next task is to design an efficient method for the computation of P^a_{temp} without going through the physical computation of $S(l_i, a_i)$ for each group. In principle this will be achieved through Bayesian experimental design (BED) to compute the response from a few informative groups for the construction of function $S(l, a)$. In the following sections, we will introduce two basic components of BED specifically in the context of the computation of P^a_{temp}: (1) an inexpensive surrogate model to obtain S based on Gaussian process regression (GPR), and (2) an acquisitive function to sequentially select the next-best sample for acceleration of the convergence to P^a_{temp}.

3.2.2 Surrogate model

As a general procedure in BED, we may construct a surrogate model for $S(l, a)$ through the Gaussian progress regression (GPR). Given a dataset $\mathcal{D} = \{(l^i, a^i), f(l^i, a^i)\}_{i=1}^{i=n}$ consisting of n inputs and the corresponding outputs, our objective is to infer the underlying function f, which can be expressed by a Gaussian process

$$f(l, a)|\mathcal{D} \sim \mathcal{N}\big(\mathbb{E}(f(l, a)|\mathcal{D}), \text{cov}(f(l, a), f(l', a')|\mathcal{D})\big), \tag{3.6}$$

with details of the mean and covariance summarized in Appendix A.

However, a direct application of the GPR on function $S(l, a)$ can be problematic, especially in the context of the computation in (3.5). By definition as the exceeding time, the function S is characterized by zero function values for majority of the input parameters (l, a) and positive for only critical ranges of input. The two regions are
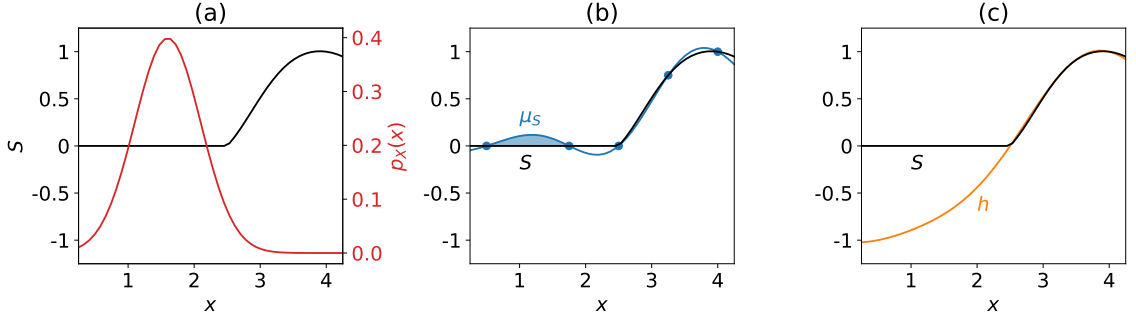
Figure 3.2: A one-dimensional (1D) demonstration of the issues regarding GPR for exceeding time. This example can be perceived with variable x as the group amplitude a with fixed group length, say $l = 1$. (a) The response function of exceeding time $S(x)$ (——) and the input probability distribution $p_x(x)$ (——); (b) The function $S(x)$ (——) and the mean of GPR prediction $\mathbb{E}(S(x)|\mathcal{D})$ (——) with five sample points (•), as well as the false positive region as the shaded area; (c) The function $S(x)$ (——) and the auxiliary function $h(x)$ (——).

associated with drastically different scales of functional variation (the former as a constant with no variation and the latter with much faster variation). In addition, the probability of the input is usually heavily placed on the region where the function value is zero (i.e., majority of groups leading to no exceedance), with the situation illustrated in a phenomenological 1D example in figure 3.2(a). If a GPR is placed on the dataset from such a function, the predicted function is sketched in figure 3.2(b), where prediction errors are inevitably associated with the part with $S = 0$. These errors are devastating for the computation of P^a_{temp} when the false positive value of S is accompanied by a large input group probability $p_{LA}(l, a)$, leading to a significant false exceeding probability (considering that the true value of P^a_{temp} is small). This situation can only be alleviated if we have a (very) large number of samples in the region of $S = 0$, but this contradicts the goal of the sequential sampling to place the emphasis in regions with large exceeding time.

To solve this problem, we define an auxiliary function $h(l, a)$ derived from $S(l, a)$, where h contains the information about exceeding time in S and is favorable for GPR to learn. Ideally, h is required to be a function of uniform scale of variation and free

of false positive value of exceedance. Considering these constraints, a function h can be defined as

$$h(l, a) = \begin{cases} \dfrac{S(l, a)}{l}, & if \ S(l, a) > 0 \\ \dfrac{r_{max}(l, a) - r_s}{r_s}, & if \ S(l, a) = 0 \end{cases}. \tag{3.7}$$

where $r_{max}(l, a) = \max_t |r(t; l, a)|$ is the group-maximum response in the group (l, a). When $S = 0$, function h takes negative values with $r_{max} - r_s$ serving as a "negative penalty" quantifying how far the response is from the threshold. In such a way, the false positive value of S can be avoided. We also normalize the two piecewise segments of h respectively by factors l and r_s so that both segments are in $O(1)$, ensuring the uniform scale of variation. Figure 3.2(c) demonstrates a typical function h in the 1D example.

After the GPR of h is available, we can recover function S by $S \equiv \mathbf{1}(h) \, h \, l$. Such recovered S is also free of false positive values because of the "negative penalty" placed on h.

3.2.3 Acquisition function

Given the function h (and S) learned from the GPR based on dataset $\mathcal{D}$, P^a_{temp} can be considered as a random variable with its randomness resulting from the uncertainty in h. The uncertainty in P^a_{temp} can therefore be estimated as

$$U(\mathcal{D}) = \frac{\sum_{i=1}^{m} \left| S^+(l_i, a_i | \mathcal{D}) - S^-(l_i, a_i | \mathcal{D}) \right|}{T_{end}}, \tag{3.8}$$

where $S^\pm(l_i, a_i | \mathcal{D}) = \mathbf{1}\big(h^\pm(l_i, a_i | \mathcal{D})\big) h^\pm(l_i, a_i | \mathcal{D}) \, l_i$, with $h^\pm$ the upper and lower bound of h estimated by (one standard deviation up and below)

$$h^\pm(l, a | \mathcal{D}) = \mathbb{E}(h(l, a) | \mathcal{D}) \pm \mathrm{std}(h(l, a) | \mathcal{D}). \tag{3.9}$$

Our purpose is to select the next sample, after adding which the uncertainty in P^a_{temp} is significantly reduced. For an efficient way to fulfill this purpose, we further formulate the uncertainty in P^a_{temp} after adding one hypothetical sample at $\tilde{l}, \tilde{a}$:

$$U(\mathcal{D}, \tilde{l}, \tilde{a}) = \frac{\sum_{i=1}^{m} \left| S^+\left(l_i, a_i | \mathcal{D}, \overline{h}(\tilde{l}, \tilde{a})\right) - S^-\left(l_i, a_i | \mathcal{D}, \overline{h}(\tilde{l}, \tilde{a})\right) \right|}{T_{end}} \qquad (3.10)$$

with $\overline{h}(\tilde{l}, \tilde{a}) = \mathbb{E}(h(\tilde{l}, \tilde{a}) | \mathcal{D})$ the mean prediction in (3.6). The computation of $S^{\pm}\left(l_i, a_i | \mathcal{D}, \overline{h}(\tilde{l}, \tilde{a})\right)$ relies on $h^{\pm}(l_i, a_i | \mathcal{D}, \overline{h}(\tilde{l}, \tilde{a}))$, with the formulation of the latter detailed in Appendix A.

The selection of the next sample can then be formulated as an optimization problem to minimize the hypothetical next-step uncertainty:

$$l^*, a^* = \mathrm{argmin}_{\tilde{l}, \tilde{a}} \, U(\mathcal{D}, \tilde{l}, \tilde{a}), \qquad (3.11)$$

which can be solved using standard optimization methods. In our work, we apply a combined brute-force grid search (with coarse grid) and a gradient-based (for fine search) method[2] in the two-dimensional space.

We note that if the (initial) dataset $\mathcal{D}$ contains only samples far from exceedance, $h^{\pm}(l_i, a_i)$ can be negative for all wave groups (see figure 3.3(a)). As a result, $S^{\pm}(l_i, a_i) = 0$ and the uncertainty defined in (3.10) vanishes (figure 3.3(c)). Under these situations, in order to robustly initiate the sequential sampling, one can temporarily apply a 'soft' indicator function in computing $S^{\pm}(l_i, a_i)$, i.e., we use a logistic function $1/(1 + e^{-ch^{\pm}})$ with $c \gg 1$ instead of $\mathbf{1}(h^{\pm})$ (figure 3.3(b)). This procedure replaces the zero values of $S^{\pm}(l_i, a_i)$ by small positive values, but with meaningful uncertainties (represented by the upper and lower bounds) to drive the next sample to regions with larger exceedance (figure 3.3(c)).

We summarize the full BED algorithm in Algorithm **2**.

---

[2] https://github.com/scipy/scipy/blob/v1.8.1/scipy/optimize/_optimize.py
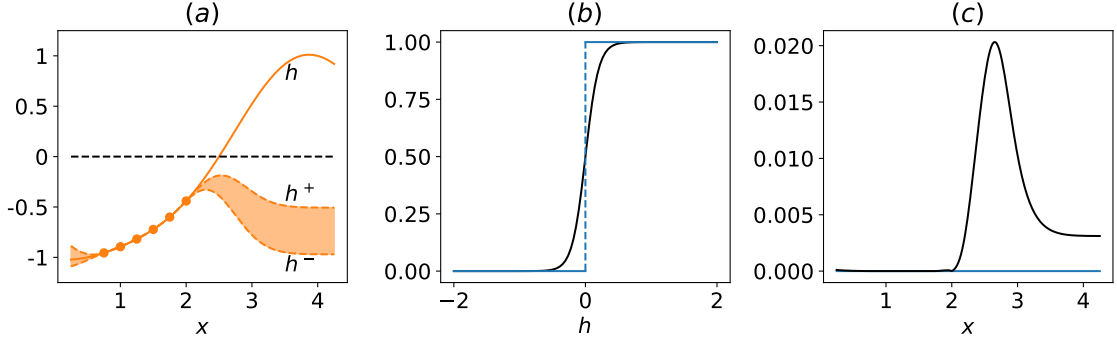
Figure 3.3: A 1D demonstration of the effect of softer indicator function. This example can be perceived with variable x as the group amplitude a with fixed group length, say $l = 1$. (a) upper and lower bounds $h^{\pm}(x)$ (- - -) generated from sample points on true function $h(x)$ (—) that are far from exceedance (- - -); (b) The original indicator function $\mathbf{1}(h)$ (—) and the soft indicator function $1/(1 + e^{-ch})$ with $c = 10$ (—); (c) The uncertainty $|S^+(x|\mathcal{D}) - S^-(x|\mathcal{D})|$ as a function of x, computed by the original (—) and soft (—) indicator functions.

Algorithm 2 Bayesian experimental design for temporal exceeding probability

---

Require: Number of initial samples n_{init} and sequential samples n_{seq}
Input: Initial dataset $\mathcal{D} = \{(l^i, a^i), h(l^i, a^i)\}_{i=1}^{n_{init}}$
    Initialization $n = n_{init}$
    while $n < n_{init} + n_{seq}$ **do**
        Train the surrogate model (3.6) with $\mathcal{D}$
        Solve (3.11) to find the next best sample (l^{n+1}, a^{n+1})
        Implement numerical simulation to get $h(l^{n+1}, a^{n+1})$
        Update the dataset $\mathcal{D} = \mathcal{D} \cup \{(l^{n+1}, a^{n+1}), h(l^{n+1}, a^{n+1})\}$
        $n = n + 1$
    end while
Output: Compute the temporal exceeding probability (3.5) based on the surrogate model

---

3.3 Validation of the Method

In this section, we compute P_{temp}^a by the proposed sequential sampling approach, with comparison to the result from the space-filling Latin hypercube (LH) sampling and the true solution of P_{temp}^a. In addition, we will discuss the invariance of P_{temp}^a, the difference between P_{temp}^a and P_{temp} and suggest improved methods to eventually capture P_{temp}. Since P_{temp}^a and P_{temp} need to be accurately evaluated, an efficient ship response simulator is required. For this work, we use a nonlinear roll equation to calculate $\xi(t)$ from a given wave signal $\eta(t)$:

$$\ddot{\xi} + \alpha_1\dot{\xi} + \alpha_2\dot{\xi}|\dot{\xi}| + (\beta_1 + \epsilon_1\cos(\chi)\eta(t))\xi + \beta_2 r^3 = \epsilon_2\sin(\chi)\eta(t), \qquad (3.12)$$

which models the ship roll response due to nonlinear resonance and parametric roll in oblique irregular waves. Empirical coefficients are set as $\alpha_1 = 0.35$, $\alpha_2 = 0.06$, $\beta_1 = 0.04$, $\beta_2 = -0.1$, $\chi = \pi/6$, $\epsilon_1 = 0.016$, and $\epsilon_2 = 0.012$. The wave field $\eta(t)$ to be decomposed into groups is generated from a narrow-band spectrum of a Gaussian form:

$$F(\omega) = \frac{H_s^2}{16}\frac{1}{\sqrt{2\pi d}}\exp(\frac{-(\omega - \omega_p)^2}{2d^2}), \qquad (3.13)$$

with ω the angular frequency, $H_s = 12m$ the significant wave height, $\omega_p = 0.067s^{-1}$ the peak (carrier) wave frequency (corresponding to peak period $T_p = 15s$), and $d = 0.02s^{-1}$ a parameter of the spectral bandwidth. In particular, we compute $\eta(t) = \sum_n a_n\cos(n\Delta\omega t + \phi_n)$ with $\Delta\omega = 0.00026s^{-1}$, n from 1 to 1024, ϕ_n being a random phase for each mode, and $a_n = \sqrt{2F(n\Delta\omega)\Delta\omega}$.

Before describing the procedure of our computation, we remark that there is no mathematical proof that the roll statistics resulted from nonlinear equations such as (3.12) is ergodic or not. Numerical studies over relatively short time intervals suggest that the roll process may not be ergodic [11], but it is not clear what the situation is for very long time series (such as our case). Furthermore, if capsize is involved in the

time series of response, then it is guaranteed that the roll process is non-ergodic. We have practically avoided this case by tuning the parameters in the model such as the damping coefficient α_1. Under this situation, our computational procedure, described below, provides a unique answer of P^a_{temp} given any time series of wave elevation that is sufficiently long (since such time series lead to the same probability distribution of wave groups). Therefore, as far as P^a_{temp} is considered, the roll process treated by our approach is assumed (or considered) ergodic.

Following procedures in §3.2.1, the wave field described by (3.13) can be reduced to a parameter space (L, A) with known probability $p_{LA}(l, a)$ (figure 3.1(c)). In computing P^a_{temp}, the response $\xi(t; l, a)$ from a group (l, a) is needed, which is calculated by simulation of (3.12) in $t \in \{-3l, \ 3l\}$ with input $\eta(t; l, a) = a \exp(-\frac{t^2}{2l^2}) \cos(\omega_p t)$ and $(0, 0)$ initial condition at $t = -3l$. We note that the constant frequency ω_p is justified by the narrow-band setting of the spectrum (3.13). The choice of $3l$ does not appreciably affect the final solution as long as the value is sufficiently large to cover the portion of the group with significant amplitude. The exceeding time in one group is then computed by $S(l, a) = \int_{-3l}^{3l} \mathbf{1}\big(|\xi(t; l, a)| - r_s\big) \, \mathrm{d}t$, and the algorithm described in §3.2.3 can be applied accordingly.

The results of P^a_{temp} for four cases with $r_s = 0.25$, 0.3, 0.35, and 0.37 radians are respectively shown in figure 3.4 (a), (b), (c), and (d). In addition, we also vary the parametric parameter in (3.12) with cases of $\epsilon_1 = 0$ and $\epsilon_1 = 0.008$ shown in (e) and (f) for $r_s = 0.3$ radians. In all cases, we present results computed by our new sequential BED sampling and standard LH sampling, along with the true solution of P^a_{temp} and P_{temp} (by brute-force calculations with a large number of groups and a long time series, with the difference in their values discussed at the end of this section). The sequential samplings are conducted with an initial data set of 8 LH samples. In order to have a fair comparison of the results, we perform 100 calculations of the sequential-sampling and LH-sampling methods with both starting from different
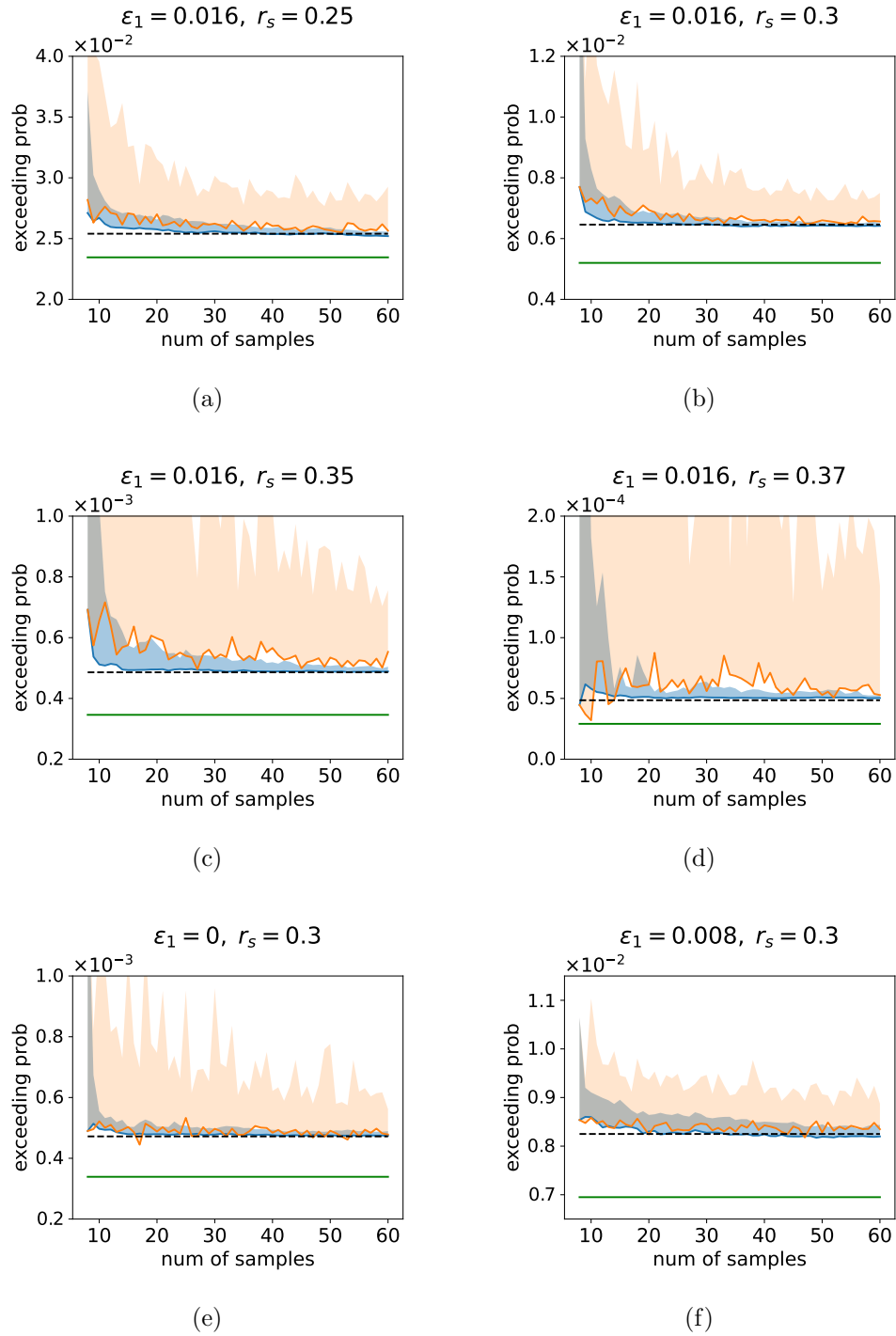
Figure 3.4: Temporal exceeding probability P_{temp}^a as a function of sampling numbers, calculated from sequential sampling (—) and LH sampling (—), for (a) $\epsilon_1 = 0.016$, $r_s = 0.25$ radians, (b) $\epsilon_1 = 0.016$, $r_s = 0.3$, (c) $\epsilon_1 = 0.016$, $r_s = 0.35$, (d) $\epsilon_1 = 0.016$, $r_s = 0.37$, (e) $\epsilon_1 = 0$, $r_s = 0.3$, and (f) $\epsilon_1 = 0.008$, $r_s = 0.3$. The shaded regions are bounded by the median and 85% of the results from 100 applications of the corresponding methods. The true solution of P_{temp}^a and P_{temp} are respectively shown by (- - -) and (—).
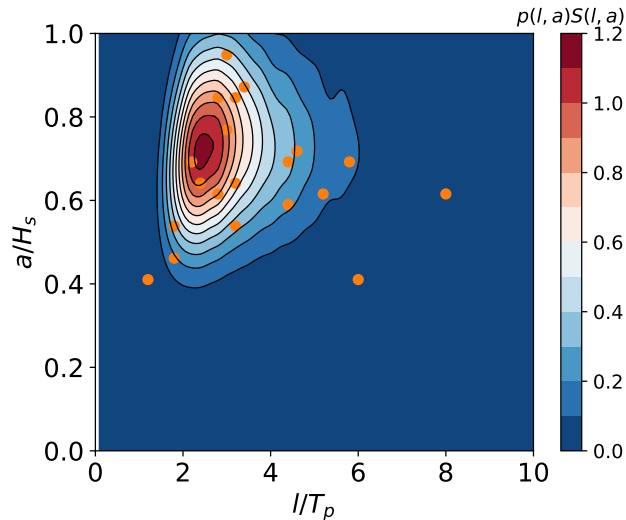
Figure 3.5: Positions of sequential samples ($\bullet$) for $\{\epsilon_1 = 0.016, r_s = 0.25\}$ in the space of $S(l,a)p_{LA}(l,a)$ with a contour plot.

initial samples, and present the median and 85% percentile of the results, with the difference shaded in figure 3.4 to represent the uncertainty of the results. In general, it is clear that the sequential-sampling results approach to the true solution of P^a_{temp} much faster than that by LH sampling, with the former also exhibiting a much smaller uncertainty. In contrast, even at the end of 60 samples, the LH-sampling results have not converged yet, reflected by a deviation of the median value from the true solution of P^a_{temp} (in most cases) and a very large uncertainty (in all cases). The excessive uncertainty level of the LH-sampling results indicate that a single experiment by LH sampling has a large chance to provide a solution that significantly deviates from the true solution of P^a_{temp}. This is especially the case when r_s is large such as those in (c) and (d).

We further examine the reason for the fast convergence of the sequential-sampling results by plotting the sample positions for case $\{\epsilon_1 = 0.016, r_s = 0.25\}$ in the space of $S(l,a)p_{LA}(l,a)$ in figure 3.5. Here the contour of $S(l,a)p_{LA}(l,a)$ provides a measure of the importance of a group (l,a) in computing P^a_{temp}, which can also be seen from

Table 3.1: Comparison of P^a_{temp} and P^a_{group} ($\epsilon_1 = 0.016$) for varying D_{thr} and r_s.

D_{thr}	P^a_{temp}			P^a_{group}		
	$r_s = 0.25$	$r_s = 0.3$	$r_s = 0.35$	$r_s = 0.25$	$r_s = 0.3$	$r_s = 0.35$
0.35	0.02566	0.006473	0.0004861	0.2632	0.09884	0.01335
0.30	0.02564	0.006471	0.0004860	0.2781	0.1047	0.01414
0.25	0.02554	0.006462	0.0004859	0.3278	0.1246	0.01688
0.20	0.02467	0.006352	0.0004825	0.4050	0.1598	0.02206
variation[1]	3.8 %	1.8 %	0.7 %	35 %	38 %	39 %

[1] The variation is computed by the relative difference (normalized by the largest values) in each column.

(3.5). As shown in the figure, most sequential samples are driven to the region with significant $S(l,a)p_{LA}(l,a)$, indicating the effectiveness of our BED method with new developments in both the surrogate model and acquisition function.

We next demonstrate the invariance of temporal exceeding probability with group detection criterion, particularly the threshold D_{thr} for (3.4). Table 3.1 lists the values of P^a_{temp} and the group-maximum exceeding probability ($P^a_{group} = \sum_{i=1}^{m} \mathbf{1}(r_{max}(l_i, a_i) - r_s)/m$) for different r_s and D_{thr} (with same $\epsilon_1 = 0.016$). For all values of r_s, P^a_{group} changes more than 30% for D_{thr} varying from 0.35 to 0.2. In contrast, P^a_{temp} remains almost a constant, with the very small variation resulting from some small/deformed groups leading to large motion that escape from the detection.

We finally discuss the difference between P^a_{temp} and P_{temp} shown in figure 3.4. For these cases, P^a_{temp} represents a $O(20 - 30\%)$ over-estimation compared to P_{temp}. As discussed in §3.2, the difference between the two quantities is mainly due to the neglect of the varying initial condition in computing P^a_{temp}. In addition, for general cases, the over-estimation cannot be guaranteed and it is desirable to develop a more sophisticated method to directly and efficiently compute P_{temp}. We believe that this can be achieved through the combination of the method in this chapter and techniques compensating the lost information in the wave group representation, which will be presented in §4.

3.4 Coupling with CFD

While our sequential BED method has been validated using a low-fidelity ship motion model in the previous section, it may be desired to couple the approach to higher-fidelity (e.g., CFD) models in practical applications. In this section, we demonstrate such an application with CFD simulations to compute the ship roll responses. For simplicity, we consider the motion of a two-dimensional (2D), square-shaped hull geometry with $40m \times 40m$ cross section and density $\rho_h = 0.5\rho_w$ with ρ_w being the water density. The input wave groups are considered as spatial groups located at a distance of about $800m$ from the ship at the initial time, and the ship response is computed as the wave group travels over $800m$ and then across the ship.

We remark that the use of spatial wave groups (instead of temporal groups in §3.3) are convenient and necessary for the CFD settings. Accordingly, in this section, we interpret t and l in (3.3) as the spatial position and spatial length of the wave groups. The probability distribution of the parameters is obtained from a wavenumber spectrum converted from (3.13), with surface elevation inside the group following the peak wavenumber $k_p = \omega_p^2/g$ (with g the gravitational constant). In each CFD simulation, the initial field is partitioned into the left and right sides, where the wave groups are located on the left side with a distance of about $800m$ from its peak to the ship at rest on the right side (see figure 3.6(a)). The velocity field of initial condition is calculated based on linear dispersion relation placed on each propagating mode of the group. The exceeding time is computed from the time series of the simulated roll motion (see figure 3.6(b) and (c)).

The CFD model is developed using the open-source code OpenFOAM [63]. The interFoam solver is used to capture the air-water interface through an algebraic volume of fluid (AVOF) method. A standard $k - \epsilon$ turbulence model is applied in conjunction with the AVOF method [93]. The 2D hull is considered as a rigid body, moving under the force exerted by flow pressure and shear stress. The motion of the hull is
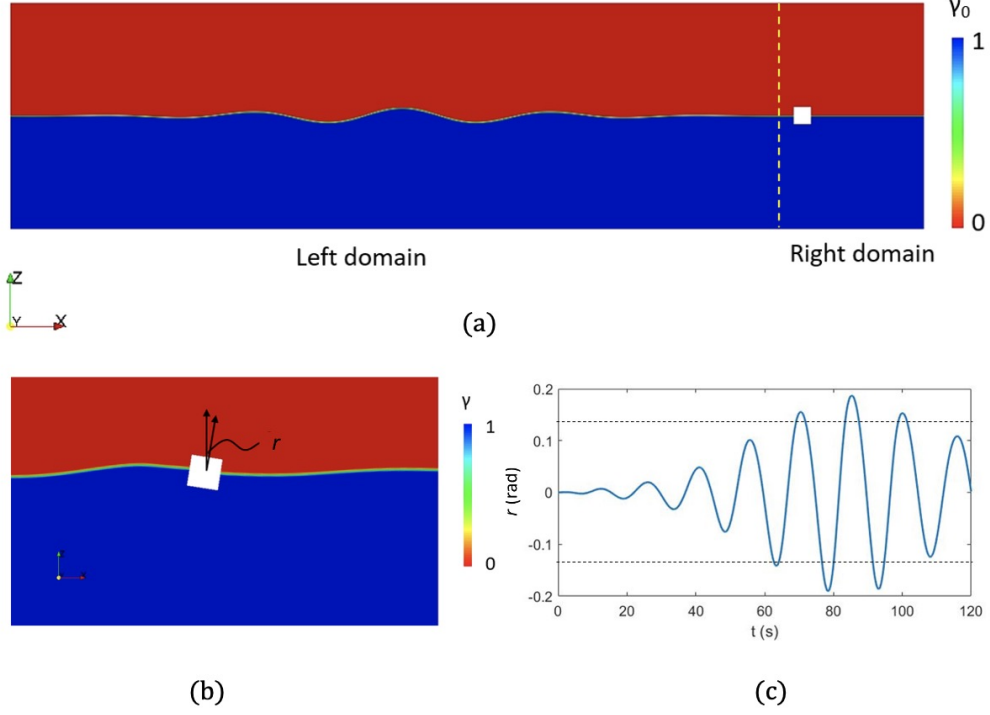
Figure 3.6: A typical CFD simulation for a wave group with $a = 8.1m$ and $l = 332.7m$. (a) initial wave field with volume fraction γ_0, with hull located on the right; (b) volume fraction γ in the process of a wave group interacting with the hull; (c) time series of $\xi(t)$ (——) with threshold r_s (- - -).
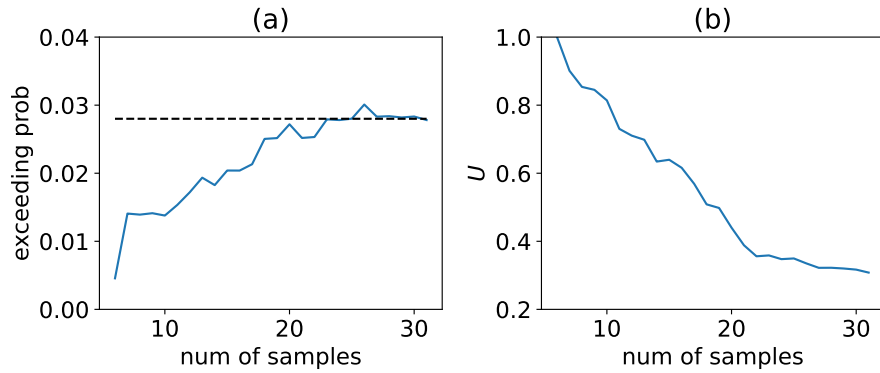
Figure 3.7: (a) Temporal exceeding probability P^a_{temp} (——) for $r_s = 0.13$ as a function of the sequential sample number, with the convergent level indicated (- - -); (b) Normalized uncertainty as a function of the sequential sample number.

calculated by numerical integration implemented by the Newmark method [99]. More details about the boundary conditions, model equations, and grid resolutions of the solver can be found in §2.5.

In computing P^a_{temp}, we set a threshold of $r_s = 0.13$ radians in a wave field of $H_s = 9m$ and otherwise the same as that in §3.3. Figure 3.7(a) shows the results of P^a_{temp} using 25 sequential samples followed by 6 LH samples. It is clear that the estimated P^a_{temp} approaches a constant level after $O(20)$ sequential samples, indicating the convergence of the result. Since the true solution of P^a_{temp} is not available for this case (unless much more computational resources can be allocated to run the CFD model, which is not feasible now), we examine the uncertainties (3.8) associated with the solution, normalized by its value at initial time. As shown in figure 3.7(b), the uncertainty decreases rapidly as P^a_{temp} approaches a constant level, demonstrating the effectiveness of our method when coupled to CFD.

3.5 Summary

In this work, we develop a computational framework to efficiently compute the temporal exceeding probability of ship responses in a random wave field, i.e., the fraction of time that the response exceeds a specified threshold. As an important complement to group-maximum exceeding probability, temporal exceeding probability provides a robust measure of the extreme motion due to its invariance regarding different group definitions and consistency with practical design metrics. To enable the computation of temporal exceeding probability, we develop a novel BED framework incorporating (1) a uniformly-varying response function resulting from negative penalty and normalization of the group exceeding time; (2) a new acquisition function focusing on sampling wave groups associated with significant exceeding time. We validate our framework in the context of a nonlinear roll equation in terms of the efficiency of

the sequential sampling and the invariance of results to wave group definitions. We finally demonstrate the coupling of our framework to CFD simulations to show its applicability to higher-fidelity models.

CHAPTER 4

Bayesian Experimental Design for Extreme-Event Probability in Stochastic Input-to-Response Systems

4.1 Introduction

In §2 and §3, we represent the irregular waves by a large number of wave groups to compute extreme ship motion statistics. While group parameterization significantly speeds up the computation, this procedure neglects information about wave phases, frequency modulation, and initial conditions of the ship when encountering the wave groups. To compensate for the lost information, it is essential to treat the input-to-response (ItR) function as a stochastic function. More generally, a stochastic ItR may originate from (a) an intrinsically stochastic dynamical system, e.g., stochastic differential equations modeling a physical diffusion process or stock prices [58]; the stochastic model of climate variability including the non-average 'weather' component as random forcing terms [57], and (b) some uncertain variables that are not easily incorporated in a low-dimensional input parameter space, in particular when dimension reduction technique is applied to a high-dimensional input space as we initially discussed. Under such situations, the probability distribution of the response is critically influenced by the randomness in the ItR (in addition to the probabil-

ity distribution of input parameters). If the randomness of the ItR is uniform for all input parameters, previous techniques [31, 15, 59, 149, 95] for deterministic ItR can be extended to handle the situation (by incorporating the uniform randomness in the Gaussian process regression). However, more often, the uncertainty of the ItR is inhomogeneous for different input parameters, e.g., due to the interaction of input parameters and random terms in stochastic equations of the aforementioned case (a) (also see [108] for more examples), or non-uniform impact of missing dimensions in case (b). This results in a heteroscedastic ItR with the variance of response not representable as a constant. To our knowledge, currently, there is no sequential BED method designed to consider heteroscedastic uncertainty in ItR, and its effect on extreme-event probability.

In this chapter, we propose a new method to quantify the probability of extreme events (defined as an observable above a given threshold) considering the ItR with heteroscedastic uncertainty. The core of our algorithm is a variational heteroscedastic Gaussian process regression (VHGPR) which approximates the ItR with sufficiently low computational cost and high accuracy. This brings major improvement upon all previous BED methods employing the standard Gaussian process regression (SGPR) which are unable to resolve the heteroscedasticity in the ItR. Accordingly, we formulate a new acquisition function for selecting the next-best sample considering both the probability distribution of inputs and uncertainty in ItR. We first demonstrate the effectiveness of our method in two synthetic problems to estimate the extreme-event probability. We show that drastically improved performance is achieved compared to existing approaches based on SGPR. Finally, we demonstrate the superiority of our method (to existing methods) in solving an engineering problem of estimating the extreme ship motion probability in irregular waves. The difficulty in this problem lies in the heteroscedastic uncertainty of the ItR resulting from the wave group parameterization which reduces the original high-dimensional wave field to a two-dimensional

parameter space. We show that the effect of this type of uncertainty on the exceeding probability can be successfully considered in our approach.

This chapter is adapted from [46]. The Python code for the proposed algorithm, named HGPextreme, is available on Github[1].

4.2   Method

4.2.1   Problem setup

We start from an ItR system with input $\mathbf{x} \in \mathbb{R}^d$ of known distribution $X \sim p_{\mathbf{x}}(\mathbf{x})$ and response $y \in \mathbb{R}$. An ItR function S directly relates $\mathbf{x}$ to y with its randomness represented by ω:

$$y(\omega) = S(\mathbf{x}, \omega), \ \omega \in \Omega. \tag{4.1}$$

To be more specific, ω is a random seed lying in the sample space Ω. For given $\mathbf{x}$, $y(\omega)$ represents a random variable, i.e., a function from sample space to real number $\Omega \to \mathbb{R}$.

Our interest is the exceeding probability of $y(\omega)$ above a threshold δ:

$$P_e \equiv \mathbb{P}(S(X, \omega) > \delta) = \int \mathbb{P}(S(X, \omega) > \delta | X = \mathbf{x}) p_{\mathbf{x}}(\mathbf{x}) \mathrm{d}\mathbf{x}$$

$$= \int \mathbb{P}(S(\mathbf{x}, \omega) > \delta) p_{\mathbf{x}}(\mathbf{x}) \mathrm{d}\mathbf{x}. \tag{4.2}$$

It is clear that both distribution $p_{\mathbf{x}}$ and uncertainty ω contribute to the exceeding probability in (4.2). Moreover, the variance of the response S (introduced by ω) is generally different for different input $\mathbf{x}$, resulting in a heteroscedastic uncertainty of the ItR. We remark that this problem setup including (4.1) and (4.2) are motivated in the discussion in §4.1, and resolving this heteroscedasticity in the ItR is critical for the success of our new method (or improvement of our method compared to all

---

[1]https://github.com/umbrellagong/HGPextreme

previous approaches) as will be discussed in §4.3

A brute-force computation of (4.2) calls for extensive Monte-Carlo samples in the probability space associated with both X and ω, e.g., [108], which is prohibitive under expensive queries of $S(\mathbf{x}, \omega)$. Therefore, we seek to develop a sampling algorithm following the sequential BED framework, where each sample is selected making use of the existing information of previous samples. Our new sampling algorithm also has to be developed in conjunction with the heteroscedastic uncertain ItR that has not been considered before. In summary, two key components in our new approach are (i) an inexpensive surrogate model based on the variational heteroscedastic Gaussian process regression (VHGPR) to approximate the heteroscedastic ItR; and (ii) an optimization based on an acquisition function to provide the next-best samples with fast convergence in computing (4.2). We next describe the two components in detail in §4.2.2 and §4.2.3 , followed by the overall algorithm.

4.2.2 Surrogate model

To introduce the surrogate model for the ItR, we first rewrite (4.1) as

$$S(\mathbf{x}, \omega) = f(\mathbf{x}) + R(\mathbf{x}, \omega), \tag{4.3}$$

where $f(\mathbf{x}) \equiv \mathbb{E}[S(\mathbf{x}, \omega)]$ is the mean of $S(\mathbf{x}, \omega)$ with respect to ω, and $R(\mathbf{x}, \omega)$ is the uncertain component with zero mean. Given a dataset (from previous samples) $\mathcal{D} = \{\mathbf{x}^i, y^i\}_{i=1}^n$, our purpose is to approximate (4.3) using Gaussian process regression as involved in many BED problems.

In standard Gaussian process regression (SGPR), as implemented in most previous

applications for extreme-event probability, one can approximate (4.3) as:

$$f(\mathbf{x})|\mathcal{D} \sim \mathcal{GP}\big(\mathbb{E}(f(\mathbf{x})|\mathcal{D}), \mathrm{cov}(f(\mathbf{x}), f(\mathbf{x}')|\mathcal{D})\big), \qquad (4.4)$$

$$R(\mathbf{x}, \omega) \sim \mathcal{N}(0, \gamma_0^2), \qquad (4.5)$$

where $\mathcal{GP}(\cdot, \cdot)$ represents a Gaussian process with the first argument as the mean and the second argument as the covariance function. The uncertain component $R(\mathbf{x}, \omega)$ is approximated by an independent normal function at all $\mathbf{x}$ with constant variance γ_0^2. Clearly, the heteroscedasticity in ItR (i.e., the dependence of R on $\mathbf{x}$) cannot be captured by the SGPR.

To incorporate the heteroscedasticity, we need to rely on the heteroscedastic Gaussian process regression (implemented as VHGPR following [72] in this chapter). In VHGPR, we are able to approximate (4.3) as

$$f(\mathbf{x})|\mathcal{D} \sim \mathcal{GP}\big(\mathbb{E}(f(\mathbf{x})|\mathcal{D}), \mathrm{cov}(f(\mathbf{x}), f(\mathbf{x}')|\mathcal{D})\big), \qquad (4.6)$$

$$R(\mathbf{x}, \omega) \sim \mathcal{N}(0, e^{g(\mathbf{x})}), \qquad (4.7)$$

$$g(\mathbf{x})|\mathcal{D} \sim \mathcal{GP}\big(\mathbb{E}(g(\mathbf{x})|\mathcal{D}), \mathrm{cov}(g(\mathbf{x}), g(\mathbf{x}')|\mathcal{D})\big), \qquad (4.8)$$

where the heteroscedastic (log) variance of the uncertain term $R(\mathbf{x}, \omega)$ is represented by another Gaussian process $g(\mathbf{x})|\mathcal{D}$. The $e^{g(\mathbf{x})}$ term in (4.7) is used to guarantee that the variance of $R(\mathbf{x}, \omega)$ is always positive for any $g(\mathbf{x})$. We remark that (4.7) implies that the distribution associated with ω in (4.1) can be approximated by a Gaussian. Although the Gaussian assumption is a standard practice in many literature [146, 80, 87, 53], we will perform a validity check of this assumption in the specific problem solved in §4.3.

Both approximations in SGPR (in terms of (4.4)) and VHGPR ((4.6) and (4.8)) are computed as posterior predictive distributions under a Bayesian framework, with

hyperparameters (say $\boldsymbol{\theta}$) determined from maximizing the likelihood function $p(\mathcal{D}|\boldsymbol{\theta})$. For SGPR, both the likelihood function and posterior (4.4) can be derived analytically, allowing a straightforward and inexpensive numerical implementation. In contrast, for heteroscedastic GPR, the introduction of the Gaussian process on $g(\mathbf{x})$ prohibits analytical results on the likelihood function and posterior, posing great challenges in the numerical computation (which involves high-dimensional integration).

In order to reduce the computational cost, variational inference is applied in VHGPR, which uses parameterized Gaussian distributions to approximate some critical distributions involved in the posterior and likelihood function. These Gaussian distributions can be determined efficiently through some optimization problems to minimize their differences from the critical distributions. As a result of this approximation, the high-dimensional integration can be reduced to analytical formulations which leads to inexpensive computations (approximations) of the posterior and the likelihood function. In particular, the computational cost of VHGPR is only twice of SGPR, alleviating the resource requirement for the computation. More details on the algorithms of the VHGPR, along with SGPR, are summarized in Appendix B.1 for completeness. The interested readers can also refer to [113, 72] for details.

In summary, the VHGPR provides us with an estimation of the ItR, $S(\mathbf{x}, \omega|f(\mathbf{x}), g(\mathbf{x}))$, where $f(\mathbf{x})|\mathcal{D}$ and $g(\mathbf{x})|\mathcal{D}$ follow distributions in (4.6) and (4.8) (Hereafter we will delete the condition on $\mathcal{D}$ for conciseness). Given realizations of $f(\mathbf{x})$ and $g(\mathbf{x})$, the intrinsic randomness in ItR is expected to be captured by ω, i.e., the heteroscedastic distribution in (4.7).

4.2.3 Acquisition function

Given the VHGPR of the ItR, the exceeding probability can be expressed as

$$\mathbb{P}(S(X, \omega|f(\mathbf{x}), g(\mathbf{x})) > \delta) = \int \mathbb{P}(S(\mathbf{x}, \omega|f(\mathbf{x}), g(\mathbf{x})) > \delta)p_{\mathbf{x}}(\mathbf{x})\mathrm{d}\mathbf{x}, \qquad (4.9)$$

which depends on the realizations of $f(\mathbf{x})$ and $g(\mathbf{x})$. The purpose here is to construct an acquisition function, based on which the next sample can be selected to minimize the variance of the estimation (4.9), i.e., $\mathrm{var}_{f,g}[\mathbb{P}(S(X,\omega|f(\mathbf{x}),g(\mathbf{x})) > \delta)]$. For this purpose, the next-best sample is selected at the value of $\mathbf{x}$ which is associated with maximum uncertainty in the integrand of (4.9) (so that the sample is expected to reduce the uncertainty of (4.9) significantly):

$$\mathbf{x}^* = \mathrm{argmax}_{\mathbf{x}}\ \mathrm{std}_{f,g}[\mathbb{P}(S(\mathbf{x},\omega|f(\mathbf{x}),g(\mathbf{x})) > \delta)]p_{\mathbf{x}}(\mathbf{x}). \tag{4.10}$$

We note that (4.10) is closely related to the so-called U criterion [31] widely used in computing the exceeding probability associated with a deterministic ItR. In general, the U criterion seeks the most 'dangerous' point (i.e., point with maximum local variance), which in our problem corresponds to $\mathbf{x}^* = \mathrm{argmax}_{\mathbf{x}}\ \mathrm{std}_{f,g}[\mathbb{P}(S(\mathbf{x},\omega|f(\mathbf{x}),g(\mathbf{x})) > \delta)]$. Therefore, our acquisition function in (4.10) can be considered as a weighted U criterion which incorporates the influence of the input distribution $p_{\mathbf{x}}(\mathbf{x})$ in computing the variance of (4.9). Furthermore, the criterion in (4.10) corresponds to the upper bound of $\mathrm{var}_{f,g}[\mathbb{P}(S(X,\omega|f(\mathbf{x}),g(\mathbf{x})) > \delta)]$, as we can show

$$\begin{aligned}
&\mathrm{var}_{f,g}[\mathbb{P}(S(X,\omega|f(\mathbf{x}),g(\mathbf{x})) > \delta)] \\
&\leq \frac{1}{2} \int \mathrm{std}_{f,g}\Big[\mathbb{P}(S(\mathbf{x},\omega|f(\mathbf{x}),g(\mathbf{x})) > \delta)\Big]p_{\mathbf{x}}(\mathbf{x})\mathrm{d}\mathbf{x}.
\end{aligned} \tag{4.11}$$

The derivation for this upper bound is shown in Appendix B.2.

In practice, we approximate the operator $\mathrm{std}_{f,g} \equiv (\mathrm{var}_{f,g})^{0.5}$ in (4.10) by the two-dimensional spherical cubature integration [145] with 4 quadrature points (although

extension to more quadrature points is straightforward):

$$\mathbb{E}_{f,g}[\mathbb{P}(S(\mathbf{x},\omega|f(\mathbf{x}),g(\mathbf{x})) > \delta)]$$

$$= \int \mathbb{P}(S(\mathbf{x},\omega|f(\mathbf{x}),g(\mathbf{x})) > \delta)p_{f,g}(f(\mathbf{x}),g(\mathbf{x}))\mathrm{d}f(\mathbf{x})\mathrm{d}g(\mathbf{x})$$

$$\approx \frac{1}{4}\sum_{i=1}^{4} \mathbb{P}(S(\mathbf{x},\omega|\mathbf{u}^{(i)}) > \delta) = m, \tag{4.12}$$

$$\mathrm{var}_{f,g}[\mathbb{P}(S(\mathbf{x},\omega|f(\mathbf{x}),g(\mathbf{x})) > \delta)]$$

$$= \int (\mathbb{P}(S(\mathbf{x},\omega|f(\mathbf{x}),g(\mathbf{x})) > \delta) - m)^2 p_{f,g}(f(\mathbf{x}),g(\mathbf{x}))\mathrm{d}f(\mathbf{x})\mathrm{d}g(\mathbf{x})$$

$$\approx \frac{1}{4}\sum_{i=1}^{4} (\mathbb{P}(S(\mathbf{x},\omega|\mathbf{u}^{(i)}) > \delta) - m)^2, \tag{4.13}$$

$$\mathbf{u}^{(1)} = \{f(\mathbf{x}) = \mathbb{E}(f(\mathbf{x})|\mathcal{D}) + \sqrt{2\,\mathrm{var}(f(\mathbf{x})|\mathcal{D})},\ g(\mathbf{x}) = \mathbb{E}(g(\mathbf{x})|\mathcal{D})\}$$

$$\mathbf{u}^{(2)} = \{f(\mathbf{x}) = \mathbb{E}(f(\mathbf{x})|\mathcal{D}) - \sqrt{2\,\mathrm{var}(f(\mathbf{x})|\mathcal{D})},\ g(\mathbf{x}) = \mathbb{E}(g(\mathbf{x})|\mathcal{D})\}$$

$$\mathbf{u}^{(3)} = \{f(\mathbf{x}) = \mathbb{E}(f(\mathbf{x})|\mathcal{D}),\ g(\mathbf{x}) = \mathbb{E}(g(\mathbf{x})|\mathcal{D}) + \sqrt{2\,\mathrm{var}(g(\mathbf{x})|\mathcal{D})}\}$$

$$\mathbf{u}^{(4)} = \{f(\mathbf{x}) = \mathbb{E}(f(\mathbf{x})|\mathcal{D}),\ g(\mathbf{x}) = \mathbb{E}(g(\mathbf{x})|\mathcal{D}) - \sqrt{2\,\mathrm{var}(g(\mathbf{x})|\mathcal{D})}\}, \tag{4.14}$$

where the quadrature points (4.14) and the corresponding weights 1/4 in (4.12) and (4.13) are selected for third-order accuracy of the scheme (see Appendix B.3). With (4.12)-(4.14) to compute the operator $\mathrm{std}_{f,g}$, (4.10) can be directly solved using standard optimization methods, say a multiple-starting L-BFGS-B quasi-Newton method [100].

We remark that the construction of acquisition function has been studied extensively in the case of deterministic ItR [31, 59, 163, 149, 15], and that there may still be room for improvement relative to (4.10) in the case of stochastic ItR. These potential improvements of acquisition function may generally take consideration of correlation between different $\mathbf{x}$ in addition to the standard deviation in (4.10). For example, techniques developed for deterministic ItR, such as using a hypothetical point [33, 105] (also see §3.2.3) and global sensitivity analysis [59], may be transferred here.

However, they may lead to significantly increased computational cost when combined with VHGPR (e.g., the re-training of the variational parameters when hypothetical points are used). These potential developments will be left to our future work.

Combining the VHGPR surrogate model ((4.6), (4.7), and (4.8)) and the optimization of acquisition function (4.10), we are able to sequentially select the next-best samples starting from an initial dataset. The final estimation of exceeding probability P_e is computed by VHGPR with predicted means to represent functions f and g:

$$P_e = \mathbb{P}\Big(S\big(X, \omega | f(\mathbf{x}) = \mathbb{E}(f(\mathbf{x})|\mathcal{D}), g(\mathbf{x}) = \mathbb{E}(g(\mathbf{x})|\mathcal{D})\big) > \delta\Big). \tag{4.15}$$

We summarize the algorithm of this sequential BED process in Algorithm **3**.

Algorithm 3 Sequential experimental design for systems with stochastic ItR

Require: Number of initial points n_{init}, number of iterations n_{iter}
Input: Initial dataset $\mathcal{D} = \{\mathbf{x}^i, y^i\}_{i=1}^{n_{init}}$
   Initialization $n = n_{init}$
   while $n < n_{init} + n_{seq}$ **do**
      Train the surrogate model ((4.6), (4.7), and (4.8)) with $\mathcal{D}$
      Maximize the acquisition function (4.10) to find the next best sample $\mathbf{x}^{n+1}$
      Implement numerical simulation to get $y^{n+1} = S(\mathbf{x}^{n+1}, \omega)$
      Update the dataset $\mathcal{D} = \mathcal{D} \cup \{\mathbf{x}^{n+1}, y^{n+1}\}$
      $n = n + 1$
   end while
Output: Compute the exceeding probability (4.15) based on the surrogate model

---

4.3 Results

In this section, we validate our approach using two synthetic problems and a realistic engineering application to quantify the extreme ship motion probability in irregular waves. The heteroscedastic randomness in the ItR are assigned artificially in the former cases, while resulted naturally from dimension reduction of the input parameter space in the latter case. For all cases, we present the results from our current

method of sequential BED with VHGPR as a surrogate model (Seq-VHGPR), as well as other methods for validation and comparison. These other methods include Monte Carlo sampling using one million samples for accurate estimation of the mean and variance of ItR (Exact-MC, which serves as the exact result to validate Seq-VHGPR); space-filling Latin hypercube (LH) sampling [89] with VHGPR as a surrogate (LH-VHGPR, which serves as a reference to demonstrate the efficiency of sequential BED); LH sampling with SGPR as a surrogate (LH-SGPR, to demonstrate the necessity of using VHGPR). We also include the asymptotic value obtained from the LH-SGPR method, i.e., the convergent result with sufficiently large number of samples. This represents the best solution that can be achieved by previous vast methods based on SGPR with constant uncertainties [59, 31, 15, 135].

4.3.1 Quadratic function

We start from a 1D synthetic problem, where the true ItR $S(x, \omega)$ (4.3) is constructed with (see figure 4.1 for an illustration)

$$f(x) = (x - 5)^2, \tag{4.16}$$

and $R(x, \omega) \sim \mathcal{N}(0, \gamma^2(x))$ with

$$\gamma(x) = 0.1 + 0.1x^2. \tag{4.17}$$

We note that $\gamma^2(x)$ is the function that we aim to approximate through $e^{g(x)}$ (in (4.7)) in VHGPR. The input X is assumed to follow a Gaussian distribution with $p_x(x) = \mathcal{N}(5, 1)$. Our objective is to estimate an exceeding probability $P_e = \mathbb{P}(S(X, \omega) > 9)$. For Seq-VHGPR, we use 40 LH samples as the initial data set, and show the results after 40 initial samples along with other methods.

Figure 4.2 plots the results P_e computed by Exact-MC, Seq-VHGPR, LH-VHGPR
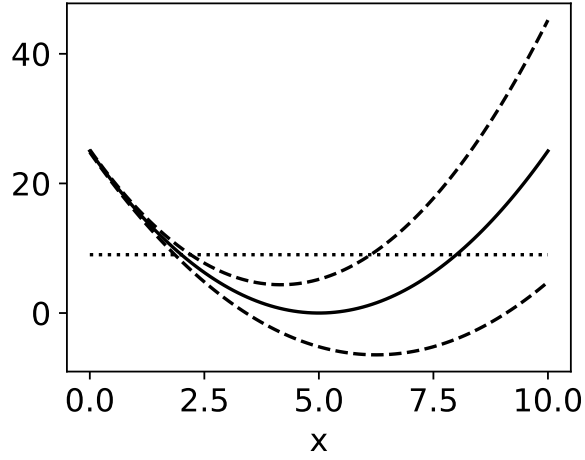
Figure 4.1: The mean $f(x)$ (—) and uncertainty bounds $f(x) \pm 2\gamma(x)$ (---) of the 1D ItR, as well as the threshold (·····) in defining the exceeding probability.
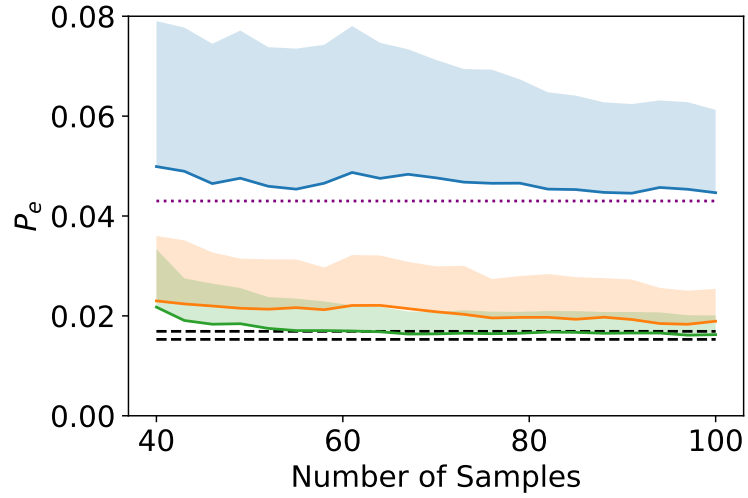
Figure 4.2: P_e in the 1D synthetic problem, computed by Seq-VHGPR(—), LH-VHGPR(—), LH-SGPR(—) with its asymptotic value (·····), Exact-MC(---) (in terms of the upper and lower 5% error bounds). The shaded region represents one standard deviation above the mean estimated from 100 applications of the each method.
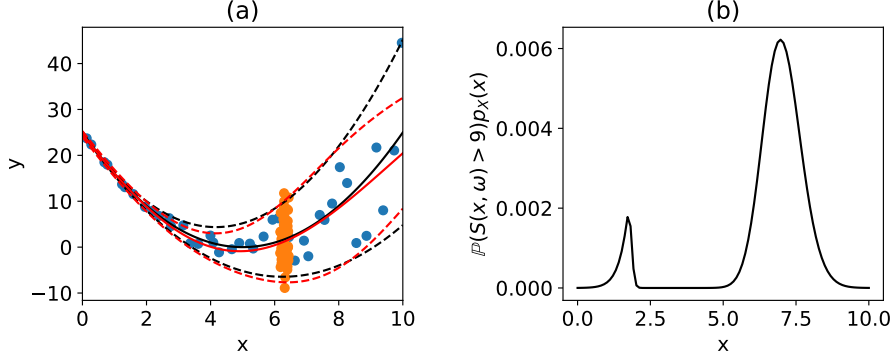
Figure 4.3: (a) Typical positions of initial 40 samples (•) and 60 sequential samples (•) in Seq-VHGPR, as well as the learned function $f(x)$(—) and $f(x) \pm 2\gamma(x)$(- - -) compared to the corresponding exact functions (—, - - -); (b) the function $\mathbb{P}(S(x, \omega) > \delta)p_x(x)$.

and LH-SGPR (where in Seq-VHGPR and LH-VHGPR, P_e is estimated by (4.15); in LH-SGPR, P_e is estimated by (4.15) with constant variance $g(x) = \log(\gamma_0^2)$). Also included in figure 4.2 are the standard deviations in Seq-VHGPR, LH-VHGPR and LH-SGPR obtained from 100 applications of the methods. (These uncertainties come from the initial sampling positions and the randomness ω of ItR in computing $S(x, \omega)$ for each query.) We see that the result from Seq-VHGPR converges rapidly to that from Exact-MC (shown in terms of the 5%-error region) in the first 20 sequential samples, with an accurate estimation of the exceeding probability. In contrast, the LH-VHGPR result converges much slower, with a non-negligible difference from the Exact-MC result at the end of 100 samples in figure 4.2 (in spite of a later convergence with about 200 samples confirmed in our test). Furthermore, the LH-SGPR result converges to an asymptotic value which is 3 times of the Exact-MC result, showing the incapability of this class of methods (i.e., most previous methods using SGPR) in estimating the exceeding probability induced by a heteroscedastic ItR. We remark that the failure of the SGPR-based methods lie in the loss of heteroscedasticity information in ItR, irrespective of the sampling approach or acquisitions used. Finally, as shown by the shaded area in figure 4.2, Seq-VHGPR leads to significantly reduced

standard deviation compared to other approaches.

We further examine the reason for the fast convergence achieved by Seq-VHGPR (relative to LH-VHGPR). Figure 4.3(a) plots the positions of 100 samples (i.e., 40 initial and 60 sequential samples) in Seq-VHGPR, as well as the learned functions $f(x)$ and $\gamma(x)$. While the initial 40 samples are randomly chosen (providing the overall trend of $f(x)$ and $\gamma(x)$), the 60 sequential samples are concentrated near $x = 6.5$, providing more accurate estimation of $f(x)$ and $\gamma(x)$ in the nearby region. This point corresponds to the maximum in $\mathbb{P}(S(x,\omega) > \delta)p_x(x)$ (the integrand in (4.2)) as shown in figure 4.3(b), leading to the largest contribution in computing the exceeding probability (4.2).

4.3.2 Four-branch function

We construct a 2D synthetic problem by setting $f(\mathbf{x})$ to be a four-branch function (that has been widely-used in estimating extreme-event probability with a deterministic ItR) [31, 59, 135, 149]:

$$f(x_1, x_2) = -\min \begin{cases} 8 + 0.1(x_1 - x_2)^2 + \dfrac{(x_1 + x_2)}{\sqrt{2}} \\ 8 + 0.1(x_1 - x_2)^2 - \dfrac{(x_1 + x_2)}{\sqrt{2}} \\ (x_1 - x_2) + \dfrac{6}{\sqrt{2}} + 5 \\ (x_2 - x_1) + \dfrac{6}{\sqrt{2}} + 5 \end{cases} .$$

To generate an uncertain ItR, we add a Gaussian randomness $R(x_1, x_2, \omega) \sim \mathcal{N}(0, \gamma^2(x_1, x_2))$ to $f(x_1, x_2)$ with standard deviation $\gamma(x_1, x_2) = f(x_1, x_2)/6$ (see figure 4.4 for $f(x_1, x_2)$ and $\gamma(x_1, x_2)$). We assume the input X to follow a Gaussian distribution $p_{X_1 X_2}(x_1, x_2) = \mathcal{N}(\mathbf{0}, \mathrm{I})$, with I being a 2×2 identity matrix, and our purpose is to estimate an exceeding probability $P_e = \mathbb{P}(S(X_1, X_2, \omega) > 5)$. For
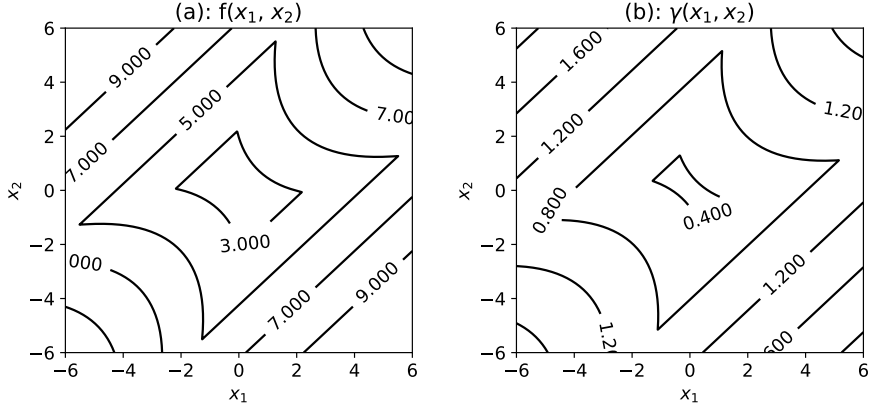
Figure 4.4: (a) $f(x_1, x_2)$ and (b) $\gamma(x_1, x_2)$ as in $R(x_1, x_2, \omega)$ in the 2D stochastic four-branch ItR.

Seq-VHGPR, 60 LH samples are used as the initial data set.

The results of the 2D problem, as shown in figure 4.5, further demonstrates the effectiveness of Seq-VHGPR, which approaches the Exact-MC solution with 5% error within the first 20 sequential samples and leads to the smallest uncertainty among all methods. The convergence rate of Seq-VHGPR is much faster than that of LH-VHGPR, where the latter fails to converge at the end of 120 samples. Compared with 1D results, the superiority of Seq-VHGPR over LH-VHGPR is more evident due to the increased sparsity of samples in the 2D case. The LH-SGPR result, on the other hand, converges to an asymptotic value which is 2.5 times of the Exact-MC result, a significant error due to the neglect of heteroscedastic randomness in ItR.

The typical positions of (60 initial and 60 sequential) samples in Seq-VHGPR are shown in figure 4.6(a)(b), as well as the the learned functions $f(x_1, x_2)$ and $\gamma(x_1, x_2)$. Similar to the 1D case, the sequential samples are expected to concentrate in regions where $\mathbb{P}(S(x_1, x_2, \omega) > \delta)p_{X_1 X_2}(x_1, x_2)$ is maximized, i.e., the four regions enclosed by -4 contour lines in figure 4.6(c). As shown in figure 4.6(a)(b), most sequential samples lie in three out of the four regions (although the situation depends on the initial samples and for some cases all four regions can be filled). The difficulty of the sequential samples transiting to all four regions within 60 samples can be anticipated,

65
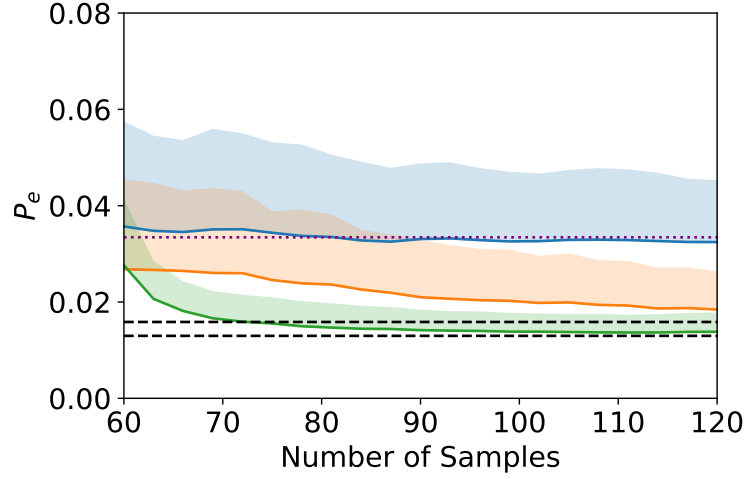
Figure 4.5: P_e in the 2D synthetic problem, computed by Seq-VHGPR(—), LH-VHGPR(—), LH-SGPR(—) with its asymptotic value (·····), Exact-MC(- - -) (in terms of the upper and lower 5% error bounds). The shaded region represents one standard deviation above the mean estimated from 100 applications of the each method.
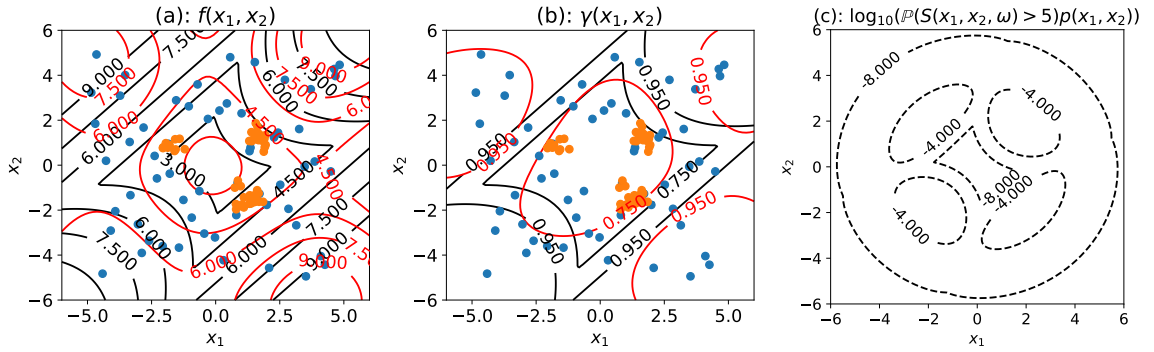
Figure 4.6: Typical positions of initial 60 samples (•) and 60 sequential samples (•) in Seq-VHGPR, as well as the learned $f(x_1, x_2)$ (—) compared to the exact function (—) in (a); and the learned $\gamma(x_1, x_2)$ (—) compared to the exact function (—) in (b); (c) the (log) function $\mathbb{P}(S(x_1, x_2, \omega) > \delta)p_{X_1 X_2}(x_1, x_2)$.

which is consistent with the observation in the case of deterministic ItR if the U criterion is used as the acquisition function [31]. While the design of better acquisition function is possible referring to the counterpart in the deterministic case [59], the current results already show the adequacy of Seq-VHGPR in estimating the exceeding probability (even if not all four regions are filled and the estimation of $\gamma(x_1, x_2)$ is relatively less accurate than that of $f(x_1, x_2)$).

We also take the opportunity of this example to show the computational efficiency of our algorithm. Using one core of Intel Xeon Gold 6154 with 2GB memory, we record the computation time of selecting one sequential sample with 50 existing samples in the dataset, with the total time splitted into (1) the training of the VHGPR and (2) optimization of the acquisition function. The two parts take respectively 1.77s and 2.91s on average. These times are generally negligible given the expensive system evaluations. In addition, since the cost of training VHGPR is comparable to that of SGPR (Appendix B.1), the computational cost of our full algorithm is similar to previous methods based on SGPR.

4.3.3 Extreme-event probability of ship motion in irregular waves

We further consider an engineering application of our method to estimate the probability of extreme ship roll motions in uni-directional irregular waves. In marine engineering, the ship motion problem can often be treated as a dynamical system where the input is a time series of wave (or surface) elevation $\eta(t)$, and the output is, say, the ship roll motion $\xi(t)$. The ItR connecting $\eta(t)$ and $\xi(t)$ can be computed by Computational Fluid Dynamics (CFD) simulations. However, the resolution of exact exceeding probability requires running expensive CFD simulations with a very long-time input $\eta(t)$ (due to the rareness of the extreme roll motion), leading to prohibitively high computational cost. Therefore, for the purpose of validating our
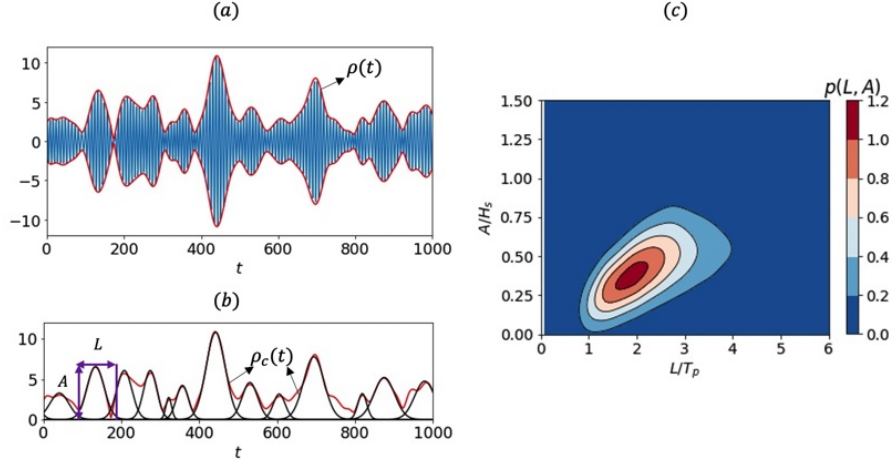
Figure 4.7: (a) surface elevation $\eta(t)$ (—) and the corresponding envelope process $\rho(t)$ (—) in a narrow-band wave field. (b) $\rho(t)$ (—) fitted by an ensemble of Gaussian wave groups $\rho_c(t)$(—) with parameters L and A. (c) $p_{LA}(L, A)$ obtained from wave fields.

approach, we use an inexpensive phenomenological nonlinear roll equation [142] to construct the ItR (with the uncertainty associated with ω introduced later)

$$\ddot{\xi} + \alpha_1 \dot{\xi} + \alpha_2 \dot{\xi}^3 + (\beta_1 + \epsilon_1 \cos(\chi)\eta(t))\xi + \beta_2 \xi^3 = \epsilon_2 \sin(\chi)\eta(t), \qquad (4.18)$$

with empirical coefficients [98] $\alpha_1 = 0.19$, $\alpha_2 = 0.06$, $\beta_1 = 0.04$, $\beta_2 = -0.1$, $\epsilon_1 = 0.020$, $\epsilon_2 = 0.004$, and $\chi = \pi/6$.

The wave elevation $\eta(t)$ is usually specified from a wave spectral process, which resides in a high-dimensional input space. A typical procedure to reduce the dimension is to describe $\eta(t)$ by an ensemble of wave groups embedded in its envelope process as in §2.2.1 (see figure 4.7(a) for an illustration). Specifically, we compute the envelope process $\rho(t)$ from $\eta(t)$ through the Hilbert transform [127], and then construct two-parameter Gaussian-like wave groups $\rho_c(x)$ which best fits $\rho(t)$:

$$\rho_c(t) \sim A \exp \frac{-(t - t_c)^2}{2L^2}, \qquad (4.19)$$

68

where t_c is the temporal location of the group, and the two parameters A (group amplitude) and L (group length) describe the geometry of the group (figure 4.7(b)). This dimension-reduction procedure allows $\eta(t)$ to be described by an ensemble of (L, A) wave groups, i.e., a two-dimensional input parameter space (see figure 4.7(c)). We can then construct an ItR with the input as (L, A) to (4.18) and the output as the maximum roll through this wave group, and consider the group-based probability.

However, the dimension reduction results in the loss of information relative to the original field $\eta(t)$, i.e., it introduces uncertainties in the ItR, including the uncertain initial conditions of $(\xi(0), \dot{\xi}(0))$ and detailed phase and frequency conditions in the wave group. As shown in [2, 130], (4.18) (and the ship roll in general) may be sensitive to the lost information such as initial conditions, and the resulted uncertainty is non-uniform for different A and L. (see figure 4.8 as an example that the uncertainty is larger for the first wave group but smaller for the second one). This creates heteroscedastic uncertainty in the ItR (associated with ω) that needs to be dealt with by our current approach Seq-VHGPR.

In the following, we show the results with input $\eta(t)$ extracted from a narrow-band Gaussian wave spectrum:

$$F(\omega) = \frac{H_s^2}{16} \frac{1}{\sqrt{2\pi}d} \exp(\frac{-(\omega - \omega_0)^2}{2d^2}), \tag{4.20}$$

with ω the angular frequency, the significant wave height $H_s = 12m$, peak (carrier) wave frequency $\omega_0 = 0.067s^{-1}$ (corresponding to peak period $T_p = 15s$), and $d = 0.02s^{-1}$. In particular, we compute $\eta(t) = \sum_n a_n \cos(n\Delta\omega t + \phi_n)$ with $\Delta\omega = 0.00026s^{-1}$, n from 1 to 1024, ϕ_n being a random phase for each mode, and $a_n = \sqrt{2F(n\Delta\omega)\Delta\omega}$.

The exact exceeding probability is computed by simulating 1500 hours (360000 T_p) of ship responses. To compute the ItR incorporating the heteroscedastic randomness
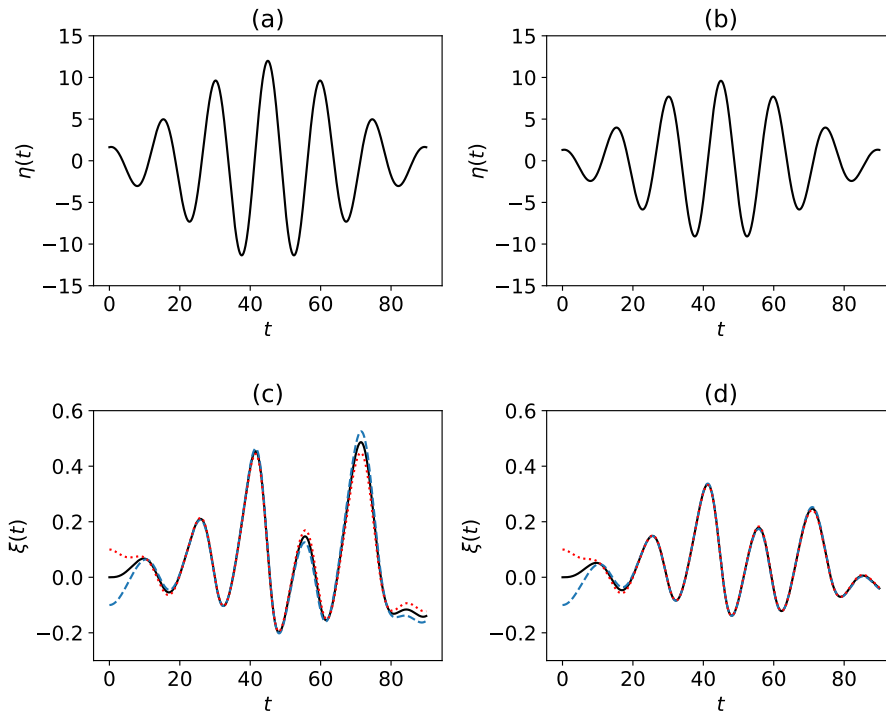
Figure 4.8: The roll responses (c) and (d) with different initial conditions $\{\dot{\xi}(0) = 0, \xi(0) = 0\}$ (——); $\{\dot{\xi}(0) = 0, \xi(0) = -0.05\}$ (- - -); $\{\dot{\xi}(0) = 0, \xi(0) = 0.05\}$ (·····), computed from (4.18) with input from a wave group with respectively (a) larger and (b) smaller amplitudes.

in ω, after a sample (L, A) is chosen, we randomly select a wave group of this (L, A) in $\rho(x)$, and simulate (4.18) starting from (on average) 3 groups ahead of the (L, A) group with a $(0, 0)$ initial condition. Since the impact of initial conditions typically decay in $O(1)$ wave group, we are able to naturally capture the true initial condition as the ship encounters the (L, A) group, as well as the phase and frequency condition in the particular group. We name this method as 'natural initial condition' and systematically compare it with other methods to incorporate the ship initial conditions in [44].

Figure 4.9 plots $P_e = \mathbb{P}(\max(\xi_{L,A}(t)) > 0.3)$ (the probability that maximum roll in a group exceeds 17 degrees) obtained from Seq-VHGPR, LH-VHGPR, LH-SGPR and the exact solution. We see that the Seq-VHGPR result converges to the exact solution within the first 30 sequential samples, much faster than the convergence of the LH-VHGPR result. The LH-SGPR result converges to an asymptotic value which is appreciably larger than the exact solution. We remark that this value represents the best result that can be achieved by all previous methods on this problem [95, 50] based on SGPR. These results, again, demonstrate the effectiveness of Seq-VHGPR in computing the exceeding probability relative to all other approaches.

Finally, the application of VHGPR in our problem assumes that the distribution of $S(\mathbf{x}, \omega)$ is approximately Gaussian (associated with ω for each $\mathbf{x} \equiv (A, L)$). While this cannot be checked in the Seq-VHGPR sampling, we provide a posterior calculation to show that this is indeed true. Figure 4.10 plots the distribution of $S(\mathbf{x}, \omega)$ for two selected values of $(A = H_s, L = 1.5T_p)$ and $(A = 0.8H_s, L = 1.5T_p)$, generated from all such groups in the time series of five million T_p. It is evident that the distributions are approximated by Gaussian distributions.
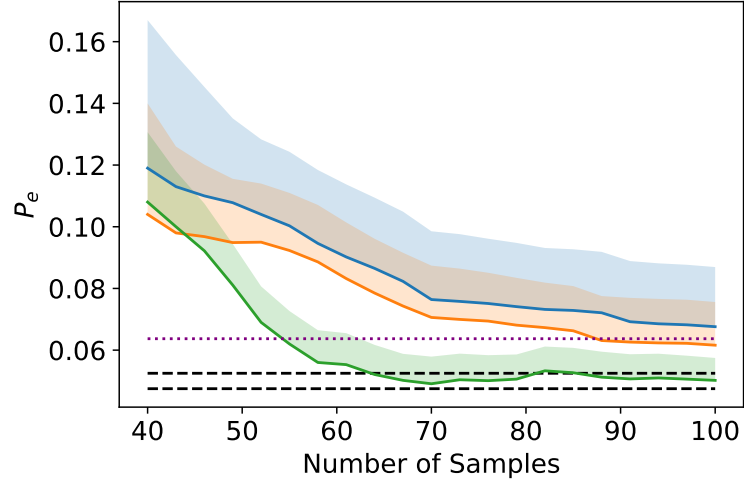
Figure 4.9: P_e in the ship roll problem, computed by Seq-VHGPR(—), LH-VHGPR(—), LH-SGPR(—) with its asymptotic value (⋯⋯), exact solution(- - -) (in terms of the upper and lower 5% error bounds). The shaded region represents one standard deviation above the mean estimated from 50 applications of the each method.
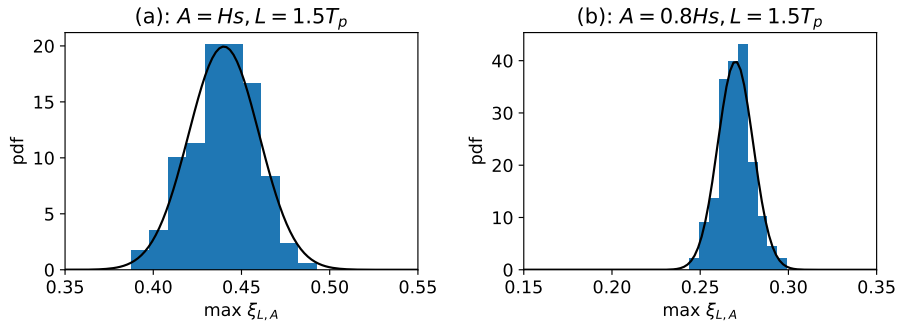
Figure 4.10: The density histograms of $S(A, L, \omega)$ for (a) $A = H_s, L = 1.5T_p$ and (b) $A = 0.8H_s, L = 1.5T_p$, generated from all such groups in the time series of five million T_p. The Gaussian fits for the histograms are shown (—).

72

4.4   Summary

In this chapter, we present a new method (Seq-VHGPR) to efficiently estimate the extreme-event probability (in terms of the exceeding probability) induced by an ItR with heteroscedastic uncertainty. The method is established in the framework of sequential Bayesian experimental design, and leverages the VHGPR as a surrogate model to estimate the uncertain ItR. A new acquisition function corresponding to the VHGPR estimation is developed to select the next-best sequential sample which leads to fast convergence of the exceeding probability. We validate our new method in two synthetic problems and one engineering application to estimate the extreme ship motion probability in irregular waves. In all cases, we find fast convergence of Seq-VHGPR to the exact solution, demonstrating its superiority to all existing methods if an ItR with heteroscedastic uncertainty is associated with the problem. This is indeed due to the effectiveness of VHGPR in estimating the ItR, although there is still room for improvement of the acquisition function to accelerate the convergence, which will be introduced in the case of deterministic ItR in §6. Finally, we remark that the present method also provides an effective way for high-dimensional BED, where the most influential dimensions can be selected as (low-dimensional) input X, with other secondary ones packaged into Ω in the ItR (as in the ship motion problem).

<div align="center">

CHAPTER 5

Multi-Fidelity Bayesian Experimental Design for Rare-Event Statistics

</div>

5.1   Introduction

In previous chapters, we have built different Bayesian experimental design (BED) frameworks for various objectives. In this chapter, we go back to the original framework in §2 for the tail part of the response PDF (where a new acquisition function has been proposed in [117, 19]) but consider an additional technique to reduce the computational cost: leveraging lower-fidelity models that calculate each response in a (small) fraction of cost (e.g., computational time or budget) of the high-fidelity model. Examples of such lower-fidelity models (as approximations to the high-fidelity counterparts) include (1) analytical models or numerical simulations as approximations to expensive physical experiments [107, 82]; (2) coarse-grid computational fluid dynamics (CFD) simulations as approximations to fine-grid CFD simulations [162]; and (3) Reynolds-averaged Navier–Stokes models as approximations to large eddy simulations for turbulent flows [152]. While the idea of using multi-fidelity information has been widely studied in standard Monte Carlo approach in the form of control variates [104] or multi-level Monte Carlo [41], here we restrict our discussion to the active learning side. Making use of the multi-fidelity Gaussian process [68, 107] or neural network [92, 91] as surrogate models, multi-fidelity sampling algorithms have been developed

<div align="center">

74

</div>

for the purpose of global optimization [131, 119], function learning [91], and contour detection [85]. In terms of the rare or extreme-event statistics, the only related work within the multi-fidelity framework, to our knowledge, is [158] which estimates the exceeding probability for reliability analysis. However, [158] employs the sub-optimal acquisition function which has been shown to be not only (much) less efficient than many improved algorithms later, but also not applicable to our purpose of obtaining the overall rare-event portion of the response PDF.

One of the key issues involved in the multi-fidelity sampling method is the determination of the fidelity level of each sample. Two types of methods have been considered regarding this issue, one to select the fidelity of the next sample adaptively based on existing samples in order to reduce the overall computational cost (though in a heuristic manner) [85, 82], and the other to follow a pre-defined fidelity hierarchy (say pre-defined ratio and sequence of high and low fidelity samples in a bi-fidelity context) [119, 162]. While the former type of method is developed in the hope of outperforming the latter one, there is not sufficient evidence to support the idea as systematic comparisons between the two types of methods are not available. It is one purpose of the current study, in the context of capturing the rare-event statistics of response PDF, to systematically compare these two types of methods in determining the fidelity level of samples, along with the identification of their improvements relative to the single-fidelity algorithm [21].

In this chapter, we develop a multi-fidelity sequential BED framework for the quantification of the response PDF of an ItR system, with emphasis on rare events. In particular, we use a multi-fidelity Gaussian process as the surrogate model, and develop an acquisition function (as a substantial extension to the single-fidelity acquisition function [21]) which allows adaptive choice of both the location in the parameter space and fidelity level of the next sample. We also construct an analytical computation of the acquisition which avoids expensive numerical integration and enables

high-dimensional implementation of the algorithm through gradient-based optimization. Our new method is mainly tested in a bi-fidelity context for a series of synthetic problems with varying dimensions, low-fidelity model accuracy and computational costs in a systematic way. We show that our bi-fidelity method outperforms the single-fidelity method in all test cases, and that our method for adaptive choice of fidelity level consistently performs among the best in all bi-fidelity runs with pre-defined fidelity hierarchy varying in a broad range. Finally, we demonstrate the coupling of our method with CFD to compute the PDF of rare-event ship roll motion in irregular ocean waves. By using CFD simulations with two different grid resolutions as high and low-fidelity models, we show that our bi-fidelity method achieves much faster convergence of the result (i.e., PDF of rare-event responses) than the previous single-fidelity method.

This chapter is adapted from [48]. The Python code for the proposed algorithm, named MFGPextreme, is available on Github[1].

5.2   Method

5.2.1   Problem setup

We consider a black-box ItR function $f(\mathbf{x}) : \mathbb{R}^d \to \mathbb{R}$, with $\mathbf{x} \in \mathbb{R}^d$ representing the input parameters with known probability distribution $p_{\mathbf{x}}(\mathbf{x})$. We assume that we have a hierarchy of models $f_i(\mathbf{x}), i \in \{1, \ldots, s\}$ (from low to high fidelity) to compute $f(\mathbf{x})$, with $f_s(\mathbf{x}) = f(\mathbf{x})$ and $f_i(\mathbf{x})$ having increasing deviations from $f(\mathbf{x})$ for $i = s - 1, s - 2, ..., 1$. In addition, the models $f_i(\mathbf{x})$ are associated with fixed computational costs c_i which increases for $i = 1, 2, ..., s$. The evaluation using these models can be conducted at given $\mathbf{x} = \mathbf{x}_j$ at fidelity level i, leading to an observation

---

[1]https://github.com/umbrellagong/MFGPextreme

$y = f_i(\mathbf{x}_j).$[2]

Our quantity of interest is the PDF of the response $p_f(f)$, focusing on the tail part. Specifically, we aim to obtain an estimation $p_{f,est}(f)$ with minimized error (see [95])

$$e = \int \Big| \log p_{f,est}(f) - \log p_f(f) \Big| \mathrm{d}f. \tag{5.1}$$

We note that the log function in (5.1) acts on the ratio $p_{f,est}(f)/p_f(f)$, which is amplified when $p_f(f)$ is small, i.e., (5.1) emphasizes on the error in the small-probability rare event portion of the PDF. We remark that while in many cases, rare events in the responses coincide with extreme events (or events of extremely large responses), these two concepts are not equivalent in general. An error metric for extreme events, in contrast to (5.1) for rare events, can be formulated through high-order moments of $p_f(f)$ as studied in §2.2.2.

To compute $p_{f,est}(f)$, we can use a sequence of samples $f_i(\mathbf{x})$ with i and $\mathbf{x}$ varying for each sample. Our objective is to find an optimized sequence, in terms of both i and $\mathbf{x}$, such that e is minimized under a given total computational cost c (i.e., summation of computational cost c_i over all members in the sequence). In general, there is no computational approach for this type of problem that can be guaranteed to provide a globally optimal solution, and the method we propose in this paper should be considered as a greedy algorithm that looks one step ahead of the existing samples. In particular, our method is based on a multi-fidelity sequential BED, which involves two basic components: (1) an inexpensive surrogate model based on the multi-fidelity Gaussian process trained by results from multi-fidelity samples; (2) a new acquisitive function measuring the benefit (i.e., reduction in e) per computational cost. The next-best sample can then be selected in terms of both i and $\mathbf{x}$ to maximize the

---

[2]In general, one can also associate the observation y with a Gaussian noise, say from the measurement error if f_i is a physical experiment. In this paper, we consider the case of zero noise since f_i in all examples represent numerical simulations, but our methodology can be naturally extended to the case with finite Gaussian noise by including additional hyperparameters in constructing the surrogate model.

acquisition function. The two components are next described in detail in §5.2.2 and §5.2.3. In addition, in §5.2.4, we develop an analytical formula to compute the acquisition function and its derivative with respect to $\mathbf{x}$, enabling the gradient-based optimization that is suitable for high-dimensional problems.

5.2.2 Surrogate model

In this section, we briefly outline the multi-fidelity Gaussian process developed in [68] as our surrogate model. Assume we have a dataset $\mathcal{D} = \{\mathcal{X}, \mathcal{Y}\}$ consisting of s levels of model outputs $\mathcal{Y} = \{\mathcal{Y}_i\}_{i=1}^s$ at input positions $\mathcal{X} = \{\mathcal{X}_i\}_{i=1}^s$ sorted by increasing fidelity. Here $\mathcal{X}_i \in \mathbb{R}^{n_i \times d}$, $\mathcal{Y}_i \in \mathbb{R}^{n_i}$, $\mathcal{X} \in \mathbb{R}^{m \times d}$, $\mathcal{Y} \in \mathbb{R}^m$, with n_i the number of samples available at fidelity level i, d the dimension of input vectors, $m = \sum_{i=1}^s n_i$ the total number of samples. In other words, $\mathcal{X}_i$ contains n_i input vectors, each with d dimensions, evaluated by models of fidelity level i with n_i outputs collected in $\mathcal{Y}_i$. In general, input vectors in $\mathcal{X}_i$ are different for different fidelity levels. $\mathcal{D} = \{\mathcal{X}, \mathcal{Y}\}$ collects all samples available at all fidelity levels. The purpose of the multi-fidelity Gaussian process is to learn the underlying relation $f_i(\mathbf{x})$ from $\mathcal{D}$. This can be achieved through an auto-regressive scheme, which models $f_i(\mathbf{x})$ as

$$f_i(\mathbf{x}) = \rho_{i-1} f_{i-1}(\mathbf{x}) + d_i(\mathbf{x}) \qquad i = 2, \ldots, s, \tag{5.2}$$

with $f_1(\mathbf{x}) \sim \mathcal{GP}(0, k_1(\mathbf{x}, \mathbf{x}'))$ and $\{d_i(\mathbf{x}) \sim \mathcal{GP}(0, k_i(\mathbf{x}, \mathbf{x}'))\}_{i=2}^s$ pairwise independent Gaussian processes, ρ_{i-1} a scaling factor to quantify the correlation between f_i and f_{i-1}. The kernels $k_i(\mathbf{x}, \mathbf{x}')$ are defined as radial-basis functions

$$k_i(\mathbf{x}, \mathbf{x}') = \tau_i^2 \exp\left(-\frac{1}{2}(\mathbf{x} - \mathbf{x}')^T \Lambda_i^{-1}(\mathbf{x} - \mathbf{x}')\right), \tag{5.3}$$

with τ_i and the diagonal matrix Λ_i respectively representing the characteristic amplitude and length scales. $\{\tau_i, \Lambda_i\}_{i=1}^s$ and $\{\rho_i\}_{i=1}^{s-1}$ are hyperparameters in the model

and can be determined by maximizing the likelihood function $p(\mathbf{Y} = \mathcal{Y})$, where $\mathbf{Y}$ is a random vector of the surrogate with input at $\mathcal{X}$, satisfying a Gaussian distribution $\mathcal{N}(\mathbf{0}, \text{cov}(\mathbf{Y}))$. Here we apply the shorthand notation $\text{cov}(\mathbf{Y}) \equiv \text{cov}(\mathbf{Y}, \mathbf{Y}) \in \mathbb{R}^{m \times m}$ to represent a covariance matrix for each pairwise random variables in $\mathbf{Y} \in \mathbb{R}^m$, which will be used throughout this chapter.

The posterior prediction $f_i(\mathbf{x})$ given the dataset $\mathcal{D}$ can then be derived as a Gaussian process

$$f_i(\mathbf{x})|\mathcal{D} \sim \mathcal{GP}\big(\mathbb{E}(f_i(\mathbf{x})|\mathcal{D}), \text{cov}(f_i(\mathbf{x}), f_i(\mathbf{x}')|\mathcal{D})\big), \quad i = 1, 2, ..., s \tag{5.4}$$

with analytically tractable mean and covariance (also defined across different fidelity levels)

$$\mathbb{E}(f_i(\mathbf{x})|\mathcal{D}) = \text{cov}(f_i(\mathbf{x}), \mathbf{Y})\text{cov}(\mathbf{Y})^{-1}\mathcal{Y}, \tag{5.5}$$

$$\text{cov}\big(f_i(\mathbf{x}), f_j(\mathbf{x}')|\mathcal{D}\big) = \text{cov}\big(f_i(\mathbf{x}), f_j(\mathbf{x}')\big) - \text{cov}\big(f_i(\mathbf{x}), \mathbf{Y}\big)\text{cov}(\mathbf{Y})^{-1}\text{cov}\big(\mathbf{Y}, f_j(\mathbf{x}')\big). \tag{5.6}$$

In (5.5) and (5.6) (as well as the likelihood function), the covariances are computed as (or can be derived from):

$$\text{cov}(f_i(\mathbf{x}), f_j(\mathbf{x}')) = \sum_{l=1}^{\min(i,j)} \pi_{ijl} k_l(\mathbf{x}, \mathbf{x}'), \tag{5.7}$$

where

$$\pi_{ijl} = \begin{cases} (\prod_{t=l}^{i-1} \rho_t)(\prod_{t=l}^{j-1} \rho_t) & l \neq \min(i,j), \\ \prod_{t=\min(i,j)}^{\max(i,j)-1} \rho_t & l = \min(i,j), i \neq j, \\ 1 & l = \min(i,j), i = j. \end{cases} \tag{5.8}$$

We finally summarize the bi-fidelity counterpart of (5.4) in Appendix C.1, which we will use in §5.3 for computation.

5.2.3 Acquisition function

Given the Gaussian process surrogate $f_s(\mathbf{x})|\mathcal{D}$ as in (5.4) of the ItR, we can estimate the response PDF $p_{f|\mathcal{D}}(f)$. Our purpose is to select the next sample in terms of the fidelity level i and the location $\tilde{\mathbf{x}}$ to significantly reduce the uncertainty in the rare-event part of the response PDF (which is expected to lead to significantly smaller e in (5.1)). In particular, this uncertainty can be estimated by (see previous work in the single-fidelity context [95])

$$U(\mathcal{D}, i, \tilde{\mathbf{x}}) = \int |\log p_{f^+|\mathcal{D}, \overline{f}_i(\tilde{\mathbf{x}})}(f) - \log p_{f^-|\mathcal{D}, \overline{f}_i(\tilde{\mathbf{x}})}(f)| \mathrm{d}f, \qquad (5.9)$$

where, $\overline{f}_i = \mathbb{E}(f_i(\tilde{\mathbf{x}})|\mathcal{D})$ is the mean response computed by the surrogate $f_i(\mathbf{x})|\mathcal{D}$ from a hypothetical location $\tilde{\mathbf{x}}$ and fidelity i, $p_{f^\pm|\mathcal{D}, \overline{f}_i(\tilde{\mathbf{x}})}(f)$ are PDF bounds generated by upper and lower bounds (say two standard deviations away from the mean) of $f|\mathcal{D}, \overline{f}_i(\tilde{\mathbf{x}})$.

Using $U(\mathcal{D}, i, \tilde{\mathbf{x}})$ directly as the acquisition, however, involves significant computational cost (e.g., building a new Gaussian process $f|\mathcal{D}, \overline{f}_i(\tilde{\mathbf{x}})$ for each hypothetical sample) even for single-fidelity problems. To address this issue, we extend the methodology developed for single-fidelity problems in [117, 21] to the multi-fidelity context. The first step is to introduce an upper bound as a proxy to U, defined as (proven in [117] for single-fidelity applications)

$$Q(\mathcal{D}, i, \tilde{\mathbf{x}}) = \int \mathrm{var}(f(\mathbf{x})|\mathcal{D}, \overline{f}_i(\tilde{\mathbf{x}}))w(\mathbf{x})\mathrm{d}\mathbf{x}, \qquad (5.10)$$

with

$$w(\mathbf{x}) = \frac{p_{\mathbf{x}}(\mathbf{x})}{p_{\overline{f}}(\overline{f}(\mathbf{x}))}, \qquad (5.11)$$

where $\overline{f}(\mathbf{x}) = \mathbb{E}[f(\mathbf{x})|\mathcal{D}]$ represents the mean prediction (5.5) with $i = s$. Q measures the model uncertainty with emphasis on positions of large w, i.e. small-probability

response regions with significant input probability. We are interested in the reduction in Q (i.e., the benefit) after adding the sample at $\tilde{\mathbf{x}}$ and i, formulated as

$$
\begin{aligned}
B(i, \tilde{\mathbf{x}}) &= Q(\mathcal{D}) - Q(\mathcal{D}, i, \tilde{\mathbf{x}}) \\
&= \int \left(\mathrm{var}(f(\mathbf{x})|\mathcal{D}) - \mathrm{var}(f(\mathbf{x})|\mathcal{D}, \overline{f}_i(\tilde{\mathbf{x}})) \right) w(\mathbf{x}) \mathrm{d}\mathbf{x} \\
&= \frac{1}{\mathrm{var}(f_i(\tilde{\mathbf{x}})|\mathcal{D})} \int \mathrm{cov}^2(f(\mathbf{x}), f_i(\tilde{\mathbf{x}})|\mathcal{D}) w(\mathbf{x}) \mathrm{d}\mathbf{x},
\end{aligned}
\tag{5.12}
$$

where $Q(\mathcal{D})$ computes the value of Q as defined in (5.10) but with the variance on the right hand side conditioning only on $\mathcal{D}$, i.e., before adding an i fidelity sample at $\tilde{\mathbf{x}}$. The derivation of the result in (5.12) is summarized in Appendix C.2, which makes use of the recursive update of the Gaussian process that is simpler than the derivation in [21] for the single-fidelity method. We note that the expensive computations of the new posterior in (5.9) and (5.10) are not involved in (5.12). Further reduction of computational cost by avoiding the numerical integration in (5.12) will be discussed later in §5.2.4.

In general, one may expect that adding a high-fidelity sample is more beneficial than adding a low-fidelity sample at the same $\mathbf{x}$. While this is indeed generally true, we note that there exist some certain special situations in which adding a low-fidelity sample becomes more beneficial according to (5.12). Such situations can occur when the function $d_i(\mathbf{x})$ in (5.2) becomes uncorrelated, with a rigorous justification provided in Appendix C.3. To select the next best sample in terms of both location and fidelity level, we need an acquisition function taking into consideration both the benefit (5.12) and cost of the sample c_i. In particular, we solve an optimization problem

$$
\mathbf{x}^*, i^* = \mathrm{argmax}_{\tilde{\mathbf{x}} \in \mathbb{R}^d, i \in \{1, 2, \dots s\}} B(i, \tilde{\mathbf{x}})/c_i.
\tag{5.13}
$$

We remark that (5.13) provides the optimal next sample in terms of the uncer-

tainty reduction (5.12) per computational cost. Nevertheless, there is no guarantee that successively applying (5.13) provides a globally optimal solution although this type of fidelity-choice algorithm has also been applied for other purposes [85, 82]. Therefore, the ultimate validity of (5.13) needs to be tested in a sufficiently wide range of examples, especially against algorithms with a fixed fidelity hierarchy, which is one of our purposes in §5.3.

In solving (5.13) as a combined discrete and continuous optimization problem, we first find the optimal location $\mathbf{x}$ for each fidelity i, i.e., $\mathbf{x}_i^* = \mathrm{argmax}_{\tilde{\mathbf{x}} \in \mathbb{R}^d} B(i, \tilde{\mathbf{x}})$ for $i = 1, 2, \ldots, s$, then we compare the benefit per cost for each fidelity level and find the optimal fidelity level, i.e., $i^* = \mathrm{argmax}_{i \in \{1,2,\ldots s\}} B(i, \mathbf{x}_i^*)/c_i$.

While the solution procedure outlined above seems straightforward, there is still difficulty in applying the method to high-dimensional problems. The reason lies in that the numerical integration in (5.12) can become prohibitively expensive for high-dimensional $\mathbf{x}$. Furthermore, the high-dimensional optimization (5.13) needs to rely on a gradient-based algorithm where the derivative of (5.12) is also expensive to compute. To address these issues, analytical formulae for (5.12) and its derivative are much preferable, which will be discussed in the next section.

5.2.4 Analytical formulae for acquisition and its derivative

To develop an analytical formula for (5.12), we first substitute the expression of covariance function (5.6) into (5.12) and obtain

$$B(i, \tilde{\mathbf{x}}) = \frac{1}{\mathrm{var}(f_i(\tilde{\mathbf{x}})|\mathcal{D})} \Big(\mathcal{K}(f_i(\tilde{\mathbf{x}}), f_i(\tilde{\mathbf{x}}))$$

$$+ \mathrm{cov}(f_i(\tilde{\mathbf{x}}), \mathbf{Y})\mathrm{cov}(\mathbf{Y})^{-1}\big(\mathcal{K}(\mathbf{Y}, \mathbf{Y})\mathrm{cov}(\mathbf{Y})^{-1}\mathrm{cov}(\mathbf{Y}, f_i(\tilde{\mathbf{x}})) - 2\mathcal{K}(\mathbf{Y}, f_i(\tilde{\mathbf{x}}))\big)\Big),$$

$$(5.14)$$

with

$$\mathcal{K}(f_i(\mathbf{x}_1), f_j(\mathbf{x}_2)) = \int \mathrm{cov}\big(f_i(\mathbf{x}_1), f(\mathbf{x})\big)\mathrm{cov}\big(f(\mathbf{x}), f_j(\mathbf{x}_2))\big)w(\mathbf{x})\mathrm{d}\mathbf{x}. \qquad (5.15)$$

We see that every term in (5.14) is analytically tractable except the $\mathcal{K}$ function in (5.15) where the numerical integration is carried out. One idea to obtain an analytical form of $\mathcal{K}$, which has been suggested in the single-fidelity cases [21], is to approximate the $w(\mathbf{x})$ with a Gaussian mixture model [51] with n_{GMM} Gaussian functions:

$$w(\mathbf{x}) \approx \sum_{t=1}^{n_{GMM}} \alpha_t \mathcal{N}(\mathbf{x}; \mu_t, \Sigma_t). \qquad (5.16)$$

This allows us to reformulate (5.15) as

$$\mathcal{K}(f_i(\mathbf{x}_1), f_j(\mathbf{x}_2)) \approx \sum_{t=1}^{n_{GMM}} \alpha_t G_t(f_i(\mathbf{x}_1), f_j(\mathbf{x}_2)), \qquad (5.17)$$

with

$$G_t(f_i(\mathbf{x}_1), f_j(\mathbf{x}_2)) = \int \mathrm{cov}\big(f_i(\mathbf{x}_1), f(\mathbf{x})\big)\mathrm{cov}\big(f(\mathbf{x}), f_j(\mathbf{x}_2))\big)\mathcal{N}(\mathbf{x}; \mu_t, \Sigma_t)\mathrm{d}\mathbf{x}. \qquad (5.18)$$

The problem now boils down to developing an analytical formula for (5.18), which involves in the integrand the multiplication of two different multi-fidelity covariance functions and a Gaussian distribution function. This situation here is more complicated than that in the single-fidelity case [21] where the problem is simplified by only involving two same single-fidelity covariance functions (e.g., analytical result from the latter case is already available [88]). We summarize the detailed derivation of the analytical form of (5.18), as well as the derivative $\partial B(i, \tilde{\mathbf{x}})/\partial\tilde{\mathbf{x}}$ that can be derived in a similar manner, in Appendix C.4.

With analytical computation of $B(i, \tilde{\mathbf{x}})$ and $\partial B(i, \tilde{\mathbf{x}})/\partial\tilde{\mathbf{x}}$ available, we can solve

the optimization (5.13) using gradient-based algorithm (which is more suitable for high-dimensional problems). In our current work, a gradient-based quasi-Newton method [100] with multiple starting points is used to solve (5.13), which completes the algorithm of the multi-fidelity BED method for rare-event statistics. We finally summarize the full algorithm in Algorithm **4** and note that the algorithm reduces to a single-fidelity BED method for $s = 1$.

Algorithm 4 Multi-fidelity Bayesian experimental design for rare-event statistics

---

Require: Number of initial samples $\{n_{init}(i)\}_{i=1}^s$, cost of each fidelity model $\{c_i\}_{i=1}^s$, total budget c_{lim}

Input: Initial dataset $\mathcal{D} = \{\mathcal{X}, \mathcal{Y}\}$ with $\mathcal{X} = \{\mathcal{X}_i\}_{i=1}^s$ and $\mathcal{Y} = \{\mathcal{Y}_i\}_{i=1}^s$

    Initialization $c_{total} = \sum_i n_{init}(i) \, c_i$

    while $c_{total} < c_{lim}$ **do**

        1. Train the surrogate model (5.2) with $\mathcal{D}$ to obtain (5.4)

        2. Compute $w(\mathbf{x})$ in (5.11) and approximate it with GMM model (5.16)

        3. Solve the optimization (5.13) to find the next-best sample $\{i^*, \mathbf{x}^*\}$

        4. Evaluate the i^*-fidelity function to get $f_{i^*}(\mathbf{x}^*)$

        5. Update the dataset $\mathcal{X}_{i^*} = \mathcal{X}_{i^*} \cup \{\mathbf{x}^*\}$ and $\mathcal{Y}_{i^*} = \mathcal{Y}_{i^*} \cup \{f_{i^*}(\mathbf{x}^*)\}$

        6. $c_{total} = c_{total} + c_{i^*}$

    end while

Output: Compute the response PDF based on the surrogate model (5.5)

5.3 Results

In this section, we test our developed method in the context of bi-fidelity problems, i.e., $s = 2$ and we use $f_1(\mathbf{x}) = f_l(\mathbf{x})$, $c_1 = c_l$ and $f_2(\mathbf{x}) = f_h(\mathbf{x})$, $c_2 = c_h$ for clarity. The tests are conducted for three synthetic problems with dimensions $d = 1, 2, 8$ (§5.3.1, §5.3.2, §5.3.3) and an engineering problem (of $d = 2$) to evaluate rare-event statistics of ship motion in irregular waves with CFD of low and high resolutions as $f_l(\mathbf{x})$ and $f_h(\mathbf{x})$ (§5.3.4). In synthetic problems, the true solution p_f involved in (5.1) is obtained from a computation using a sufficiently large number of high-fidelity samples (since we assume $f(\mathbf{x}) = f_h(\mathbf{x})$ discussed in §5.2.1).

    In all cases, we compare the results from our method (bi-fidelity optimal sampling

in both location and fidelity level as in Algorithm **4**, hereafter termed "BF-O") to those from single-fidelity optimal sampling [21] (Algorithm **4** with $s = 1$, hereafter termed "SF"). In addition, for the synthetic problems with $d = 2$ and 8, we further include the results from the bi-fidelity model with fixed ratio n (as well as sequence) of low and high fidelity samples with locations optimized (Algorithm **4** but with optimization $\mathbf{x}_i^* = \text{argmax}_{\tilde{\mathbf{x}} \in \mathbb{R}^d} B(i, \tilde{\mathbf{x}})$ solved for fixed fidelity level i on each sample, hereafter termed "BF-Fn"). Specifically, we repeatedly use one high-fidelity sample followed by n low-fidelity samples, with each sample location $\mathbf{x}$ chosen through optimization regarding acquisition function B. The cost weight, although not included in the acquisition, is still considered in computation of the total cost c_{total}. The purpose to compare the results from BF-Fn and BF-O is to understand the benefit of the BF-O scheme to automatically choose the fidelity level. We shall vary the number n, the ratio c_h/c_l and the low-fidelity accuracy level in a broad range, to assess the performance of our method BF-O in various situations for cases of intermediate to relatively high complexity.

For all the methods in comparison, an initial data set is required to start the sequential BED procedure. To create a fair situation for comparison, we keep the cost to generate the initial dataset the same for all methods. In particular, the initial set is generated by a space-filling Latin Hypercube sampling method [89], with $4d$ high-fidelity samples for SF method, and $2d$ high-fidelity and $2d$ c_h/c_l low-fidelity samples for BF-Fn and BF-O methods. Since the initial sample locations are not fixed, we will present the results from SF, BF-O and BF-Fn methods in terms of the average over 100 implementations with different initial datasets for the synthetic cases in §5.3.1, §5.3.2, and §5.3.3.
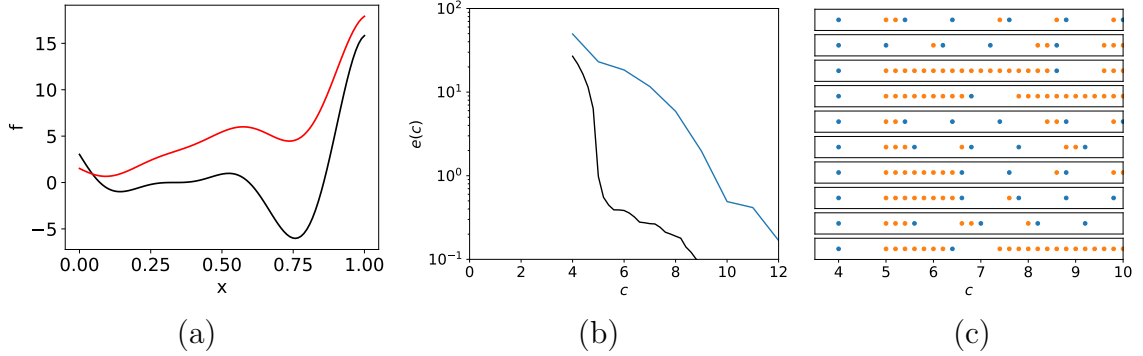
Figure 5.1: (a) The low fidelity function $f_l(x)$ (—) and high-fidelity function $f_h(x)$ (—) in the 1D problem. (b) The corresponding error $e(c)$ computed by BF-O(—) and SF(—). (c) The sequence of high-fidelity (·) and low-fidelity (·) samples in ten experiments of BF-O.

5.3.1 Forrester function

We start the method validation from a one-dimensional (1D) Forrester function $f(\mathbf{x})$ that has been previously used to demonstrate the multi-fidelity global optimization [38]. The high-fidelity $(f_h(\mathbf{x}) = f(\mathbf{x}),\, c_h = 1)$ and low-fidelity $(f_l(\mathbf{x}),\, c_l = 0.2)$ models are constructed as (see figure 5.1(a))

$$f_h(x) = (6x - 2)^2 \sin(12x - 4), \tag{5.19}$$

$$f_l(x) = 0.5 f_h(x) + 10x, \tag{5.20}$$

where the input x is assumed to follow a Gaussian distribution with $p_x(x) = \mathcal{N}(0.5, 0.1)$. The results of BF-O and SF methods are shown in figure 5.1(b) in terms of the error e (as in (5.1)) as a function of the total cost c (i.e., summation of c_l and c_h for all samples). The BF-O method clearly outperforms the SF method to a large extent. For example, at $c = 6$, the BF-O method achieves a value of e that is nearly two orders of magnitude smaller than the SF method. In figure 5.1(c), we show 10 examples of the sequence of fidelity levels in samples by BF-O. While the sequences vary for different initial datasets, the BF-O algorithm exclusively selects a

86

high-fidelity sample as the first sequential sample. This is consistent with an intuitive understanding that the algorithm tends to use three high-fidelity samples (2 in the initial dataset and 1 as selected) to learn the linear difference term $10x$ in (5.20) using a Gaussian process, which results in a significant reduction in e at $c \approx 5$ in figure 5.1(b).

While this simple example demonstrates the advantage of using BF-O method, it is only for a single case with fixed accuracy level of the low-fidelity model (in terms of (5.20)) and the cost ratio c_h/c_l. We will next use a two-dimensional (2D) case to test a much broader range of situations in the next section.

5.3.2   Stochastic oscillator

We consider a 2D function constructed from the solution of a stochastic oscillator equation, which has been previously used for testing the single-fidelity BED method (i.e., SF method) in [95, 21]. In particular, the oscillator equation is formulated as

$$\ddot{u} + \delta \dot{u} + F(u) = \xi(t), \tag{5.21}$$

where $u(t)$ is the state variable, F is a nonlinear restoring force defined by:

$$F(u) = \begin{cases} \alpha u & if\ 0 \leq |u| \leq u_1 \\ \alpha u_1 & if\ u_1 \leq |u| \leq u_2 \\ \alpha u_1 + \beta(u - u2)^3 & if\ u_2 \leq |u| \end{cases} \tag{5.22}$$

The stochastic process $\xi(t)$, with a correlation function $\sigma_\xi^2 e^{-\tau^2/(2l_\xi^2)}$, is approximated by a two-term Karhunen-Loeve expansion

$$\xi(t) \approx \sum_{i=1}^{2} x_i \lambda_i \phi(t), \tag{5.23}$$

87
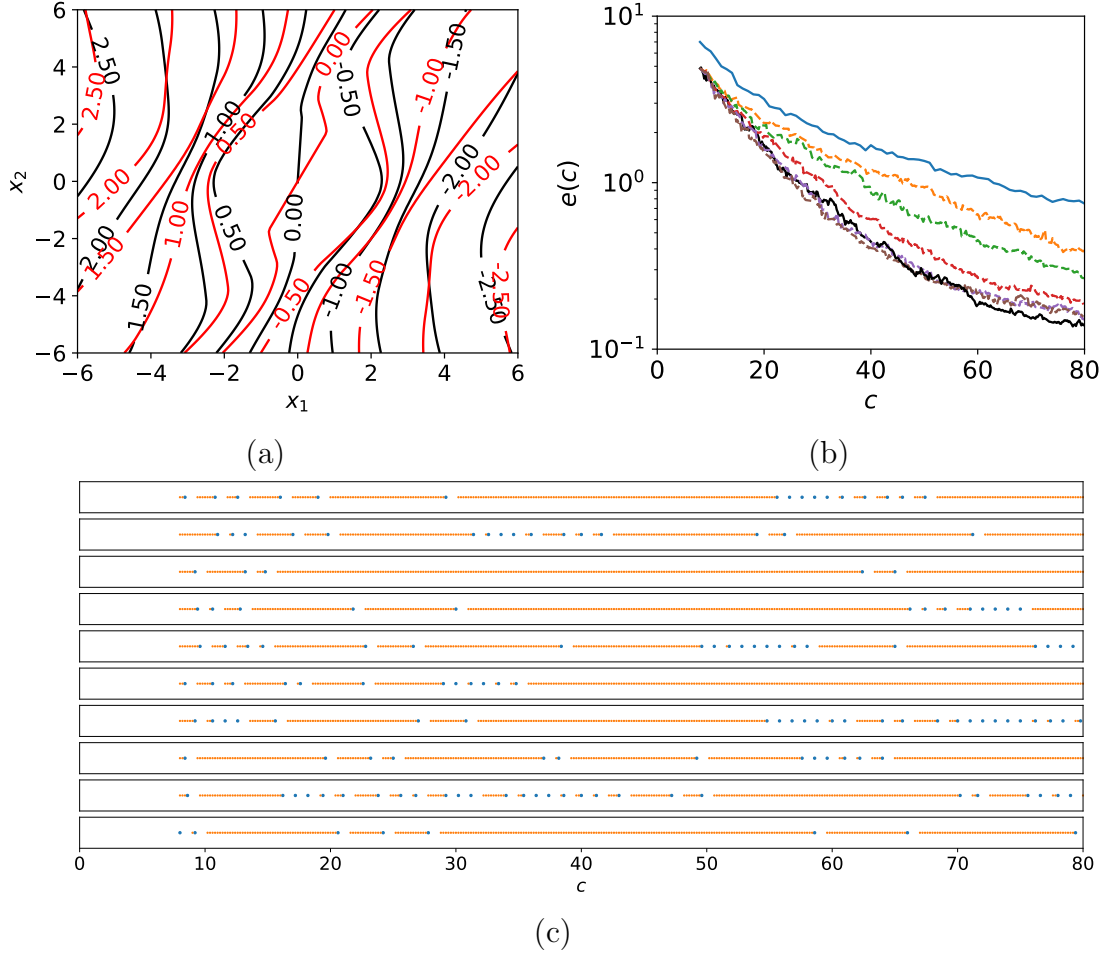
Figure 5.2: (a) The low-fidelity (—) and high-fidelity (—) functions with a linear difference in the 2D oscillator problem. (b) The corresponding error $e(c)$ computed by BF-O(—), BF-F1(- - -), BF-F2(- - -), BF-F5(- - -), BF-F10(- - -), BF-F15 (- - -) and SF(—). (c) The sequence of high-fidelity (•) and low-fidelity (•) samples in ten experiments of BF-O.

with λ_i and $\phi(t)$ respectively the eigenvalue and eigenfunction of the correlation function, $\mathbf{x} \equiv (x_1, x_2)$ is a standard normal variable as the input to the system, satisfying $p_{\mathbf{x}}(\mathbf{x}) = \mathcal{N}(\mathbf{0}, \mathrm{I})$ with I being a 2×2 identity matrix. The values of the parameters are kept the same as those in the single-fidelity work [21][1].

The response of the system is considered as the mean value of $u(t; \mathbf{x})$ in the interval

---

[1]δ=1.5, α=1, β=0.1, u_1=0.5, u_2=1.5, σ_ξ^2=0.1, l_ξ=4.

$[0, 25]$, which serves as our high-fidelity model:

$$f_h(\mathbf{x}) = \frac{1}{25} \int_0^{25} u(t; \mathbf{x})\mathrm{d}t. \tag{5.24}$$

For our low-fidelity model, we construct a function $f_l(\mathbf{x})$ (to be varied later in this section) with a difference $d(\mathbf{x})$ from (5.24):

$$f_l(\mathbf{x}) = \rho f_h(\mathbf{x}) + d(\mathbf{x}), \tag{5.25}$$

with $\rho = 1$ and $d(\mathbf{x})$ chosen as a linear function $0.05(x_1 + x_2)$ in this case. Both the $f_h(\mathbf{x})$ and $f_l(\mathbf{x})$ functions are shown in figure 5.2(a) to illustrate the functional forms and their difference. In this case, we use $c_l = 0.2$ and $c_h = 1$ as the computational cost of low and high fidelity models.

The error $e(c)$ is plotted in figure 5.2(b) for SF, BF-O and BF-Fn with n varying from 1 to 15. We see that all bi-fidelity methods (BF-O and BF-Fn) achieve acceleration on the error reduction (to different extents) compared to the SF method. For the BF-Fn method, faster convergence is observed for larger n in the test range of $n \in [1, 15]$, but with much less benefit for n increasing from 10 to 15. The BF-O method provides the best result, in terms of the error e at cost $c = 80$, although the BF-O result is somewhat less accurate than the BF-F15 result for smaller c in the range of [25,50]. Accounting for all the sequence of fidelity levels (with 10 examples shown in 5.2(c)), the average ratio of high and low fidelity samples selected by BF-O is approximately 19.06, close to the value of $n = 15$ which is found to be the best in BF-Fn in the test range.

To further test the performance of BF-O (as well as BF-Fn) in more diversified situations, in the following, we construct additionally 9 cases with different computational costs c_l and low-fidelity functions $f_l(\mathbf{x})$ as summarized in Table 5.1. In addition to the linear difference function used in the previous case (now case 1 in Table 5.1),

89

Table 5.1: Setting of cases in the 2D oscillator problem.

no.	$d(\mathbf{x})$	ρ	δ	c_h/c_l	rank of BF-O at cost 80	n_l/n_h in BF-O	best n in BF-Fn
1	$\delta(x_1+x_2)$		0.05		1	19.06	15
2			0.02	5	2	11.95	5
3			0.05		1	7.19	5
4				2	1	0.47	1
5		1	0.1	5	1	4.74	5
6	$\delta\sin(x_1+x_2)$			8	1	5.93	5
7				10	1	6.85	10
8			0.2		2	3.19	5
9			0.4	5	2	2.16	5
10		1.5	0.1		1	6.91	5

we also consider a nonlinear difference function $d(\mathbf{x}) = \delta\sin(x_1+x_2)$ which is known to be a difficult situation to approximate by a Gaussian process [113].

Figure 5.3 shows the results for varying the low-fidelity accuracy level with fixed $c_h/c_l = 5$, i.e., cases 2, 3, 5, 8-10 in Table 5.1 with varying ρ and δ, with the range of δ corresponding to difference terms up to 20% of the maximum response $f_h(\mathbf{x})$ for $|\mathbf{x}| < 4$. Together with figure 5.2, we see that the BF-O scheme consistently performs among the best of all tested methods. More specifically, with the increase of complexity in the difference function (i.e., increasing δ but not much for increasing ρ), the performance of BF-O can deteriorate for smaller number of sequential samples (e.g., figure 5.3(f)), but still behaves close to the optimal at larger number of samples with cost $c \approx 80$. These cases correspond to the situation where the difference terms are difficult to learn, which takes more cost to make low-fidelity samples useful for the final results.

Figure 5.4 shows the results for varying c_l, i.e., $c_h/c_l = 2, 5, 8, 10$ with fixed $c_h = 1$ as in cases 4-7 in Table 5.1. Similar to the results above, we see that the BF-O method consistently performs among the best. This indicates that the acquisition function (5.12) employing the ratio between benefit $B(i, \mathbf{x})$ (uncertainty reduction) and cost
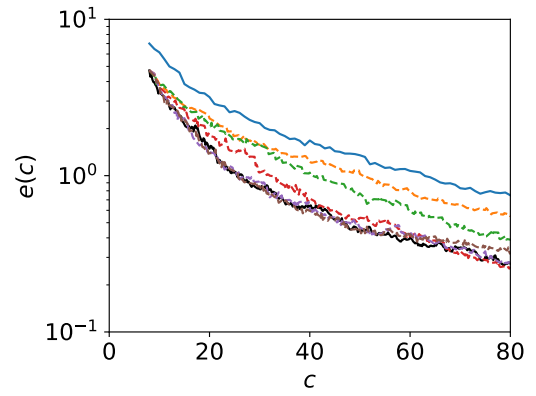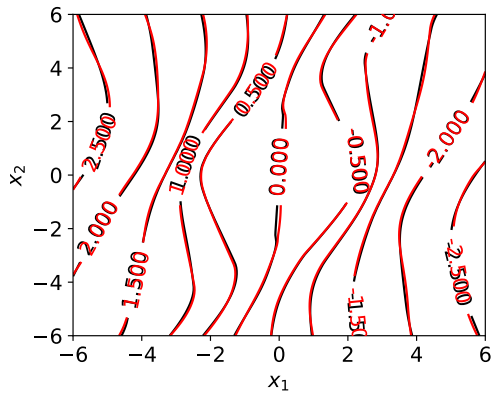
(a)
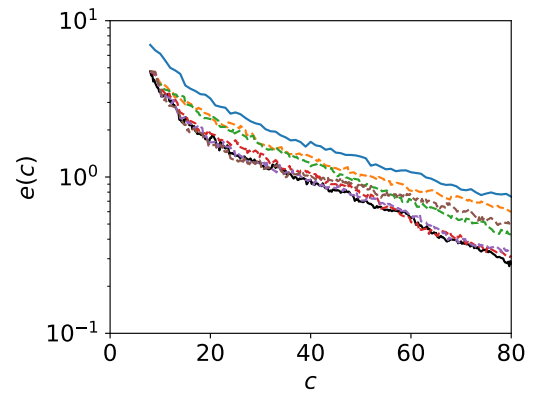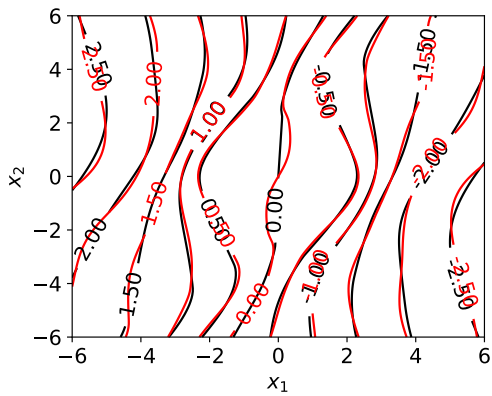
(b)
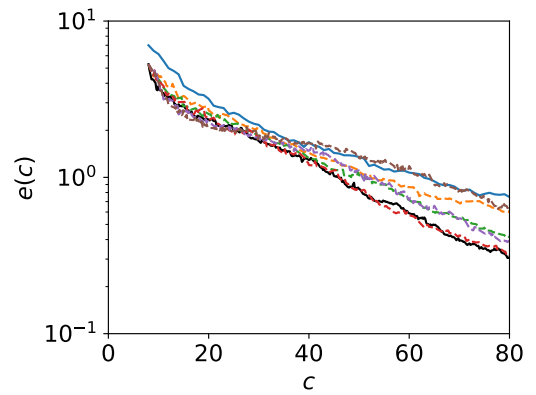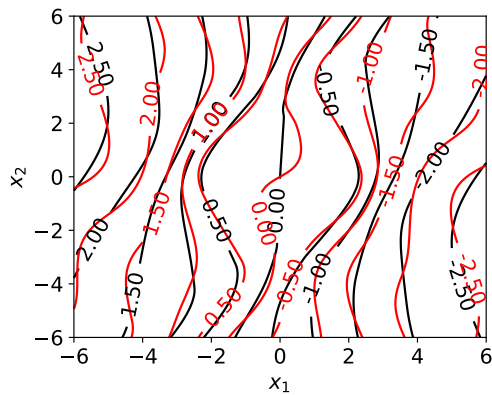
(c)

See next page for caption.
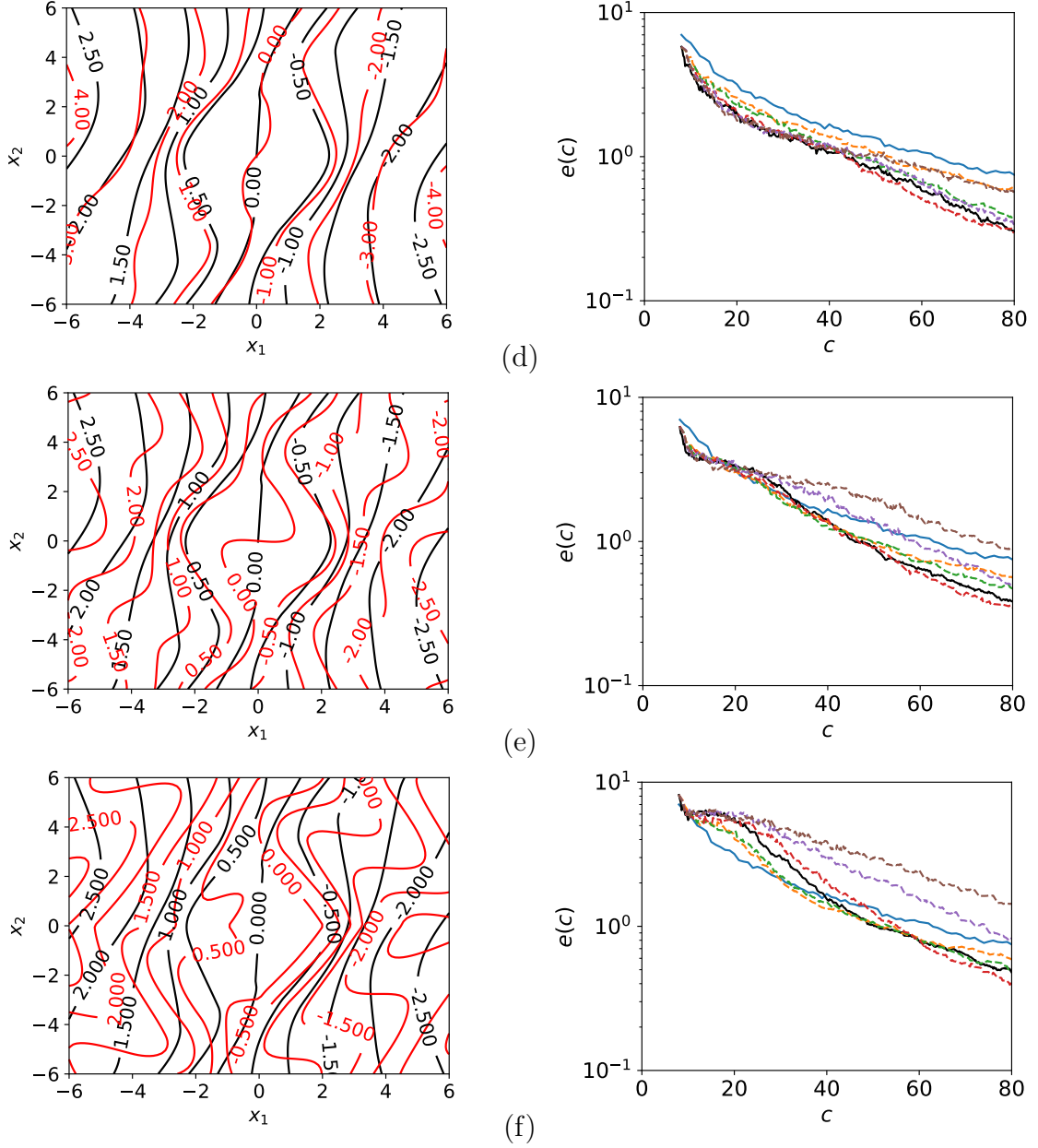
Figure 5.3: Results of (a) case 2 $\{\delta = 0.02, \rho = 1\}$, (b) case 3 $\{\delta = 0.05, \rho = 1\}$, (c) case 5 $\{\delta = 0.1, \rho = 1\}$, (d) case 10 $\{\delta = 0.1, \rho = 1.5\}$, (e) case 8 $\{\delta = 0.2, \rho = 1\}$, and (f) case 9 $\{\delta = 0.4, \rho = 1\}$ for low-fidelity function (5.25) with nonlinear difference in the 2D oscillator problem. Left: the low fidelity function (——) and high-fidelity function (——). Right: the error $e(c)$ computed by BF-O(——), BF-F1(- - -), BF-F2(- - -), BF-F5(- - -), BF-F10(- - -), BF-F15 (- - -) and SF(——).
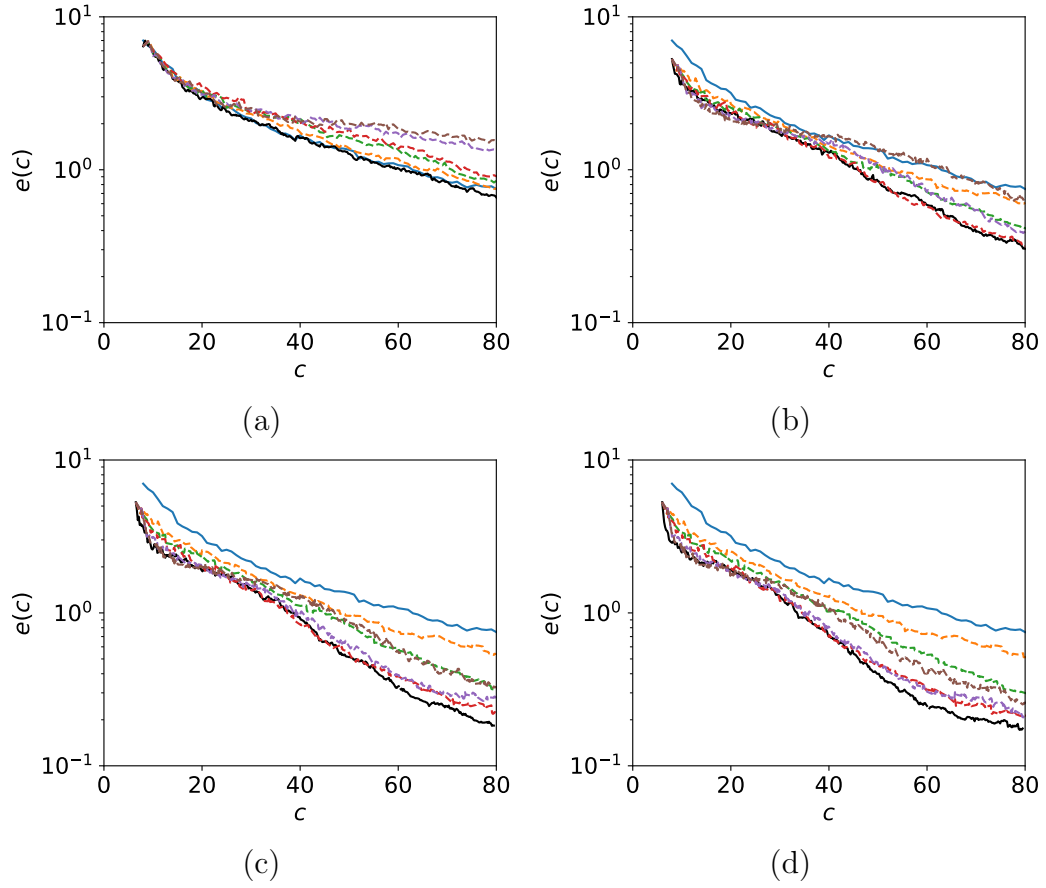
Figure 5.4: Errors in the cases of $\{\delta = 0.1, \rho = 1\}$ with (a) case 4, $c_h/c_l = 2$, (b) case 5, $c_h/c_l = 5$, (c) case 6, $c_h/c_l = 8$, and (d) case 7, $c_h/c_l = 10$ computed by BF-O(—), BF-F1(- - -), BF-F2(- - -), BF-F5(- - -), BF-F10(- - -), BF-F15 (- - -) and SF(—).

c_i effectively captures the optimal that balances the two factors.

The performance of the BF-O scheme, in terms of its ranking in all schemes, is summarized in Table 5.1. It is clear that the BF-O method consistently provides accurate results of the rare-event response PDF (ranking the 1st or 2nd among all methods for all cases). Moreover, we include in the table the average ratio of low and high fidelity sample numbers n_l/n_h in the BF-O method, as well as the value of n corresponding to the BF-Fn method with the best performance at $c = 80$. While the optimal n in the BF-Fn method does not necessarily correspond to the n_l/n_h in BF-O method, we find that the two numbers are close in most cases, with the latter automatically captured by the algorithm.

5.3.3 Borehole hydrological model

We next consider an eight dimensional borehole hydrological model, which has been used as an example of high-dimensional problems to quantify the rare response statistics using single-fidelity methods [56, 20, 32]. The model physically computes the flow rate through a borehole, formulated as

$$f_h(\mathbf{x}) = \frac{2\pi T_u(H_u - H_l)}{\ln(r_i/r_b)}[1 + \frac{2LT_u}{\ln(r_i/r_b)r_b^2 K_w} + \frac{T_u}{T_l}], \qquad (5.26)$$

with an eight-dimensional input $\mathbf{x} = \{r_b, r_i, T_u, H_u, T_l, H_l, L, K_w\}$, including their distributions, detailed in Table 5.2. We further construct a low-fidelity model as

$$f_l(\mathbf{x}) = f_h(\mathbf{x}) + 7.5\sin(x_1) + 0.75\sin(\sum_{i=2}^{8} x_i), \qquad (5.27)$$

where we use a nonlinear (sinusoidal) form of the difference function and put more weights on the parameter $x_1 = r_b$, as it is the most influential factor to the response $f_h(\mathbf{x})$ [56]. In particular, the coefficient 7.5 is chosen such that the term of $\sin(x_1)$ corresponds to 5% of the maximum response. We keep the computational cost ratio

Table 5.2: The input parameters and their distributions of the borehole function.

	Definition	Range	Distribution
r_b	radius of borehole	$[0.05; 0.15]\ m$	normal[1]
r_i	radius of influence	$[100; 50000]\ m$	logNormal[2]
T_u	transmissivity of upper aquifer	$[63070; 115600]\ m^2/yr$	uniform
H_u	potentiometric head of upper aquifer	$[990; 1110]\ m$	uniform
T_l	transmissivity of lower aquifer	$[63.1; 116]\ m^2/yr$	uniform
H_l	potentiometric head of lower aquifer	$[700; 820]\ m$	uniform
L	length of borehole	$[1120; 1680]\ m$	uniform
k_w	hydraulic conductivity of borehole	$[9855; 12045]\ m/yr$	uniform

[1] r_b with mean 0.10, stand deviation 0.0161812.
[2] $\ln(r_i)$ with mean 7.71, stand deviation 1.0056.
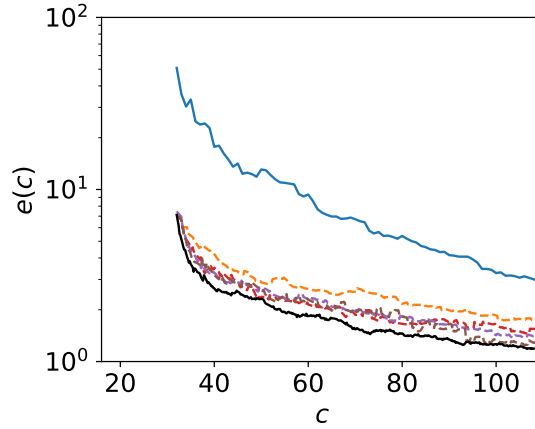
Figure 5.5: The error $e(c)$ computed by BF-O(—), BF-F1(---), BF-F5(---), BF-10(---), BF-F15 (---) and SF(—) for the high-dimensional borehole problem.

as $c_h/c_l = 5$ for this case.

The results from the SF, BF-O and BF-Fn methods are shown in figure 5.5. Comparing to the SF result, it is clear that the benefit of using bi-fidelity models is more evident than the low-dimensional cases, even for the initial data set without sequential samples. The reason is that the "value" of a high-fidelity sample becomes compromised with the increase of dimensions. In addition, the BF-O method again performs the best among all BF-Fn methods with varying n.

We finally remark that the computation in this eight-dimensional case is only en-
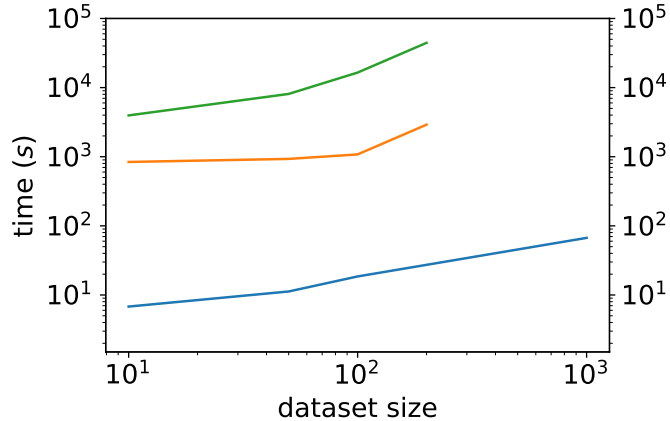
Figure 5.6: The computation time for selecting one sequential sample in the borehole problem by using the U criterion in (5.9)(—), the B criterion in (5.12) with numerical integration (—) and analytical formula (—) as objective functions for different sizes of the existing dataset. In the former two computations, we assume that 10 times (a conservative number) of acquisition evaluations are needed compared to the gradient-based optimization in the third case, as a common practice found in [86].

abled because of (i) the development of (5.12) which avoids the construction of a new Gaussian process for each hypothetical sample, and (ii) the development of analytical formula through GMM which avoids the numerical integration in (5.12) and enables the gradient computation. To illustrate this point, we show in figure 5.6 the computation time for solving (5.13) as a function of the number of samples in the existing dataset on a single core of Intel Xeon Gold 6154 CPU (specifically we use $n_{GMM} = 2$ with (5.16) computed by 10^6 quadrature points, and 10 starting points in the quasi-Newton method). For comparison, we also include in figure 5.6 the computation time of using (5.9) (with a new Gaussian process for each hypothetical point) and (5.12) (with numerical integration) as objective functions in optimization. In both cases, not only are the computation for the objective functions expensive, these computations also need to be repeated many more times in gradient-free optimization than that in the gradient-based method, resulting in prohibitive computational costs for large number of samples. In contrast, using the analytical formula combined with the GMM model, the computation takes only $O(100)s$ even for a dataset of 1000 sample

points, which is supposed to be negligible compared to the evaluation of the output of the high-fidelity physical model.

5.3.4 Coupling with CFD to compute rare-event statistics of ship motion in irregular waves

We further consider an application of our method to evaluate the rare-event PDF of ship roll motion in irregular waves (other applications of multi-fidelity methods in naval engineering can be found in [122, 144, 143] for design optimization purpose). More specifically, we study the motion of a two-dimensional, square-shaped hull geometry with $40m \times 40m$ cross section and density $\rho_h = 0.5\rho_w$ (with ρ_w the water density) subject to beam waves. The input to this problem is considered as $\mathbf{x} = \{L, A\}$, with L and A the wave group amplitude and length, as a reduced-order description of an uni-directional irregular wave field [26]. Figure 5.7 shows an example to evaluate parameters L and A from a given wave field as well as the resulted probability distribution $p_\mathbf{x}(L, A)$. The irregular wave field is described by a Gaussian spectrum in the form

$$F(k) \sim \exp \frac{-(k - k_0)^2}{2\mathcal{K}^2},$$

(5.28)

with the significant wave height $H_s = 12m$, peak (carrier) wavenumber $k_0 = 0.018m^{-1}$ (corresponding to peak period $T_p = 15s$), and $\mathcal{K} = 0.05k_0$. The response in this case is considered as the maximum ship roll r_{max} in a wave group, i.e., we consider a response function $r_{max}(L, A)$.

The high and low fidelity models in this case are constructed by CFD models with high and low resolutions, both developed using the open-source code OpenFOAM [64]. In particular, the grid resolutions for both cases are shown in figure 5.8, where the low-fidelity model uses half number of grids in both horizontal and vertical directions relative to the high-fidelity model. The setting of the fluid solvers, other than the
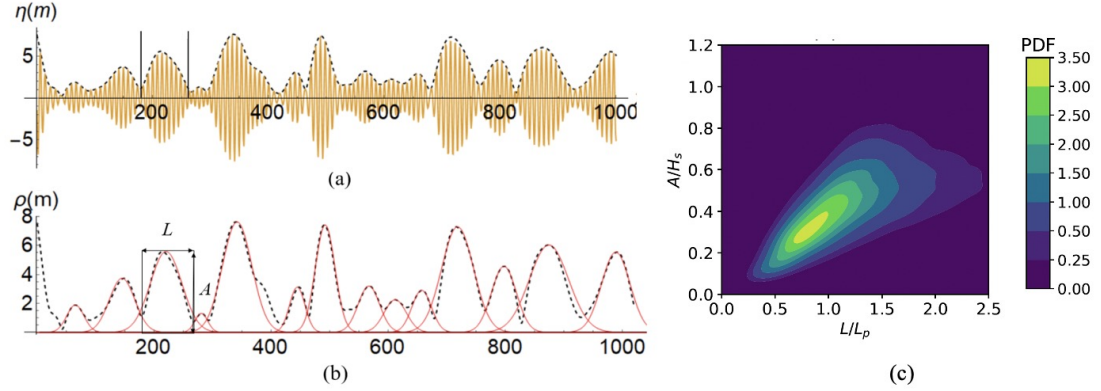
Figure 5.7: An example of a wave field with (a) wave elevation η (——) and (b) envelop ρ (- - -) approximated by a sequence of wave groups (——) with group amplitude parameter A and length parameter L. (c) Joint PDF of L (normalized by the spectral peak wavelength L_p) and A (normalized by the significant wave height H_s).
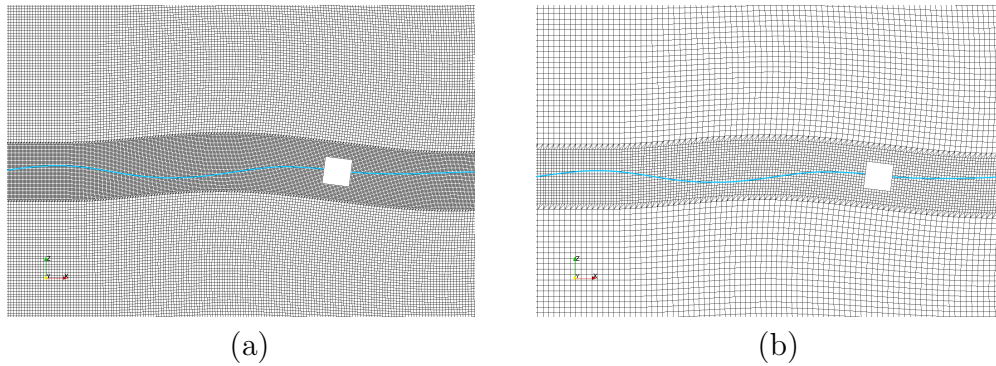
Figure 5.8: (a) Fine and (b) coarse grids of the CFD simulations, with free surface indicated by a cyan line, and hull by a white box in each case.

resolution, is the same for both models, with details presented in §2.5. The average wall time of high and low-fidelity simulations with 40 cores (Intel Xeon Gold 6154 CPU) are calibrated as 0.67 and 0.20 hours (considering adaptive time step size and parallel efficiency), leading to $c_h/c_l = 3.36$ as the value we use in the sequential BED method. The time series of ship roll motion computed from the high and low fidelity models are shown in figure 5.9 for two examples with different input wave parameters, showing that the difference of the results from the two models (in general) increases with the group amplitude A.

98
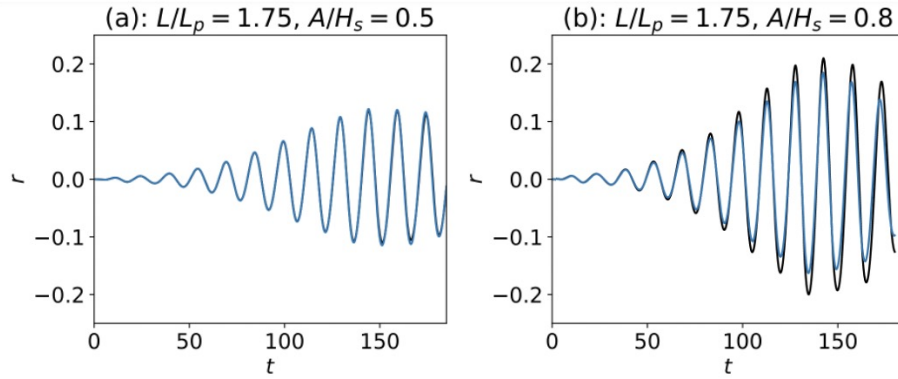
Figure 5.9: Results of high-fidelity (—) and low-fidelity (—) simulations for wave groups of (a) $L/L_p = 1.75$, $A/H_s = 0.5$ and (b) $L/L_p = 1.75$, $A/H_s = 0.8$.
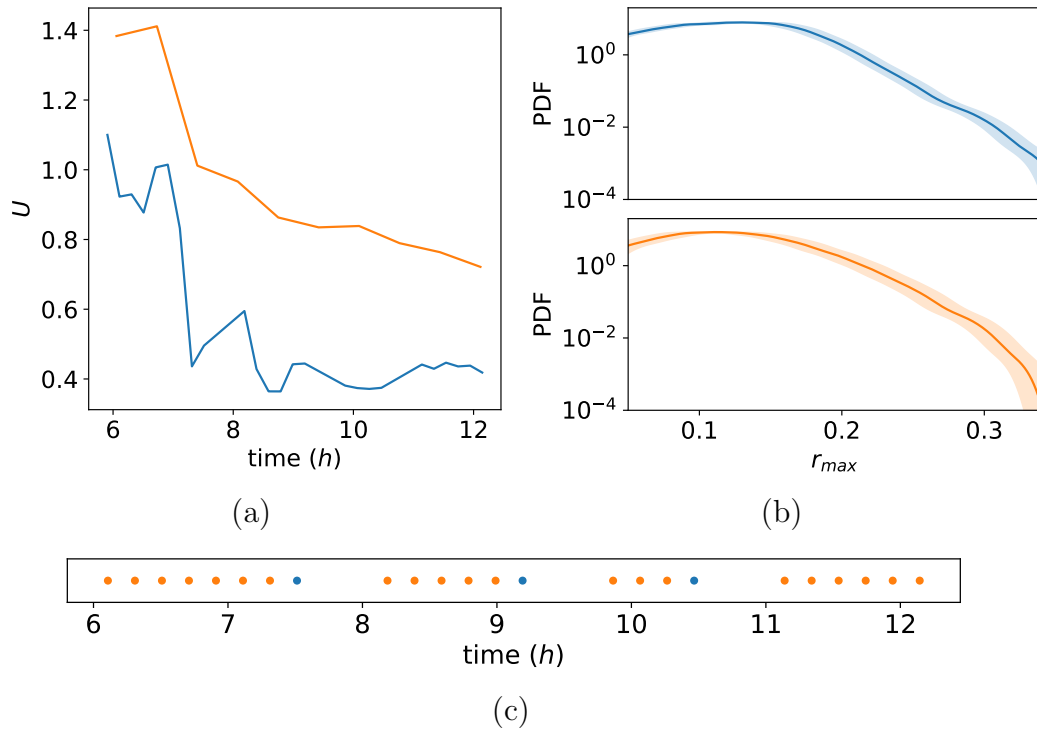
Figure 5.10: (a) The uncertainty level U computed by (5.9) for BF-O (—) and SF (—) as a function of the total computation time. (b) The final PDF computed by BF-O (—, upper panel) and SF (—, lower panel) at approximately 12 hours of computation time, with two standard deviations marked by the shaded regions. (c) The sequence of high-fidelity (•) and low-fidelity (•) CFD simulations.

We use as the initial dataset 4 high-fidelity and 16 low-fidelity samples in the BF-O method, and 9 high-fidelity samples in the SF method (resulting in almost the same

cost). Since the exact PDF in this case is unknown, we directly use the uncertainty level of the rare-event portion of PDF, i.e., U as in (5.9), as an evaluation of the quality of results, which is plotted in figure 5.10(a) as a function of the computational cost measured by computation time in hours (only CFD times). It can be seen that the BF-O method results in a faster convergence compared to the SF method. This point is further illustrated in figure 5.10(b), which plots the PDFs from SF and BF-O methods for the same computation time of 12 hours together with the upper and lower bounds (in terms of two standard deviations two-sided from the mean). It is clear that the PDF from the BF-O method is associated with appreciably lower uncertainty. Finally, the high and low fidelity sample sequence in BF-O is shown in 5.10(c) with $n_l/n_h = 7$ in this case.

5.4 Summary

In this chapter, we develop a multi-fidelity sequential Bayesian experimental design framework for efficient evaluation of the response PDF emphasizing the rare-event portion. Our method leverages the multi-fidelity Gaussian process as a surrogate model, and a new acquisition function which allows the selection of the next-best sample in terms of both the location and fidelity level. We also construct an analytical formula for the acquisition function, which enables implementation of the method (e.g., gradient-based optimization) for high-dimensional problems. Our new method is first tested in a bi-fidelity context for a series of synthetic problems. With a broad range of low-fidelity accuracy level and computational cost, we show that the bi-fidelity method always outperforms the single-fidelity method, and that the BF-O method consistently shows advantage over the BF-Fn method, i.e., the bi-fidelity method with pre-defined fidelity hierarchy. We finally demonstrate the effectiveness of our BF-O method (relative to the single-fidelity method) in an engineering problem

to evaluate the rare-event PDF of ship roll motion in irregular waves, where CFD of two resolutions serve as the high and low fidelity models.

We note that an assumption to use the multi-fidelity scheme is that the 'difficulty' of constructing the low-fidelity function, difference functions, and (directly) the high-fidelity function is approximately the same. If that is not true, e.g., when learning the low-fidelity function or the difference function is much harder than the original high-fidelity function, the multi-fidelity scheme may have the opposite effect. Under this circumstance, one can try a nonlinear relation between low and high-fidelity functions [107, 73], improve the quality of low-fidelity samples, or directly use a single-fidelity scheme.

We finally remark on two points regarding the acquisition (5.10) and (5.12). First, these acqusitions are both designed to effectively minimize the error regarding the overall PDF tail according to (5.1). Such criteria do not guarantee the minimization of error in exceeding probability with any given threshold, and vice versa. For the application of multi-fidelity sequential sampling to problems of exceeding probability, we refer the readers to the next chapter §6. Second, in minimization of error (5.1), (5.10) and (5.12) are expected to be effective only if the predicted response PDF (by surrogate model) is sufficiently close to the true PDF (see [95, 118] for a derivation). In other words, if the mean prediction by the surrogate model happens to miss some rare-event responses of interest in some regions of the input parameter space, these acquisitions may not drive sequential samples to the rare event region. The key to addressing this unfavorable issue is to endow the acquisition function with necessary exploration power, a topic covered in §7.

CHAPTER 6

Multi-Fidelity Bayesian Experimental Design for Safety Analysis of Connected and Automated Vehicles

6.1 Introduction

In §5, a multi-fidelity Bayesian experimental design framework has been developed to resolve response PDF with emphasis on the tail part. In this chapter, we adapted the multi-fidelity framework with a new acquisition function for reliability analysis, i.e., estimating exceeding probability above/below a threshold. Specifically, we focus on the safety of connected and automated vehicles (CAVs) and start this chapter by introducing its background and existing literature.

CAVs have attracted increasing attention due to their potential to improve mobility and safety while reducing the energy consumed. As we discussed in §1, the converged statistics of their accident rate may require hundreds of millions of miles for each configuration of CAVs [66]. To reduce the testing cost, scenario-based approaches have been developed where the scenarios (and their distribution) describing certain traffic environments are parameterized from the Naturalistic Driving Data (NDD). The performance of CAVs is then evaluated for given scenarios as the input, and the accident rate of CAVs is quantified considering the distribution of scenarios.

In order to efficiently compute the accident rate under the scenario-based approach, many methods have been developed to reduce the number of scenario evaluations. One category of methods relies on importance sampling, where samples are selected from a proposal distribution to stress the critical input regions (leading to most accidents). Different ways to construct the proposal distribution have been developed in [161, 160, 37, 35, 34], leading to significant acceleration compared to standard Monte Carlo method.

Another category of methods in safety analysis is based on adaptive sampling enabled by Bayesian experimental design (or more broadly active learning), which was first developed for structural reliability analysis [136, 31, 59, 15, 135, 149] and have recently been introduced to the CAV field [96, 134, 79, 62, 34]. To provide more details, two acquisition functions are proposed in [96], respectively designed for Gaussian process regression and k-nearest neighbors as surrogate models, in order to better resolve performance boundaries between accidents and safe scenarios. These acquisition functions combine exploration and exploitation under some heuristic consideration of the surrogate models. The approach in [96] is extended in [79] by clustering samples into different groups that allow a parallel search of optimal samples to accelerate the overall algorithm. In [134], the authors develop two acquisition functions applicable to six different surrogate models, which favor samples expected to respectively produce (i) poor performance, and (ii) performance close to accident threshold, and in the meanwhile, far from existing samples (for exploration). In these works, the proposed acquisition functions are rather empirical and cannot guarantee optimal convergence of the accident rate. In addition, an acquisition function that directly targets the accident rate (in which sense similar to what we develop) is proposed in [62, 61], but their method is not sufficiently supported by numerical tests provided in their papers. In viewing the state-of-the-art methods in the field, it is clear that large room exists for further improvement of the sampling efficiency (i.e., reduction of the

required number of samples) through a more rigorous information-theoretic approach to develop the acquisition. Such developments are not only desired to reduce the cost of CAV safety evaluation but are also valuable to the general field of reliability analysis.

The cost in the evaluation of CAV accident rate can also be reduced by leveraging low-fidelity models applied in conjunction with the high-fidelity model. In principle, the low-fidelity models can provide useful information on the surrogate model (e.g., the general trend of the function) although their own predictions may be associated with considerable errors. For example, low-fidelity models have been used to generate the proposal distribution for importance sampling [35, 34]. It needs to be emphasized that almost all existing works (in the CAV field) assume that the low-fidelity models are associated with negligible cost, i.e., the low-fidelity map from scenario space to CAV performance can be considered as a known function. However, in practical situations, the cost ratio between high and low-fidelity models may not be that drastic. Typical cases include (i) CARLA [30] simulator versus SUMO simulator [78], (ii) the same simulator with fine versus coarse-time resolutions. For these cases, a new adaptive-sampling algorithm considering the cost ratio is needed, which is expected to be able to select both the model (i.e., fidelity level) and scenario for the next-best sample to reduce the overall cost in the evaluation of the accident rate. Such methods are not yet available for CAV testing.

To fill these gaps, in this chapter, we develop an adaptive sampling algorithm in the Bayesian experimental design framework for safety testing and evaluations of CAVs. The novelty of our method lies in the development of an information-theoretic-based acquisition function that leads to very high sampling efficiency and can be extended to bi-fidelity contexts in a relatively straightforward manner. In particular, our method is applied to two situations: (i) the single-fidelity context where only a high-fidelity model is available; and (ii) the bi-fidelity context where the high-to-low

model cost ratio is finite and fixed. We note that for case (ii), our method needs to be established by using a bi-fidelity Gaussian process as the surrogate model and an acquisition function to select the next sample (in terms of both model fidelity and traffic scenario) which maximizes information gain per cost. Both applications of our method are tested in a widely-considered two-dimensional cut-in problem for CAVs, with the high-fidelity model taken as the Intelligent Driving Model (IDM) with fine time resolution. The low-fidelity model is constructed by a coarser-time-resolution IDM model in application (ii). We compare the performance of our method with the state-of-the-art approaches in the CAV field for the same problem and find that even the single-fidelity approach can considerably outperform the existing approaches. The method in application (ii) can further reduce the computational cost by at least a factor of 2.

This chapter is adapted from [42]. The Python code for the proposed algorithm, named MFGPreliability, is available on Github[1].

6.2 Method

6.2.1 Problem setup

We consider a black-box function $f_h(\mathbf{x}) : \mathbb{R}^d \to \mathbb{R}$ with input $\mathbf{x}$ a d-dimensional decision variable of a driving scenario and output a measure of the CAV performance. A subscript h is used here to denote that the function needs to be evaluated by an expensive high-fidelity model. Taking the cut-in problem (figure 6.1) as an example, the input can be formulated as $\mathbf{x} = (R_0, \dot{R}_0)$ where R_0 and $\dot{R}_0$ denote the initial range and range rate between the CAV and background vehicle (BV) at the cut-in moment $t = 0$ (more details in §6.3.1). The output is the minimum range between the two vehicles during their speed adjustment process for $t \geq 0$.

---

[1]https://github.com/umbrellagong/MFGPreliability
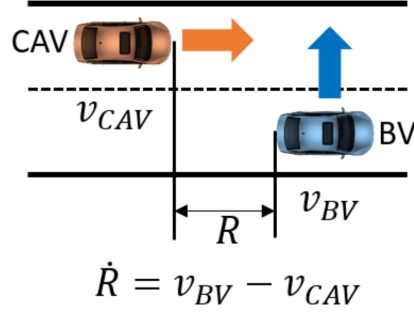
Figure 6.1: Illustration of the cut-in scenario [34]. R and $\dot{R}$ respectively denote the range and range rate between CAV and BV.

The probability of the input $\mathbf{x} \sim p_{\mathbf{x}}(\mathbf{x})$ is assumed to be known from the naturalistic driving data (NDD). Our objective is the evaluation of accident rate, i.e., probability of the output smaller than some threshold δ (or range between CAV and BV smaller than δ):

$$P_a = \int \mathbf{1}_\delta(f_h(\mathbf{x}))p_{\mathbf{x}}(\mathbf{x})\mathrm{d}\mathbf{x}, \tag{6.1}$$

where

$$\mathbf{1}_\delta(f_h(\mathbf{x})) = \begin{cases} 1, & \text{if } f_h(\mathbf{x}) < \delta \\ 0, & \text{o.w.} \end{cases} \tag{6.2}$$

A brute-force computation of P_a calls for a large number of Monte Carlo samples in the space of $\mathbf{x}$, which may become computationally prohibitive (considering the expensive evaluation of f_h and the small P_a). In this chapter, we seek to develop an adaptive sampling framework based on Bayesian experimental design, where samples are selected optimally to accelerate the convergence of the computed value of P_a. We will present algorithms for (i) single-fidelity cases, where only one model f_h is available, and (ii) bi-fidelity cases. For case (ii), we consider a practical situation that a low-fidelity model f_l with a lower but finite cost is also available to us that can provide a certain level of approximation to f_h. Making use of f_l, as will be demonstrated, can further reduce the cost of computing P_a.

106

6.2.2 Single fidelity method

We consider the single-fidelity context where only the model f_h is available. Two basic components of our Bayesian experimental design are presented below: (i) an inexpensive surrogate model based on the standard Gaussian process; (ii) a new acquisitive function to select the next-best sample.

6.2.2.1 surrogate model by GPR

Gaussian process regression (GPR) is a probabilistic machine learning approach [113] widely used for Bayesian experimental design. Consider the task of inferring f_h from $\mathcal{D} = \{\mathbf{X}, \mathbf{Y}\}$, which consists of n inputs $\mathbf{X} = \{\mathbf{x}^i \in \mathbb{R}^d\}_{i=1}^{i=n}$ and the corresponding outputs $\mathbf{Y} = \{f_h(\mathbf{x}^i) \in \mathbb{R}\}_{i=1}^{i=n}$. In GPR, a prior, representing our beliefs over all possible functions we expect to observe, is placed on f_h as a Gaussian process $f_h(\mathbf{x}) \sim \mathcal{GP}(0, k(\mathbf{x}, \mathbf{x}'))$ with zero mean and RBF kernel k defined in Appendix A.

Following the Bayes' theorem, the prediction for f_h given the dataset $\mathcal{D}$ can be derived to be another Gaussian process:

$$f_h(\mathbf{x})|\mathcal{D} \sim \mathcal{GP}\big(\mathbb{E}(f_h(\mathbf{x})|\mathcal{D}), \mathrm{cov}(f_h(\mathbf{x}), f_h(\mathbf{x}')|\mathcal{D})\big), \tag{6.3}$$

where formulae of posterior mean and covariance are summarized in Appendix A.

6.2.2.2 acquisition function

Given the GPR surrogate $f_h(\mathbf{x})|\mathcal{D}$, the accident rate $P_a|\mathcal{D}$ becomes a random variable with its randomness coming from the uncertainty of the GPR. The principle of finding the next-best sample is to provide most information to the quantity of interest P_a. This can be achieved in two ways: (i) through an information-theoretic perspective for the next sample to maximize the information gain, i.e., the K-L divergence between the current estimation and the hypothetical next-step estimation of P_a; (ii) through a

more intuitive and efficient approach for the next sample to minimize the uncertainty level associated with the distribution of P_a. In this paper, we describe the algorithm for (ii) in the main text and the algorithm for (i) in Appendix D.1. The results from the two approaches are equivalent after simplification of the results from (i) under reasonable assumptions presented in Appendix D.1.

For approach (ii), we need to formulate the uncertainty of P_a (measured by the variance of its distribution) after adding a hypothetical sample at $\tilde{\mathbf{x}}$:

$$\text{var}(P_a|\mathcal{D}, \overline{f}_h(\tilde{\mathbf{x}})) = \text{var}\Big(\int \mathbf{1}_\delta\big(f_h(\mathbf{x})|\mathcal{D}, \overline{f}_h(\tilde{\mathbf{x}})\big)p_{\mathbf{x}}(\mathbf{x})\mathrm{d}\mathbf{x}\Big), \tag{6.4}$$

where $\overline{f}_h(\tilde{\mathbf{x}}) = \mathbb{E}(f_h(\tilde{\mathbf{x}})|\mathcal{D})$ the CAV performance computed as the mean prediction from the current GPR. Our purpose is to find $\tilde{\mathbf{x}}$ so that (6.4) is minimized. However, the computation of (6.4) is very expensive since the variance operator involves the sampling of an integral (i.e., integral to be computed many times with expensive sampling of $f_h(\mathbf{x})|\mathcal{D}, \overline{f}_h(\tilde{\mathbf{x}})$). The computational cost of (6.4) can be significantly reduced by considering an upper bound of (6.4), following approaches developed in §4. With the detailed derivation presented in Appendix D.2, the upper bound of (6.4) gives

$$U(\mathcal{D}, \overline{f}_h(\tilde{\mathbf{x}})) = \int \text{var}^{\frac{1}{2}}\Big(\mathbf{1}_\delta\big(f_h(\mathbf{x})|\mathcal{D}, \overline{f}_h(\tilde{\mathbf{x}})\big)\Big)p_{\mathbf{x}}(\mathbf{x})\mathrm{d}\mathbf{x}, \tag{6.5}$$

where the variance function in (6.5) can be analytically evaluated (since the indicator function simply follows a Bernoulli distribution for each $\mathbf{x}$) as

$$\begin{aligned}
\text{var}\Big(\mathbf{1}_\delta\big(f_h(\mathbf{x})|\mathcal{D}, \overline{f}_h(\tilde{\mathbf{x}})\big)\Big) &= \Big(1 - \Phi\big(\frac{\mathbb{E}\big(f_h(\mathbf{x})|\mathcal{D}, \overline{f}_h(\tilde{\mathbf{x}})\big) - \delta}{\text{var}^{\frac{1}{2}}\big(f_h(\mathbf{x})|\mathcal{D}, \overline{f}_h(\tilde{\mathbf{x}})\big)}\big)\Big) \\
&\quad * \Phi\big(\frac{\mathbb{E}\big(f_h(\mathbf{x})|\mathcal{D}, \overline{f}_h(\tilde{\mathbf{x}})\big) - \delta}{\text{var}^{\frac{1}{2}}\big(f_h(\mathbf{x})|\mathcal{D}, \overline{f}_h(\tilde{\mathbf{x}})\big)}\big),
\end{aligned} \tag{6.6}$$

with Φ the cumulative distribution function of a standard Gaussian. It is clear that in evaluating (6.6), no sampling for $f_h(\mathbf{x})|\mathcal{D}, \overline{f}_h(\tilde{\mathbf{x}})$ is needed and the integration only needs to be evaluated once, leading to a much cheaper computation compared to

(6.4). Furthermore, while (6.6) seems to involve an updated GPR conditioning on $\{\mathcal{D}, \overline{f}_h(\tilde{\mathbf{x}})\}$, the relevant quantities can be efficiently computed using the currently available GPR conditioning on $\mathcal{D}$:

$$\mathbb{E}(f_h(\mathbf{x})|\mathcal{D}, \overline{f}_h(\tilde{\mathbf{x}})) = \mathbb{E}(f_h(\mathbf{x})|\mathcal{D}) + \Big(\frac{\mathrm{cov}(f_h(\mathbf{x}), f_h(\tilde{\mathbf{x}})|\mathcal{D})}{\mathrm{var}(f_h(\tilde{\mathbf{x}})|\mathcal{D})} * \big(\overline{f}_h(\tilde{\mathbf{x}}) - \mathbb{E}(f_h(\tilde{\mathbf{x}})|\mathcal{D})\big)\Big)$$

$$= \mathbb{E}(f_h(\mathbf{x})|\mathcal{D}), \tag{6.7}$$

$$\mathrm{var}\big(f_h(\mathbf{x})|\mathcal{D}, \overline{f}_h(\tilde{\mathbf{x}}))\big) = \mathrm{var}(f_h(\mathbf{x})|\mathcal{D}) - \frac{\mathrm{cov}(f_h(\mathbf{x}), f_h(\tilde{\mathbf{x}})|\mathcal{D})^2}{\mathrm{var}(f_h(\tilde{\mathbf{x}})|\mathcal{D})}. \tag{6.8}$$

Up to this point, the algorithm for single-fidelity method can be considered complete, and one simply needs to find $\tilde{\mathbf{x}}$ to minimize (6.5). However, for the purpose of convenience in developing the bi-fidelity method later, it is more desirable to formulate an equivalent acquisition through the reduction of the variance, i.e., the benefits, of adding a hypothetical sample. This can be expressed as

$$B(\tilde{\mathbf{x}}) = U(\mathcal{D}) - U(\mathcal{D}, \overline{f}_h(\tilde{\mathbf{x}})), \tag{6.9}$$

where $U(\mathcal{D})$ is defined as (6.5) conditioning on $\mathcal{D}$ only. The next-best sample can then be selected through the solution of an optimization problem.

$$\mathbf{x}^* = \mathrm{argmax}_{\tilde{\mathbf{x}} \in \mathbb{R}^d} B(\tilde{\mathbf{x}}), \tag{6.10}$$

which can be directly solved using standard global optimization methods, e.g., multiple-starting L-BFGS-B quasi-Newton method [100] used in our study. The optimization (6.10) is repeated for each sequential sample until reaching a user-defined number of samples n_{lim} which needs to be practically chosen balancing the computational budget and required accuracy of the result.

We finally summarize the full algorithm in Algorithm **5**.

Algorithm 5 Single-fidelity method for CAV safety analysis

Require: Number of initial samples n_{init}, limit of number of samples n_{lim}
Input: Initial dataset $\mathcal{D} = \{\mathbf{X}, \mathbf{Y}\}$
   Initialization $n_{total} = n_{init}$
   while $n_{total} < n_{lim}$ **do**
     1. Train the surrogate model with $\mathcal{D}$ to obtain (6.3)
     2. Solve the optimization (6.10) to find the next-best sample $\boldsymbol{x}^*$
     3. Evaluate the function f_h to get $f_h(\mathbf{x}^*)$
     4. Update the dataset $\mathcal{D}$ with $\mathbf{X} = \mathbf{X} \cup \{\boldsymbol{x}^*\}$ and $\mathbf{Y} = \mathbf{Y} \cup f_h(\mathbf{x}^*)$
     5. $n_{total} = n_{total} + 1$
   end while
Output: Compute P_a according to (6.1) based on the surrogate model (6.3)

6.2.3 Bi-fidelity method

We consider the situation that, in addition to the high-fidelity model f_h, we also have a low-fidelity model f_l with lower computational cost. The model f_l can be considered to provide an approximation to f_h with a relation

$$f_h(\mathbf{x}) = f_l(\mathbf{x}) + d(\mathbf{x}), \tag{6.11}$$

where $d(\mathbf{x})$ is an unknown difference function to be determined.

We further assume that the cost for an evaluation using f_h is c_h, and that for f_l is c_l, with $c_h/c_l > 1$. Since f_l is associated with finite cost, we cannot assume that the full low-fidelity map is available to us, in contrast to the situation in [35, 37]. The adaptive sampling algorithm for this bi-fidelity application is required to find a sequence of samples with optimal fidelity level and location, i.e., $f_i(\mathbf{x})$ with $i = h$ or l and $\mathbf{x}$ varying for each sample. For this purpose, the algorithm for the single-fidelity method needs to be extended in two aspects: (i) construction of the surrogate model through a bi-fidelity Gaussian process; and (ii) a more comprehensive acquisitive function measuring the benefit per computational cost for each sample, allowing the next-best sample to be selected in terms of both fidelity level and sampling position.

6.2.3.1 surrogate model by BFGPR

Bi-fidelity Gaussian process regression (BFGPR) [68] is a direct extension of GPR to infuse bi-fidelity data. Given a dataset $\mathcal{D} = \{\mathbf{X}, \mathbf{Y}\}$ consisting of two levels of model outputs $\mathbf{Y} = \{\mathbf{Y}_h, \mathbf{Y}_l\}$ at input positions $\mathbf{X} = \{\mathbf{X}_h, \mathbf{X}_l\}$, the purpose of the bi-fidelity Gaussian process is to learn the underlying relation $f_{h,l}(\mathbf{x})$ from $\mathcal{D}$. This can be achieved through an auto-regressive scheme, which models $f_h(\mathbf{x})$ in (6.11) by two independent Gaussian processes $f_l(\mathbf{x}) \sim \mathcal{GP}(0, k_l(\mathbf{x}, \mathbf{x}'))$ and $d(\mathbf{x}) \sim \mathcal{GP}(0, k_d(\mathbf{x}, \mathbf{x}'))$. The posterior prediction $f_{h,l}(\mathbf{x})$ given the dataset $\mathcal{D}$ can then be derived as a Gaussian process:

$$\begin{bmatrix} f_h(\mathbf{x}) \\ f_l(\mathbf{x}') \end{bmatrix} \mid \mathcal{D} \sim \mathcal{N}\Big(\mathbb{E}\big(\begin{bmatrix} f_h(\mathbf{x}) \\ f_l(\mathbf{x}') \end{bmatrix} \mid \mathcal{D}\big), \mathrm{cov}\big(\begin{bmatrix} f_h(\mathbf{x}) \\ f_l(\mathbf{x}') \end{bmatrix} \mid \mathcal{D}\big)\Big), \tag{6.12}$$

with posterior mean and covariance detailed in §C.1.

We note that $f_h(\mathbf{x})|\mathcal{D}$ in (6.12), as the major prediction in BFGPR, provides the high-fidelity function infusing both high and low fidelity samples $\{\mathbf{X}, \mathbf{Y}\}$ (instead of only $\{\mathbf{X}_h, \mathbf{Y}_h\}$). This is achieved, intuitively, through the two Gaussian processes on $f_l(\mathbf{x})$ and $d(\mathbf{x})$ which rely on all data. The prediction $f_h(\mathbf{x})|\mathcal{D}$ will be used as the surrogate model for the computation of P_a and the development of acquisition function.

6.2.3.2 bi-fidelity acquisition function

In the bi-fidelity context, the next-best sample needs to be determined in terms of both its location and fidelity level. Given a total cost budget, the principle to select the next-best sample is to maximize its benefit per cost. Accordingly, we consider the optimization of an acquisition function which captures both the benefit and cost of a sample $\tilde{\mathbf{x}}$:

$$\mathbf{x}^*, i^* = \mathrm{argmax}_{\tilde{\mathbf{x}} \in \mathbb{R}^d, i \in \{h,l\}} B_i(\tilde{\mathbf{x}})/c_i. \tag{6.13}$$

Following the formulations in the single-fidelity problem, the benefit of adding an i-fidelity hypothetical sample at $\tilde{\mathbf{x}}$, $B_i(\tilde{\mathbf{x}})$, can be expressed as

$$B_i(\tilde{\mathbf{x}}) = U(\mathcal{D}) - U(\mathcal{D}, \overline{f}_i(\tilde{\mathbf{x}})), \quad i = h, l, \tag{6.14}$$

with

$$U(\mathcal{D}, \overline{f}_i(\tilde{\mathbf{x}})) = \int \mathrm{var}^{\frac{1}{2}}\Big(\mathbf{1}_\delta\big(f_h(\mathbf{x})|\mathcal{D}, \overline{f}_i(\tilde{\mathbf{x}})\big)\Big) p_{\mathbf{x}}(\mathbf{x})\mathrm{d}\mathbf{x}. \tag{6.15}$$

The computation of (6.15) can be conducted following (6.7) and (6.8) adapted to the bi-fidelity context using the BFGPR surrogate model. In solving (6.13) as a combined discrete and continuous optimization problem, we first find the optimal location $\mathbf{x}$ for each fidelity i, i.e., $\mathbf{x}_i^* = \mathrm{argmax}_{\tilde{\mathbf{x}} \in \mathbb{R}^d}\ B_i(\tilde{\mathbf{x}})$ for $i = h, l$, then we compare the benefit per cost $B_i(\mathbf{x}_i^*)/c_i$ between $i = h$ and $i = l$ and find the optimal fidelity level i^*, i.e., $i^* = \mathrm{argmax}_{i \in \{h,l\}}\ B_i(\mathbf{x}_i^*)/c_i$. We further remark that this idea of maximizing benefit per cost has been systematically tested with a different benefit function $B_i(\tilde{\mathbf{x}})$ in §5.

We finally summarize the full algorithm in Algorithm **6**.

Algorithm 6 Bi-fidelity method for CAV safety analysis

Require: Number of initial samples $\{n_h^{init}, n_l^{init}\}$, cost of each fidelity model $\{c_h, c_l\}$, total cost budget c_{lim}
Input: Initial dataset $\mathcal{D} = \{\mathbf{X}, \mathbf{Y}\}$ with $\mathbf{X} = \{\mathbf{X}_h, \mathbf{X}_l\}$ and $\mathbf{Y} = \{\mathbf{Y}_h, \mathbf{X}_l\}$
    Initialization $c_{total} = n_h^{init}\, c_h + n_l^{init}\, c_l$
    while $c_{total} < c_{lim}$ **do**
        1. Train the surrogate model with $\mathcal{D}$ to obtain (6.12)
        2. Solve the optimization (6.13) to find the next-best sample $\{i^*, \boldsymbol{x}^*\}$
        3. Evaluate the i^*-fidelity function to get $f_{i^*}(\mathbf{x}^*)$
        4. Update the dataset $\mathcal{D}$ with $\mathbf{X}_{i^*} = \mathbf{X}_{i^*} \cup \{\boldsymbol{x}^*\}$ and $\mathbf{Y}_{i^*} = \mathbf{Y}_{i^*} \cup \{f_{i^*}(\mathbf{x}^*)\}$
        5. $c_{total} = c_{total} + c_{i^*}$
    end while
Output: Compute the P_a according to (6.1) based on the surrogate model (6.12)

6.3 Cut-in case analysis

In this section, we demonstrate the application of our proposed (single and bi-fidelity) methods to the cut-in problem, starting with a more detailed description of the setup of the case. Since our method is new to the general reliability analysis field, we also document its favorable performance for two widely-used benchmark problems in reliability analysis in Appendix D.3.

6.3.1 Case setup

The cut-in situation is illustrated in figure 6.1 where a BV makes a line change in front of a CAV. We assume that the BV moves in a constant speed $u_{BV} = 20m/s$ after the cut-in moment, so that (given the CAV model) the performance of the CAV only depends on $\mathbf{x} = (R_0, \dot{R}_0)$, the initial range and range rate at the cut-in moment $t = 0$ (time t in unit of seconds hereafter). The probability of $\mathbf{x}$ is generated from the naturalistic driving data (NDD) of the Safety Pilot Model Deployment (SPMD) at the University of Michigan [14]. A total number of 414,770 qualified cut-in events are analyzed with joint distribution of $\mathbf{x}$ shown in figure 6.2(a).

The model output for the problem is the minimum range between two vehicles during their speed adjustment process for $t \geq 0$. In this chapter, we use the the Intelligent Driving Model (IDM) which describes the speed of CAV by an ordinary differential equation

$$\frac{du_{cav}(t)}{dt} = \alpha\Big(1 - \big(\frac{u_{cav}(t)}{\beta}\big)^c - \big(\frac{s(u_{cav}(t), \dot{R}(t))}{R(t) - L}\big)^2\Big), \tag{6.16}$$

$$s(u_{cav}(t), \dot{R}(t)) = s_0 + u_{cav}(t)T + \frac{u_{cav}(t)\dot{R}(t)}{2\sqrt{\alpha b}}, \tag{6.17}$$

where α, β, c, s_0, L, b, and T are constant parameters and values in [115] are used

here[2]. We integrate (6.16) in time using forward Euler method starting from initial condition $u_{cav}(t = 0) = u_{BV} - \dot{R}_0$, and accordingly find the range R and range rate $\dot{R}$ for $t \geq 0$. In addition, we constrain the velocity and acceleration of the CAV to be $2 \leq u_{cav}(t) \leq 40 \, m/s$ and $-4 \leq du_{cav}(t)/dt \leq 2 \, m/s^2$. The minimum range R obtained for $0 \leq t \leq 10$ is taken as the model output. We use a number of different time resolutions in integrating (6.16), with the result obtained for $\Delta_t = 0.2$ as the high-fidelity model output (f_h) that is plotted in figure 6.2(b) for visualization. The results for $\Delta_t = 0.5, 1, 2,$ and 5 are taken as the outputs from low-fidelity models (f_l) with different fidelity levels, i.e., we will consider the bi-fidelity context as f_h plus one of the low-fidelity options f_l. The cost ratio c_h/c_l is inversely proportional to the ratio of Δ_t's.

In order to evaluate the performance of our methods in §6.3.2, we compute accurate (ground truth) values of P_a for both $\delta = 0$ and $\delta = 3$ (see (6.2)) according to high-fidelity model f_h for all 414,770 available events. The former will be used for validating the single-fidelity method (so that its performance can be compared to existing cases in literature) and the latter for the bi-fidelity method (to have a considerable difference between the results from f_h and f_l).

6.3.2 Results

6.3.2.1 single-fidelity results

We first apply the single-fidelity method (Algorithm **5**) to the context where only f_h (IDM with $\Delta_t = 0.2$) is available to us. The computation starts from 16 initial random samples, with the following 104 adaptive samples obtained from Algorithm **5**. The results of P_a (estimated from (6.1) in the sampling process) as a function of the number of samples are shown in figure 6.3 together with the 10% error bounds of the ground truth. Since the value of P_a depends on the locations of the initial samples,

---

[2] $\alpha = 2$, $\beta = 18$, $c = 4$, $s_0 = 2$, $L = 4$, $b = 3$, and $T = 1$.
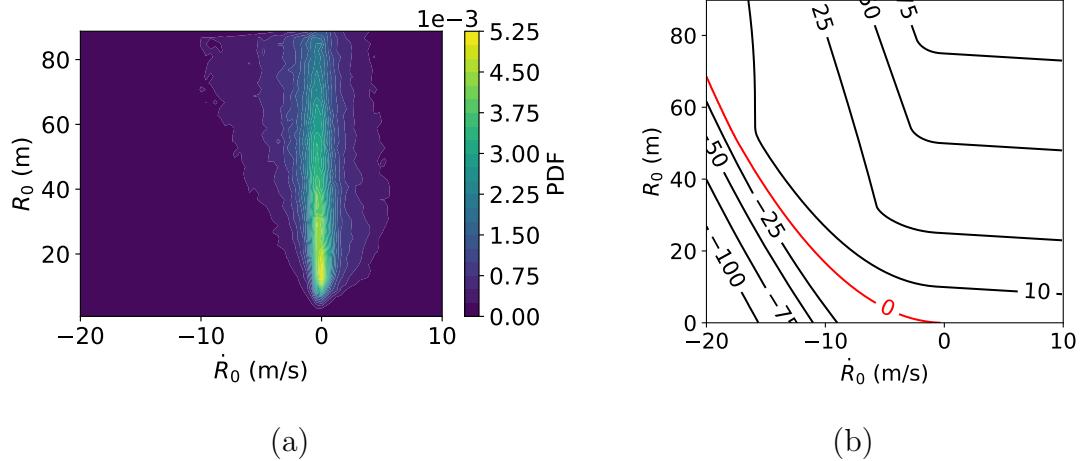
Figure 6.2: (a) probability distribution and (b) output from f_h for input parameters R_0 and $\dot{R}_0$. The limiting state $\{\mathbf{x} : f_h(\mathbf{x}) = 0\}$ is marked in (b) by a red line.

we quantify its uncertainty by plotting both the median value as well as 15% and 85% percentiles in figure 6.3 obtained from 200 applications of our algorithm starting from different initial samples. The percentile concept is used here because the distribution of P_a over different experiments is not guaranteed to be Gaussian, and the 15% and 85% percentiles are used for the convenience of a fair comparison with other results discussed below.

From figure 6.3 we see that it only takes 83 samples (or 67 adaptive samples) for the upper and lower percentiles to converge into the 10% error bounds of the ground truth. In comparison, to reach convergence with a similar criterion[3], it takes 121 samples for the importance sampling method presented in earlier work [34]. It should also be emphasized that the method in [34] requires a pre-known low-fidelity map (e.g., f_l with negligible cost) to guide the proposal distribution in importance sampling. This extra component is not required at all in our method. Therefore, for this validation case, it can be concluded that our approach takes about 2/3 number

---

[3]Our convergence criterion means that within every 100 experiments, 70 of them provide results within 10% error bounds. This is, by definition, equivalent to the relative half-width of the 70% confidence interval falling below 0.1. The latter, as described in figure 9 of [34] (presented in terms of ± 2 standard deviation which is close to two times of the 70% confidence interval for Gaussian estimation in importance sampling), takes 121 samples.

115
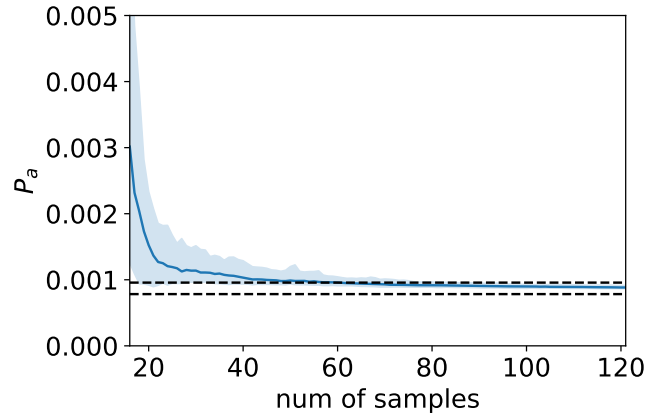
Figure 6.3: Results of P_a from the single-fidelity method, presented by the median value (—) as well as the 15% and 85% percentiles (shaded region) from 200 experiments. The true solution of P_a (- - -) is shown in terms of the 10% error bounds.
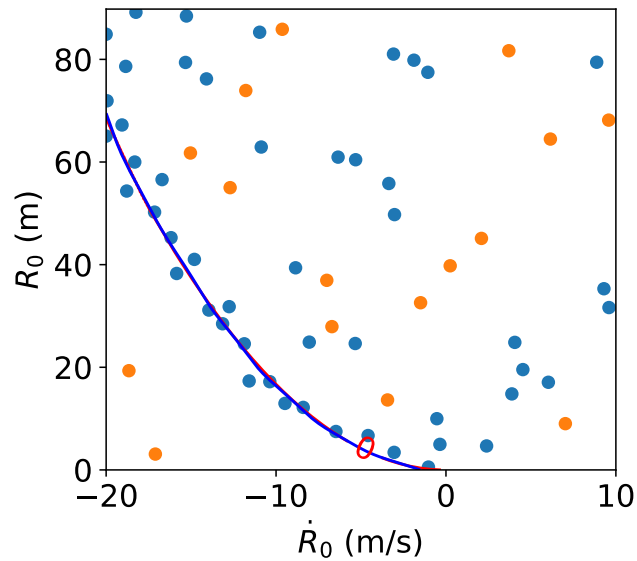
Figure 6.4: Positions of 16 initial samples (•) and 50 adaptive samples (•) from a typical experiment of our method, as well as the learned limiting state (—) compared to the exact one (—).

of samples to achieve the same accuracy as [34] based on the current criterion, and can be conducted in a much simpler setting. Another possible source of comparison is [134], which performs the adaptive sampling with a simplified/empirical acquisition function, without needing a low-fidelity model, to a slightly more complicated 3D car-following problem. It takes $O(500)$ samples to obtain convergent result of accident rate, which is based on one experiment result without analyzing the uncertainty bounds as we do here. It is desirable to apply our method to the same case, but the information provided in [134] is not sufficient for us to do so (e.g., no code or input probability data is provided).

Finally, we plot in figure 6.4 the sampling positions in the input space $\mathbf{x}$ for a typical case out of the 200 experiments. After 16 initial random samples, we see that most adaptive samples are located close to the limiting state to better resolve P_a. The limiting state $\{\mathbf{x} : f_h(\mathbf{x}) = 0\}$ estimated from the GPR constructed by only 50 adaptive samples is also included in the figure to show its proximity to the true state.

6.3.2.2 Bi-fidelity results

We next consider the bi-fidelity application where the high-fidelity model f_h is used together with a low-fidelity model f_l. Different choices of f_l are considered, which are obtained from IDM model with coarser time resolutions of $\Delta_t = 0.5, 1, 2$, and 5, in contrast to $\Delta_t = 0.2$ for f_h. As shown in figure 6.5, these low-fidelity models provide different levels of approximation to f_h, especially regarding the limiting state $\{\mathbf{x} : f_h(\mathbf{x}) = 3\}$ considered here. While the f_l of $\Delta_t = 0.5$ provides a close estimation of the limiting state, the f_l of $\Delta_t = 5$ provides a poor estimation even including an extra region that leads to false accidents. The values of P_a computed solely by different choices of f_l are listed in Table 6.1, which shows $8 - 217\%$ relative difference with the ground-truth value (computed by f_h) for f_l with $\Delta_t = 0.5 - 5$. In particular, results obtained with $\Delta_t = 5$ show an abnormally large P_a because of the false-
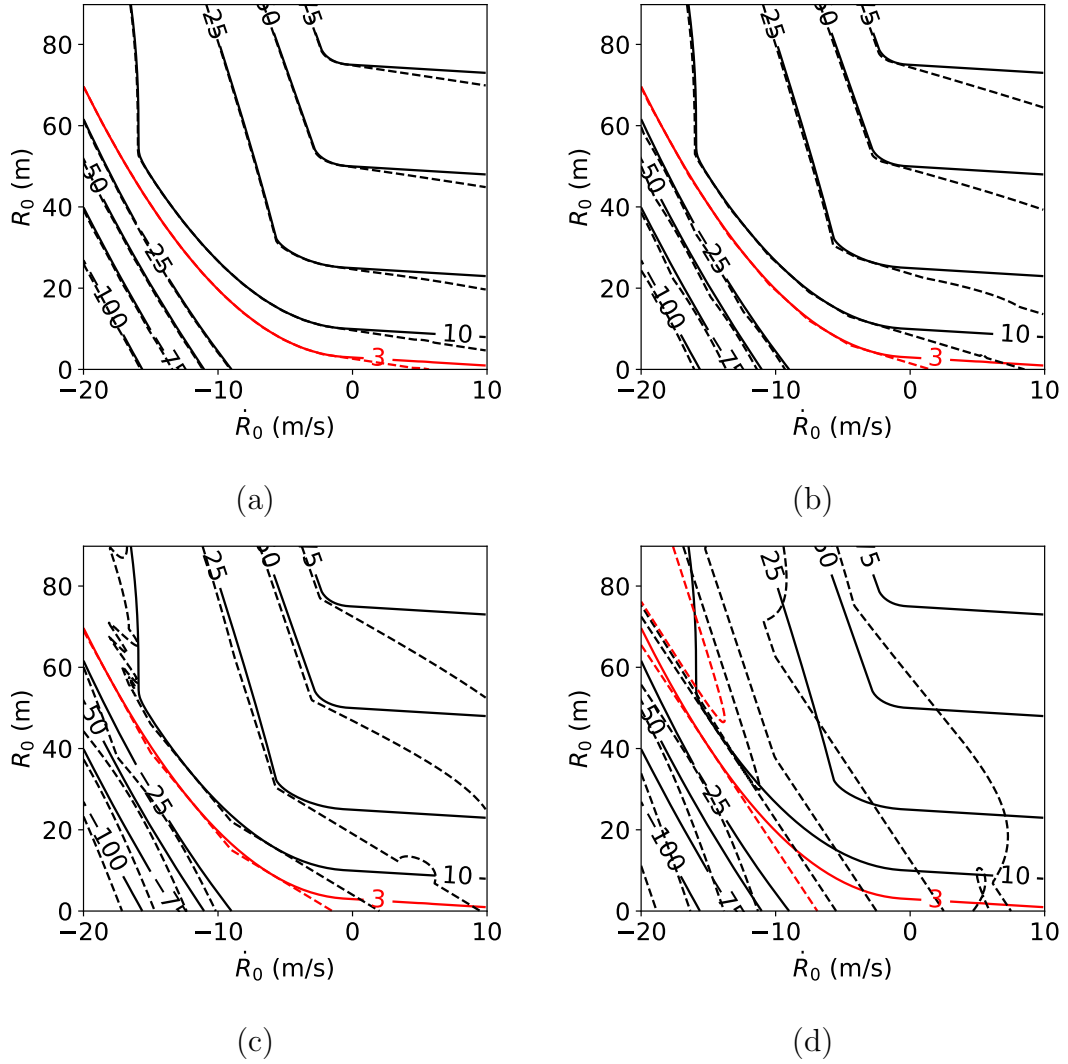
Figure 6.5: Output contour from f_l (- - -) of (a) $\Delta_t = 0.5$, (b) $\Delta_t = 1$, (c) $\Delta_t = 2$, (d) $\Delta_t = 5$ compared with contour from f_h (—) of $\Delta_t = 0.2$. The limiting states $\{\mathbf{x} : f_{h,l}(\mathbf{x}) = 3\}$ of f_h and f_l are respectively marked by (—) and (- - -).

(a)

(b)

(c)

(d)
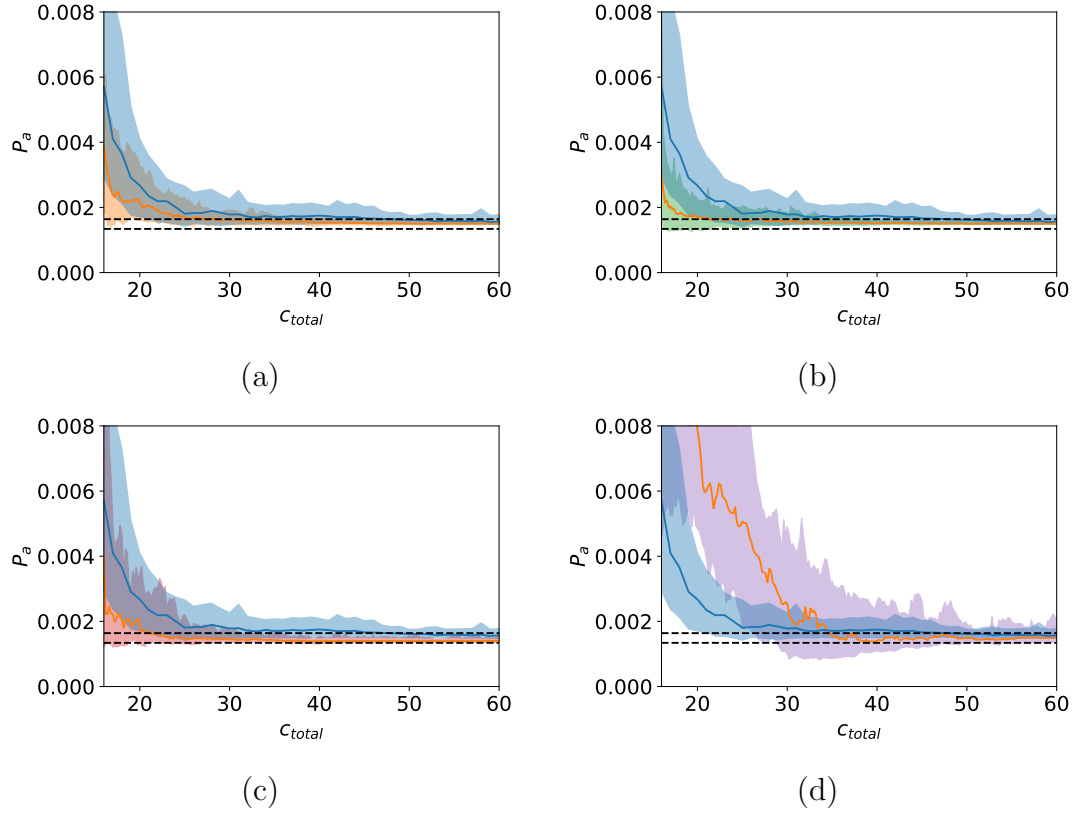
Figure 6.6: Results of P_a from the single-fidelity method with f_h (—) and bi-fidelity method with both f_h and f_l ($\Delta_t = 0.5$: (—) in (a); $\Delta_t = 1$: (—) in (b), $\Delta_t = 2$: (—) in (c);, $\Delta_t = 5$: (—) in (d)), presented by the median value (solid lines) as well as the 15% and 85% percentiles (shaded region) from 50 experiments. The true solution of P_f (- - -) is shown in terms of the 10% error bounds.

Table 6.1: Accident rates computed by models with different time resolutions

Δ_t	0.2	0.5	1	2	5
P_a	0.001489	0.001375 (8%)[1]	0.001202 (19%)	0.001114 (25%)	0.004725 (217%)

[1] Relative differences are computed with respect to the result of $\Delta_t = 0.2$.

Table 6.2: Summary of the performance of single-fidelity (No. 1) and bi-fidelity (No. 2-5) cases

No.	$\Delta_t(f_h)$	$\Delta_t(f_l)$	c_h/c_l	n_l/n_h[1]	percentile[2]	median[2]
1	0.2	—	—	-	82	47
2	0.2	0.5	2.5	6.63	42	30
3	0.2	1	5	6.22	38	22
4	0.2	2	10	8.15	39	23
5	0.2	5	25	4.1	> 60	35

[1] n_l/n_h gives the ratio of low and high-fidelity sampling numbers in the adaptive sampling process.
[2] "percentiles" and "median" columns respectively give the units of cost for the (15% and 85%) percentiles and median of P_a to converge to the 10% error bounds of the ground truth.

accident region in figure 6.5(d). In terms of computational cost, we consider that one application of f_h takes one unit of cost, and one application of f_l takes $0.2/\Delta_t$ units of cost.

The results of our bi-fidelity method with respect to total cost c_{total} in the sampling process are shown in figure 6.6 for four cases, with (a)-(d) corresponding to f_l of $\Delta_t = 0.5, 1, 2$, and 5. In each bi-fidelity case, the application starts with 8 high-fidelity samples (8 units of cost) and 8 c_h/c_l low-fidelity samples (8 units of cost) as the initial dataset, followed by adaptive samples with 44 units of cost. Also shown in figure 6.6 are single-fidelity results using f_h, starting from 16 high-fidelity random samples (i.e., 16 units of cost), as well as 10% error bounds of the ground-truth P_a. Similar to §6.3.2.1, the results are presented including the median value as well as

15% and 85% percentiles obtained from 50 applications. We see that in general, the bi-fidelity approach converges much faster than the single-fidelity approach, especially regarding the cases with f_l constructed by $\Delta_t = 0.5, 1$, and 2 in (a), (b) and (c). For these cases, the percentiles of P_a converge into the 10% error bound in O(40) units of cost, whereas the single-fidelity approach takes more than 60 units (82 units upon extended test not shown in the figure) to reach the same convergence criterion. For the bi-fidelity case with f_l of $\Delta_t = 5$, the percentiles of P_a converge with a comparable (but slightly slower) speed relative to the single-fidelity case within 60 units of cost as seen in (d). However, even in this case, the median of P_a in the bi-fidelity approach converges to the 10% error bound in O(35) samples that is faster than O(50) samples in the single-fidelity approach. It must also be kept in mind that for this bi-fidelity case, P_a starts with a much less accurate value at the beginning of the adaptive sampling process, due to the initial $8c_h/c_l = 200$ low-fidelity samples that are very misleading in constructing the initial surrogate model. Shall a different allocation of high/low fidelity samples are used for the initial dataset, it is possible to further improve the performance of this bi-fidelity application.

We further summarize the performance of all 5 cases (one single-fidelity and four bi-fidelity cases) in Table 6.2. In addition to the exact units of cost consumed for convergence, a notable information is provided by the column of n_l/n_h that gives the ratio of low and high-fidelity sampling numbers in the adaptive sampling process. It is clear that n_l/n_h does not monotonically increase with the increase of c_h/c_l, i.e., the algorithm does not select more low-fidelity samples just because they are cheaper but instead considers the benefit per cost of each sample formulated in (6.13). This is most evident when c_h/c_l increases from 10 to 25 and meanwhile, n_l/n_h drops from 8.15 to 4.1, mainly because a sample by f_l with $\Delta_t = 5$ does not provide much useful information to the computation of P_a.

Finally, we plot in figure 6.7 the positions of the high and low-fidelity (initial and

121
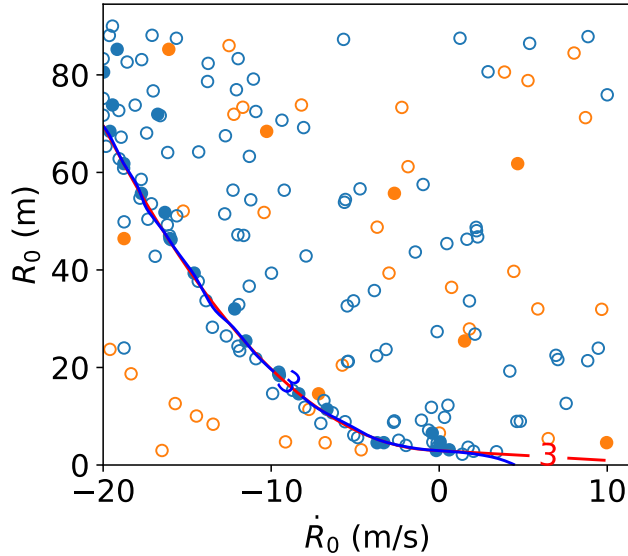
Figure 6.7: Positions of initial high-fidelity samples (•), initial low-fidelity samples (○), adaptive high-fidelity samples (•) and adaptive low-fidelity samples (○) from a typical experiment with f_l of $\Delta_t = 1$, as well as the learned limiting state (——) compared to the exact one (——).

adaptive) samples for a typical experiment of bi-fidelity algorithm for the case with f_l of $\Delta_t = 1$. It can be observed that while the adaptive low-fidelity samples are scattered in the space of $\mathbf{x}$, the adaptive high-fidelity samples are most concentrated at the limiting state $\{\mathbf{x} : f_{h,l}(\mathbf{x}) = 3\}$. This is because the limiting state represents the location where the difference between f_h and f_l critically affects the computed value of P_a, and it necessarily takes high-fidelity samples to resolve this important region. The resolved limiting state at 60 units of cost in adaptive sampling is also shown in the figure to demonstrate its proximity to the ground truth.

6.4   Summary

In this chapter, we develop an adaptive sampling framework to efficiently evaluate the accident rate P_a of connected and automated vehicles (CAVs) in scenario-based tests. The core components of our approach include a surrogate model by Gaussian process

regression and a novel acquisition function to select the next-best sample maximizing its benefit (uncertainty reduction) to P_a. The framework can be applied to both single-fidelity and bi-fidelity contexts, where the latter involves a low-fidelity model to help construct the function of the CAV performance. Accordingly, for the latter the two components of our approach need to be extended to the bi-fidelity Gaussian process regression and an acquisition function, allowing the choice of both fidelity level and sampling location. We demonstrate the effectiveness of the framework in a widely-considered two-dimensional cut-in problem, with the low and high-fidelity models constructed by IDM with different resolutions in time. It is shown that the single-fidelity method already outperforms the state-of-the-art method for the same problem, and the bi-fidelity method further accelerates the convergence by a factor of about two (i.e., with half of the computational cost) for most settings that are tested.

We finally remark that the method we develop in this chapter is new to the entire field of reliability analysis according to our knowledge, and its application to other fields may prove equally fruitful. For example, it may be applied to evaluate the ship capsizing probability in ocean engineering [49, 48], structural safety analysis [59, 149], probability of extreme pandemic spikes for public health [109] and many other physical, engineering and societal problems. Within the CAV field, our method is certainly not limited to the IDM models used in this paper as demonstrations. It can be connected to a broad range of CAV evaluation tools across on-road tests, closed-facility tests, simulations based on various kinds of simulators (e.g. Google/Waymo's Car-Craft9 [67], Intel's CARLA6 [30], Microsoft's Air-Sim7 [123], NVIDIA's Drive Constellation [101]). Among these examples, we would like to emphasize the possible benefit of our method to the augmented-reality test environment combing a real test vehicle on road and simulated background vehicles [36]. Due to the bi-fidelity capability of our method, it also becomes beneficial to combine two different tools in the above list to further improve the testing efficiency.

CHAPTER 7

A Generalized Likelihood-Weighted Acquisition for Rare-Event Statistics

7.1 Introduction

In §5, we introduced the likelihood-weighted (LW) algorithm [117, 118, 19] for quantifying rare-event statistics and extended it to multi-fidelity scenarios. The idea of using LW weight has later been adapted to many different applications beyond its original purpose of rare-event statistics evaluation, including rare-event forecasting [109, 116], Bayesian optimization [20], robot path planning [18], multi-arm bandit [157], causal inference [159]. In this chapter, we will take a deep dive into the LW algorithm for its derivation, weaknesses, and corresponding improvements.

Given an ItR function $f(\mathbf{x})$ with known input probability $p_{\mathbf{x}}(\mathbf{x})$, the next optimal sample $\mathbf{x}^*$ in LW algorithm, based existing dataset $\mathcal{D}$, is selected as the one that maximizes an acquisition function

$$acq_{LW}(\mathbf{x}) = \mathrm{var}(f(\mathbf{x})|\mathcal{D})w(\mathbf{x}). \tag{7.1}$$

The LW factor $w(\mathbf{x}) = p_{\mathbf{x}}(\mathbf{x})/p_{\hat{f}}(\hat{f}(\mathbf{x}))$ is the ratio of input probability to predicted output probability with $\hat{f} \equiv \mathbb{E}(f|\mathcal{D})$ the surrogate model. The effectiveness of the LW acquisition can be understood from the LW factor $w(\mathbf{x})$ in (7.1). It is argued

in [117] that due to $w(\mathbf{x})$, the next sample is chosen in favor of $\mathbf{x}$ with larger input probability $p_{\mathbf{x}}(\mathbf{x})$ (thus contributing more to $p_f(f)$) and smaller predicted response probability $p_{\hat{f}}(\hat{f}(\mathbf{x}))$ (thus associated with rare events). Such samples are more likely to contribute more to the rare-event (or tail) portion of the response PDF $p_f(f)$. While this interpretation is plausible, (7.1) is clearly not an optimal sampling criterion. To see this, let us consider any one-dimensional monotonic function $f(x)$, say a logistic function $f(x) = 1/(1 + e^{-x})$ (figure 7.1(a)) and assume no difference between surrogate $\hat{f}(x)$ and ground-truth $f(x)$. The critical weighting factor is now reduced simply to $w(x) = f'(x)$, which peaks at $x = 0$ (figure 7.1(b)). It is clear not only that the input leading to large (and usually rare) responses is not stressed, but also that $w(x)$ has nothing to do with the input and response probability, violating the above claims made in [117]. The failure of (7.1) in the above example lies in the fact that the simple ratio in $w(\mathbf{x})$ is not necessarily optimal in sampling to resolve the tail of $p_f(f)$. Indeed, while larger $p_{\mathbf{x}}(\mathbf{x})$ and smaller $p_{\hat{f}}(\hat{f}(\mathbf{x}))$ helps, there is no guarantee that the optimal form is their direct ratio. In addition, another more severe issue regarding (7.1) is that $\hat{f}$ may have a significant deviation from f, which makes small $p_{\hat{f}}(\hat{f}(\mathbf{x}))$ a poor estimation of the rarity of response. This is especially the case when function $f(\mathbf{x})$ is complex, given the limited number of samples that can be afforded. In such cases, the rare-event statistics provided by sampling through (7.1) may become misleading since regions associated with small $p_f(f(\mathbf{x}))$ may never be explored (if the corresponding $p_{\hat{f}}(\hat{f}(\mathbf{x}))$ is large).

In the following, we propose a generalized LW acquisition targeting the two limitations mentioned above. Our new acquisition takes a generalized form of (7.1) with two additional parameters, by varying which one can achieve (i) optimal deployment of $p(\mathbf{x})$ and $p_{\hat{f}}(\hat{f}(\mathbf{x}))$ in the LW factor, and (ii) a much more effective guidance of sample exploration when $\hat{f}$ is very different from f. The generalized LW acquisition also shares the theoretical property of (7.1) in terms of its derivation from the first
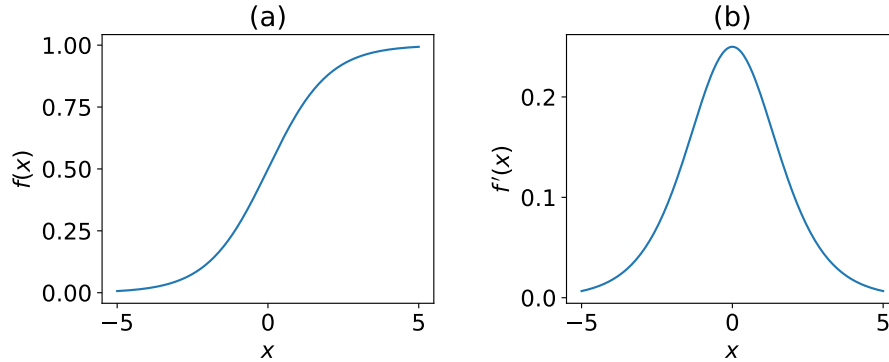
Figure 7.1: Plots of (a) a logistic function $f(x) = 1/(1 + e^{-x})$ and (b) its derivative.

principle. In addition, we point out an acceleration in Monte-Carlo discrete optimization regarding the acquisition, achieving orders of magnitude speedup compared to existing algorithms used in [117, 19, 20]. The superior performance of our new acquisition is consistently demonstrated in a number of test cases, including a stochastic oscillator [18, 118], a pandemic spike model [109] and cases with arbitrary complex functions $f(\mathbf{x})$ generated by kernels. We finally show the application of the new acquisition in an engineering example of quantifying the rare-event roll-motion statistics of a ship in a random sea.

This chapter is adapted from [47]. The Python code for the proposed algorithm, named GPextreme, is available on Github[1].

7.2 Method

7.2.1 Problem setup

We consider an ItR system described by a response function $f(\mathbf{x}) : \mathbb{R}^d \to \mathbb{R}$ with input $\mathbf{x}$ a d-dimensional decision variable over a compact set and response an observable of the system. The input probability $p_{\mathbf{x}}(\mathbf{x})$ is assumed to be known, and our quantity

---
[1]https://github.com/umbrellagong/GPextreme

of interest is the probability density function (PDF) of the response $p_f(f)$ with an emphasis on the tail part (to be precisely defined later in (7.2)). While $p_f(f)$ can be directly evaluated via standard Monte-Carlo method, an accurate resolution of its tail part is extremely expensive considering the expensiveness of system evaluations and the rareness of samples contributing to the PDF tail.

To reduce the computational cost, we make use of surrogate modeling with f approximated by a learned surrogate (regressor) $\hat{f}$, achieved through Gaussian process regression (GPR). Assume we have a dataset $\mathcal{D} = \{\mathbf{X}, \mathbf{Y}\}$ consisting of n inputs $\mathbf{X} = \{\mathbf{x}^i \in \mathbb{R}^d\}_{i=1}^n$ and the corresponding outputs $\mathbf{Y} = \{f(\mathbf{x}^i) \in \mathbb{R}\}_{i=1}^n$. In GPR, the underlining function is inferred as a posterior Gaussian process $f(\mathbf{x})|\mathcal{D} \sim \mathcal{GP}\big(\mathbb{E}(f(\mathbf{x})|\mathcal{D}), \mathrm{cov}(f(\mathbf{x}), f(\mathbf{x}')|\mathcal{D})\big)$ with the mean as the surrogate, i.e., $\hat{f} \equiv \mathbb{E}(f|\mathcal{D})$ (see Appendix A or [113] for detailed formulae). With $\hat{f}$ available, the response PDF can be estimated as $p_{\hat{f}}(f)$ via evaluating $\hat{f}$ on a large number of samples (say standard Monte-Carlo samples), and our objective is to minimize the estimation error defined as (see [95])

$$\epsilon = \int_\Omega \Big| \log p_{\hat{f}}(f) - \log p_f(f) \Big| \mathrm{d}f, \tag{7.2}$$

where the integral is computed over a finite domain $\Omega = \mathrm{supp}(p_f) \cup \mathrm{supp}(p_{\hat{f}})$. We note that the log function in (7.2) acts on the ratio $p_{\hat{f}}(f)/p_f(f)$, which is amplified when $p_f(f)$ is small, i.e., the tail part of the PDF is reached.

Our goal is to construct $\hat{f}$ with a limited number of samples, i.e., to choose the most informative samples to learn $\hat{f}$ which facilitates the convergence of (7.2). To achieve this goal, we use the idea of sequential Bayesian experimental design (or more broadly active learning) where the next sample is selected optimally based on the existing data $\mathcal{D}$. Specifically in the general form, the next-best sample is sequentially determined based on the optimization of an acquisition function:

$$\mathbf{x}^* = \mathrm{argmax}_{\tilde{\mathbf{x}}}\, acq(\tilde{\mathbf{x}}; f(\mathbf{x})|\mathcal{D}), \tag{7.3}$$

with the overall algorithm detailed in Algorithm **7**. We will discuss the form of the acquisition function in (7.3) in this paper, as the core of the algorithm.

Algorithm 7 Bayesian experimental design for rare-event statistics

---

Require: Number of initial samples n_{init} and sequential samples n_{seq}
Input: Initial dataset $\mathcal{D} = \{\mathbf{x}^i, f(\mathbf{x}^i)\}_{i=1}^{n_{init}}$
  Initialization $n = n_{init}$
  while $n < n_{seq} + n_{init}$ **do**
    Train GPR $f(\mathbf{x})|\mathcal{D}$
    Solve (7.3) to find the next-best sample location $\mathbf{x}^{n+1}$
    Implement simulation/experiment to get $f(\mathbf{x}^{n+1})$
    Update the dataset $\mathcal{D} = \mathcal{D} \cup \{\mathbf{x}^{n+1}, f(\mathbf{x}^{n+1})\}$
    $n = n + 1$
  end while
Output: Compute the response PDF $p_{\hat{f}}$ based on the surrogate model $\hat{f}$

7.2.2   Likelihood-weighted acquisition

In using (7.2) as the error metric to guide the next sample, an issue comes up since the true function $f(\mathbf{x})$, thus $p_f(f)$, is unknown. To overcome this issue, [95] proposed an effective proxy to (7.2) as

$$\epsilon_L(\tilde{\mathbf{x}}) = \int |\log p_{f^+|\mathcal{D},\hat{f}(\tilde{\mathbf{x}})}(f) - \log p_{f^-|\mathcal{D},\hat{f}(\tilde{\mathbf{x}})}(f)|\mathrm{d}f, \tag{7.4}$$

where $\tilde{\mathbf{x}}$ is the hypothetical location of the next sample, $p_{f^\pm|\mathcal{D},\hat{f}(\tilde{\mathbf{x}})}(f)$ are PDF bounds generated by upper and lower bounds (say two standard deviations away from the mean) of GPR $f|\mathcal{D}, \hat{f}(\tilde{\mathbf{x}}) \sim \mathcal{GP}\big(\mathbb{E}(f(\mathbf{x})|\mathcal{D}, \hat{f}(\tilde{\mathbf{x}})), \mathrm{cov}(f(\mathbf{x}), f(\mathbf{x}')|\mathcal{D}, \hat{f}(\tilde{\mathbf{x}}))\big)$. However, in minimizing (7.4) with the next sample $\tilde{\mathbf{x}}$, one needs to compute $p_{f^\pm|\mathcal{D},\hat{f}(\tilde{\mathbf{x}})}(f)$ for many $\tilde{\mathbf{x}}$, which is an expensive operation. To avoid this high computational cost, [117] and [118] further constructed an upper bound of ϵ_L (up to a constant factor)

$$\epsilon_{LW}(\tilde{\mathbf{x}}) = \int \mathrm{var}(f(\mathbf{x})|\mathcal{D}, \hat{f}(\tilde{\mathbf{x}})) \frac{p_{\mathbf{x}}(\mathbf{x})}{p_{\hat{f}}(\hat{f}(\mathbf{x}))} \mathrm{d}\mathbf{x}, \tag{7.5}$$

where $w(\mathbf{x}) \equiv p_{\mathbf{x}}(\mathbf{x})/p_{\hat{f}}(\hat{f}(\mathbf{x}))$ is the LW factor with its significance reviewed in §7.1[2].

Unlike the situation in (7.4), the predicted response PDF $p_{\hat{f}}(f)$ only needs to be evaluated once in (7.5).

The derivation of (7.5) assumes that the surrogate $\hat{f}$ is sufficiently close to the true f. Under this assumption, an asymptotic form of ϵ_L can be first derived as (see Appendix E.1 for a summary of the derivation following [95] and [118] but with clarifications of some critical procedures)

$$\epsilon_L(\tilde{\mathbf{x}}) \leq \int \mathrm{std}(f(\mathbf{x})|\mathcal{D}, \hat{f}(\tilde{\mathbf{x}})) \frac{p_{\mathbf{x}}(\mathbf{x})|p'_{\hat{f}}(\hat{f}(\mathbf{x}))|}{p_{\hat{f}}^2(\hat{f}(\mathbf{x}))} \mathrm{d}\mathbf{x}, \qquad (7.6)$$

where std denotes standard deviation. With Cauchy-Schwarz inequality, ϵ_L can be further formulated as (see (3.4) in [118])

$$\epsilon_L(\tilde{\mathbf{x}}) \leq [\int \frac{p_{\mathbf{x}}(\mathbf{x})p'^2_{\hat{f}}(\hat{f}(\mathbf{x}))}{p_{\hat{f}}^3(\hat{f}(\mathbf{x}))} \mathrm{d}\mathbf{x}]^{1/2}[\int \mathrm{var}(f(\mathbf{x})|\mathcal{D}, \hat{f}(\tilde{\mathbf{x}})) \frac{p_{\mathbf{x}}(\mathbf{x})}{p_{\hat{f}}(\hat{f}(\mathbf{x}))} \mathrm{d}\mathbf{x}]^{1/2}. \qquad (7.7)$$

In (7.7), the first term reduces to a constant and the second term squared leads to (7.5). While the above derivation starts from the first principles, (7.5) can also be seen as a continuous counterpart of the *precision* metric in classification in a sense that inputs associated with each predictive output as a whole contribute *equally* to the total error. We detail this interpretation in Appendix E.2.

The next sample can be selected by minimizing (7.5), i.e., to construct the acquisition function in (7.3) as $-\epsilon_{LW}(\tilde{\mathbf{x}})$. Alternatively, a more inexpensive but almost equally effective way (as tested in [19]) is to choose the next sample at $\mathbf{x}$ which maximizes the integrand of (7.5) without using the hypothetical sample $\tilde{\mathbf{x}}$, since getting sample there is supposed to contribute most significantly in reducing (7.5). Under the latter approach, we solve an optimization problem $\mathbf{x}^* = \mathrm{argmax}_{\mathbf{x}}\, acq_{LW}(\mathbf{x})$, with

---

[2]Prior information about extreme events can also be encoded in $w(\mathbf{x})$, e.g., by modifying $p_{\mathbf{x}}(\mathbf{x})$.

the acquisition function constructed as

$$acq_{LW}(\mathbf{x}) = \text{var}(f(\mathbf{x})|\mathcal{D})\frac{p_{\mathbf{x}}(\mathbf{x})}{p_{\hat{f}}(\hat{f}(\mathbf{x}))}, \tag{7.8}$$

which is exactly (7.1) in §7.1. (7.8) can also be considered as the standard uncertainty sampling acquisition with a weighting factor inspired by (7.5). We further note that another advantage of (7.8) over (7.5) is that if neural networks are used to construct the surrogate model, (7.5) involves excessive computational cost since $\text{var}(f(\mathbf{x})|\mathcal{D}, \hat{f}(\tilde{\mathbf{x}}))$ does not have an analytical formulation as in GPR and needs to be re-trained for each $\tilde{\mathbf{x}}$ [109, 54]. Given the simplicity and effectiveness of (7.8), we will establish most of our analysis based on (7.8), but will discuss the derivation leading to (7.5) (that inspires (7.8)) in the subsequent parts of the paper.

7.2.3 Proposed generalization of the likelihood-weighted acquisition

The LW acquisition (7.8) outperforms the other existing acquisitions in rare-event statistics quantification in several cases presented in [19, 117, 118], and it has a theoretical foundation outlined in §7.2.2. However, the insufficiency of (7.8) discussed in §7.1 (e.g., discussion regarding figure 7.1) is also intuitively true. How can we reconcile these two views on the LW acquisition (7.8)?

In fact, the two limitations discussed in §7.1 roots exactly from the derivation of (7.5). First, from (7.6) the Cauchy-Schwarz inequality can be applied in many different ways, and (7.7) is not necessarily the unique form of the upper bound. To be more specific, the integrand of (7.6) can be distributed into two factors in many different ways, resulting in the fact that the second term in (7.7) may yield arbitrary powers on $p_{\mathbf{x}}(\mathbf{x})$ and $p_{\hat{f}}(\hat{f}(\mathbf{x}))$. Indeed, from this derivation itself, any of these resultant forms can serve as an upper bound to (7.6) and none of them is unique.

This is consistent with our intuitive argument in §7.1 that the direct ratio between $p_{\mathbf{x}}$ and $p_{\hat{f}}$ is not necessarily the optimal. Second, the derivation leading to (7.6) relies on the assumption that $\hat{f} \approx f$. As discussed in §7.1, this is not necessarily true especially for complex function f, considering limited number of samples that can be placed. In case that $\hat{f}$ misses a region of large (usually rare) responses of interest, this region may never get explored by using (7.8) since the associated $p_{\hat{f}}$ is not small.

To address the above two limitations, we propose a generalization of LW acquisition (7.8), in the form of

$$acq_{GLW}(\mathbf{x}) = \mathrm{var}(f(\mathbf{x}|\mathcal{D}))\big(w_G(\mathbf{x}, t, 0) + w_G(\mathbf{x}, t, \alpha) + w_G(\mathbf{x}, t, -\alpha)\big), \qquad (7.9)$$

where

$$w_G(\mathbf{x}, t, \alpha) = p_{\mathbf{x}}(\mathbf{x})/p_{\hat{f}_\alpha}(\hat{f}_\alpha(\mathbf{x}))^t, \qquad (7.10)$$

$$\hat{f}_\alpha(\mathbf{x}) = \hat{f}(\mathbf{x}) + \alpha \, \mathrm{std}(f(\mathbf{x})|\mathcal{D}), \qquad (7.11)$$

which contains two additional parameters t and α. Regarding the first limitation, the parameter t controls the level of emphasis on small $p_{\hat{f}}$ in the LW factor, and provides flexibility in balancing the need to sample at large-$p_{\mathbf{x}}$ and small-$p_{\hat{f}}$ region. With $t = 1$, the first term in (7.9) reduces to the original LW acquisition (7.8). For $t > 1$ and $t < 1$, (7.9) places respectively more and less emphasis on small $p_{\hat{f}}$, i.e., the rarity of predicted response. We note that larger value of t (i.e., more emphasis on small $p_{\hat{f}}$) does not mean better performance, since the performance needs to be eventually judged by the error metric (7.2). Regarding the second limitation, the second and third terms in (7.9) provide more exploration power for the acquisition function. If $\hat{f}$ misses some large responses at $\mathbf{x}$, the GPR at these $\mathbf{x}$ is certain to be associated with large variance, so that either f_α or $f_{-\alpha}$ captures the large responses and plays an active role in (7.9).

While the inclusion of t and α in (7.9) provides flexibility in addressing the limitations in (7.8), the optimal values of these parameters cannot be theoretically determined (at least from the theoretical framework reviewed in this paper) and must depend on specific features of the function $p_{\mathbf{x}}(\mathbf{x})$ and $f(\mathbf{x})$. Therefore, the optimal t and α values can only be empirically obtained through numerical tests as we will discuss in §7.3.

7.2.4 Acceleration in optimization of the acquisitions

In solving the optimization problem regarding the acquisition (7.8) (and thus the generalized form (7.9)), the Monte Carlo discrete optimization (MCDO) method has been considered as an effective approach, which is tested to be superior to gradient-based method due to the non-convexity of the acquisitions in many cases [109]. In MCDO method, a large number of candidate samples located at $\mathbf{X}_{mc} \in \mathbb{R}^{n_{mc}*d}$ (usually from space-filling L-H sampling) are created, with $n_{mc} \gg n$ (with n the number of samples in $\mathcal{D}$), from which one selects the candidate that returns a maximum in the acquisition. Such procedure allows all $acq(\mathbf{X}_{mc})$ to be evaluated in one vector operation that saves much computational cost than other global or gradient-based optimization methods that rely on iterations. We also note that in optimization regarding (7.5), function evaluation on pre-selected Monte-Carlo samples $\mathbf{X}_{mc}$ is also needed in evaluation of the integral, as conducted in [117, 19, 20]. Therefore, the acceleration method we introduce below applies equally to the optimization problems regarding (7.5), (7.8), and (7.9).

In computing $acq(\mathbf{X}_{mc})$, say with (7.8), one needs to evaluate a new GPR with $\hat{f}(\mathbf{X}_{mc}) = \mathbb{E}(\mathbf{X}_{mc}|\mathcal{D})$ and $\mathrm{var}(\mathbf{X}_{mc}|\mathcal{D})$, with the former needed to calculate the function $p_{\hat{f}}(f)$. In obtaining these quantities, [19, 20] have suggested to apply the recursive formula such that the new GPR can be built recursively with new data $\mathbf{x}_n$ leveraging previous GPR based on $\mathcal{D}_{n-1}$, instead of a brute-force retraining taking all $\mathcal{D}$. In do-

ing so, the previous works argued that the computational cost can be much reduced compared to brute-force retraining. However, in the context of MCDO method, a careful analysis conducted in Appendix E.3 shows that the retraining process (in particular the inverse of covariance on $\mathcal{D}$) only constitutes a very small portion of the total computational cost considering $n_{mc} \gg n$. Therefore, the bottleneck of the computation in fact comes from the prediction step, that is the generation of the covariance matrix and multiplication of matrices involving n_{mc} rows/columns. In order to overcome this major part of the computational cost, we develop a matrix re-grouping technique (that is in analogy to the regrouping technique used in many adjoint methods) and apply the idea of memory-time tradeoff on top of the recursive formula. With details and test cases presented in Appendix E.3, we show that the original computational complexity $O(n_{mc} * n^2)$ (which holds with or without simply applying the recursive formula) can be reduced to $O(n_{mc} * n)$. This is a significant reduction considering $n \gtrsim O(100)$ in many applications.

7.3   Results

In this section, we test the performance of generalized LW acquisition acq_{GLW} in (7.9) with variations of α and t, to show its advantage over acq_{LW}. The test cases are organized as follows: In §7.3.1, we choose two models with simple response functions that were previously used for demonstrating the effectiveness of acq_{LW} in [19, 118] and [109]. We shall show that using acq_{GLW} (especially with appropriate t) achieves additional significant benefits in reducing the error defined in (7.2). In §7.3.2, we use as response functions a large number $O(1000)$ of synthetic functions from realizations of Gaussian processes, with most functions complex with multi-modes. We demonstrate the advantage of acq_{GLW} over acq_{LW} especially with appropriate value of α. In §7.3.3, we consider the application of acq_{GLW} to an engineering problem of evaluating

133

the rare-event statistics of ship motion in a random sea.

7.3.1 Two test cases in existing works

7.3.1.1 Stochastic oscillator

The first case we choose consists of a 2D response function constructed from the solution of a nonlinear oscillator under stochastic excitation, which is studied in [19, 118] and §5.3.2.

In this case, the oscillator equation is formulated as

$$\ddot{u}(t) + \delta\dot{u}(t) + F(u) = \xi(t), \tag{7.12}$$

where $u(t)$ is the state variable, F is a nonlinear restoring force defined as

$$F(u) = \begin{cases} \alpha u & if\ 0 \leq |u| \leq u_1 \\ \alpha u_1 & if\ u_1 \leq |u| \leq u_2 \\ \alpha u_1 + \beta(u - u_2)^3 & if\ u_2 \leq |u| \end{cases} . \tag{7.13}$$

The stochastic process $\xi(t)$, with a correlation function $\sigma_\xi^2 e^{-\tau^2/(2l_\xi^2)}$, is approximated by a two-term Karhunen-Loeve expansion

$$\xi(t) \approx \sum_{i=1}^{2} x_i \lambda_i \phi_i(t), \tag{7.14}$$

with λ_i and $\phi_i(t)$ respectively the eigenvalue and eigenfunction of the correlation function, $\mathbf{x} \equiv (x_1, x_2)$ is a standard normal variable as the input to the system (see figure 7.2(a)), satisfying $p_{\mathbf{x}}(\mathbf{x}) = \mathcal{N}(\mathbf{0}, I_2)$ with I_2 being a 2×2 identity matrix. The values of the parameters are kept the same as those in the existing works[3].

The response of the system is considered as the mean value of $u(t; \mathbf{x})$ in the interval

---

[3]δ=1.5, α=1, β=0.1, u_1=0.5, u_2=1.5, σ_ξ^2=0.1, l_ξ=4.
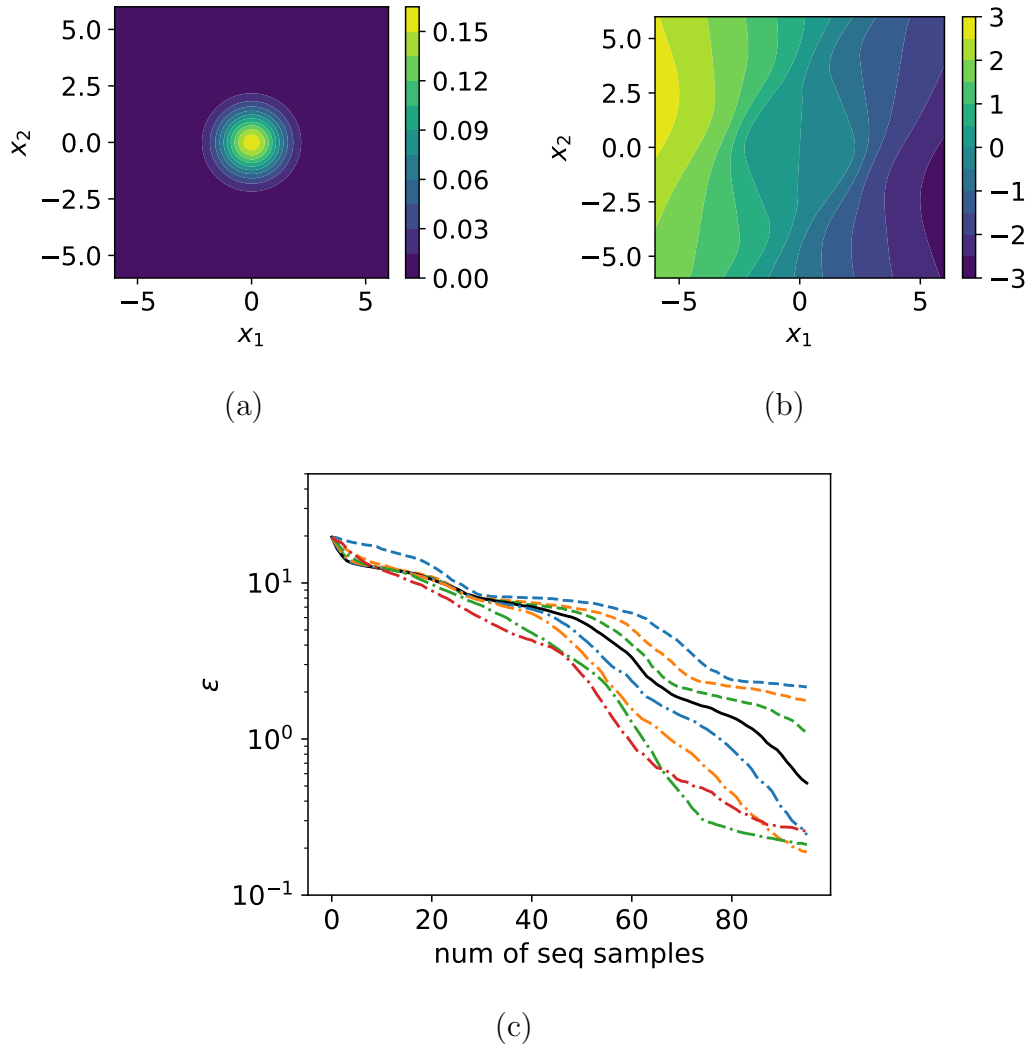
Figure 7.2: (a) input probability distribution and (b) response function of the stochastic oscillator example. (c) error ϵ as functions of sample numbers with $\alpha = 0$ and varying $t = 0.6$ (‑ ‑ ‑), 0.8(‑ ‑ ‑), 0.9(‑ ‑ ‑), 1(——), 1.1(‑·‑·‑), 1.2(‑·‑·‑), 1.4(‑·‑·‑), 1.6(‑·‑·‑).
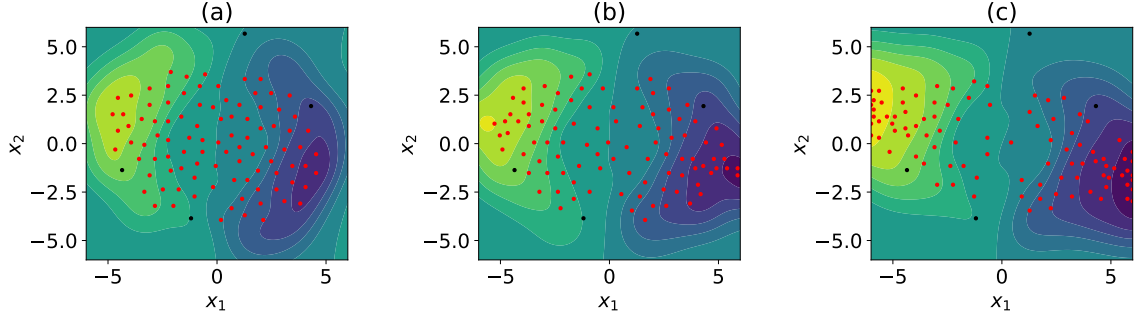
Figure 7.3: Predicted response functions and sequential sampling locations (•) with $\alpha = 0$ and (a) $t = 0.6$, (b) $t = 1$, (c) $t = 1.4$ in the stochastic oscillator example, starting from the same initial samples (•).

$[0, 25]$:

$$f(\mathbf{x}) = \frac{1}{25} \int_0^{25} u(t; \mathbf{x}) \mathrm{d}t, \tag{7.15}$$

with contour shown in figure 7.2(b).

In our computation, we use 4 initial samples followed by 96 sequential samples with the error metric ϵ in (7.2) calculated after each sample. Considering the randomness of initial samples, all results show below are average from 100 different initializations unless otherwise specified. Figure 7.2(c) shows ϵ as a function of sample number for different values of t in acq_{GLW}, including the case of $t = 1$ for which $acq_{GLW} = acq_{LW}$[4]. We see that the optimal performance of acq_{GLW} is achieved for t roughly in $[1.2, 1.6]$, where ϵ is about half an order of magnitude smaller than that with $t = 1$ close to the end of sampling. The favorable performance with $t \in [1.2, 1.6]$ can be further understood from the sample locations shown in figure 7.3. As expected, when t is increased from 0.6 to 1.4, more samples are allocated in the input space with extreme-value responses, leading to a smaller error ϵ characterizing the accuracy of the tail of the response PDF.

---

[4]We note that our result with $t = 1$ is different from that in [19]. This is because [19], for some reason, sets a floor value of e^{-16} for $p_f(f)$ in their calculation, which is unnecessarily high for double precision. We instead set a floor value of 10^{-16} that is consistent with double precision.

136

7.3.1.2 Pandemic spike

We consider another case used in [109], where the response function is constructed from the evolution of infections in a pandemic. In particular, the evolution of infections is simulated by Susceptible, Infected, Recovered (SIR) model developed in [69] and [4]

$$\begin{aligned}
\frac{dS}{dt} &= -\beta IS + \delta R \\
\frac{dI}{dt} &= \beta IS - \gamma I \\
\frac{dR}{dt} &= \gamma I - \delta R,
\end{aligned} \tag{7.16}$$

with $S(t)$, $I(t)$, and $R(t)$ respectively state variables representing the number of susceptible, infectious and recovered individuals. δ, γ, and β are immunity loss rate, recovery rate, and infection rate. The parameter β is endowed with a two-term K-L expansion of the stochastic process: $\beta(t) \approx \beta_0(\sum_{i=1}^{2} x_i \lambda_i \phi_i(t) + \phi_0)$ with $\phi_0 > 0$ and $\phi_i(t)$, λ_i determined from the correlation function $\sigma_\beta^2 e^{-\tau^2/(2l_\beta^2)}$. We keep all parameter values and initial conditions to (7.16) the same as in [109][5]. The input variable $\mathbf{x} \equiv (x_1, x_2)$ is a standard normal variable with $p_{\mathbf{x}}(\mathbf{x}) = \mathcal{N}(\mathbf{0}, I_2)$ (see figure 7.4(a)). We are interested in, as the response of the system, the infections at $t = 20$:

$$f(\mathbf{x}) = I(t = 20; \mathbf{x}), \tag{7.17}$$

with its contour shown in figure 7.4(b).

Our computation starts from 4 initial samples, followed by 46 sequential samples employing acq_{GLW}. The results with $\alpha = 0$ and varying t from 0.6 to 1.6 are plotted in figure 7.4(c) as a function of the number of sequential samples. We see a similar pattern as in §7.3.1.1 where the optimal performance occurs with t roughly in $[1.2, 1.6]$, for which the error ϵ at majority of sample numbers is about half an order of magnitude smaller than that obtained in the case with $t = 1$. Furthermore, the

---

[5]$\delta = 0$, $\gamma = 0.1$, $\beta_0 = 3 * 19^{-9}$, $\phi_0 = 2.55$, $\sigma_\beta^2 = 0.1$, $l_\beta = 4$, $S(0) = 10^8$, $I(0) = 50$, $R(0) = 0$.

(a)

(b)

(c)
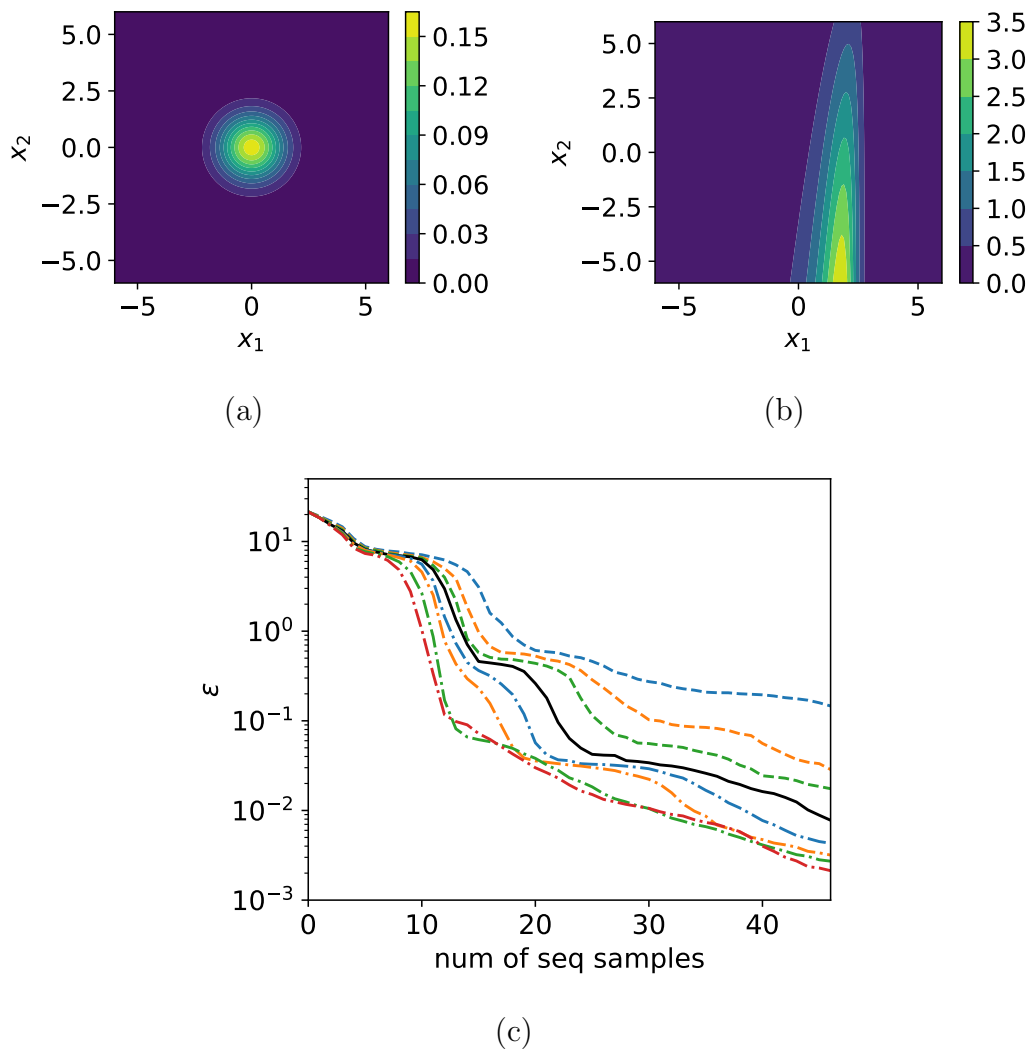
Figure 7.4: (a) input probability distribution and (b) response function of the pandemic spike example. (c) error ϵ as functions of sample numbers with $\alpha = 0$ and varying $t = 0.6$ (- - -), 0.8(- - -), 0.9(- - -), 1(——), 1.1(-·-·-), 1.2(-·-·-), 1.4(-·-·-), 1.6(-·-·-).
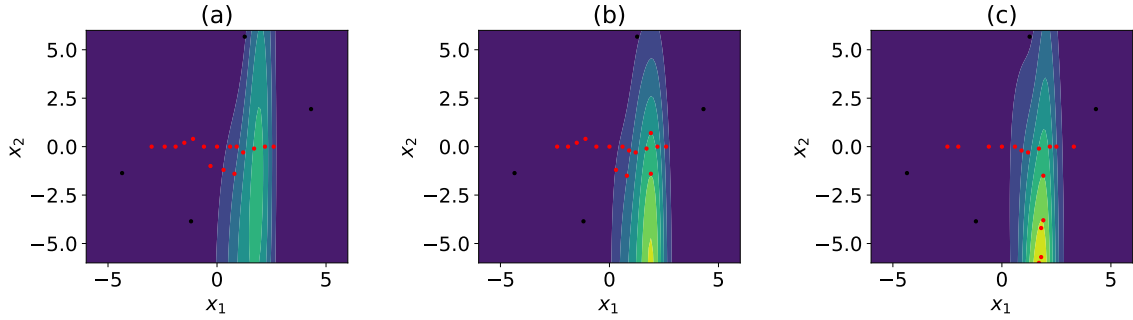
Figure 7.5: Predicted response functions and sequential sampling locations (•) with $\alpha = 0$ and (a) $t = 0.6$, (b) $t = 1$, (c) $t = 1.4$ in the pandemic spike example, starting from the same initial samples (•).

sample locations for $t = 0.6$, 1, and 1.4 plotted in figure 7.5 again show that the increase of t pushes more samples toward rare-event regions in the input space.

We note that for the above two cases (and other cases with relatively simple response functions), the variation of α can also have an impact on the performance of sequential sampling. In particular, we have observed some cases with $\alpha > 0$ that produce somewhat better results than those with $\alpha = 0$. However, the mechanism associated with α is much more subtle than that with t for these simple response functions, and we will not elaborate it in this paper. The impact of α on the sampling performance is most evident for complex (multi-modal) response functions, which we discuss in detail in §7.3.2.

7.3.2 Complex response functions generated by kernels

In this section, we test the performance of acq_{GLW} for a large number of arbitrarily-generated complex response functions. These functions are constructed as random realizations of Gaussian processes with RBF kernel and Matern kernel[6], hereafter referred to as RBF and Matern functions for simplicity. Examples of such functions in the 2D case are shown in figure 7.6, which illustrates the complex and multi-modal

---

[6]The hyperparameters of kernels are set as $\tau = 2$ and $\Lambda = I_d$, and for Matern kernel the additional parameter ν is fixed as 1.5 (see Appendix E.4 for the definition of kernels and parameters).
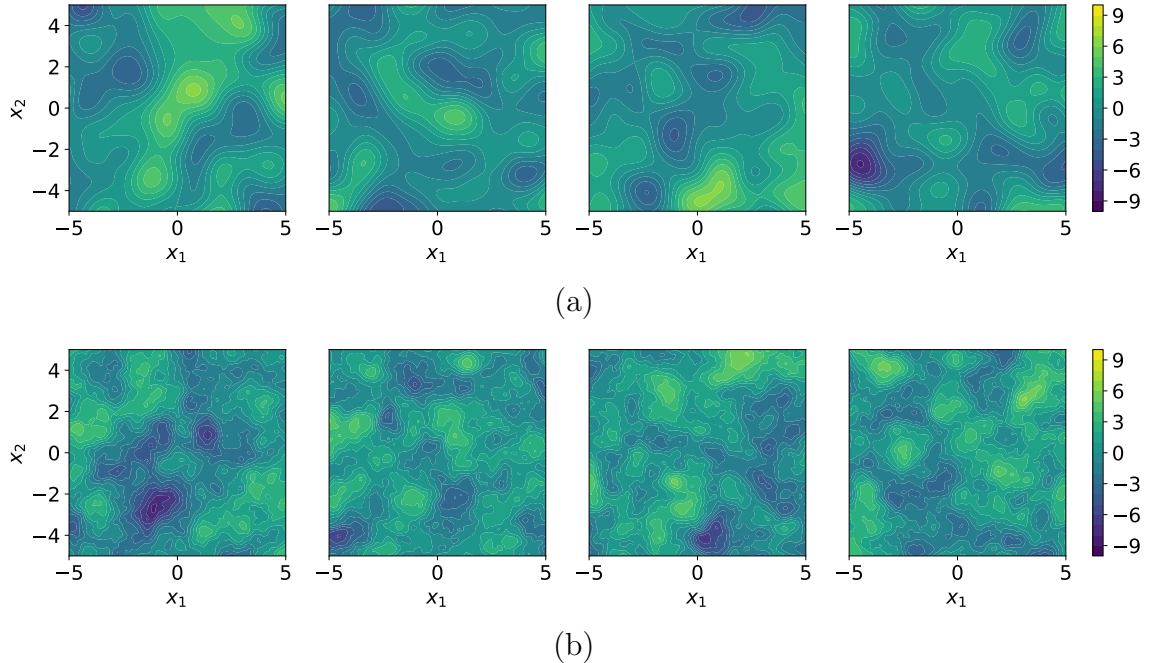
Figure 7.6: Examples of two-dimensional (a) RBF and (b) Matern functions.

features of these functions (especially for Matern functions which exhibits more small-scale variations). In the tests presented below, we consider both 2D ($d = 2$) and 3D ($d = 3$) cases, with the input set as a standard normal $p_{\mathbf{x}}(\mathbf{x}) = \mathcal{N}(\mathbf{0}, \mathrm{I}_d)$. For each kernel and dimension, the results presented in terms of error ϵ are averaged over 200 function realizations of the random process and 20 different realizations of initial samples (for each function), i.e., over 4000 cases in total. Due to this massive average, the improved results from acq_{GLW} presented below is statistically significant. For clarity of the presentation, we will show results for the RBF functions in the main text, and leave results for the Matern functions that lead to similar conclusions in Appendix E.5.

7.3.2.1 Two-dimensional (2D) RBF functions

We first consider 2D RBF functions with examples plotted in figure 7.6, showing much stronger variations (i.e., higher complexities) than the cases presented in §7.3.1. Figure 7.7(a) shows the error ϵ with increase of number of samples (4 initial samples

140

(a)

(b)

(c)
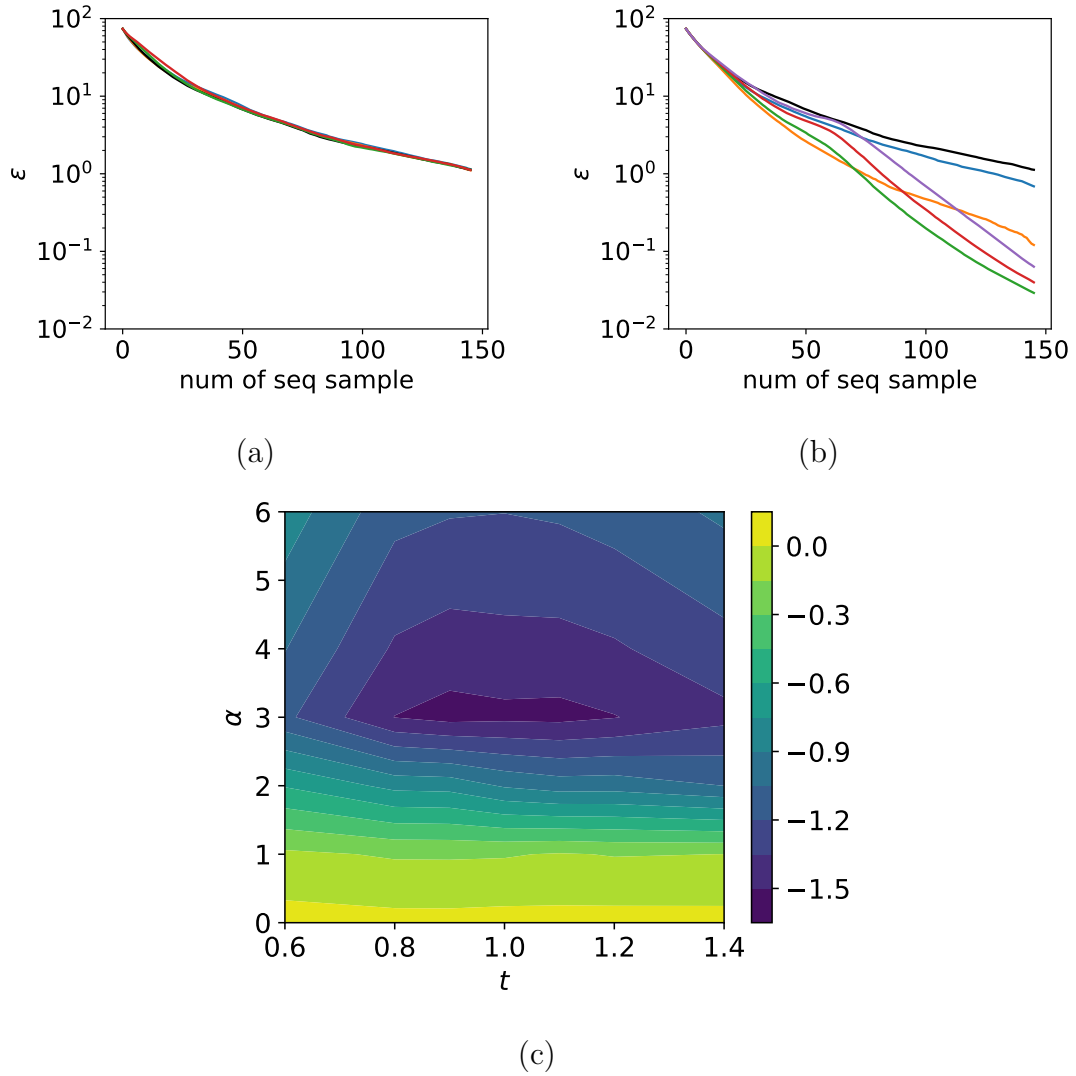
Figure 7.7: Results for two-dimensional RBF functions. Error ϵ as function of number of samples for (a) $\alpha = 0$ and varying $t = 0.6$ (——), 0.8(——), 1(——), 1.2(——), 1.4(——), (b) $t = 1$ and varying $\alpha = 0$(——), 1 (——), 2(——), 3(——), 4(——), 6(——); (c) contour plot of $\log_{10} \epsilon$ at 146 sequential samples for varying t and α.
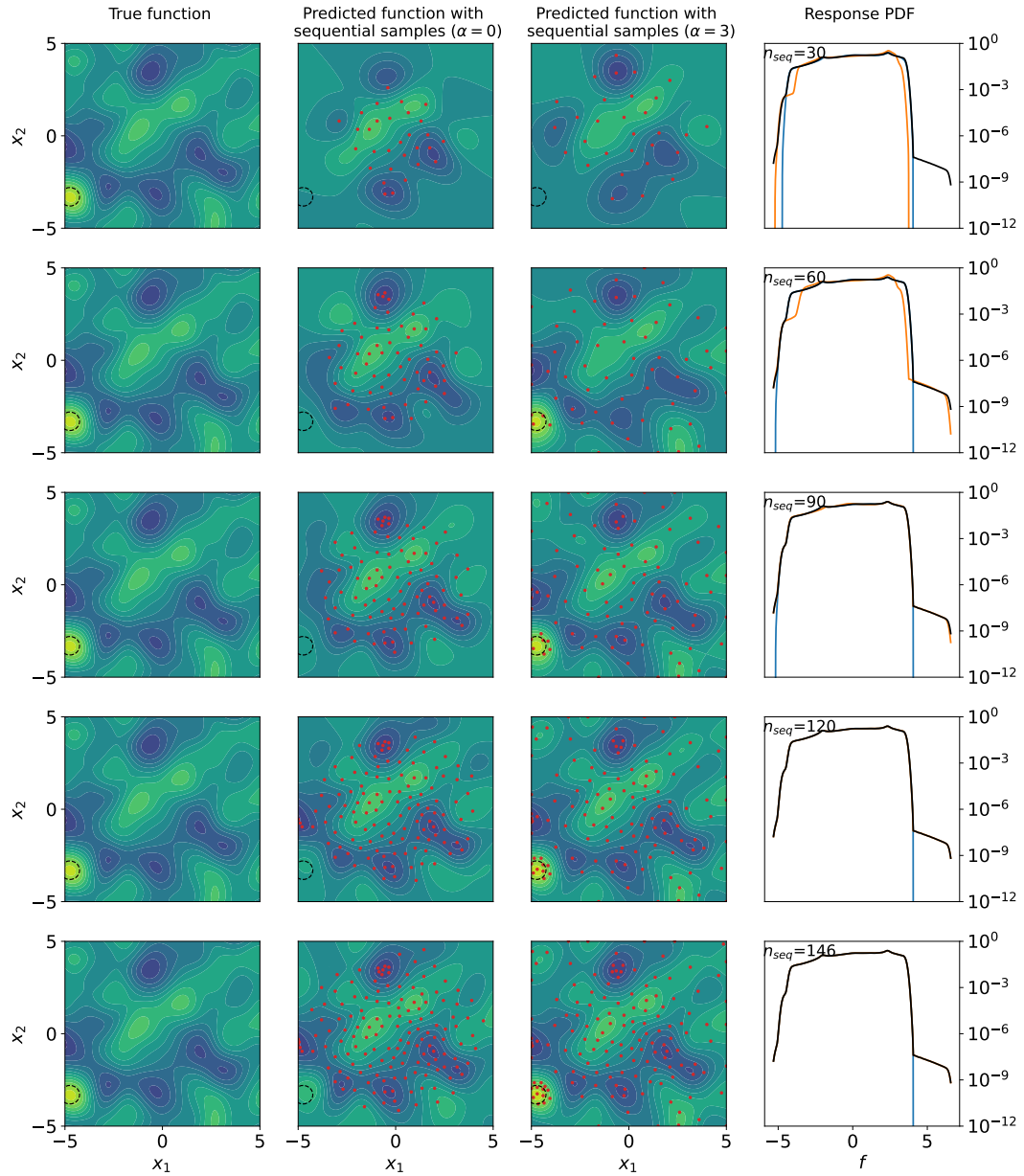
Figure 7.8: First column: true response RBF function as a reference; second column: sequential samples (•) with $\alpha = 0$ on the predicted response function; third column: sequential samples (•) with $\alpha = 3$ on the predicted response function; fourth column: predicted PDF $p_{\hat{f}}(f)$ with $\alpha = 0$ (——) and $\alpha = 3$ (——) compared with the true PDF $p_f(f)$ (——). The top-to-bottom rows correspond to situations with number of sequential samples $n_{seq} = [30, 60, 90, 120, 146]$. The black circles shown in columns 1-3 mark the rare-event region around $(-4.7, -3.3)$ that is missed by sequential samples with $\alpha = 0$ but captured with $\alpha = 3$.

followed by 146 sequential samples) for t varying from 0.6 and 1.4 and fixed $\alpha = 0$. Unlike the cases in §7.3.1, the variation of parameter t almost does not affect the performance of sequential sampling using acq_{GLW}, with results indistinguishable for the selected range of t. On the other hand, variation of α leads to a much stronger impact on the error ϵ, as shown in figure 7.7(b). One can see from the figure that $\alpha = 3$ produces the best result in the tested range, with error at 146 sequential samples about two orders of magnitude smaller than that with $\alpha = 0$ (i.e., the original acq_{LW} acquisition). Figure 7.7(c) further shows a contour plot of the error ϵ at 146 sequential samples as a function of t and α. We see that $\alpha \approx 3$ and $t \approx 1$ is indeed close to the global optimal among all choices tested here.

Considering the behavior observed in figure 7.7, it is clear that the improved performance associated with larger α comes from the increased exploration power of the acquisition that captures more rare-event regions in the input space. Such exploration is not achievable by the variation of t, at least in the tested range. To demonstrate this reasoning, we plot in figure 7.8 the evolution of sampling locations, predicted response functions, and predicted response PDFs from 30 to 146 sequential samples with $\alpha = 0$ and $\alpha = 3$ ($t = 1$ fixed) for a typical RBF function. It is clear that with $\alpha = 0$ the rare-event region near $\mathbf{x} = (-4.7, -3.3)$ (which happens to be the global maximum) is not captured, leading to a failure in resolving the right tail of the response PDF. More specifically, the missing of this important rare-event region is due to the fact that the predicted response $\hat{f}$ fails to capture the large response in this region with limited number of samples, together with the lack of exploration power with $\alpha = 0$. In contrast, for $\alpha = 3$, the region near $\mathbf{x} = (-4.7, -3.3)$ is identified within 60 sequential samples, leading to a much more accurate resolution of the right tail of the PDF. We encourage the readers to also take a look at figure E.4 for Matern response functions where such behavior is more evident due to the increased complexity of the function.
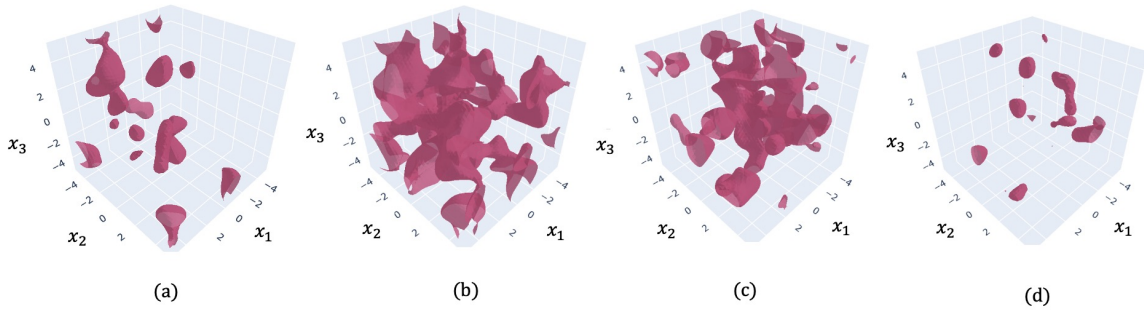
Figure 7.9: Level sets of a typical 3D RBF function. From (a) to (d), $\{\mathbf{x} : f(\mathbf{x}) = -4, -2, 2, 4\}$

7.3.2.2 Three-dimensional (3D) RBF functions

We further test the performance of acq_{GLW} on 3D RBF functions, with one example of these functions shown in figure 7.9 visualized through level sets of the function. It is clear that the multi-modal feature is still present in the 3D case, which needs to be captured in sampling to resolve the tail of the response PDF. For these 3D functions, our computations start from 8 initial samples followed by 392 sequential samples.

Like in 2D cases, the variation of t with $\alpha = 0$ does not affect the performance of acq_{GLW} (figure 7.10(a)). The major improvement in performance is achieved by increasing α to 3, for which the error ϵ after 392 sequential samples is about one order of magnitude smaller than that from $\alpha = 0$, corresponding to acq_{LW} (figure 7.10(b)). Finally, figure 7.10(c) shows that $\alpha \approx 3$ and $t \approx 1$ still provides the globally optimal results for these 3D functions.

The mechanism underlying the improved performance with $\alpha = 3$ is also similar to the 2D cases. To illustrate this, we consider an example of the RBF function with the global minimum of value -7.26 at $\mathbf{x} = (2.6, -3.4, 2.5)$. The predicted response functions with $\alpha = 0$ and $\alpha = 3$ after 392 samples are shown in figure 7.11, visualized respectively on the cross-section at $x_3 = 2.5$ in (a) and in terms of level set of value -6 in (b). It is clear that the global minimum of the function is only captured with $\alpha = 3$ and completely missed with $\alpha = 0$.

144

(a)

(b)

(c)
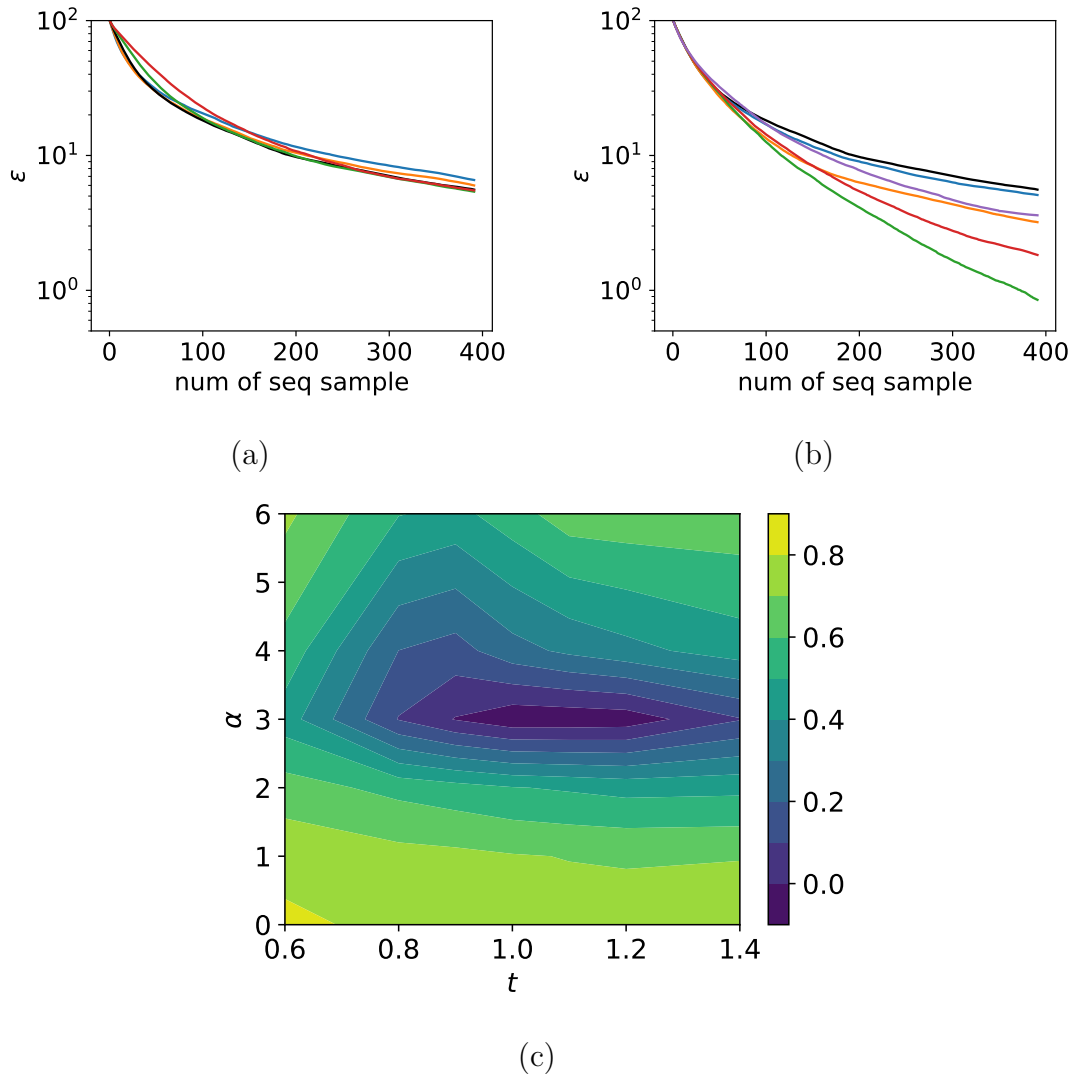
Figure 7.10: Results for three-dimensional RBF functions. Error ϵ as function of number of samples for (a) $\alpha = 0$ and varying $t = 0.6$ (——), 0.8(——), 1(——), 1.2(——), 1.4(——), (b) $t = 1$ and varying $\alpha = 0$(——), 1 (——), 2(——), 3(——), 4(——), 6(——); (c) contour plot of $\log_{10} \epsilon$ at 392 sequential samples for varying t and α.
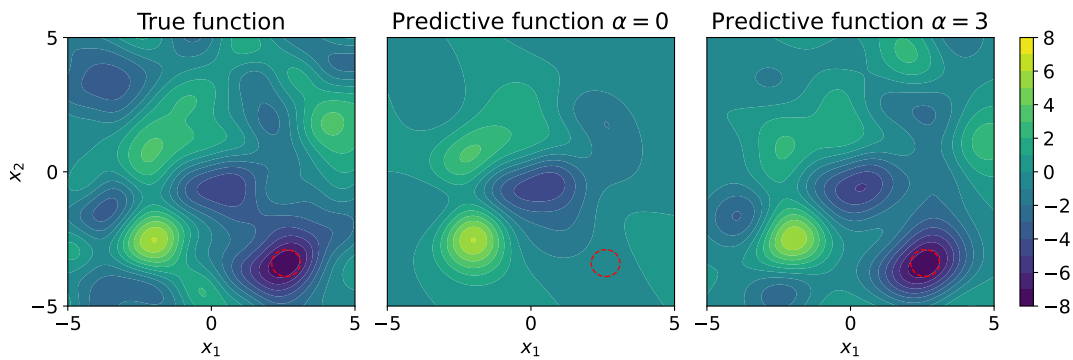
(a) cross-section at $x_3 = 2.5$
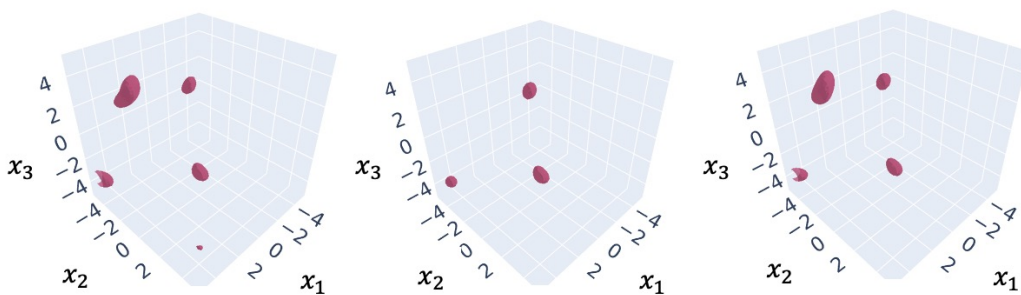
(b) level set $f = -6$

Figure 7.11: Results for a typical 3D RBF function after 396 sequential samples. (Left column) True function, (middle column) predicted function with $\alpha = 0$ and (right column) predicted function with $\alpha = 3$, visualized (a) on a cross-section at $x_3 = 2.5$ and (b) in terms of level set $f = -6$. The global minimum of the function around (2.6, -3.4, 2.5) is circled in (a).

7.3.3 Rare-event statistics of ship motion in a random sea

We finally consider an application of our method to an engineering problem of esti-
mating the rare-event statistics of ship roll motion in a random sea. To simulate the
ship roll response in waves, we use a phenomenological nonlinear roll equation that
is widely used in marine engineering [142, 132, 50, 46, 49]

$$\ddot{\xi} + \alpha_1 \dot{\xi} + \alpha_2 \dot{\xi}|\dot{\xi}| + (\beta_1 + \epsilon_1 \sin(\chi)\eta(t))\xi + \beta_2 \xi^3 = \epsilon_2 \cos(\chi)\eta(t), \qquad (7.18)$$

where $\xi(t)$ is the time series of roll motion excited by waves with elevation $\eta(t)$, χ is the
angle between ship heading direction and the wave crest. The empirical coefficients
in (7.18) are set as $\alpha_1 = 0.1$, $\alpha_2 = 0.1$, $\beta_1 = 1$, $\beta_2 = 0.1$, $\epsilon_1 = 1$, $\epsilon_2 = 1$.

Since large ship motions are usually excited by wave groups at sea, we consider
$\eta(t)$ modeled by wave groups with Gaussian envelop

$$\eta(t) = \exp(-\frac{1}{2}(\frac{t - 5T}{2T})^2) \sin(\frac{2\pi}{T}t), \qquad (7.19)$$

with T the period of each individual wave in the group. In a random sea, we further
consider two independent random parameters (T, χ) as the input space, satisfying
the distribution of $T \sim \mathcal{N}(T_p, (T_p/4)^2)$ with $T_p = 15s$ and $\chi \sim \mathcal{N}(\chi_p, (\chi_p/4)^2)$ with
$\chi_p = \pi/2$. Our quantity of interest is the maximum roll over the time window $[0, 10T]$:

$$f(\mathbf{x}) = \max_{t \in [0, 10T]} |\xi(t; \mathbf{x})|, \qquad (7.20)$$

with the response function plotted in figure 7.12(a) clearly showing the multi-modal
feature.

In computation, we only consider half of the input space due to symmetry of (7.18)
with $\chi = \pi/2$. The results from sequential sampling with acq_{GLW} are shown in figure
7.12(b), comparing the cases with $\alpha = 3$ and $\alpha = 0$. For such a multi-modal response

147
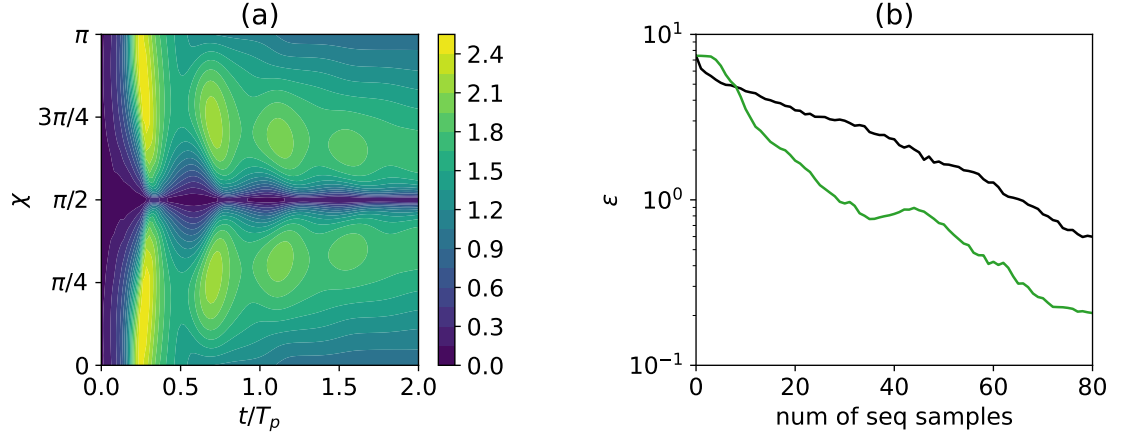
Figure 7.12: (a) Contour plot of the true response function calculated by (7.18) and (b) results with $\alpha = 3$ (——) and $\alpha = 0$ (——) for comparison, both with $t = 1$.

(a)   (b)
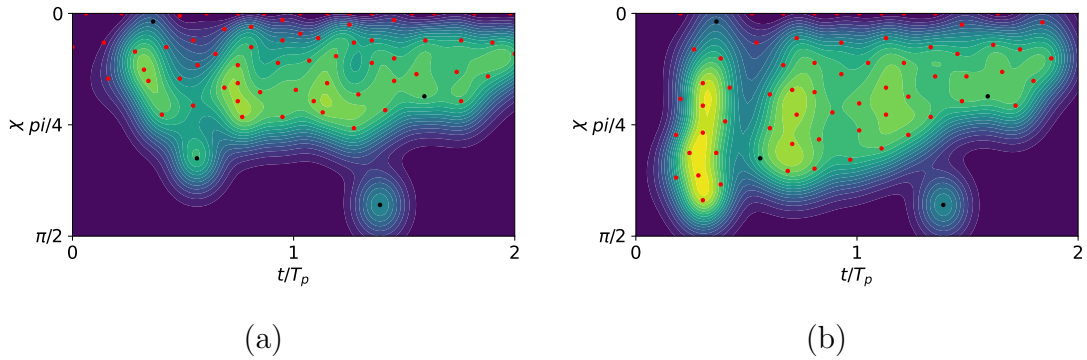
Figure 7.13: Predicted response functions and sequential sampling locations (•) with $t = 1$ and (a) $\alpha = 0$, (b) $\alpha = 3$ in the ship motion example, starting from the same initial samples (•).

function, it is clear that the result from $\alpha = 3$ (empirically determined as optimal in §7.3.2) is much better than that from $\alpha = 0$, with the error ϵ from the former half an order of magnitude smaller than the latter in majority of the sampling process. The sampling location plotted in figure 7.13 further demonstrates the effectiveness of $\alpha = 3$ to explore rare-event regions in the full input space.

7.4 Summary

In this chapter, we develop a new acquisition function acq_{GLW} in sequential sampling to efficiently quantify the rare-event statistics in the response of an ItR system. Our new acquisition takes a generalized form of the existing likelihood-weighed acquisition acq_{LW} [117, 118] and contains two additional parameters α and t. By varying α and t, acq_{GLW} is able to (i) place different level of emphasis on rare-event regions in sampling, and (ii) remedy the situation when the predicted ItR function has a large discrepancy from the true function. We demonstrate the advantage of acq_{GLW} over acq_{LW} in a number of test cases with empirically optimal values of α and t identified. It is suggested in these cases that if the response function is relatively simple, using $\alpha = 1$ and $t \in [1.2, 1.6]$ in acq_{GLW} produces consistently better results than that from acq_{LW}, due to the more appropriate emphasis on the known rare-event region. If the response function is complex with multi-modal structures, using $\alpha \approx 3$ and $t \approx 1$ is critical since it allows more exploration in sampling to identify multiple rare-event regions in the input space.

CHAPTER 8

Non-Myopic Batch Bayesian Experimental Design for Statistical Expectation

8.1 Introduction

In previous chapters, our sampling algorithms were exclusively designed to select one sample per iteration in a myopic way. In this chapter, we will discuss an ongoing work on non-myopic batch sampling, where several samples can be selected and evaluated together. Instead of extreme-event statistics, our effort for batch sampling algorithm will start from the statistical expectation due to its central role in uncertainty quantification and some favorable properties which will be covered soon in the following.

For almost a century, the fundamental method to estimate statistical expectation has been Monte Carlo with the core idea of learning a system by many random samples [104]. Although the convergence of Monte Carlo is guaranteed by the *law of large numbers*, its convergence rate—inversely proportional to the square root of the number of samples—is notoriously slow. To increase the convergence rate under a limited number of samples, a sequential Bayesian experimental design framework targeting statistical expectation was developed in [105]. Specifically, they use Gaussian process regression (GPR) as the surrogate and select the next-best sample which maximizes the information gain of adding one sample, i.e., the K-L divergence between the current and next-step estimation. Later, we proved that maximizing this

information-theoretic acquisition is equivalent to minimizing the variance of the next-step estimation [43]. This finding gives us a concise form with a much more intuitive interpretation.

Although [105] shows that the proposed method works in several synthetic and practical cases, its sequential nature does bring two drawbacks (affecting not only statistical expectation but also other quantities such as extreme-event statistics). Firstly, the samples need to be evaluated one by one, making the duration of the whole process remarkably long. In contrast, the standard Monte Calor determines all samples in the beginning which can maximally utilize the parallel computational resources in evaluating samples. Secondly, the determination of samples only focuses on the benefits of the immediate next step rather than the long-term objective, for example, the convergence after a certain number of samples. It might be presumed that a batch design, aimed at long-term collective benefits, would easily achieve faster convergence compared to sequential sampling. However, the reality is more complex. A sequential algorithm can update the surrogate after each sample, making the selection of the next sample based on a more accurate model (although in a myopic way). In other words, the sampling efficiency of the batch algorithm needs to be evaluated in light of the benefits of long-term perspective and the disadvantages of less frequent model updates.

The above-mentioned problem of batch sampling is generally true for many algorithms where the sample responses appear in the acquisition function (including all previous acquisitions in this dissertation), thus the evaluation of the previous sample will directly influence the selection of the following samples. For statistical expectation, we noticed that in [43] and [105], the responses of samples affect the following sampling only by the hyperparameters (prior) of the GPR (which will be demonstrated in §8.2). In practice, the hyperparameters that control the global smoothness and variability of functions can be known in advance or fixed after a few samples,

significantly mitigating the adverse effects of batch sampling. Hence, batch sampling has a great chance to converge faster than or on par with its sequential counterpart. This potential for improved efficiency has led us to develop and explore the batch algorithm further.

In this chapter, we develop a non-myopic batch Bayesian experimental design for statistical expectation. The next batch of samples are selected which maximizes the long-term information gain (as the acquisition) when they are added together. To solve the resulting high-dimensional optimization problem, we formulate an analytic approximation for the acquisition with the gradient computed through automatic differentiation via PyTorch (rather than theoretical derivation as in §5). The superior performance of the proposed algorithm, in terms of wall time saving and a faster or matched convergence rate than sequential sampling, is demonstrated in a case with arbitrary complex functions generated by kernels and another case using a stochastic oscillator.

Part of this chapter is adapted from [43]. The Python code for the algorithm, named gpexpectation, is available on Github[1].

8.2 Method

8.2.1 Problem setup

We consider an ItR system described by a response function $f(\mathbf{x}) : \mathbb{R}^d \to \mathbb{R}$ with $\mathbf{x}$ a d-dimensional random input. The input probability $p_\mathbf{x}(\mathbf{x})$ is assumed to be known and our objective is the statistical expectation defined as:

$$q = \int f(\mathbf{x})p_\mathbf{x}(\mathbf{x})\mathrm{d}\mathbf{x}. \tag{8.1}$$

---

[1]https://github.com/umbrellagong/gpexpectation

While q can be directly evaluated via the standard Monte Carlo method, i.e., $q \approx \frac{1}{m} \sum_{i=1}^{m} f(\mathbf{x}^i)$ with $\mathbf{x}^i$ sampling from $p_{\mathbf{x}}(\mathbf{x})$, its convergence within a limited number of samples is unsatisfactory. Instead, we take a Bayesian perspective by placing f a Gaussian process prior $f \sim \mathcal{GP}(0, k(\mathbf{x}, \mathbf{x}'))$ where k is covariance function with hyperparameters $\boldsymbol{\theta}$. Given a dataset $\mathcal{D}_n = \{\mathbf{X}_n, \mathbf{Y}_n\}$ consisting of n inputs $\mathbf{X}_n = \{\mathbf{x}^i \in \mathbb{R}^d\}_{i=1}^{n}$ and the corresponding outputs $\mathbf{Y}_n = \{f(\mathbf{x}^i) \in \mathbb{R}\}_{i=1}^{n}$, the underling relation f is predicted as a posterior Gaussian process $f(\mathbf{x})|\mathcal{D}_n \sim \mathcal{GP}(m_n(\mathbf{x}), k_n(\mathbf{x}, \mathbf{x}'))$ with formulae of posterior mean m_n and covariance k_n detailed in Appendix A. The statistical expectation $q|\mathcal{D}_n$ then becomes a random variable with randomness coming from the epistemic uncertainties of $f(\mathbf{x})|\mathcal{D}_n$. Our goal is to choose the most informative batch of samples by optimizing the acquisition function that facilitates convergence of q. In the following, we will discuss the form of the acquisition function as the core of the algorithm.

8.2.2   Acquisition function

For selecting the next samples, a popular way is to maximize the information gain (measured by K-L divergence) between the current estimation $q|\mathcal{D}_n$ and hypothetical next estimation $q|\mathcal{D}_n, \tilde{\mathbf{X}}_s, \tilde{\mathbf{Y}}_s$ after adding s number of samples $\tilde{\mathbf{X}}_s$ with responses $\tilde{\mathbf{Y}}_s$ (see [105] for a sequential version):

$$
\begin{aligned}
\mathbf{X}_s^* &= \operatorname{argmax}_{\tilde{\mathbf{X}}_s} \mathbb{E}\Big[\int \mathrm{KL}\big(p(q|\mathcal{D}_n, \tilde{\mathbf{X}}_s, \tilde{\mathbf{Y}}_s) \,\|\, p(q|\mathcal{D}_n))\big) \Big], \\
&\equiv \operatorname{argmax}_{\tilde{\mathbf{X}}_s} \int \mathrm{KL}\big(p(q|\mathcal{D}_n, \tilde{\mathbf{X}}_s, \tilde{\mathbf{Y}}_s) \,\|\, p(q|\mathcal{D}_n))\big) \, p(\tilde{\mathbf{Y}}_s|\tilde{\mathbf{X}}_s, \mathcal{D}_n) \mathrm{d}\tilde{\mathbf{Y}}_s, \quad (8.2)
\end{aligned}
$$

where $\tilde{\mathbf{Y}}_s$ is chosen based on the current surrogate $f(\mathbf{x})|\mathcal{D}_n$ following a distribution of $\mathcal{N}(\tilde{\mathbf{Y}}_s; m_n(\tilde{\mathbf{X}}_s), k_n(\tilde{\mathbf{X}}_s, \tilde{\mathbf{X}}_s))$. Another way, which is more intuitive, is to minimize

the uncertainty (measured by variance) of $q|\mathcal{D}_n, \tilde{\mathbf{X}}_s, \tilde{\mathbf{Y}}_s$:

$$\mathbf{X}_s^* = \text{argmin}_{\tilde{\mathbf{X}}_s} \ \mathbb{E}\Big[\int \text{var}(q|\mathcal{D}_n, \tilde{\mathbf{X}}_s, \tilde{\mathbf{Y}}_s)\Big]$$

$$\equiv \text{argmin}_{\tilde{\mathbf{X}}_s} \int \text{var}(q|\mathcal{D}_n, \tilde{\mathbf{X}}_s, \tilde{\mathbf{Y}}_s) \, p(\tilde{\mathbf{Y}}_s|\tilde{\mathbf{X}}_s, \mathcal{D}_n)\mathrm{d}\tilde{\mathbf{Y}}_s. \tag{8.3}$$

Indeed, these two ways are equivalent for estimating the statistical expectation. In the following, we will show their detailed formulae and the equivalence (see a similar conclusion for sequential design in [43]).

We first notice that $q|\mathcal{D}_n$ follows a Gaussian distribution with mean μ_1 and variance σ_1^2:

$$p(q|\mathcal{D}_n) = \mathcal{N}(q; \mu_1, \sigma_1^2), \tag{8.4}$$

$$\mu_1 = \mathbb{E}\Big[\int f(\mathbf{x})p_{\mathbf{x}}(\mathbf{x})\mathrm{d}\mathbf{x}|\mathcal{D}_n\Big]$$

$$= \int m_n(\mathbf{x})p_{\mathbf{x}}(\mathbf{x})\mathrm{d}\mathbf{x}, \tag{8.5}$$

$$\sigma_1^2 = \mathbb{E}\Big[\Big(\int f(\mathbf{x})p_{\mathbf{x}}(\mathbf{x})\mathrm{d}\mathbf{x}\Big)^2|\mathcal{D}_n\Big] - \mathbb{E}\Big[\Big(\int f(\mathbf{x})p_{\mathbf{x}}(\mathbf{x})\mathrm{d}\mathbf{x}\Big)|\mathcal{D}_n\Big]^2$$

$$= \iint k_n(\mathbf{x}, \mathbf{x}')p_{\mathbf{x}}(\mathbf{x})p_{\mathbf{x}}(\mathbf{x}')\mathrm{d}\mathbf{x}'\mathrm{d}\mathbf{x}. \tag{8.6}$$

After adding s hypothetical samples $\{\tilde{\mathbf{X}}_s, \tilde{\mathbf{Y}}_s\}$, f follows an updated distribution $f(\mathbf{x})|\mathcal{D}_n, \tilde{\mathbf{X}}_s, \tilde{\mathbf{Y}}_s \sim \mathcal{GP}(m_{n+s}(\mathbf{x}), k_{n+s}(\mathbf{x}, \mathbf{x}'))$ with

$$m_{n+s}(\mathbf{x}) = m_n(\mathbf{x}) + \mathbf{k}_n(\mathbf{x}, \tilde{\mathbf{X}}_s)\mathbf{K}_n(\tilde{\mathbf{X}}_s, \tilde{\mathbf{X}}_s)^{-1}(\tilde{\mathbf{Y}}_s - m_n(\tilde{\mathbf{X}}_s)), \tag{8.7}$$

$$k_{n+s}(\mathbf{x}, \mathbf{x}') = k_n(\mathbf{x}, \mathbf{x}') - \mathbf{k}_n(\mathbf{x}, \tilde{\mathbf{X}}_s)\mathbf{K}_n(\tilde{\mathbf{X}}_s, \tilde{\mathbf{X}}_s)^{-1}\mathbf{k}_n(\tilde{\mathbf{X}}_s, \mathbf{x}'). \tag{8.8}$$

The quantity $q|\mathcal{D}_n, \tilde{\mathbf{X}}_s, \tilde{\mathbf{Y}}_s$ can then be represented by another Gaussian with mean

μ_2 and variance σ_2^2:

$$p(q|\mathcal{D}_n, \tilde{\mathbf{X}}_s, \tilde{\mathbf{Y}}_s) = \mathcal{N}(q; \mu_2(\tilde{\mathbf{X}}_s, \tilde{\mathbf{Y}}_s), \sigma_2^2(\tilde{\mathbf{X}}_s)), \tag{8.9}$$

$$\mu_2(\tilde{\mathbf{X}}_s, \tilde{\mathbf{Y}}_s) = \int m_{n+s}(\mathbf{x}) p_{\mathbf{x}}(\mathbf{x}) \mathrm{d}\mathbf{x}$$

$$= \mu_1 + \int \mathbf{k}_n(\mathbf{x}, \tilde{\mathbf{X}}_s) p_{\mathbf{x}}(\mathbf{x}) \mathrm{d}\mathbf{x} \, \mathbf{K}_n(\tilde{\mathbf{X}}_s, \tilde{\mathbf{X}}_s)^{-1} (\tilde{\mathbf{Y}}_s - m_n(\tilde{\mathbf{X}}_s)), \tag{8.10}$$

$$\sigma_2^2(\tilde{\mathbf{X}}_s) = \iint k_{n+s}(\mathbf{x}, \mathbf{x}') p_{\mathbf{x}}(\mathbf{x}) p(\mathbf{x}') \mathrm{d}\mathbf{x}' \mathrm{d}\mathbf{x}$$

$$= \sigma_1^2 - \int \mathbf{k}_n(\mathbf{x}, \tilde{\mathbf{X}}_s) p_{\mathbf{x}}(\mathbf{x}) \mathrm{d}\mathbf{x} \, \mathbf{K}_n(\tilde{\mathbf{X}}_s, \tilde{\mathbf{X}}_s)^{-1} \int \mathbf{k}_n(\tilde{\mathbf{X}}_s, \mathbf{x}) p_{\mathbf{x}}(\mathbf{x}) \mathrm{d}\mathbf{x}. \tag{8.11}$$

Substitute (8.4) and (8.9) into (8.2), one can simplify the objective function in (8.2):

$$\int \mathrm{KL}\big(p(q|\mathcal{D}_n, \tilde{\mathbf{X}}_s, \tilde{\mathbf{Y}}_s) \,\|\, p(q|\mathcal{D}_n)\big) \, p(\tilde{\mathbf{Y}}_s|\tilde{\mathbf{X}}_s, \mathcal{D}_n) \mathrm{d}\tilde{\mathbf{Y}}_s.$$

$$= \iint p(q|\mathcal{D}_n, \tilde{\mathbf{X}}_s, \tilde{\mathbf{Y}}_s) \log \frac{p(q|\mathcal{D}_n, \tilde{\mathbf{X}}_s, \tilde{\mathbf{Y}}_s)}{p(q|\mathcal{D}_n)} \mathrm{d}q \, p(\tilde{\mathbf{Y}}_s|\tilde{\mathbf{X}}_s, \mathcal{D}_n)) \mathrm{d}\tilde{\mathbf{Y}}_s$$

$$= \int \Big(\log(\frac{\sigma_1}{\sigma_2(\tilde{\mathbf{X}}_s)}) + \frac{\sigma_2^2(\tilde{\mathbf{X}}_s)}{2\sigma_1^2} + \frac{(\mu_2(\tilde{\mathbf{X}}_s, \tilde{\mathbf{Y}}_s) - \mu_1)^2}{2\sigma_1^2} - \frac{1}{2} \Big) p(\tilde{\mathbf{Y}}_s|\tilde{\mathbf{X}}_s, \mathcal{D}_n) \mathrm{d}\tilde{\mathbf{Y}}_s$$

$$= \log(\frac{\sigma_1}{\sigma_2(\tilde{\mathbf{X}}_s)}) + \frac{1}{2\sigma_1^2} \Big(\sigma_2^2(\tilde{\mathbf{X}}_s) - \sigma_1^2 + \int (\mu_2(\tilde{\mathbf{X}}_s, \tilde{\mathbf{Y}}_s) - \mu_1)^2 p(\tilde{\mathbf{Y}}_s|\tilde{\mathbf{X}}_s, \mathcal{D}_n) \mathrm{d}\tilde{\mathbf{Y}}_s \Big)$$

$$= \log(\frac{\sigma_1}{\sigma_2(\tilde{\mathbf{X}}_s)}) + \frac{1}{2\sigma_1^2} \big(\sigma_2^2(\tilde{\mathbf{X}}_s) - \sigma_1^2$$

$$+ \int \mathbf{k}_n(\mathbf{x}, \tilde{\mathbf{X}}_s) p_{\mathbf{x}}(\mathbf{x}) \mathrm{d}\mathbf{x} \, \mathbf{K}_n(\tilde{\mathbf{X}}_s, \tilde{\mathbf{X}}_s)^{-1} \int \mathbf{k}_n(\tilde{\mathbf{X}}_s, \mathbf{x}) p_{\mathbf{x}}(\mathbf{x}) \mathrm{d}\mathbf{x} \big)$$

$$= \log(\frac{\sigma_1}{\sigma_2(\tilde{\mathbf{X}}_s)}). \tag{8.12}$$

Since σ_1 does not depend on $\tilde{\mathbf{X}}_s$, (8.2) can be reformulated as

$$\mathbf{X}_s^* = \mathrm{argmin}_{\tilde{\mathbf{X}}_s} \, \sigma_2^2(\tilde{\mathbf{X}}_s) \tag{8.13}$$

$$= \mathrm{argmax}_{\tilde{\mathbf{X}}_s} \int \mathbf{k}_n(\mathbf{x}, \tilde{\mathbf{X}}_s) p_{\mathbf{x}}(\mathbf{x}) \mathrm{d}\mathbf{x} \, \mathbf{K}_n(\tilde{\mathbf{X}}_s, \tilde{\mathbf{X}}_s)^{-1} \int \mathbf{k}_n(\tilde{\mathbf{X}}_s, \mathbf{x}) p_{\mathbf{x}}(\mathbf{x}) \mathrm{d}\mathbf{x}, \tag{8.14}$$

where (8.13) is exactly (8.3) with $\mathrm{var}(q|\mathcal{D}_n, \tilde{\mathbf{X}}_s, \tilde{\mathbf{Y}}_s) = \sigma_2^2(\tilde{\mathbf{X}}_s)$ a constant for $\tilde{\mathbf{Y}}_s$. The final optimization problem (8.14) is obtained by substituting (8.11) into (8.13). Compared with sequential design with only one sample to be determined, the solution

of (8.14), likely a high-dimensional optimization problem, is much more difficult. To alleviate the difficulties, we develop an analytic approximation for acquisition, while the gradient is computed through automatic differentiation to enable a gradient-based optimization. We will introduce them in the following section.

8.2.3 Analytical formulae for acquisition and its derivative

In computing the right-hand side of (8.14), the heaviest computation involved is the integral $\int k_n(\tilde{\mathbf{x}}, \mathbf{x}) p_{\mathbf{x}}(\mathbf{x}) \mathrm{d}\mathbf{x}$. Expanding k_n with (A.4), we have:

$$\int k_n(\tilde{\mathbf{x}}, \mathbf{x}) p_{\mathbf{x}}(\mathbf{x}) \mathrm{d}\mathbf{x} = \mathcal{K}(\tilde{\mathbf{x}}) - \mathbf{k}(\tilde{\mathbf{x}}, \mathbf{X}_n) \mathbf{K}(\mathbf{X}_n, \mathbf{X}_n)^{-1} \mathcal{K}(\mathbf{X}_n), \tag{8.15}$$

with

$$\mathcal{K}(\mathbf{x}) = \int k(\mathbf{x}, \mathbf{x}') p_{\mathbf{x}}(\mathbf{x}') \mathrm{d}\mathbf{x}', \tag{8.16}$$

where k is the RBF kernel with characteristic amplitude τ and scales Λ ($\boldsymbol{\theta} = \tau, \Lambda$) defined in Appendix A. For RBF kernel, if the input $\mathbf{x}$ is Gaussian with mean $\mathbf{w}$ and covariance Σ, (8.16) has analytical expression:

$$\int k(\mathbf{x}, \mathbf{x}') \mathcal{N}(\mathbf{x}; \mathbf{w}, \Sigma) \mathrm{d}\mathbf{x}' = |\Sigma \Lambda^{-1} + \mathrm{I}|^{-\frac{1}{2}} k(\mathbf{x}, \mathbf{w}; \Sigma + \Lambda). \tag{8.17}$$

To make $\mathcal{K}$ analytically tractble for arbitrary $p_{\mathbf{x}}(\mathbf{x})$, we approximate $p_{\mathbf{x}}(\mathbf{x})$ with the Gaussian mixture model (as in §5.2.4):

$$p_{\mathbf{x}}(\mathbf{x}) \approx \sum_{i=1}^{n_{GMM}} \alpha_i \mathcal{N}(\mathbf{x}; \mathbf{w}_i, \Sigma_i). \tag{8.18}$$

(8.16) can then be formulated as:

$$\mathcal{K}(\mathbf{x}) \approx \sum_{i=1}^{n_{GMM}} \alpha_i \int k(\mathbf{x}, \mathbf{x}') \mathcal{N}(\mathbf{x}'; \mathbf{w}_i, \Sigma_i) \mathrm{d}\mathbf{x}'$$

$$= \sum_{i=1}^{n_{GMM}} \alpha_i |\Sigma_i \Lambda^{-1} + \mathrm{I}|^{-\frac{1}{2}} k(\mathbf{x}, \mathbf{w}_i; \Sigma_i + \Lambda). \tag{8.19}$$

With an analytical formula, a gradient-based optimization method is still needed to solve (8.14) considering the high dimensionality of the optimization domain $(s * d)$. One way to get the gradient is by manual derivation as in [21] and §5, which is tedious and vulnerable to any changes. Here, we take a new paradigm utilizing the automatic differentiation (AD) by PyTorch[2]. With AD, we only need to define the forward process, i.e., coding the right-hand side of (8.14), and its gradient will be evaluated collaterally when executing the forward code.

Algorithm 8 Batch Bayesian experimental design for statistical expectation
---
Require: Number of initial samples n_{init}, number of batches t, number of samples in each batch $s(\cdot)$
Input: Initial dataset $\mathcal{D}_{n_{init}} = \{\mathbf{x}^i, f(\mathbf{x}^i)\}_{i=1}^{n_{init}}$
    Initialization $i = 0$
    while $i < t$ **do**
        Train $f(\mathbf{x})|\mathcal{D}_n$
        Solve (8.14) to find the next-best samples location $\mathbf{X}^*_{s(i)}$
        Implement simulation/experiment to get $f(\mathbf{X}^*_{s(i)})$
        Update the dataset $\mathcal{D}_{n+s(i)} = \mathcal{D}_n \cup \{\mathbf{X}^*_{s(i)}, f(\mathbf{X}^*_{s(i)})\}$
        $i = i + 1$, $n = n + s(i)$
    end while
Output: Compute the statistical expectation based on the surrogate.

We finally show the overall algorithm in Algorithm 8. In each iteration, the number of samples to be selected is specified by $s(i)$ with i the index of iterations. Setting $s(i) = 1$ reduces to the sequential algorithm in [105] and [43]. In this algorithm, one might wonder why we don't schedule all samples initially, considering that the sample responses do not directly appear in (8.14). Regarding this, we note that the

---
[2]https://github.com/pytorch/pytorch

sample responses do influence (8.14) implicitly via hyperparameters $\boldsymbol{\theta}$ (specifying the prior we give to f). What we updated in each iteration is actually the newly-learned hyperparameters. Should we know the hyperparameters in the beginning, we can determine all samples in one batch without worrying about discarding the benefits of the iterative model updates. As we consider all samples together, the batch design can perform better than sequential sampling, which will be demonstrated in the next section.

8.2.4   Potential improvement on sampling efficiency

In this section, we show the improvement of batch sampling efficiency over sequential sampling achieved through known hyperparameters (the prior). As discussed in §8.1, although we do not have function evaluations everywhere, it is not rare to know other properties of the function, for example, discontinuity or variability along axes expressed as hyperparameters. In this way, we can generate the samples in one batch in the beginning:

$$\mathbf{X}_s^* = \mathrm{argmax}_{\tilde{\mathbf{X}}_s} \int \mathbf{k}(\mathbf{x}, \tilde{\mathbf{X}}_s) p_\mathbf{x}(\mathbf{x}) \mathrm{d}\mathbf{x} \, \mathbf{K}(\tilde{\mathbf{X}}_s, \tilde{\mathbf{X}}_s)^{-1} \int \mathbf{k}(\tilde{\mathbf{X}}_s, \mathbf{x}) p_\mathbf{x}(\mathbf{x}) \mathrm{d}\mathbf{x}, \qquad (8.20)$$

where the posterior covariance k_n in (8.14) is replaced by the prior covariance k.

In figure 8.1, we plot the standard deviation of $q|\mathcal{D}_n$ with batch design (one batch) and sequential design for standard Gaussian input and known hyperparameters $\boldsymbol{\theta} = \{4, \mathrm{I}_2\}$ with I_2 being a 2×2 identity matrix. It shows the batch design performs much better than the sequential design which is anticipated as we get the 'free lunch'—the benefits of a long-term perspective without any side effects from fewer model updates. The sampling positions of sequential design and batch design are plotted in figure 8.2. The batch samples show beautiful symmetric structures fitting the symmetric input and hyperparameters. In contrast, sequential samples show a strong greedy pattern.
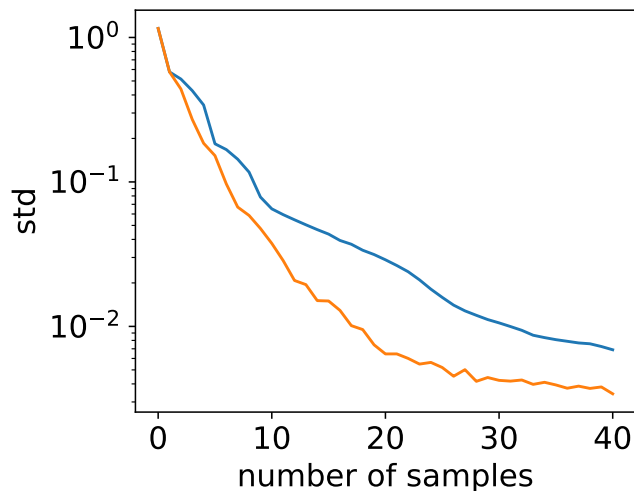
Figure 8.1: The standard deviation of $q|\mathcal{D}_n$ computed by sequential design (—) and batch design (—) for Gaussian input $\mathbf{x} \sim \mathcal{N}(\mathbf{0}, I_2)$ and known hyperparameters $\boldsymbol{\theta} = \{4, I_2\}$.

For example, when we have three samples, the sequential samples clearly favor one direction while batch samples form an equilateral triangle.

8.3 Results

In this section, we test the performance of the proposed batch design algorithm in two cases: (1) a larger number of complex functions from realizations of Gaussian processes in §8.3.1, and (2) a stochastic oscillator in §8.3.2. In each case, we compare the results of batch design (*batch-design*) with sequential design (*seq-design*), direct random sampling (*random*), and random sampling with Gaussian process surrogate (*random-gpr*). For *random*, the expectation is directly computed as the mean of samples, while for *random-gpr* the expectation is computed with a surrogate learned from random samples. The comparison between *random-gpr* and *random* highlights the impact of imposing a prior for f, while the advantage of choosing optimal samples over random samples is evidenced in the contrast between *seq-design* and *random-gpr*.
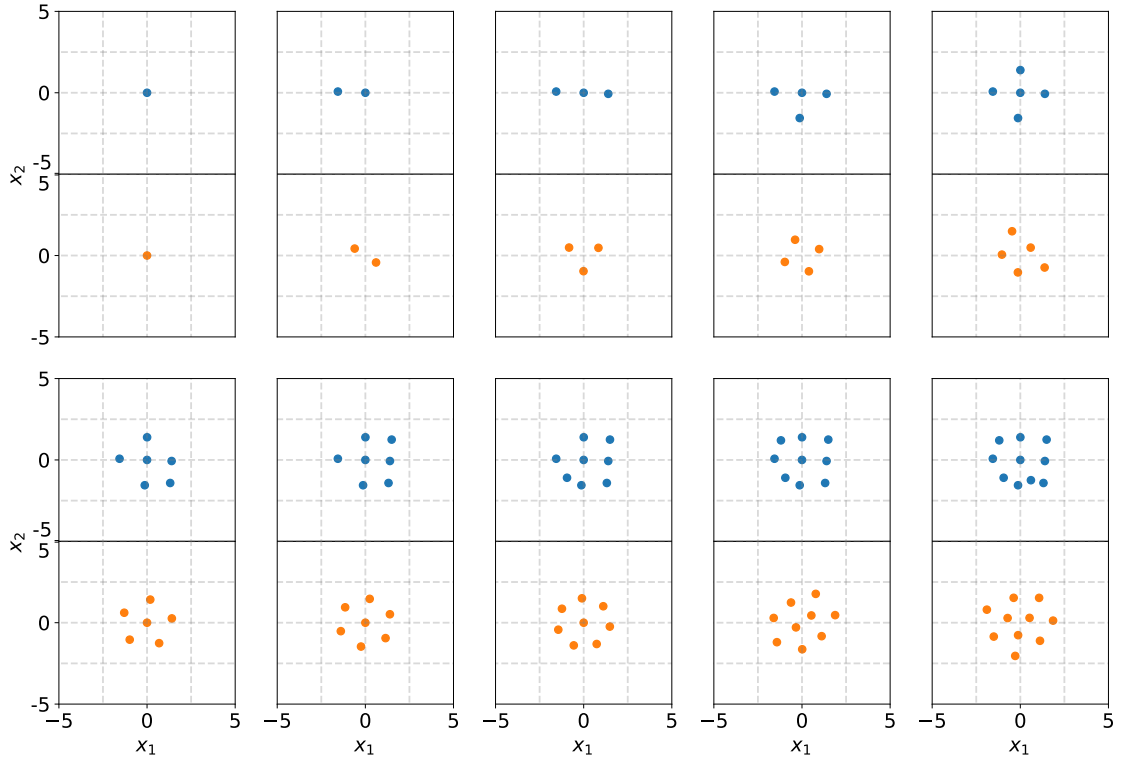
Figure 8.2: Sampling position of sequential design (•) and batch design (•) for Gaussian input $\mathbf{x} \sim \mathcal{N}(\mathbf{0}, \mathrm{I}_2)$ and known hyperparameters $\boldsymbol{\theta} = \{4, \mathrm{I}_2\}$.
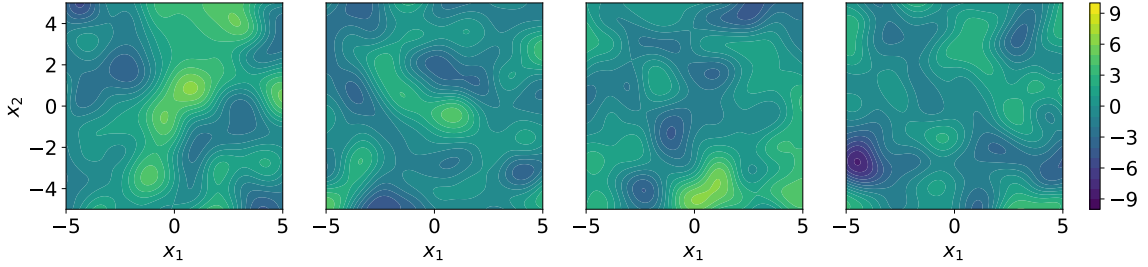
Figure 8.3: Examples of two-dimensional RBF functions.

<div style="text-align:center">(a)   (b)</div>
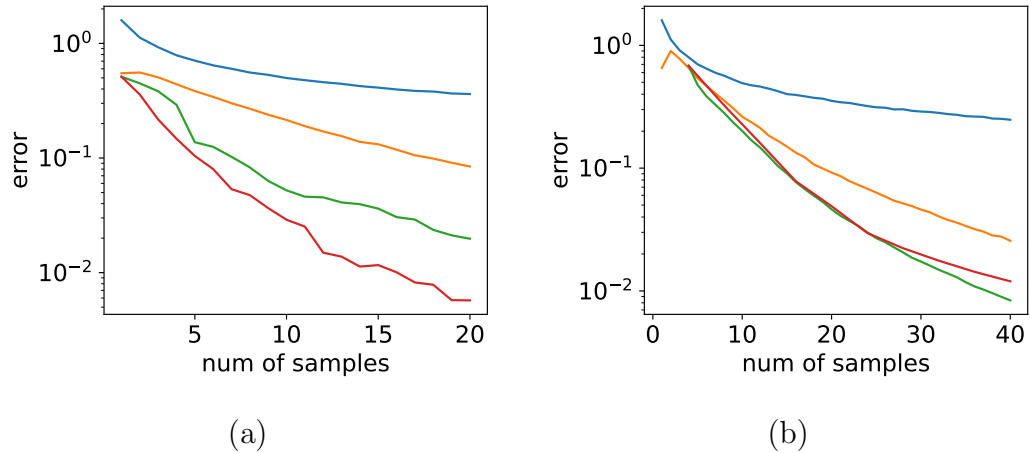
Figure 8.4: Results of RBF functions with (a) known hyperparameters and (b) learned hyperparameters: *random* (——), *random-gpr* (——), *seq-design* (——), and *batch-design* (——) ($s = 4$).

Finally, the difference between *batch-design* and *seq-design* measures the effectiveness of picking a group of samples simultaneously instead of a single sample during each iteration.

8.3.1 RBF functions

We test the proposed algorithm in 100 two-dimensional functions constructed from RBF kernel. The hyperparameters for generating these functions are $\boldsymbol{\theta} = \{4, I_2\}$ with examples shown in figure 8.3.

The results for a standard Gaussian input $p_{\mathbf{x}}(\mathbf{x})$ with the assumption of known hyperparameters are demonstrated in figure 8.4(a). Considering there are 100 dif-

ferent functions, we average the error across all functions where, in each function, the error is computed in a mean-squared form of 50 runs considering the randomness in drawing samples. For *seq-design* and *batch-design*, the sampling position is fixed, so we will directly take the fixed error. For *batch-design*, we sample only one batch in the beginning as we assume the hyperparameters are known. From figure 8.4(a), we can see that methods are ranked in an increasing performance from *random* to *random-gpr* to *seq-design* and finally *batch-design*. That means the prior information is useful and a careful design would also improve the performance. Regarding the design method, the batch design is better than the sequential design as it optimizes all samples as a whole.

We further consider situations where the hyperparameters are unknown with results shown in figure 8.4(b). For both *batch-design* and *seq-design*, we use 4 initial samples, and the error of each function is also computed in a mean squared form across different initializations. The *batch-design* with $s(i) = 4$ performs almost the same with *seq-design*, meaning the pro of a non-myopic design is actually offset by the con of fewer hyperparameters updates. But we note that the wall computational time of *batch-design* is only a quarter of *seq-design*.

8.3.2 Stochastic oscillator

We next consider a stochastic oscillator used in §5.3.2 and §7.3.1.1 with a standard Gaussian input and contour of ItR shown in figure 8.5(a). We plot the results for different methods in figure 8.5(b). All results are mean-squared errors with randomness in *random* and *random-gpr* coming from random sampling and randomness in *seq-design* and *batch-design* coming from initializations. For *batch-design*, we test both $s(i) = 2$ and $s(i) = 4$. It demonstrates that *seq-design* performs best among all while *batch-design* with $s(i) = 2$ is almost on par with *seq-design* albeit slightly less efficient.
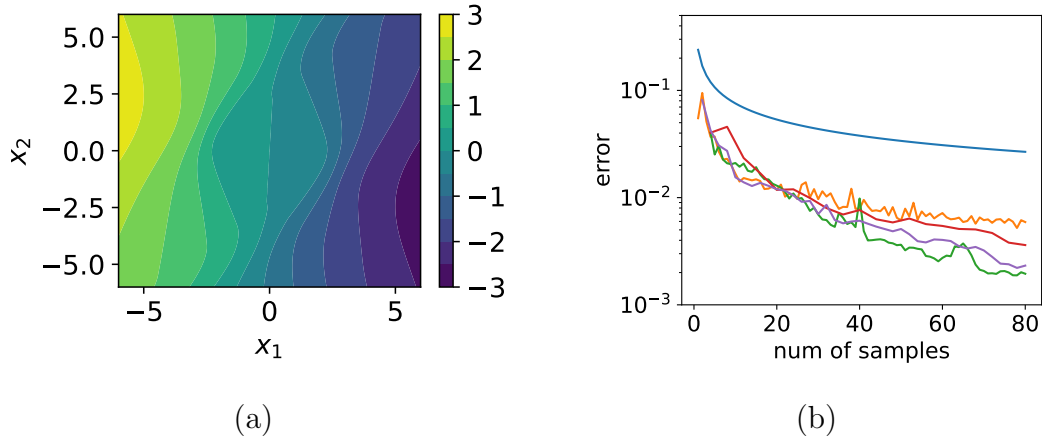
Figure 8.5: (a) response function of the stochastic oscillator. (b) results of *random* (——), *random-gpr* (——), *seq-design* (——), *batch-design* with $s(i) = 4$ (——), *batch-design* with $s(i) = 2$ (——).

8.4   Summary

In this chapter, we develop a non-myopic batch Bayesian experimental design algorithm for statistical expectation, where the next batch of samples are selected to maximize the information gained (or equivalently to minimize the estimation uncertainty) when they are added together. We apply the results in two test cases, showing that if the hyperparameters (prior) are known, the batch design algorithm converges much faster than the sequential design. For more common situations requiring learned hyperparameters, the batch design algorithm performs marginally below sequential design, however, with a big saving on wall time. We emphasize that this conclusion is only for statistical expectation. The performance of batch sampling targeting extreme-event statistics requires further developments and tests.

CHAPTER 9

Conclusion and Future Work

9.1 Conclusion

In this dissertation, we have presented a set of methods to efficiently compute the extreme-event statistics of nonlinear systems with stochastic input. These methods follow a general framework consisting of two engines, dimension reduction and surrogate modeling achieved through active learning.

In §2, we introduced the computational framework in the context of quantifying the extreme-event statistics of the ship response in a random wave field. In addition to some algorithmic improvements in wave group parameterization and sequential sampling following existing works, we also incorporate full system dynamics through nonlinear wave simulation and ship responses CFD simulation. In §3, we further adapted the computational framework to compute the temporal exceeding probability of ship responses, i.e., the fraction of time that the response exceeds a specified threshold. As an important complement to group-maximum statistics in §2, temporal exceeding probability contains more information about 'severity' and provides a robust measure of the extreme motion due to its invariance regarding different group definitions. In §4 we presented a new method to efficiently estimate the extreme-event probability induced by a random ItR with heteroscedastic uncertainty that could arise from the dimension reductions, e.g., group parameterization in §2 and §3.

The method leverages VHGPR as a surrogate model and a new acquisition function considering randomness from both input and ItR.

In §5, we added another tool to our toolkit to leverage low-fidelity models, for instance, a low-resolution CFD simulation whose cost is only a certain fraction of its high-resolution counterpart. In detail, we employ the multi-fidelity Gaussian process regression as the surrogate model and design a new acquisition function with its analytic approximation that allows the fast selection of both the sampling location and fidelity level. In §6, an information-theoretic acquisition to estimate the accident rate of connected and automated vehicles was developed and further extended to multi-fidelity context following §5. In §7, we generalized the likelihood-weighed algorithm, a widely-used one in recent years (also in §5), with two additional parameters. These parameters are able to (i) place different levels of emphasis on rare-event regions in sampling, and (ii) remedy the situation where the surrogate model has a significant deviation from the true ItR function. Our last work is §8 where a batch sampling strategy was designed to reduce the wall computational time while simultaneously improving (or keeping) the sampling efficiency.

9.1.1 Contributions

We highlight our contributions in the following three points:

1. Establishment of an Efficient Computational Framework: We developed a computational framework that integrates dimension reduction and active learning for quantifying extreme ship responses in random waves. Special emphasis was placed on:

- considering full nonlinear dynamics of wave and wave-ship interactions,

- resolving an improved statistical measure—the temporal exceeding probability,

- addressing uncertainties arising from dimension reductions.

165

2. Advancement in Multi-Fidelity Active Learning Methods: Our pioneering work in multi-fidelity active learning techniques is directed towards:

- quantifying extreme response PDF including a comparative analysis of two strategies for selecting sample fidelity,

- evaluating the exceeding probability with an information-theoretic acquisition in the context of safety analysis of automated vehicles.

3. Enhancements in Sampling Methods: We substantially refined the sampling strategy specifically by:

- generalizing the likelihood-weighted acquisition for extreme response PDF with two additional parameters,

- designing a batch sampling strategy to reduce computational time while maintaining or improving sampling efficiency for statistical expectation.

9.2 Limitation and future work

9.2.1 Broadband wave field

In this dissertation, we operate under the assumption of narrow-band wave fields. This assumption facilitates the decomposition of the wave field into consecutive Gaussian groups. Our final goal is to expand the current method to more realistic broadband wave fields, where groups may no longer live in a Gaussian shape. One preliminary thought is to define a dangerous group as a series of consecutive single waves with the majority of amplitudes above a given threshold. Each group is characterized by its length and maximum amplitude. Such simplification paves the road for a feasible low-dimensional sampling. However, a complete neglect of reduced dimensions will certainly lead to an inaccurate result considering the intrinsic high-dimensionality of the broadband wave field. For example, groups with the same characteristic length

166

and height have significantly different detailed shapes, as demonstrated in [45]. These detailed shapes, together with the initial condition of the group responses, lead to a random exceeding time for groups with the same parameters. To deal with this randomness, the method developed in §4 will be a critical component.

9.2.2   Choice of surrogates

Considering a moderate number of input dimensions and samples, we employ Gaussian process regression as our surrogate. The Gaussian process regression, however, suffers from a scale issue with a cubic complexity of the number of samples. To deal with a large number of samples, many approximation algorithms are available to reduce the scale, for example, Nyström approximation [153] or variational learning (i.e., sparse Gaussian process) [138]. From an optimization perspective, we can also efficiently train the Gaussian process by conjugate gradient descent with major operation matrix-matrix-multiplication, enabling the utilizing of GPU hardware [39], or by stochastic gradient descent enabling mini-batch updates. Regarding high dimensions, the Gaussian process encounters inherent limitations due to its non-parametric nature [16], usually accompanied by a dimension reduction algorithm as an initial stage. For 'big data problems', neural networks may be a better choice. Various methods have been proposed to quantify the uncertainty of neural networks, including Bayesian neural networks or ensemble networks. A comparison of these methods on scientific datasets can be found in [54]. Other common choices of the probabilistic surrogate in active learning include stochastic radial basis functions [122] and polynomial chaos expansion [154, 60], also deserving further research.

Additionally, it is worth noting that rather than learning the ItR function and treating the system as a black box, we can dive into the inside of the system if the evolution of its states is (partially) observable. This allows us to build a surrogate to learn the dynamics behind the system evolution, where the inputs are state variables

of the system. Such surrogate can be developed in a fully data-driven way [146] or a hybrid way to correct reduced-order dynamics from Galerkin projection [147] or a low-fidelity model [84]. The selection of different surrogate models and targets depends on both the complexity of the system and the availability of data, warranting additional research.

9.2.3 Pre-trained sampling policy

We also remark that acquisitions in this dissertation are all developed from the first principles with necessary assumptions made. Although they perform better than the existing ones, the optimality of their format cannot be guaranteed. A similar situation happens for the two parameters α and t in §7 which optimal values cannot be theoretically determined but suggested based on empirical tests. While these rule-of-thumb values are helpful, it may be more desirable to develop other advanced methods. That motivates us to consider a data-driven approach. One idea is to apply reinforcement learning to train (1) a policy network directly for the next-best sample $\mathbf{x}^* = \pi(\mathcal{D})$ given the existing dataset in a fully data-driven way, or (2) a policy network for underdetermined parameters in an acquisition, e.g., $(\alpha^*, t^*) = \pi(\mathcal{D})$ in a hybrid data-driven and first-principle way. In such a manner, the format or parameters of acquisitions can also vary in the sampling process for a given case, achieving even better performance. Some pioneering work has been done in terms of global optimization (meta-learning) [23, 24] and experiment design [124].

9.2.4 Derivative as observations

Finally, all observations in this dissertation are restricted to the values of ItR, i.e., $f(\mathbf{x})$. In many cases, we may have access to the derivative $f'(\mathbf{x})$ that greatly facilitates the learning of the surrogate [112]. For numerical simulations, the derivative can be computed from adjoint method [106, 86] or automatic differentiation like JAX-Fluids

[13]. With derivative observations, the acquisition function should be re-designed to consider the benefits and costs of the derivative. We leave this task to our future work.

APPENDIX A

Gaussian Process Regression

In this section, we briefly introduce the Gaussian process regression (GPR) [113], which is a probabilistic machine learning approach. Consider the task of inferring the underline relatioin f from dataset $\mathcal{D} = \{\mathbf{x}^i, y^i = f(\mathbf{x}^i)\}_{i=1}^n$ consisting of n inputs $\mathbf{X} = \{\mathbf{x}^i \in \mathbb{R}^d\}_{i=1}^{i=n}$ and the corresponding outputs $\mathbf{y} = \{y^i \in \mathbb{R}\}_{i=1}^{i=n}$. In GPR, a prior, representing our beliefs over all possible functions we expect to observe, is placed on f as a Gaussian process $f(\mathbf{x}) \sim \mathcal{GP}(0, k(\mathbf{x}, \mathbf{x}'))$ with zero mean and covariance function k (usually defined by a radial-basis-function (RBF) kernel):

$$k(\mathbf{x}, \mathbf{x}') = \tau^2 \exp(-\frac{1}{2}((\mathbf{x} - \mathbf{x}')^T \Lambda^{-1}(\mathbf{x} - \mathbf{x}'))), \tag{A.1}$$

where τ (characteristic amplitude) and diagonal matrix Λ (characteristic length scales) are hyperparameters determined by maximizing the likelihood $p(\mathbf{y})$.

Following the Bayes' theorem, the posterior prediction for f given the dataset $\mathcal{D}$ can be derived to be another Gaussian:

$$f(\mathbf{x})|\mathcal{D} \sim \mathcal{GP}\big(\mathbb{E}(f(\mathbf{x})|\mathcal{D}), \text{cov}(f(\mathbf{x}), f(\mathbf{x}')|\mathcal{D})\big), \tag{A.2}$$

with mean and covariance respectively:

$$\mathbb{E}(f(\mathbf{x})|\mathcal{D}) = k(\mathbf{x}, \mathbf{X})\mathrm{K}(\mathbf{X}, \mathbf{X})^{-1}\mathbf{y}, \tag{A.3}$$

$$\mathrm{cov}(f(\mathbf{x}), f(\mathbf{x}')|\mathcal{D}) = k(\mathbf{x}, \mathbf{x}') - k(\mathbf{x}, \mathbf{X})\mathrm{K}(\mathbf{X}, \mathbf{X})^{-1}k(\mathbf{X}, \mathbf{x}'), \tag{A.4}$$

where matrix element $\mathrm{K}(\mathbf{X}, \mathbf{X})_{ij} = k(\mathbf{x}^i, \mathbf{x}^j)$. Suppose we have a new sample $\{\tilde{\mathbf{x}}, \tilde{y}\}$, we can update (A.2) via recursive formula:

$$\mathbb{E}(f(\mathbf{x})|\mathcal{D}, \tilde{\mathbf{x}}, \tilde{y}) = \mathbb{E}(f(\mathbf{x})|\mathcal{D}) + \frac{\mathrm{cov}(f(\mathbf{x}), f(\tilde{\mathbf{x}})|\mathcal{D})}{\mathrm{var}(f(\tilde{\mathbf{x}})|\mathcal{D})}\big(\tilde{y} - \mathbb{E}(f(\tilde{\mathbf{x}})|\mathcal{D})\big), \tag{A.5}$$

$$\mathrm{cov}(f(\mathbf{x}), f(\mathbf{x}')|\mathcal{D}, \tilde{\mathbf{x}}, \tilde{y}) = \mathrm{cov}(f(\mathbf{x}), f(\mathbf{x}')|\mathcal{D}) - \frac{\mathrm{cov}(f(\mathbf{x}), f(\tilde{\mathbf{x}})|\mathcal{D})^2}{\mathrm{var}(f(\tilde{\mathbf{x}})|\mathcal{D})}. \tag{A.6}$$

APPENDIX B

Appendix of Bayesian Experimental Design for Extreme-Event Probability in Stochastic Input-to-Response Systems

B.1 Gaussian process regression for stochastic functions

We consider the task of inferring the input to response (ItR) function from a dataset $\mathcal{D} = \{\mathbf{x}^i, y^i\}_{i=1}^{i=n}$ consisting of n inputs $\mathbf{X} = \{\mathbf{x}^i \in \mathbb{R}^d\}_{i=1}^{i=n}$ and the corresponding outputs $\mathbf{y} = \{y^i \in \mathbb{R}\}_{i=1}^{i=n}$.

B.1.1 Standard Gaussian process regression (SGPR)

SGPR assumes the function to be a sum of a mean $f(\mathbf{x})$ and a Gaussian randomness with constant variance γ_0^2 (at all $\mathbf{x}$):

$$y = f(\mathbf{x}) + R \quad R \sim \mathcal{N}(0, \gamma_0^2), \tag{B.1}$$

A prior, representing our beliefs over all possible functions we expect to observe, is placed on f as a Gaussian process $f(\mathbf{x}) \sim \mathcal{GP}(0, k_f(\mathbf{x}, \mathbf{x}'))$ with zero mean and

covariance function k_f (usually defined by a radial-basis-function kernel):

$$k_f(\mathbf{x}, \mathbf{x}') = \tau^2 \exp(-\frac{1}{2}((\mathbf{x} - \mathbf{x}')^T \Lambda^{-1}(\mathbf{x} - \mathbf{x}'))), \tag{B.2}$$

where τ and diagonal matrix Λ, together with γ_0, are hyperparameters $\boldsymbol{\theta} = \{\tau, \Lambda, \gamma_0\}$ in SGPR.

Following the Bayesian formula, the posterior prediction for f given the dataset $\mathcal{D}$ can be derived to be another Gaussian:

$$f(\mathbf{x})|\mathcal{D} \sim \mathcal{GP}\big(\mathbb{E}(f(\mathbf{x})|\mathcal{D}), \text{cov}(f(\mathbf{x}), f(\mathbf{x}')|\mathcal{D})\big), \tag{B.3}$$

with analytically tractable mean and covariance:

$$\mathbb{E}(f(\mathbf{x})|\mathcal{D}) = k_f(\mathbf{x}, \mathbf{X})(\text{K}_f(\mathbf{X}, \mathbf{X}) + \gamma_0^2 \text{I})^{-1}\mathbf{y}, \tag{B.4}$$

$$\text{cov}(f(\mathbf{x}), f(\mathbf{x}')|\mathcal{D}) = k_f(\mathbf{x}, \mathbf{x}') - k_f(\mathbf{x}, \mathbf{X})(\text{K}_f(\mathbf{X}, \mathbf{X}) + \gamma_0^2 \text{I})^{-1}k_f(\mathbf{X}, \mathbf{x}'), \tag{B.5}$$

where matrix element $\text{K}_f(\mathbf{X}, \mathbf{X})_{ij} = k_f(\mathbf{x}^i, \mathbf{x}^j)$. The hyperparameters $\boldsymbol{\theta}$ are determined which maximizes the likelihood function $p(\mathcal{D}|\boldsymbol{\theta}) \equiv p(\mathbf{y}|\boldsymbol{\theta}) = \mathcal{N}(0, \text{K}_f(\mathbf{X}, \mathbf{X}) + \gamma_0^2 \text{I})$.

B.1.2 Variational heteroscedastic Gaussian process regression (VHGPR)

In this section, we briefly outline the algorithm of VHGPR. The purpose is to provide the reader enough information to understand the logic behind VHGPR. For the conciseness of the presentation, some details in the algorithm have to be omitted. We recommend the interested readers to read [72] and §10 in [17] for details.

In VHGPR, the function of ItR is considered as the sum of a mean and a Gaussian

term (independent at all $\mathbf{x}$) with heteroscedastic uncertainty:

$$y = f(\mathbf{x}) + R \quad R \sim \mathcal{N}(0, e^{g(\mathbf{x})}). \tag{B.6}$$

Two Gaussian priors are placed on the mean and the (log) variance function:

$$f(\mathbf{x}) \sim \mathcal{GP}(0, k_f(\mathbf{x}, \mathbf{x}')), \tag{B.7}$$

$$g(\mathbf{x}) \sim \mathcal{GP}(\mu_0, k_g(\mathbf{x}, \mathbf{x}')), \tag{B.8}$$

where k_f and k_g are respectively the radial-basis-function kernels for f and g, defined similarly as (B.2). μ_0 is the prior mean for g. The hyperparameters in VHGPR can then be defined as $\boldsymbol{\theta} = (\mu_0, \boldsymbol{\theta}_f, \boldsymbol{\theta}_g)$ where $\boldsymbol{\theta}_{f,g}$ includes the amplitudes and length scales involved in k_f or k_g. Moreover, we assume f is independent with g.

The increased expressive power with the heteroscedastic variance is at the cost of analytically intractable likelihood (for determination of hyperparameters) and posterior (prediction). Let $\mathbf{f} = f(\mathbf{X})$ and $\mathbf{g} = g(\mathbf{X})$ denote the realizations of mean and variance at training inputs $\mathbf{X}$ following the distributions in (B.7) and (B.8) . The likelihood and prediction are formulated as:

$$p(\mathcal{D}|\boldsymbol{\theta}) \equiv p(\mathbf{y}|\boldsymbol{\theta}) = \iint p(\mathbf{y}|\mathbf{f}, \mathbf{g}) p(\mathbf{g}|\boldsymbol{\theta}) p(\mathbf{f}|\boldsymbol{\theta}) \, \mathrm{d}\mathbf{f} \mathrm{d}\mathbf{g}, \tag{B.9}$$

$$p(f(\mathbf{x}), g(\mathbf{x})|\mathcal{D}) = \iint p(f(\mathbf{x}), g(\mathbf{x})|\mathbf{f}, \mathbf{g}) p(\mathbf{f}, \mathbf{g}|\mathbf{y}) \, \mathrm{d}\mathbf{f} \mathrm{d}\mathbf{g}. \tag{B.10}$$

Since analytical integration cannot be achieved for (B.9) and (B.10), numerical evaluations of the integrals are needed for their computations. However, the dimension of integration (w.r.t $\mathbf{f}$ and $\mathbf{g}$) is the same as number of data points n, which can be prohibitively high for a direct integration, say, using quadrature methods. While the Monte-Carlo method (e.g., MCMC) offers some advantages in computational cost, its

application is still too expensive for most practical problems. For these problems, the VHGPR leveraging variational inference is the only method which provides practical solutions with low computational cost and sufficient accuracy.

The key distribution in computing (B.10) and (B.9) is $p(\mathbf{f}, \mathbf{g}|\mathbf{y})$ (directly involved in (B.10) and related to (B.9) due to (B.11) that will be discussed), which is however expensive to compute directly. The key idea in VHGPR is to approximate $p(\mathbf{f}, \mathbf{g}|\mathbf{y})$ by $q(\mathbf{f}, \mathbf{g})$, where the latter is assumed to have a Gaussian distribution with parameters (multi-dimensional means and covariance) denoted here by $\boldsymbol{\theta}_q$. Through minimizing the KL divergence [71] between $p(\mathbf{f}, \mathbf{g}|\mathbf{y})$ and $q(\mathbf{f}, \mathbf{g})$, the parameters $\boldsymbol{\theta}_q$ can be determined and both the posterior and likelihood can be evaluated accordingly as discussed below.

For the posterior, (B.10) becomes a linear Gaussian model (an integration of the exponential of quadratic function of $\mathbf{f}$ and $\mathbf{g}$ with Gaussian weights) which has an analytical formulation. For the likelihood, we avoid directly using (B.9) but decompose $\log p(\mathbf{y}|\boldsymbol{\theta})$ as a summation of the evidence lower bound (ELBO) $\mathcal{L}(q(\mathbf{f}, \mathbf{g}))$ and the K-L divergence between $q(\mathbf{f}, \mathbf{g})$ and $p(\mathbf{f}, \mathbf{g}|\mathbf{y})$[1]:

$$\log p(\mathbf{y}|\boldsymbol{\theta}) = \mathcal{L}(q(\mathbf{f}, \mathbf{g})) + \text{KL}(q(\mathbf{f}, \mathbf{g})|p(\mathbf{f}, \mathbf{g}|\mathbf{y})), \tag{B.11}$$

where

$$\mathcal{L}(q(\mathbf{f}, \mathbf{g})) = \iint q(\mathbf{f}, \mathbf{g}) \log \frac{p(\mathbf{y}, \mathbf{f}, \mathbf{g})}{q(\mathbf{f}, \mathbf{g})} \mathrm{d}\mathbf{f}\mathrm{d}\mathbf{g}, \tag{B.12}$$

$$\text{KL}(q(\mathbf{f}, \mathbf{g})|p(\mathbf{f}, \mathbf{g}|\mathbf{y})) = \iint q(\mathbf{f}, \mathbf{g}) \log \frac{q(\mathbf{f}, \mathbf{g})}{p(\mathbf{f}, \mathbf{g}|\mathbf{y})} \mathrm{d}\mathbf{f}\mathrm{d}\mathbf{g}. \tag{B.13}$$

We then formulate an optimization problem of $\mathcal{L}$ (where $\mathcal{L}$, as a weighted integration

---

[1]This decomposition can be derived from manipulation of (B.13) to be

$$\text{KL}(q(\mathbf{f}, \mathbf{g})|p(\mathbf{f}, \mathbf{g}|\mathbf{y})) = -\iint q(\mathbf{f}, \mathbf{g}) \log \frac{p(\mathbf{y}, \mathbf{f}, \mathbf{g})}{q(\mathbf{f}, \mathbf{g})} \mathrm{d}\mathbf{f}\mathrm{d}\mathbf{g} + \log p(\mathbf{y}).$$

of the exponential function of $\mathbf{f}$ and $\mathbf{g}$, has an analytical expression for Gaussian weights $q(\mathbf{f}, \mathbf{g})$ [17]) to determine both the parameters in $q(\mathbf{f}, \mathbf{g})$ ($\boldsymbol{\theta}_q$) and the hyper-parameters ($\boldsymbol{\theta}$):

$$\boldsymbol{\theta}_q^*, \boldsymbol{\theta}^* = \operatorname{argmax}_{\boldsymbol{\theta}_q, \boldsymbol{\theta}} \mathcal{L}(\boldsymbol{\theta}_q, \boldsymbol{\theta}, \mathbf{y}). \tag{B.14}$$

We remark that (B.14) can be conceived as two sequential optimization problems with respect to $\boldsymbol{\theta}_q$ and $\boldsymbol{\theta}$. For the former, maximizing $\mathcal{L}$ is equivalent to minimizing the KL divergence (as an aforementioned goal) since $\log p(\mathbf{y}|\boldsymbol{\theta})$ in (B.11) is not a function of $\boldsymbol{\theta}_q$. For the latter, since the KL divergence has been minimized, the ELBO gives a good approximation of likelihood $\log p(\mathbf{y}|\boldsymbol{\theta})$. Therefore, the solution of (B.14) simultaneously provides the optimal $\boldsymbol{\theta}_q$ leading to a good approximation of $p(\mathbf{f}, \mathbf{g}|\mathbf{y})$ by $q(\mathbf{f}, \mathbf{g})$, as well as the optimal $\boldsymbol{\theta}$ leading to a maximized likelihood.

However, (B.14) with respect to $\boldsymbol{\theta}_q$ is still a prohibitively expensive optimization with dimensions $2n + 2n(2n+1)/2$ (i.e., number of unique elements in the mean and co-variance matrix of $q(\mathbf{f}, \mathbf{g})$). To reduce the number of dimensions, a key procedure employed in VHGPR is to assume that the function $q(\mathbf{f}, \mathbf{g})$ is separable in $\mathbf{f}$ and $\mathbf{g}$, i.e., $q(\mathbf{f}, \mathbf{g}) = q_f(\mathbf{f}) q_g(\mathbf{g})$. This brings two benefits: First, one can show that the maximized solution of $\mathcal{L}$ involves a relation between $q_f(\mathbf{f})$ and $q_g(\mathbf{g})$, i.e., $q_f(\mathbf{f})$ can be represented as a function of $q_g(\mathbf{g})$ so that the parameters in $\boldsymbol{\theta}_q$ is reduced to the mean $\boldsymbol{\mu}$ and covariance Σ of $q_g(\mathbf{g})$ with dimension $n + n(n+1)/2$ (see (10.6) in [17] for proof). Second, the stationary point of $\mathcal{L}$ with respect to $\boldsymbol{\mu}$ and Σ (by making $\partial \mathcal{L}/\partial \boldsymbol{\mu} = \mathbf{0}$ and $\partial \mathcal{L}/\partial \Sigma = \mathbf{0}$) leads to an analytical form

$$\boldsymbol{\mu} = \mathrm{K}_\mathrm{g}(\mathbf{X}, \mathbf{X})(\mathbf{V} - \frac{1}{2}\mathrm{I})\mathbf{1} + \mu_0 \mathbf{1}, \tag{B.15}$$

$$\Sigma = (\mathrm{K}_\mathrm{g}(\mathbf{X}, \mathbf{X})^{-1} + \mathbf{V})^{-1}, \tag{B.16}$$

with $\mathrm{K}_g(\mathbf{X}, \mathbf{X})_{ij} = k_g(\mathbf{x}^i, \mathbf{x}^j)$, $\mathbf{V}$ being a diagonal matrix involving n unknown parameters and $\mathbf{1}$ being a vector with all elements 1. Therefore the optimization (B.14)

is finally reduced to

$$\mathbf{V}^*, \boldsymbol{\theta}^* = \mathrm{argmax}_{\mathbf{V},\boldsymbol{\theta}}\ \mathcal{L}(\boldsymbol{\mu}(\mathbf{V}), \Sigma(\mathbf{V}), \boldsymbol{\theta}, \mathbf{y}), \tag{B.17}$$

with only n parameters in $\boldsymbol{\theta}_q$ (or $\mathbf{V}$). This can be solved by gradient-based method, with the major computational cost in computing $n \times n$ matrix inversions in $\mathcal{L}$. The computational cost of each iteration in (B.17) is only approximately twice as that in SGPR ($O(n^3)$), which is significantly lower than the complexity $O(n^3 m)$ for a direct computation of (B.9) and (B.10) with generally $m \gg n$ quadrature points. To further reduce the complexity, particularly for a large dataset, the sparse VHGPR [76] (as a combination of VHGPR with the Sparse Gaussian process) with $O(nu^2)$ can be leveraged, with $u \ll n$ the number of pseudo points.

With $q_f(\mathbf{f})$, $q_g(\mathbf{g})$ (computed from $\mathbf{V}^*$), and $\boldsymbol{\theta}$ available, the posterior predictions for f and g in (B.10) are:

$$
\begin{aligned}
p(f(\mathbf{x}), g(\mathbf{x})|\mathcal{D}) &= \iint p(f(\mathbf{x}), g(\mathbf{x})|\mathbf{f}, \mathbf{g}) p(\mathbf{f}, \mathbf{g}|\mathbf{y}) \mathrm{d}\mathbf{f} \mathrm{d}\mathbf{g} \\
&\approx \iint p(f(\mathbf{x}), g(\mathbf{x})|\mathbf{f}, \mathbf{g}) q_f(\mathbf{f}) q_g(\mathbf{g}) \mathrm{d}\mathbf{f} \mathrm{d}\mathbf{g} \\
&= \iint p(f(\mathbf{x})|\mathbf{f}) p(g(\mathbf{x})|\mathbf{g}) q_f(\mathbf{f}) q_g(\mathbf{g}) \mathrm{d}\mathbf{f} \mathrm{d}\mathbf{g} \\
&\qquad \text{(independence of } f(\mathbf{x}) \text{ and } g(\mathbf{x})) \\
&= \int p(f(\mathbf{x})|\mathbf{f}) q_f(\mathbf{f}) \mathrm{d}\mathbf{f} \int p(g(\mathbf{x})|\mathbf{g}) q_g(\mathbf{g}) \mathrm{d}\mathbf{g} \\
&\approx p(f(\mathbf{x})|\mathcal{D}) p(g(\mathbf{x})|\mathcal{D}), \tag{B.18}
\end{aligned}
$$

where:

$$p(f(\mathbf{x})|\mathcal{D}) = \mathcal{N}\big(\mathbb{E}(f(\mathbf{x})|\mathcal{D}), \mathrm{cov}(f(\mathbf{x}), f(\mathbf{x}')|\mathcal{D})\big), \tag{B.19}$$

$$p(g(\mathbf{x})|\mathcal{D}) = \mathcal{N}\big(\mathbb{E}(g(\mathbf{x})|\mathcal{D}), \mathrm{cov}(g(\mathbf{x}), g(\mathbf{x}')|\mathcal{D})\big), \tag{B.20}$$

$$\mathbb{E}(f(\mathbf{x})|\mathcal{D}) \approx k_f(\mathbf{x}, \mathbf{X})(\mathrm{K}_f(\mathbf{X}, \mathbf{X}) + Z)^{-1}\mathbf{y}, \tag{B.21}$$

$$\mathrm{cov}(f(\mathbf{x}), f(\mathbf{x}')|\mathcal{D}) \approx k_f(\mathbf{x}, \mathbf{x}') - k_f(\mathbf{x}, \mathbf{X})(\mathrm{K}_f(\mathbf{X}, \mathbf{X}) + Z)^{-1}k_f(\mathbf{X}, \mathbf{x}'), \tag{B.22}$$

$$\mathbb{E}(g(\mathbf{x})|\mathcal{D}) \approx k_g(\mathbf{x}, \mathbf{X})(\mathbf{V} - \frac{1}{2}\mathrm{I})\mathbf{1} + \mu_0, \tag{B.23}$$

$$\mathrm{cov}(g(\mathbf{x}), g(\mathbf{x}')|\mathcal{D}) \approx k_g(\mathbf{x}, \mathbf{x}') - k_g(\mathbf{x}, \mathbf{X})(\mathrm{K}_g(\mathbf{X}, \mathbf{X}) + \mathbf{V}^{-1})^{-1}k_g(\mathbf{X}, \mathbf{x}'), \tag{B.24}$$

with $Z_{ii} = e^{\mu_i - \Sigma_{ii}/2}$ being a diagonal matrix and $\mathrm{K}_f(\mathbf{X}, \mathbf{X})_{ij} = k_f(\mathbf{x}^i, \mathbf{x}^j)$.

B.2 The upper bound of the estimation variance

Here we show the construction of an upper bound of the estimation variance (For convenience we use $\hat{p}(\mathbf{x})$ to represent $\mathbb{P}(S(\mathbf{x}, \omega | f(\mathbf{x}), g(\mathbf{x})) > \delta)$):

$$
\begin{aligned}
& \mathrm{var}_{f,g}[\mathbb{P}(S(X, \omega | f(\mathbf{x}), g(\mathbf{x})) > \delta)] \\
&= \mathrm{var}_{f,g}\Big[\int \hat{p}(\mathbf{x}) p_{\mathbf{x}}(\mathbf{x}) \mathrm{d}\mathbf{x}\Big] \\
&= \mathbb{E}_{f,g}\Big[\Big(\int \hat{p}(\mathbf{x}) p_{\mathbf{x}}(\mathbf{x}) \mathrm{d}\mathbf{x}\Big)^2\Big] - \Big(\mathbb{E}_{f,g}\Big[\int \hat{p}(\mathbf{x}) p_{\mathbf{x}}(\mathbf{x}) \mathrm{d}\mathbf{x}\Big]\Big)^2 \\
&= \mathbb{E}_{f,g}\Big[\int \hat{p}(\mathbf{x}) p_{\mathbf{x}}(\mathbf{x}) \mathrm{d}\mathbf{x} \int \hat{p}(\mathbf{x}') p_X(\mathbf{x}') \mathrm{d}\mathbf{x}'\Big] \\
&\quad - \Big(\mathbb{E}_{f,g}\Big[\int \hat{p}(\mathbf{x}) p_{\mathbf{x}}(\mathbf{x}) \mathrm{d}\mathbf{x}\Big]\Big)\Big(\mathbb{E}_{f,g}\Big[\int \hat{p}(\mathbf{x}') p_X(\mathbf{x}') \mathrm{d}\mathbf{x}'\Big]\Big) \\
&= \iint \mathbb{E}_{f,g}\Big[\hat{p}(\mathbf{x})\hat{p}(\mathbf{x}')\Big] p_{\mathbf{x}}(\mathbf{x}) p_X(\mathbf{x}') \mathrm{d}\mathbf{x}\mathrm{d}\mathbf{x}' \\
&\quad - \iint \mathbb{E}_{f,g}\Big[\hat{p}(\mathbf{x})\Big]\mathbb{E}_{f,g}\Big[\hat{p}(\mathbf{x}')\Big] p_{\mathbf{x}}(\mathbf{x}) p_X(\mathbf{x}') \mathrm{d}\mathbf{x}\mathrm{d}\mathbf{x}' \\
&= \iint \mathrm{cov}_{f,g}\Big[\hat{p}(\mathbf{x}), \hat{p}(\mathbf{x}')\Big] p_{\mathbf{x}}(\mathbf{x}) p_X(\mathbf{x}') \mathrm{d}\mathbf{x}d\mathbf{x}' \\
&= \int \mathrm{std}_{f,g}\Big[\hat{p}(\mathbf{x})\Big]\Big(\int \mathrm{std}_{f,g}\Big[\hat{p}(\mathbf{x}')\Big]\rho\Big[\hat{p}(\mathbf{x}), \hat{p}(\mathbf{x}')\Big] p_X(\mathbf{x}') d\mathbf{x}'\Big) p_{\mathbf{x}}(\mathbf{x}) \mathrm{d}\mathbf{x} \\
&\leq 0.5 \int \mathrm{std}_{f,g}\Big[\hat{p}(\mathbf{x})\Big] p_{\mathbf{x}}(\mathbf{x}) \mathrm{d}\mathbf{x} \quad\quad\quad\quad\quad\quad\quad\quad\quad\text{(B.25)}
\end{aligned}
$$

where $\rho[\cdot, \cdot]$ denotes the correlation coefficient. The last inequality comes from $\mathrm{std}[\hat{p}_\omega(\mathbf{x}')] \leq 0.5$ and $\rho[\hat{p}_\omega(\mathbf{x}), \hat{p}_\omega(\mathbf{x}')] \leq 1$. The equality in (B.25) holds when $\hat{p}_\omega$ degenerates to a Bernoulli random variable with equal probability 0.5 at both $\hat{p}_\omega = 0$ and $\hat{p}_\omega = 1$ and $\rho[\hat{p}_\omega(\mathbf{x}), \hat{p}_\omega(\mathbf{x}')] = 1$.

B.3   Spherical cubature integration in equation (4.12)-(4.14)

For given $\mathbf{x}$, both equations (4.12) and (4.13) can be considered as two-dimensional ($d = 2$) Gaussian weighted integrals. Let $\mathbf{u}$ denote $\{f(\mathbf{x}), g(\mathbf{x})\}$ and $h(\mathbf{u})$ denote $\mathbb{P}(S(\mathbf{x}, \omega | \mathbf{u}) > \delta)$ or $(\mathbb{P}(S(\mathbf{x}, \omega | \mathbf{u}) > \delta) - m)^2$. Both (4.12) and (4.13) can be rewritten in the following form:

$$\int h(\mathbf{u}) \mathcal{N}(\mathbf{u}; \boldsymbol{\mu}, \Sigma) \mathrm{d}\mathbf{u}, \tag{B.26}$$

where

$$\boldsymbol{\mu} = [\mathbb{E}(f(\mathbf{x}) | \mathcal{D}) \ \ \mathbb{E}(g(\mathbf{x}) | \mathcal{D})]^T, \tag{B.27}$$

$$\Sigma = \begin{bmatrix} \mathrm{var}(f(\mathbf{x}) | \mathcal{D}) & 0 \\ 0 & \mathrm{var}(g(\mathbf{x}) | \mathcal{D}) \end{bmatrix}. \tag{B.28}$$

We further define a standard Gaussian random vector $\tilde{\mathbf{u}} = \sqrt{\Sigma}^{-1}(\mathbf{u} - \boldsymbol{\mu})$, where

$$\sqrt{\Sigma} = \begin{bmatrix} \mathrm{std}(f(\mathbf{x}) | \mathcal{D}) & 0 \\ 0 & \mathrm{std}(g(\mathbf{x}) | \mathcal{D}) \end{bmatrix}, \tag{B.29}$$

or more general, the Cholesky decomposition of Σ. Then (B.26) can be transformed to

$$\int h(\boldsymbol{\mu} + \sqrt{\Sigma}\tilde{\mathbf{u}}) \mathcal{N}(\tilde{\mathbf{u}}; \mathbf{0}, \mathrm{I}) \mathrm{d}\tilde{\mathbf{u}} = \int \hat{h}(\tilde{\mathbf{u}}) \mathcal{N}(\tilde{\mathbf{u}}; \mathbf{0}, \mathrm{I}) \mathrm{d}\tilde{\mathbf{u}}, \tag{B.30}$$

where we have defined $h(\boldsymbol{\mu} + \sqrt{\Sigma}\tilde{\mathbf{u}}) = \hat{h}(\tilde{\mathbf{u}})$ for simplicity.

The spherical cubature integration aims to approximate (B.30) with $2d$ points [120]:

$$\int \hat{h}(\tilde{\mathbf{u}}) \mathcal{N}(\tilde{\mathbf{u}}; \mathbf{0}, \mathrm{I}) \mathrm{d}\tilde{\mathbf{u}} \approx w \sum_{i=1}^{2d=4} \hat{h}(c\tilde{\mathbf{u}}^{(i)}), \tag{B.31}$$

where $\tilde{\mathbf{u}}^{(i)} \in \{\{1,0\},\{0,1\},\{-1,0\},\{0,-1\}\}$, and w and c are coefficients determined by satisfying the following conditions for third order accuracy:

$$\int \mathcal{N}(\tilde{\mathbf{u}};\mathbf{0},\mathrm{I})\mathrm{d}\tilde{\mathbf{u}} = w\sum_{i=1}^{4} 1, \tag{B.32}$$

$$\int \tilde{u}_j\mathcal{N}(\tilde{\mathbf{u}};\mathbf{0},\mathrm{I})\mathrm{d}\tilde{\mathbf{u}} = w\sum_{i=1}^{4}(c\tilde{u}_j^{(i)}), \tag{B.33}$$

$$\int \tilde{u}_j^2\mathcal{N}(\tilde{\mathbf{u}};\mathbf{0},\mathrm{I})\mathrm{d}\tilde{\mathbf{u}} = w\sum_{i=1}^{4}(c\tilde{u}_j^{(i)})^2, \quad j = 1,2. \tag{B.34}$$

According to [120], this yields $w = 1/4$ and $c = \sqrt{2}$.

Finally, the values of $\tilde{\mathbf{u}}^{(i)}$ can be transformed back to $\mathbf{u}^{(i)} = \boldsymbol{\mu} + \sqrt{\Sigma}\tilde{\mathbf{u}}^{(i)}$, corresponding to (4.14) in §4.2.3

APPENDIX C

Appendix of Multi-Fidelity Bayesian Experimental Design for Rare-Event Statistics

C.1 Bi-fidelity Gaussian process

The bi-fidelity Gaussian process conditioned on $\mathcal{D} = \{\{\mathcal{X}_1, \mathcal{X}_2\}, \{\mathcal{Y}_1, \mathcal{Y}_2\}\}$ can be considered as the counterpart of (5.4) for $s = 2$, with its mean and covariance formulated as

$$\mathbb{E}\left(\begin{bmatrix} f_1(\mathbf{x}) \\ f_2(\mathbf{x}) \end{bmatrix} \Big| \mathcal{D}\right) = \text{cov}\left(\begin{bmatrix} f_1(\mathbf{x}) \\ f_2(\mathbf{x}) \end{bmatrix}, \begin{bmatrix} \mathbf{Y}_1 \\ \mathbf{Y}_2 \end{bmatrix}\right)\text{cov}\left(\begin{bmatrix} \mathbf{Y}_1 \\ \mathbf{Y}_2 \end{bmatrix}\right)^{-1}\begin{bmatrix} \mathcal{Y}_1 \\ \mathcal{Y}_2 \end{bmatrix}, \tag{C.1}$$

$$\text{cov}\left(\begin{bmatrix} f_1(\mathbf{x}) \\ f_2(\mathbf{x}') \end{bmatrix} \Big| \mathcal{D}\right) = \text{cov}\left(\begin{bmatrix} f_1(\mathbf{x}) \\ f_2(\mathbf{x}') \end{bmatrix}\right) - \left(\text{cov}\left(\begin{bmatrix} f_1(\mathbf{x}) \\ f_2(\mathbf{x}') \end{bmatrix}, \begin{bmatrix} \mathbf{Y}_1 \\ \mathbf{Y}_2 \end{bmatrix}\right)\right) \tag{C.2}$$

$$\text{cov}\left(\begin{bmatrix} \mathbf{Y}_1 \\ \mathbf{Y}_2 \end{bmatrix}\right)^{-1}\text{cov}\left(\begin{bmatrix} \mathbf{Y}_1 \\ \mathbf{Y}_2 \end{bmatrix}, \begin{bmatrix} f_1(\mathbf{x}) \\ f_2(\mathbf{x}') \end{bmatrix}\right)\right), \tag{C.3}$$

where

$$\text{cov}\left(\begin{bmatrix} \mathbf{Y}_1 \\ \mathbf{Y}_2 \end{bmatrix}\right) = \begin{bmatrix} k_1(\mathcal{X}_1, \mathcal{X}_1) & \rho_1 k_1(\mathcal{X}_1, \mathcal{X}_2) \\ \rho_1 k_1(\mathcal{X}_2, \mathcal{X}_1) & \rho_1^2 k_1(\mathcal{X}_2, \mathcal{X}_2) + k_2(\mathcal{X}_2, \mathcal{X}_2) \end{bmatrix}, \qquad \text{(C.4)}$$

$$\text{cov}\left(\begin{bmatrix} f_1(\mathbf{x}) \\ f_2(\mathbf{x}') \end{bmatrix}, \begin{bmatrix} \mathbf{Y}_1 \\ \mathbf{Y}_2 \end{bmatrix}\right) = \begin{bmatrix} k_1(\mathbf{x}, \mathcal{X}_1) & \rho_1 k_1(\mathbf{x}, \mathcal{X}_2) \\ \rho_1 k_1(\mathbf{x}', \mathcal{X}_1) & \rho_1^2 k_1(\mathbf{x}', \mathcal{X}_2) + k_2(\mathbf{x}', \mathcal{X}_2) \end{bmatrix}, \qquad \text{(C.5)}$$

$$\text{cov}\left(\begin{bmatrix} f_1(\mathbf{x}) \\ f_2(\mathbf{x}') \end{bmatrix}\right) = \begin{bmatrix} k_1(\mathbf{x}, \mathbf{x}) & \rho_1 k_1(\mathbf{x}, \mathbf{x}') \\ \rho_1 k_1(\mathbf{x}', \mathbf{x}) & \rho_1^2 k_1(\mathbf{x}', \mathbf{x}') + k_2(\mathbf{x}', \mathbf{x}') \end{bmatrix}. \qquad \text{(C.6)}$$

C.2 Derivation of (5.12) using recursive update

For the derivation of (5.12), we consider the following Bayes' theorem:

$$p(f(\mathbf{x})|\mathcal{D}, \overline{f}_i(\tilde{\mathbf{x}})) = \frac{p(f(\mathbf{x}), \overline{f}_i(\tilde{\mathbf{x}})|\mathcal{D})}{p(\overline{f}_i(\tilde{\mathbf{x}})|\mathcal{D})}, \qquad \text{(C.7)}$$

where $f(\mathbf{x})|\mathcal{D}, \overline{f}_i(\tilde{\mathbf{x}})$ can be seen as the posterior of $f(\mathbf{x})|\mathcal{D}$ (as a prior) after adding one sample $\overline{f}_i(\tilde{\mathbf{x}})$. One can then get the mean and variance of $f(\mathbf{x})|\mathcal{D}, \overline{f}_i(\tilde{\mathbf{x}})$ using the standard conditional Gaussian formula:

$$\mathbb{E}(f(\mathbf{x})|\mathcal{D}, \overline{f}_i(\tilde{\mathbf{x}})) = \mathbb{E}(f(\mathbf{x})|\mathcal{D}) + \frac{\text{cov}(f(\mathbf{x}), f_i(\tilde{\mathbf{x}})|\mathcal{D})(\overline{f}_i(\tilde{\mathbf{x}}) - \mathbb{E}(f_i(\tilde{\mathbf{x}})|\mathcal{D}))}{\text{var}(f_i(\tilde{\mathbf{x}})|\mathcal{D})}, \qquad \text{(C.8)}$$

$$\text{var}(f(\mathbf{x})|\mathcal{D}, \overline{f}_i(\tilde{\mathbf{x}})) = \text{var}(f(\mathbf{x})|\mathcal{D}) - \frac{\text{cov}(f(\mathbf{x}), f_i(\tilde{\mathbf{x}})|\mathcal{D})^2}{\text{var}(f_i(\tilde{\mathbf{x}})|\mathcal{D})}. \qquad \text{(C.9)}$$

The formula in (5.12) is a direct result of (C.9).

C.3 Example of a more informative low-fidelity sample

We show a special case in Theorem 1 where a high-fidelity sample is less informative than a low-fidelity sample in a bi-fidelity model ($s = 2$).

Theorem 1: Assume $\rho_1 = 1$ (see (5.2)) and noiseless observations with $i = 1, 2$. The benefits of adding a high-fidelity sample $B(2, \tilde{\mathbf{x}})$ is smaller than $B(1, \tilde{\mathbf{x}})$ for $\|\Lambda_2\| \to 0$.

Proof: Based on (5.12), (5.6), (5.7), we can compute the benefits as

$$
\begin{aligned}
B(1, \tilde{\mathbf{x}}) &= Q(\mathcal{D}) - Q(\mathcal{D}, 1, \tilde{\mathbf{x}}) \\
&= \frac{1}{\text{var}(f_1(\tilde{\mathbf{x}})|\mathcal{D})} \int \text{cov}^2(f(\mathbf{x}), f_1(\tilde{\mathbf{x}})|\mathcal{D}) w(\mathbf{x}) d\mathbf{x} \\
&= \int \frac{(\text{cov}(f(\mathbf{x}), f_1(\tilde{\mathbf{x}})) - \text{cov}(f(\mathbf{x}), \mathbf{Y})\text{cov}(\mathbf{Y})^{-1}\text{cov}(\mathbf{Y}, f_1(\tilde{\mathbf{x}})))^2}{\text{cov}(f_1(\tilde{\mathbf{x}})) - \text{cov}(f_1(\tilde{\mathbf{x}}), \mathbf{Y})\text{cov}(\mathbf{Y})^{-1}\text{cov}(\mathbf{Y}, f_1(\tilde{\mathbf{x}}))} w(\mathbf{x}) d\mathbf{x} \\
&= \int \frac{\left(\begin{array}{c} k_1(\mathbf{x}, \tilde{\mathbf{x}}) - [k_1(\mathbf{x}, \mathcal{X}_1), k_1(\mathbf{x}, \mathcal{X}_2) + k_2(\mathbf{x}, \mathcal{X}_2)] \\ \text{cov}(\mathbf{Y})^{-1}[k_1(\tilde{\mathbf{x}}, \mathcal{X}_1), k_1(\tilde{\mathbf{x}}, \mathcal{X}_2)]^T \end{array} \right)^2}{\begin{array}{c} k_1(\tilde{\mathbf{x}}, \tilde{\mathbf{x}}) - [k_1(\tilde{\mathbf{x}}, \mathcal{X}_1), k_1(\tilde{\mathbf{x}}, \mathcal{X}_2)] \\ \text{cov}(\mathbf{Y})^{-1}[k_1(\tilde{\mathbf{x}}, \mathcal{X}_1), k_1(\tilde{\mathbf{x}}, \mathcal{X}_2)]^T \end{array}} w(\mathbf{x}) d\mathbf{x}, \quad \text{(C.10)}
\end{aligned}
$$

$$
\begin{aligned}
B(2, \tilde{\mathbf{x}}) &= Q(\mathcal{D}) - Q(\mathcal{D}, 2, \tilde{\mathbf{x}}) \\
&= \frac{1}{\text{var}(f_2(\tilde{\mathbf{x}})|\mathcal{D})} \int \text{cov}^2(f(\mathbf{x}), f_2(\tilde{\mathbf{x}})|\mathcal{D}) w(\mathbf{x}) d\mathbf{x} \\
&= \int \frac{(\text{cov}(f(\mathbf{x}), f_2(\tilde{\mathbf{x}})) - \text{cov}(f(\mathbf{x}), \mathbf{Y})\text{cov}(\mathbf{Y})^{-1}\text{cov}(\mathbf{Y}, f_2(\tilde{\mathbf{x}})))^2}{\text{cov}(f_2(\tilde{\mathbf{x}})) - \text{cov}(f_2(\tilde{\mathbf{x}}), \mathbf{Y})\text{cov}(\mathbf{Y})^{-1}\text{cov}(\mathbf{Y}, f_2(\tilde{\mathbf{x}}))} w(\mathbf{x}) d\mathbf{x} \\
&= \int \frac{\left(\begin{array}{c} k_1(\mathbf{x}, \tilde{\mathbf{x}}) + k_2(\mathbf{x}, \tilde{\mathbf{x}}) - [k_1(\mathbf{x}, \mathcal{X}_1), k_1(\mathbf{x}, \mathcal{X}_2) + k_2(\mathbf{x}, \mathcal{X}_2)] \\ \text{cov}(\mathbf{Y})^{-1}[k_1(\tilde{\mathbf{x}}, \mathcal{X}_1), k_1(\tilde{\mathbf{x}}, \mathcal{X}_2) + k_2(\tilde{\mathbf{x}}, \mathcal{X}_2)]^T \end{array} \right)^2}{\begin{array}{c} k_1(\tilde{\mathbf{x}}, \tilde{\mathbf{x}}) + k_2(\tilde{\mathbf{x}}, \tilde{\mathbf{x}}) - [k_1(\tilde{\mathbf{x}}, \mathcal{X}_1), k_1(\tilde{\mathbf{x}}, \mathcal{X}_2) + k_2(\tilde{\mathbf{x}}, \mathcal{X}_2)] \\ \text{cov}(\mathbf{Y})^{-1}[k_1(\tilde{\mathbf{x}}, \mathcal{X}_1), k_1(\tilde{\mathbf{x}}, \mathcal{X}_2) + k_2(\tilde{\mathbf{x}}, \mathcal{X}_2)]^T \end{array}} w(\mathbf{x}) d\mathbf{x}.
\end{aligned}
$$

$$\text{(C.11)}$$

If $\|\Lambda_2\| \to 0$, we have $k_2(\mathbf{x}, \mathbf{x}'; \Lambda_2) \to 0$ for $\mathbf{x} \neq \mathbf{x}'$ (see (5.3)). Thus

$$\lim_{\|\Lambda_2\| \to 0} B(1, \tilde{\mathbf{x}}) = \int \frac{(k_1(\mathbf{x}, \tilde{\mathbf{x}}) - k_1(\mathbf{x}, \mathcal{X})\mathrm{cov}(\mathbf{Y})^{-1}k_1(\tilde{\mathbf{x}}, \mathcal{X})^T)^2}{k_1(\tilde{\mathbf{x}}, \tilde{\mathbf{x}}) - k_1(\tilde{\mathbf{x}}, \mathcal{X})\mathrm{cov}(\mathbf{Y})^{-1}k_1(\tilde{\mathbf{x}}, \mathcal{X})^T} w(\mathbf{x})\mathrm{d}\mathbf{x}, \qquad \text{(C.12)}$$

$$\lim_{\|\Lambda_2\| \to 0} B(2, \tilde{\mathbf{x}}) = \int \frac{(k_1(\mathbf{x}, \tilde{\mathbf{x}}) - k_1(\mathbf{x}, \mathcal{X})\mathrm{cov}(\mathbf{Y})^{-1}k_1(\tilde{\mathbf{x}}, \mathcal{X})^T)^2}{k_1(\tilde{\mathbf{x}}, \tilde{\mathbf{x}}) + k_2(\tilde{\mathbf{x}}, \tilde{\mathbf{x}}) - k_1(\tilde{\mathbf{x}}, \mathcal{X})\mathrm{cov}(\mathbf{Y})^{-1}k_1(\tilde{\mathbf{x}}, \mathcal{X})^T} w(\mathbf{x})\mathrm{d}\mathbf{x},$$

$$\text{(C.13)}$$

and

$$\lim_{\|\Lambda_2\| \to 0} B(1, \tilde{\mathbf{x}}) - B(2, \tilde{\mathbf{x}}) > 0, \qquad \text{(C.14)}$$

which complete the proof.

C.4   Analytic computation of the acquisition

We first present the formula of $G_t(f_i(\mathbf{x}_1), f_j(\mathbf{x}_2))$ in (5.18), which is the key part towards an analytical acquisition and its derivative. Substituting the covariance function (5.7) to (5.18), one obtains

$$\begin{aligned}
&G_t(f_i(\mathbf{x}_1), f_j(\mathbf{x}_2)) \\
&= \int \mathrm{cov}(f_i(\mathbf{x}_1), f(\mathbf{x}))\mathrm{cov}(f(\mathbf{x}), f_j(\mathbf{x}_2)))\mathcal{N}(\mathbf{x}; \mu_t, \Sigma_t)\mathrm{d}\mathbf{x} \\
&= \int \Big(\sum_{l=1}^{i} \pi_{isl}k_l(\mathbf{x}_1, \mathbf{x})\Big)\Big(\sum_{r=1}^{j} \pi_{jsr}k_r(\mathbf{x}, \mathbf{x}_2)\Big)\mathcal{N}(\mathbf{x}; \mu_t, \Sigma_t)\mathrm{d}\mathbf{x} \\
&= \sum_{l=1}^{i}\sum_{r=1}^{j} \pi_{isl}\pi_{jsr}\mathcal{I}_{l,r,t}(\mathbf{x}_1, \mathbf{x}_2), \qquad\qquad \text{(C.15)}
\end{aligned}$$

where

$$\mathcal{I}_{l,r,t}(\mathbf{x}_1, \mathbf{x}_2) = \int k_l(\mathbf{x}_1, \mathbf{x}; \Lambda_l)k_r(\mathbf{x}_2, \mathbf{x}; \Lambda_r)\mathcal{N}(\mathbf{x}; \mu_t, \Sigma_t)\mathrm{d}\mathbf{x} \qquad \text{(C.16a)}$$

$$= |\Sigma_t\mathrm{M}^{-1} + \mathrm{I}|^{-\frac{1}{2}}k_l(\mathbf{x}_1, \mathbf{x}_2; \Lambda_l + \Lambda_r)k_r(\mathbf{m}, \mu_t; \Sigma_t + \mathrm{M}), \qquad \text{(C.16b)}$$

185

with

$$\mathbf{m} = \Lambda_r(\Lambda_l + \Lambda_r)^{-1}\mathbf{x}_1 + \Lambda_l(\Lambda_l + \Lambda_r)^{-1}\mathbf{x}_2, \tag{C.17}$$

$$\mathrm{M} = \Lambda_l\Lambda_r(\Lambda_l + \Lambda_r)^{-1}. \tag{C.18}$$

We note that from (C.16a) to (C.16b) one needs to use the formulae for transferring kernels to Gaussian functions as well as the multiplication of Gaussian functions [113]:

$$k(\mathbf{x}, \mathbf{x}_1; \Lambda, \tau) = \tau^2(2\pi)^{d/2}|\Lambda|^{1/2}\mathcal{N}(\mathbf{x}; \mathbf{x}_1, \Lambda), \tag{C.19}$$

$$\mathcal{N}(\mathbf{x}; \mathbf{x}_1, \Sigma_1)\mathcal{N}(\mathbf{x}; \mathbf{x}_2, \Sigma_2) = \mathcal{N}(\mathbf{x}_1; \mathbf{x}_2, \Sigma_1 + \Sigma_2)\mathcal{N}(\mathbf{x}; \mathrm{C}(\Sigma_1^{-1}\mathbf{x_1} + \Sigma_2^{-1}\mathbf{x_2}), \mathrm{C}), \tag{C.20}$$

with $\mathrm{C} = (\Sigma_1^{-1} + \Sigma_2^{-1})^{-1}$.

Specifically, the detailed derivation of (C.16b) using (C.19) and (C.20) is shown below:

$$\mathcal{I}_{l,r,t}(\mathbf{x}_1, \mathbf{x}_2) = \int k_l(\mathbf{x}_1, \mathbf{x}; \Lambda_l, \tau_l)k_r(\mathbf{x}_2, \mathbf{x}; \Lambda_r, \tau_r)\mathcal{N}(\mathbf{x}; \mu_t, \Sigma_t)\mathrm{d}\mathbf{x} \tag{C.21a}$$

$$= \tau_l^2\tau_r^2(2\pi)^d|\Lambda_l|^{1/2}|\Lambda_r|^{1/2}\int \mathcal{N}(\mathbf{x}; \mathbf{x}_1, \Lambda_l)\mathcal{N}(\mathbf{x}; \mathbf{x}_2, \Lambda_r)\mathcal{N}(\mathbf{x}; \mu_t, \Sigma_t)\mathrm{d}\mathbf{x} \tag{C.21b}$$

$$= \tau_l^2\tau_r^2(2\pi)^d|\Lambda_l|^{1/2}|\Lambda_r|^{1/2}\mathcal{N}(\mathbf{x}_1; \mathbf{x}_2, \Lambda_l + \Lambda_r)\int \mathcal{N}(\mathbf{x}; \mathbf{m}, \mathrm{M})\mathcal{N}(\mathbf{x}; \mu_t, \Sigma_t)\mathrm{d}\mathbf{x} \tag{C.21c}$$

$$= \tau_l^2\tau_r^2(2\pi)^d|\Lambda_l|^{1/2}|\Lambda_r|^{1/2}\mathcal{N}(\mathbf{x}_1; \mathbf{x}_2, \Lambda_l + \Lambda_r)\mathcal{N}(\mathbf{m}; \mu_t, \mathrm{M} + \Sigma_t) \tag{C.21d}$$

$$= |\Sigma_t\mathrm{M}^{-1} + \mathrm{I}|^{-\frac{1}{2}}k_l(\mathbf{x}_1, \mathbf{x}_2; \Lambda_l + \Lambda_r)k_r(\mathbf{m}, \mu_t; \Sigma_t + \mathrm{M}). \tag{C.21e}$$

With the analytical formula for $G_t(f_i(\mathbf{x}_1), f_j(\mathbf{x}_2))$ available in (C.15), the benefit $B(i, \tilde{\mathbf{x}})$ in (5.14) can be derived accordingly. To further obtain the analytical derivative of $B(i, \tilde{\mathbf{x}})$, we rewrite $B(i, \tilde{\mathbf{x}}) = T(i, \tilde{\mathbf{x}})/\mathrm{var}(f_i(\tilde{\mathbf{x}})|\mathcal{D})$ (comparing to (5.12)), and

derive the derivative of T as

$$\frac{\partial T(i, \tilde{\mathbf{x}})}{\partial \tilde{\mathbf{x}}} = \frac{\partial \mathcal{K}(f_i(\tilde{\mathbf{x}}), f_i(\tilde{\mathbf{x}}))}{\partial \tilde{\mathbf{x}}} + 2\frac{\partial \text{cov}(\mathbf{Y}, f_i(\tilde{\mathbf{x}}))}{\partial \tilde{\mathbf{x}}}\text{cov}(\mathbf{Y})^{-1}\mathcal{K}(\mathbf{Y}, \mathbf{Y})\text{cov}(\mathbf{Y})^{-1}\text{cov}(\mathbf{Y}, f_i(\tilde{\mathbf{x}}))$$

$$-2\frac{\partial \text{cov}(\mathbf{Y}, f_i(\tilde{\mathbf{x}}))}{\partial \tilde{\mathbf{x}}}\text{cov}(\mathbf{Y})^{-1}\mathcal{K}(\mathbf{Y}, f_i(\tilde{\mathbf{x}})) - 2\frac{\partial \mathcal{K}(\mathbf{Y}, f_i(\tilde{\mathbf{x}}))}{\partial \tilde{\mathbf{x}}}\text{cov}(\mathbf{Y})^{-1}\text{cov}(\mathbf{Y}, f_i(\tilde{\mathbf{x}})).$$

$$(\text{C.22})$$

The analytical computation of (C.22) requires the formula for the derivative of the covariance function and $\mathcal{K}$. For the former, we have:

$$\frac{\partial \text{cov}(f_i(\mathbf{x}_1), f_j(\mathbf{x}_2))}{\partial \mathbf{x}_1} = \sum_{l=1}^{\min(i,j)} \pi_{ijl}\frac{\partial k_l(\mathbf{x}_1, \mathbf{x}_2)}{\partial \mathbf{x}_1}, \qquad (\text{C.23})$$

with

$$\frac{\partial k(\mathbf{x}_1, \mathbf{x}_2; \Lambda)}{\partial \mathbf{x}_1} = k(\mathbf{x}_1, \mathbf{x}_2; \Lambda)\Lambda^{-1}(\mathbf{x}_2 - \mathbf{x}_1). \qquad (\text{C.24})$$

For the latter, we have:

$$\frac{\partial \mathcal{K}(f_i(\mathbf{x}_1), f_i(\mathbf{x}_1))}{\partial \mathbf{x}_1} = \sum_{t=1}^{n_{GMM}} \alpha_t \frac{\partial G_t(f_i(\mathbf{x}_1), f_i(\mathbf{x}_1))}{\partial \mathbf{x}_1}$$

$$= \sum_{t=1}^{n_{GMM}} \alpha_t \sum_{l=1}^{i}\sum_{r=1}^{i} \pi_{isl}\pi_{isr}\frac{\partial \mathcal{I}_{l,r,t}(\mathbf{x}_1, \mathbf{x}_1)}{\partial \mathbf{x}_1}, \qquad (\text{C.25})$$

$$\frac{\partial \mathcal{K}(f_i(\mathbf{x}_1), f_j(\mathbf{x}_2))}{\partial \mathbf{x}_1} = \sum_{t=1}^{n_{GMM}} \alpha_t \frac{\partial G_t(f_i(\mathbf{x}_1), f_j(\mathbf{x}_2))}{\partial \mathbf{x}_1}$$

$$= \sum_{t=1}^{n_{GMM}} \alpha_t \sum_{l=1}^{i}\sum_{r=1}^{j} \pi_{isl}\pi_{jsr}\frac{\partial \mathcal{I}_{l,r,t}(\mathbf{x}_1, \mathbf{x}_2)}{\partial \mathbf{x}_1}. \qquad (\text{C.26})$$

Finally, (C.25) and (C.26) require the derivatives of $\mathcal{I}_{i,j,t}$, which can be obtained by combing (C.16b) and (C.24):

$$\frac{\partial \mathcal{I}_{l,r,t}(\mathbf{x}_1, \mathbf{x}_1)}{\partial \mathbf{x}_1} = \mathcal{I}_{l,r,t}(\mathbf{x}_1, \mathbf{x}_1)(\Sigma_t + M)^{-1}(\mu_t - \mathbf{x}_1), \qquad (\text{C.27})$$

$$\frac{\partial \mathcal{I}_{l,r,t}(\mathbf{x}_1, \mathbf{x}_2)}{\partial \mathbf{x}_1} = \mathcal{I}_{l,r,t}(\mathbf{x}_1, \mathbf{x}_2)\Big((\Lambda_l + \Lambda_r)^{-1}(\mathbf{x}_2 - \mathbf{x}_1) + \Lambda_r(\Lambda_l + \Lambda_r)^{-1}(\Sigma_t + M)^{-1}(\mu_t - \mathbf{m})\Big).$$

$$(\text{C.28})$$

187

APPENDIX D

Appendix of Multi-Fidelity Bayesian Experimental Design for Safety Analysis of Connected and Automated Vehicles

D.1 Information gain and its approximation

Let us start from the situation that we have a surrogate model $f_h(\mathbf{x})|\mathcal{D}$ and the associated distribution of P_a. From an information-theoretical perspective, the next-best sample $\tilde{\mathbf{x}}$ should be chosen to maximize the information gain for P_a, i.e., to maximize the expected KL divergence defined as

$$
\begin{aligned}
G(\tilde{\mathbf{x}}) &= \mathbb{E}_{f_h}\Big(\mathrm{KL}\big(p(P_a|\mathcal{D}, f_h(\tilde{\mathbf{x}})) \,\|\, p(P_a|\mathcal{D})\big)\Big) \\
&= \iint p(P_a|\mathcal{D}, f_h(\tilde{\mathbf{x}})) \log \frac{p(P_a|\mathcal{D}, f_h(\tilde{\mathbf{x}}))}{p(P_a|\mathcal{D})} \, \mathrm{d}P_a \, p(f_h(\tilde{\mathbf{x}})|\mathcal{D}) \, \mathrm{d}f_h(\tilde{\mathbf{x}}). \quad \text{(D.1)}
\end{aligned}
$$

While a direct computation of (D.1) is extremely expensive, three assumptions can be made to simplify its expression, which are summarized below.

1. Instead of a random $f_h(\tilde{\mathbf{x}})$, we assume that it can be approximated by the mean prediction from the current GPR: $\overline{f}_h(\tilde{\mathbf{x}}) = \mathbb{E}(f_h(\tilde{\mathbf{x}})|\mathcal{D})$. Thus (D.1) becomes:

$$
G(\tilde{\mathbf{x}}) \approx \int p(P_a|\mathcal{D}, \overline{f}_h(\tilde{\mathbf{x}})) \log \frac{p(P_a|\mathcal{D}, \overline{f}_h(\tilde{\mathbf{x}}))}{p(P_a|\mathcal{D})} \mathrm{d}P_a. \quad \text{(D.2)}
$$

2. We assume P_a follows Gaussian distributions with $P_a|\mathcal{D} \sim \mathcal{N}(\mu_1, \sigma_1^2)$ and $P_a|\mathcal{D}, \overline{f}_h(\tilde{\mathbf{x}}) \sim \mathcal{N}(\mu_2(\tilde{\mathbf{x}}), \sigma_2^2(\tilde{\mathbf{x}}))$. Substitution of these distributions into (D.2) gives

$$G(\tilde{\mathbf{x}}) \approx \log(\frac{\sigma_1}{\sigma_2(\tilde{\mathbf{x}})}) + \frac{\sigma_2^2(\tilde{\mathbf{x}})}{2\sigma_1^2} + \frac{(\mu_2(\tilde{\mathbf{x}}) - \mu_1)^2}{2\sigma_1^2} - \frac{1}{2}. \qquad (D.3)$$

3. The difference of μ_1 and μ_2 is much smaller than the standard deviation of P_a i.e. $|\mu_2(\tilde{\mathbf{x}}) - \mu_1| \ll \sigma_2(\tilde{\mathbf{x}})$ (which is generally true unless P_a has been estimated very well). This leads to

$$G(\tilde{\mathbf{x}}) \approx \log(\frac{\sigma_1}{\sigma_2(\tilde{\mathbf{x}})}) + \frac{\sigma_2^2(\tilde{\mathbf{x}})}{2\sigma_1^2} - \frac{1}{2}. \qquad (D.4)$$

It can be shown that (D.4) monotonously increases with the decrease of σ_2/σ_1 in the range of $\sigma_2/\sigma_1 < 1$. Since a sample always provides information to P_a, the condition $\sigma_2/\sigma_1 < 1$ is always satisfied. Therefore, the maximization of (D.4) is equivalent to the minimization of σ_2, that is consistent the minimization of (6.4).

D.2 The upper bound of (6.4)

Here we show the derivation of the upper bound (6.5) from (6.4):

$$\text{var}[P_a|\mathcal{D}, \overline{f}_h(\tilde{\mathbf{x}})]$$

$$= \text{var}\left[\int \mathbf{1}_\delta\big(f_h(\mathbf{x})|\mathcal{D}, \overline{f}_h(\tilde{\mathbf{x}})\big)p_\mathbf{x}(\mathbf{x})\mathrm{d}\mathbf{x}\right]$$

(hereafter we write $\mathbf{1}_\delta\big(f_h(\mathbf{x})|\mathcal{D}, \overline{f}_h(\tilde{\mathbf{x}})\big)$ as $\hat{\mathbf{1}}(\mathbf{x})$)

$$= \mathbb{E}\left[\left(\int \hat{\mathbf{1}}(\mathbf{x})p_\mathbf{x}(\mathbf{x})\mathrm{d}\mathbf{x}\right)^2\right] - \left(\mathbb{E}\left[\int \hat{\mathbf{1}}(\mathbf{x})p_\mathbf{x}(\mathbf{x})\mathrm{d}\mathbf{x}\right]\right)^2$$

$$= \mathbb{E}\left[\int \hat{\mathbf{1}}(\mathbf{x})p(\mathbf{x})\mathrm{d}\mathbf{x} \int \hat{\mathbf{1}}(\mathbf{x}')p(\mathbf{x}')\mathrm{d}\mathbf{x}'\right]$$

$$- \left(\mathbb{E}\left[\int \hat{\mathbf{1}}(\mathbf{x})p(\mathbf{x})\mathrm{d}\mathbf{x}\right]\right)\left(\mathbb{E}\left[\int \hat{\mathbf{1}}(\mathbf{x}')p(\mathbf{x}')\mathrm{d}\mathbf{x}'\right]\right)$$

$$= \iint \mathbb{E}\left[\hat{\mathbf{1}}(\mathbf{x})\hat{\mathbf{1}}(\mathbf{x}')\right]p(\mathbf{x})p(\mathbf{x}')\mathrm{d}\mathbf{x}\mathrm{d}\mathbf{x}'$$

$$- \iint \mathbb{E}\left[\hat{\mathbf{1}}(\mathbf{x})\right]\mathbb{E}\left[\hat{\mathbf{1}}(\mathbf{x}')\right]p(\mathbf{x})p(\mathbf{x}')\mathrm{d}\mathbf{x}\mathrm{d}\mathbf{x}'$$

$$= \iint \text{cov}\left[\hat{\mathbf{1}}(\mathbf{x}), \hat{\mathbf{1}}(\mathbf{x}')\right]p(\mathbf{x})p(\mathbf{x}')\mathrm{d}\mathbf{x}dx' \tag{D.5}$$

$$= \int \text{std}\left[\hat{\mathbf{1}}(\mathbf{x})\right]\left(\int \text{std}\left[\hat{\mathbf{1}}(\mathbf{x}')\right]\rho\left[\hat{\mathbf{1}}(\mathbf{x}), \hat{\mathbf{1}}(\mathbf{x}')\right]p(\mathbf{x}')dx'\right)p(\mathbf{x})\mathrm{d}\mathbf{x}$$

$$\leq 0.5 \int \text{std}\left[\hat{\mathbf{1}}(\mathbf{x})\right]p(\mathbf{x})\mathrm{d}\mathbf{x}, \tag{D.6}$$

where $\rho[\cdot, \cdot]$ denotes the correlation coefficient. The last inequality comes from $\text{std}\left[\hat{\mathbf{1}}(\mathbf{x})\right] \leq 0.5$ and $\rho\left[\hat{\mathbf{1}}(\mathbf{x}), \hat{\mathbf{1}}(\mathbf{x}')\right] \leq 1$. The former inequality is due to that $\hat{\mathbf{1}}(\mathbf{x})$ is a Bernoulli variable whose standard deviation is maximized to be 0.5 when the two branches $\hat{\mathbf{1}}(\mathbf{x}) = 0$ and $\hat{\mathbf{1}}(\mathbf{x}) = 1$ both have probability of 0.5.

We can also show that (6.4) can be directly approximated as $\sim \int \text{var}\left[\hat{\mathbf{1}}(\mathbf{x})\right]p(\mathbf{x})\mathrm{d}\mathbf{x}$ if only variance terms are kept in (D.5). This form is somewhat similar to the upper-bound result (D.6) when used as an acquisition function. We apply the upper bound (D.6) in the current paper.

D.3   Benchmark tests of the framework

In this section, we test our (single-fidelity) framework in two benchmark problems to demonstrate the efficiency of our new acquisition function. In each problem, we report the median value as well as 15% and 85% percentiles of the exceeding probability (defined below) obtained from 100 applications of our method with different initial samples.

The first example we consider is a multi-modal function (see figure D.1 for a contour plot) that is also studied in [59] and [158]:

$$f_h(x_1, x_2) = \frac{((1.5 + x_1)^2 + 4)(1.5 + x_2)}{20} - \sin(\frac{7.5 + 5x_1}{2}) - 2. \qquad (D.7)$$

The input $\mathbf{x}$ follows a Gaussian distribution $p(x_1, x_2) = \mathcal{N}(\mathbf{0}, I)$, with I being a 2×2 identity matrix. We are interested in the exceeding probability $P_e = \int \mathbf{1}_{f_h > 0}(f_h(\mathbf{x}))p_{\mathbf{x}}(\mathbf{x})d\mathbf{x}$. In the application of our method, we use 8 random initial samples followed by 22 adaptive samples, with the computed P_e plotted in figure D.2 as a function of the sample numbers. Also shown in figure D.2 is the 3% error bounds of ground truth. We see that percentiles of P_e converge to the error bounds in 18 samples. The sampling positions of our method are plotted in figure D.3 as well as the computed limiting state after 30 samples with comparison to the ground truth.

Other existing approaches were also tested for the same problem for the number of samples leading to convergence, but usually with less strict (or different) criterion defined for convergence. The number of samples are 19 [59], 31−44 [31], 36−69 [15].

191
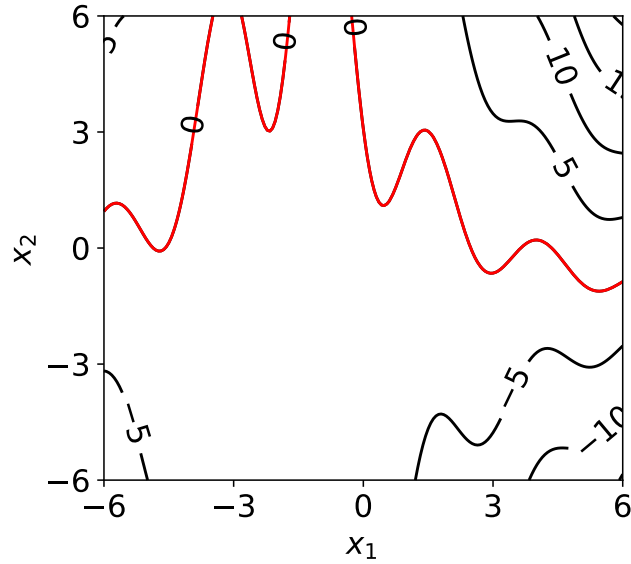
Figure D.1: $f_h(x_1, x_2)$ of the multi-modal function with the limiting state $\{\mathbf{x} : f_h(x_1, x_2) = 0\}$ (——).
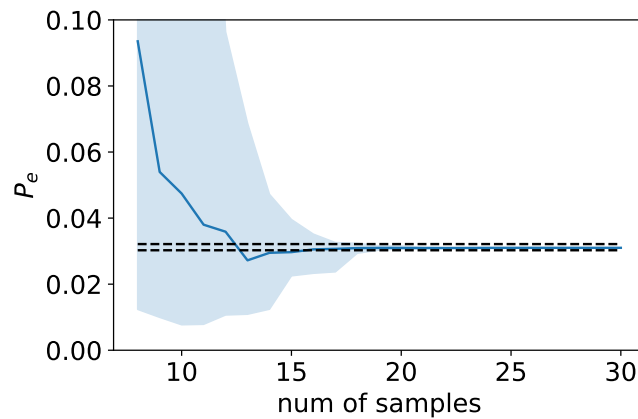
Figure D.2: Results of single-fidelity method for the problem of multi-modal function, presented by the median value (——) as well as the 15% and 85% percentiles (shaded region) from 100 experiments. The ground-truth of P_e is shown (- - -) in terms of the 3% error bounds.
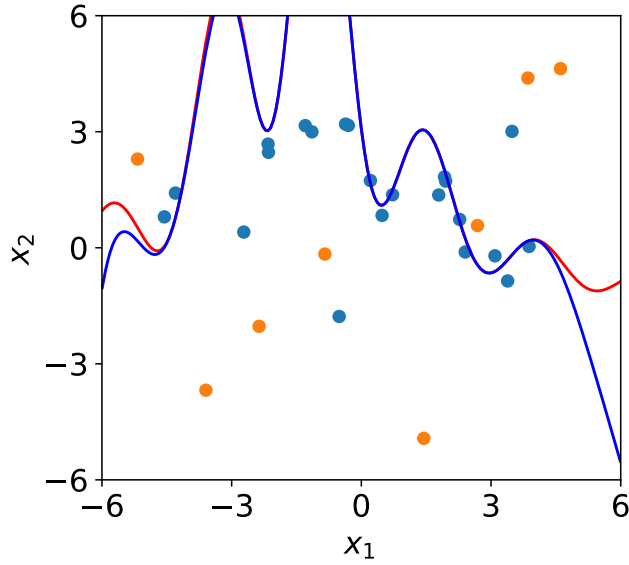
Figure D.3: Typical positions of 8 initial samples (•) and 22 adaptive samples (•) for the problem of multi-modal function, as well as the learned limiting state (—) compared to the exact one (—).

The second example is the four-branch function (figure D.4):

$$
f(x_1, x_2) = -\min \begin{cases} 3 + 0.1(x_1 - x_2)^2 + \dfrac{(x_1 + x_2)}{\sqrt{2}} \\ 3 + 0.1(x_1 - x_2)^2 - \dfrac{(x_1 + x_2)}{\sqrt{2}} \\ (x_1 - x_2) + \dfrac{6}{\sqrt{2}} \\ (x_2 - x_1) + \dfrac{6}{\sqrt{2}} \end{cases}.
$$

The input $\mathbf{x}$ follows a Gaussian distribution $p(x_1, x_2) = \mathcal{N}(\mathbf{0}, \mathrm{I})$. Our method is applied with 12 random initial samples followed by 68 adaptive samples, with P_e plotted in figure D.5 as a function of sample numbers. Also shown in figure D.5 is the 3% error bounds of ground truth. We see that percentiles of P_e converge to the error bounds in 42 samples. The sampling positions of our method are plotted in figure D.6 as well as the computed limiting state after 80 samples with comparison to the ground truth.
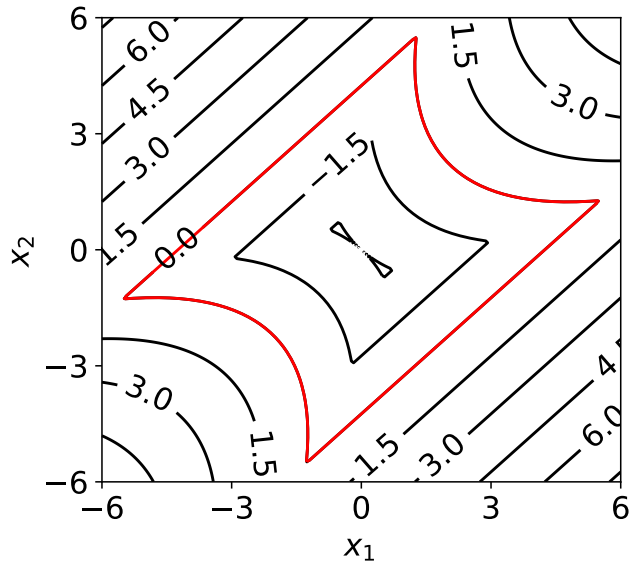
Figure D.4: $f_h(x_1, x_2)$ of the four-branch function with the limiting state $\{\mathbf{x} : f_h(x_1, x_2) = 0\}$ (——).

With less strict (or different) convergence criterion, the number of samples leading to the convergence in existing works are 36 [59], 38 [135], 78 [121], 65−126 [31], 68−124 [15], and 167 [83].
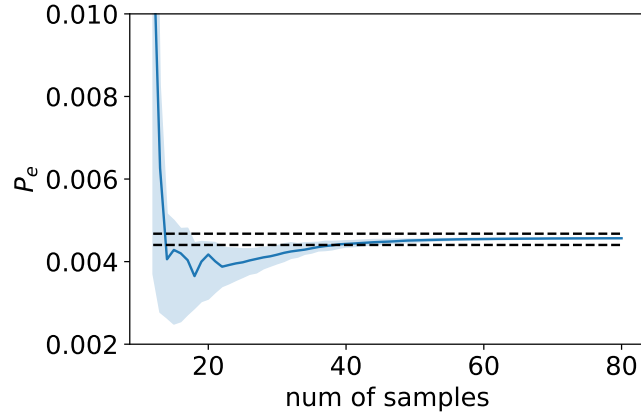
Figure D.5: Results of single-fidelity method for the problem of four-branch function, presented by the median value (—) as well as the 15% and 85% percentiles (shaded region) from 100 experiments. The ground-truth of P_e is shown (---) in terms of the 3% error bounds.
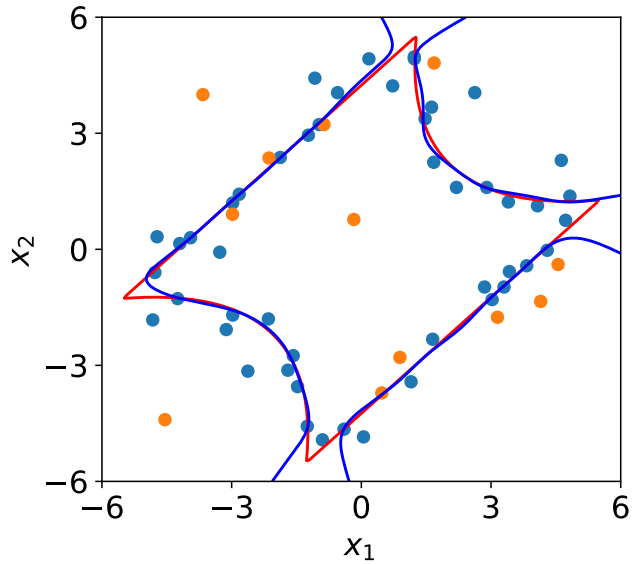
Figure D.6: Typical positions of 12 initial samples (•) and 68 adaptive samples (•) for the problem of four-branch function, as well as the learned limiting state (—) compared to the exact one (—).

APPENDIX E

Appendix of a Generalized Likelihood-Weighted Acquisition for Rare-Event Statistics

E.1 Derivation of (7.6)

The derivation of (7.6) is built on the Theorem 2 in [95], restated here with slight change of notations in the context of the current paper.

Theorem: *Let $p_{f^\pm|\mathcal{D},\hat{f}(\tilde{\mathbf{x}})}(f)$ be PDF bounds generated by upper and lower bounds of GPR $f|\mathcal{D}, \hat{f}(\tilde{\mathbf{x}})$. Assume $\mathrm{std}(f(\mathbf{x})|\mathcal{D}, \hat{f}(\tilde{\mathbf{x}}))$ is sufficiently small (thus $p_{f^\pm}(f)$ are close enough). The integration of log difference between $p_{f^\pm}(f)$ in (7.4) is then given by*

$$
\begin{aligned}
\epsilon_L(\tilde{\mathbf{x}}) &= \int |\log p_{f^+|\mathcal{D},\hat{f}(\tilde{\mathbf{x}})}(s) - \log p_{f^-|\mathcal{D},\hat{f}(\tilde{\mathbf{x}})}(s)| \mathrm{d}s \\
&\approx \int \left| \frac{\frac{\mathrm{d}}{\mathrm{d}s} \int \mathrm{std}(f(\mathbf{x})|\mathcal{D}, \hat{f}(\tilde{\mathbf{x}})) p_{\mathbf{x}}(\mathbf{x})\delta(s - \hat{f}(\mathbf{x}))\mathrm{d}\mathbf{x}}{p_{\hat{f}}(s)} \right| \mathrm{d}s.
\end{aligned} \tag{E.1}
$$

Let $g(s,\tilde{\mathbf{x}}) \equiv \int \mathrm{std}(f(\mathbf{x})|\mathcal{D}, \hat{f}(\tilde{\mathbf{x}})) p_{\mathbf{x}}(\mathbf{x})\delta(s - \hat{f}(\mathbf{x}))\mathrm{d}\mathbf{x}$ and denote $\partial g(s,\tilde{\mathbf{x}})/\partial s$ as

$g'(s, \tilde{\mathbf{x}})$, ϵ_L in (E.1) can be further computed as

$$\epsilon_L(\tilde{\mathbf{x}}) \approx \int_{g'(s,\tilde{\mathbf{x}})>0} \frac{g'(s,\tilde{\mathbf{x}})}{p_{\hat{f}}(s)} \mathrm{d}s - \int_{g'(s,\tilde{\mathbf{x}})<0} \frac{g'(s,\tilde{\mathbf{x}})}{p_{\hat{f}}(s)} \mathrm{d}s$$
$$\approx B + \int_{g'(s,\tilde{\mathbf{x}})>0} \frac{p'_{\hat{f}}(s)}{p_{\hat{f}}^2(s)} g(s,\tilde{\mathbf{x}}) \mathrm{d}s - \int_{g'(s,\tilde{\mathbf{x}})<0} \frac{p'_{\hat{f}}(s)}{p_{\hat{f}}^2(s)} g(s,\tilde{\mathbf{x}}) \mathrm{d}s, \qquad \text{(E.2)}$$

where we have used integration by parts, with all boundary terms collected in B. We note that since there are only finite number of boundary terms, B is guaranteed to be bounded.

Noticing that $g(s, \tilde{\mathbf{x}}) > 0$ always, we further have from (E.2)

$$e_L(\tilde{\mathbf{x}}) \leq B + \int \frac{|p'_{\hat{f}}(s)|}{p_{\hat{f}}^2(s)} g(s,\tilde{\mathbf{x}}) \mathrm{d}s,$$
$$\leq C \int \frac{|p'_{\hat{f}}(s)|}{p_{\hat{f}}^2(s)} \int \mathrm{std}(f(\mathbf{x})|\mathcal{D}, \hat{f}(\tilde{\mathbf{x}})) p_{\mathbf{x}}(\mathbf{x}) \delta(s - \hat{f}(\mathbf{x})) \mathrm{d}\mathbf{x}\, \mathrm{d}s$$
$$= C \iint \frac{|p'_{\hat{f}}(s)|}{p_{\hat{f}}^2(s)} \mathrm{std}(f(\mathbf{x})|\mathcal{D}, \hat{f}(\tilde{\mathbf{x}})) p_{\mathbf{x}}(\mathbf{x}) \delta(s - \hat{f}(\mathbf{x})) \mathrm{d}s\mathrm{d}\mathbf{x}$$
$$= C \int \mathrm{std}(f(\mathbf{x})|\mathcal{D}, \hat{f}(\tilde{\mathbf{x}})) \frac{p_{\mathbf{x}}(\mathbf{x})|p'_{\hat{f}}(\hat{f}(\mathbf{x}))|}{p_{\hat{f}}^2(\hat{f}(\mathbf{x}))} \mathrm{d}\mathbf{x}, \qquad \text{(E.3)}$$

where in the 2nd line we absorb B into another constant C since the two terms in the 1st line are bounded (from above and below). In the third line we have applied the Fubini's theorem and in the fourth line we have integrated out the delta function. (E.3) is exactly (7.6) up to a constant.

We note that our derivation outlined above is different from that in [118] (in particular their proof of theorem 3.2) which is at least not well understood by the authors.

E.2 ϵ_{LW} as a continuous counterpart of precision in classification

To demonstrate ϵ_{LW} as a continuous counterpart of precision, we first introduce the precision metric in (imbalanced) classification. After that, we analyze the properties of ϵ_{LW} from an output viewpoint followed by an explanation of their relation.

E.2.1   Properties of precision for classification

In classification problems, we have ItR $g(\mathbf{x}) : \mathbb{R} \to \{1, 2, \cdots, k\}$ with k the number of classes. Our objective is to train a classifier $\hat{g}$ that achieves the best performance for certain metrics (errors) in an evaluation dataset with probability $p_{\mathbf{x}}(\mathbf{x})$. The most widely used metric, accuracy of the trained classifier $\hat{g}$ is defined as:

$$\mathrm{acc}(\hat{g}(\mathbf{x})) = \frac{\text{number of right prediction}}{\text{number of total evaluations}} = \int \mathbf{1}_{\hat{g}(\mathbf{x})=g(\mathbf{x})}(\mathbf{x})\, p_{\mathbf{x}}(\mathbf{x})\mathrm{d}\mathbf{x}, \qquad \text{(E.4)}$$

where $\mathbf{1}_{\hat{g}(\mathbf{x})=g(\mathbf{x})}(\mathbf{x}) = 1$ if the prediction is correct otherwise 0. However, when one of the classes is the minority, the classifier can have a very high accuracy even if the learning of the minority is totally wrong. One solution for such imbalanced classification is to use precision and recall, respectively defined as:

$$\mathrm{precision}(\hat{g}(\mathbf{x})) = \sum_{i=1}^{k} \frac{\text{number of correct predictions with predicted label } i}{\text{number of evaluations with predicted label } i\ (\hat{g}(\mathbf{x}) = i)}$$

$$= \sum_{i=1}^{k} \int \mathbf{1}_{\hat{g}(\mathbf{x})=g(\mathbf{x})}(\mathbf{x})\, p_{\mathbf{x}|\hat{g}(\mathbf{x})=i}(\mathbf{x})\mathrm{d}\mathbf{x}, \qquad \text{(E.5)}$$

$$\mathrm{recall}(\hat{g}(\mathbf{x})) = \sum_{i=1}^{k} \frac{\text{number of correct predictions with true label } i}{\text{number of evaluations with true label } i\ (g(\mathbf{x}) = i)}$$

$$= \sum_{i=1}^{k} \int \mathbf{1}_{\hat{g}(\mathbf{x})=g(\mathbf{x})}(\mathbf{x})\, p_{\mathbf{x}|g(\mathbf{x})=i}(\mathbf{x})\mathrm{d}\mathbf{x}, \qquad \text{(E.6)}$$

which measures the correct proportion of each class with the class divided by prediction or ground truth. The complement of (E.4), (E.5), and (E.6), specifically replacing $\mathbf{1}_{\hat{g}(\mathbf{x})=g(\mathbf{x})}(\cdot)$ with $\mathbf{1}_{\hat{g}(\mathbf{x})\neq g(\mathbf{x})}(\cdot)$, are the corresponding error metrics where each class (discrete output) contribute equally to the total error.

E.2.2 Properties of ϵ_{LW} in (7.5)

In this section, we interpret ϵ_{LW} from an output viewpoint that each output contributes equally to the total error. For convenience, we will neglect the hypothetical sample $\tilde{\mathbf{x}}$.

We first note that ϵ_{LW} can be seen as a variation of the mean-variance metric defined as:

$$\epsilon_V = \int \mathrm{var}(f(\mathbf{x})|\mathcal{D})p_{\mathbf{x}}(\mathbf{x})\mathrm{d}\mathbf{x}, \tag{E.7}$$

which is actually the mean-squared error (MSE) in regression problem.[1] From an output viewpoint, (E.7) can be rewritten as:

$$\begin{aligned} \epsilon_V &= \int \mathrm{var}(f(\mathbf{x})|\mathcal{D})p_{\mathbf{x}}(\mathbf{x})\mathrm{d}\mathbf{x} \\ &= \int p_{\hat{f}}(f) \int \mathrm{var}(f(\mathbf{x})|\mathcal{D})p_{\mathbf{x}|\hat{f}(\mathbf{x})=f}(\mathbf{x})\mathrm{d}\mathbf{x} \ \mathrm{d}f, \end{aligned} \tag{E.8}$$

where derivation is based on the smoothing theorem of conditional expectation [52]. In (E.8), the inner integration $\int \mathrm{var}(f(\mathbf{x})|\mathcal{D})p_{\mathbf{x}|\hat{f}(\mathbf{x})=f}(\mathbf{x})\mathrm{d}\mathbf{x}$ represents the mean of variance associated with the predicted output f. Rare events, thus, are endowed with

---

[1]In supervised learning, the mean-squared error between prediction $\hat{f}$ and the true function f is defined $\int (\hat{f}(\mathbf{x})-f(\mathbf{x}))^2 p_{\mathbf{x}}(\mathbf{x})\mathrm{d}\mathbf{x}$. With the posterior mean as the prediction and posterior probability as our belief of true function, we have the expected mean-squared error:

$$\begin{aligned} \mathbb{E}_{f|\mathcal{D}}\Big(\int (\hat{f}(\mathbf{x})-f(\mathbf{x}))^2 p_{\mathbf{x}}(\mathbf{x})\mathrm{d}\mathbf{x}\Big) &= \int \int (\mathbb{E}(f(\mathbf{x})|\mathcal{D})-f(\mathbf{x}))^2 p(\mathbf{x})\mathrm{d}\mathbf{x} \ p_{f|\mathcal{D}}(f(\mathbf{x}))\mathrm{d}f(\mathbf{x}) \\ &= \int \mathrm{var}(f(\mathbf{x})|\mathcal{D})p_{\mathbf{x}}(\mathbf{x})\mathrm{d}\mathbf{x}, \end{aligned}$$

which is exactly (E.7).

low weights, i.e., their probability densities $p_{\hat{f}}(f)$ in outer integration in assembling the total error.

In comparison, following the same procedure, ϵ_{LW} becomes (see a similar derivation for neural network training in [116]):

$$
\begin{aligned}
\epsilon_{LW}(\tilde{\mathbf{x}}) &= \int \text{var}(f(\mathbf{x})|\mathcal{D}) \frac{p_{\mathbf{x}}(\mathbf{x})}{p_{\hat{f}}(\hat{f}(\mathbf{x}))} \mathrm{d}\mathbf{x} \\
&= \int p_{\hat{f}}(f) \int \frac{1}{p_{\hat{f}}(\hat{f}(\mathbf{x}))} \text{var}(f(\mathbf{x})|\mathcal{D}) p_{\mathbf{x}|\hat{f}(\mathbf{x})=f}(\mathbf{x}) \mathrm{d}\mathbf{x} \, \mathrm{d}f \\
&= \int \int \text{var}(f(\mathbf{x})|\mathcal{D}) p_{\mathbf{x}|\hat{f}(\mathbf{x})=f}(\mathbf{x}) \mathrm{d}\mathbf{x} \, \mathrm{d}f.
\end{aligned} \tag{E.9}
$$

Compared with (E.8), each continuous output f contributes equally in computing the total error regardless of $p_{\hat{f}}(f)$. To illustrate this (in a discretized form for convenience), we plot a 2D function partitioned into different levels by contour lines ($\Delta f = 1$) in figure E.1. Across each level, for example, region shaded by red points as a whole ($\mathbf{x} : f(\mathbf{x}) \in [5,6]$) and region shaded by black points ($\mathbf{x} : f(\mathbf{x}) \in [-3,-2]$), ϵ_{LW} puts equal emphasis on them, i.e., $\int_{\mathbf{x}:f(\mathbf{x}) \in [f,f+\Delta f]} w(\mathbf{x}) \mathrm{d}\mathbf{x}$ is constant. Within each level, the relative importance of each $\mathbf{x}$ is determined by their input probability distribution.

It is clear that (E.5) and (E.9) share the same form in terms of equalizing each output (class) in assembling total error (summation for discrete output and integration for continuous output). Thus ϵ_{LW} can be treated as a continuous analogy of the precision where the former applies squared error (variance), and the latter applies zero-one error.

With this relation, we can think one step further. In classification, precision and recall will usually be considered equally (e.g., by F1 score) if there is no preference for Type I and Type II errors. Similarly, the predicted PDF and true PDF are indistinguishable in computing (7.2). However, the current acquisition only considers
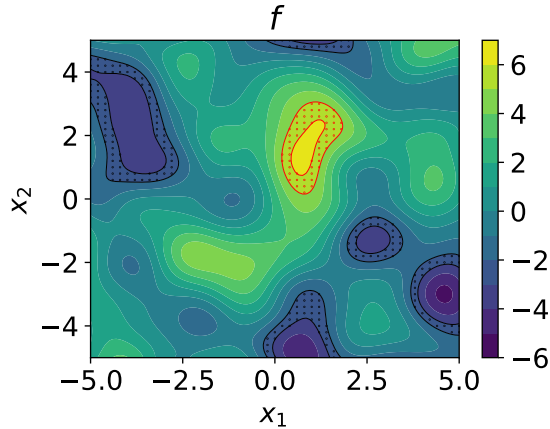
Figure E.1: Illustration of ϵ_{LW} from output viewpoint. Each level bounded by contour lines contributes equally to the total error regardless of their rarity.

precision. Intuitively, we also hope to include the recall, $\int \mathrm{var}(f(\mathbf{x})|\mathcal{D}) \frac{p_{\mathbf{x}}(\mathbf{x})}{p_f(f(\mathbf{x}))} \mathrm{d}\mathbf{x}$, in selecting the next sample. Given that we do not know f in the sampling process, one way is to use the mean of recall based on current GPR belief, which requires a cumbersome sampling process. In practice, f can be approximated by the upper and lower bounds of the GPR as shown in (7.9).

E.3 Acceleration in MCDO regarding the acquisitions

In MCDO method, we pre-select a large number of candidate samples located at $\mathbf{X}_{mc} \in \mathbb{R}^{n_{mc}*d}$ (usually from space-filling L-H sampling), where $n_{mc} \gg n$ with n the number of samples in the existing dataset $\mathcal{D}$. The optimization problem is then approximated by a discrete optimization

$$\mathbf{x}^* = \mathrm{argmax}_{\mathbf{x} \in \mathbf{X}_{mc}} \; acq(\mathbf{x}). \tag{E.10}$$

In the following, we will take (7.8) as the acquisition function in presenting the algorithm, but the algorithm applies equally to (7.9) and (7.5) as discussed in §7.2.4. In computing $acq(\mathbf{X}_{mc})$, one needs to evaluate a new GPR with $\hat{f}(\mathbf{X}_{mc}) = \mathbb{E}(\mathbf{X}_{mc}|\mathcal{D})$ and $\text{var}(\mathbf{X}_{mc}|\mathcal{D})$, with the former needed to calculate the function $p_{\hat{f}}(f)$ and its arguments $\hat{f}(\mathbf{x})$. A direct (brute-force) computation following (A.3) and (A.4) can be conducted as

$$\mathbb{E}(\mathbf{X}_{mc}|\mathcal{D}) = \text{K}(\mathbf{X}_{mc}, \mathbf{X})\text{K}(\mathbf{X}, \mathbf{X})^{-1}\mathbf{y}, \tag{E.11a}$$

$$\text{var}(\mathbf{X}_{mc}|\mathcal{D}) = \text{diag}\big(\text{K}(\mathbf{X}_{mc}, \mathbf{X}_{mc})\big) - \text{diag}\big(\text{K}(\mathbf{X}_{mc}, \mathbf{X})\text{K}(\mathbf{X}, \mathbf{X})^{-1}\text{K}(\mathbf{X}, \mathbf{X}_{mc})\big). \tag{E.11b}$$

The computational complexity of (E.11) consists of three major parts: (i) the Cholesky decomposition of $\text{K}(\mathbf{X}, \mathbf{X}) \in \mathbb{R}^{n^2}$ for computing its inverse $\text{K}(\mathbf{X}, \mathbf{X})^{-1}$, with complexity $O(n^3)$, (ii) obtaining each element in $\text{K}(\mathbf{X}_{mc}, \mathbf{X}) \in \mathbb{R}^{n_{mc}*n}$ with $O(n_{mc}*n)$, (iii) the Cholesky solve of $\text{K}(\mathbf{X}_{mc}, \mathbf{X})\text{K}(\mathbf{X}, \mathbf{X})^{-1}$ based on results of (i), with complexity $O(n_{mc}*n^2)$. Since $n_{mc} \gg n$, (iii) yields the highest computational complexity among the three procedures, instead of (i) (part of the re-training procedure) which is most computationally intensive for many other applications. In practice, for $n = 300$ and $n_{mc} = 10^5$ (typical sizes of problems in this paper), only 1% of the total computational time is spent on (i), while (ii) and (iii), on the other hand, contribute approximately equally to the remaining 99% computational time (note that (ii) has a large pre-factor in front of the Big O operator due to the need to compute covariance for each element). Therefore, alleviating the computational cost regarding (ii) and (iii) are most important in developing a fast computational approach.

Our developed approach leverages the recursive update of GPR used in [19, 20, 18, 49] with additional techniques of memory-time tradeoff and matrix multiplication strategy. To start, we employ the recursive update of the mean and variance building on that of the last step $f|\mathcal{D}_{n-1}$ (subscript $n-1$ refers to the dataset before the n-th

sample $\mathbf{x}_n$ is added):

$$\mathbb{E}(\mathbf{X}_{mc}|\mathcal{D}_{n-1}, \mathbf{x}_n) = \mathbb{E}(\mathbf{X}_{mc}|\mathcal{D}_{n-1}) + \frac{\text{cov}(\mathbf{X}_{mc}, \mathbf{x}_n|\mathcal{D}_{n-1})}{\text{var}(\mathbf{x}_n|\mathcal{D}_{n-1})}\big(f(\mathbf{x}_n) - \mathbb{E}(\mathbf{x}_n|\mathcal{D}_{n-1})\big), \quad \text{(E.12a)}$$

$$\text{var}(\mathbf{X}_{mc}|\mathcal{D}_{n-1}, \mathbf{x}_n) = \text{var}(\mathbf{X}_{mc}|\mathcal{D}_{n-1}) - \frac{\text{cov}(\mathbf{X}_{mc}, \mathbf{x}_n|\mathcal{D}_{n-1})^2}{\text{var}(\mathbf{x}_n|\mathcal{D}_{n-1})}, \quad \text{(E.12b)}$$

with:

$$\text{cov}(\mathbf{X}_{mc}, \mathbf{x}_n|\mathcal{D}_{n-1}) = k(\mathbf{X}_{mc}, \mathbf{x}_n) - \text{K}(\mathbf{X}_{mc}, \mathbf{X}_{n-1})\text{K}(\mathbf{X}_{n-1}, \mathbf{X}_{n-1})^{-1}k(\mathbf{X}_{n-1}, \mathbf{x}_n). \quad \text{(E.13)}$$

In (E.12), we can reuse $\mathbb{E}(\mathbf{X}_{mc}|\mathcal{D}_{n-1})$ and $\text{var}(\mathbf{X}_{mc}|\mathcal{D}_{n-1})$ from last iteration, and the major computational cost lies on (E.13). Here we note that a direct computation of (E.13), as conducted in [19, 20, 18] (judged by their uploaded codes in Github[2]), scales similarly as in (E.11). This is because the Cholesky solve step of $(\mathbf{X}_{mc}, \mathbf{X}_{n-1})\text{K}(\mathbf{X}_{n-1}, \mathbf{X}_{n-1})^{-1}$ costs $O(n_{mc} * (n-1)^2)$ that is similar as (iii) for the brute-force GPR formula (E.11). In order to reduce this part of the computational cost, we can compute (E.13) by *parenthesizing in a different way* (see [133] for applications in other contexts): first computing $\text{K}(\mathbf{X}_{n-1}, \mathbf{X}_{n-1})^{-1}k(\mathbf{X}_{n-1}, \mathbf{x}_n) \in \mathbb{R}^{(n-1)*1}$ and then multiplying the result by $\text{K}(\mathbf{X}_{mc}, \mathbf{X}_{n-1})$. In this way, the original $O(n_{mc}*n^2)$ complexity in (iii) is reduced to $O(n_{mc}*n)$. Regarding (ii), we now do not need to construct $\text{K}(\mathbf{X}_{mc}, \mathbf{X}_n)$ in (E.11), since only $\text{K}(\mathbf{X}_{mc}, \mathbf{X}_{n-1})$ is involved in (E.13) that can be taken from last iteration. The only new construction regards $k(\mathbf{X}_{mc}, \mathbf{x}_n) \in \mathbb{R}^{n_{mc}*1}$ only takes $O(n_{mc})$ complexity. This is a standard memory-time tradeoff idea where we save $\text{K}(\mathbf{X}_{mc}, \mathbf{X}_{n-1})$ in the memory, with the advantage of greatly reducing the computational requirement.

To illustrate the superiority of the developed computational method, we show in figure E.2 the computation time using the developed approach and direct computation as in (E.12) for $n_{mc} = 2 * 10^5$ and varying n on four cores of Intel Xeon Gold 6154

---

[2]https://github.com/ablancha/gpsearch
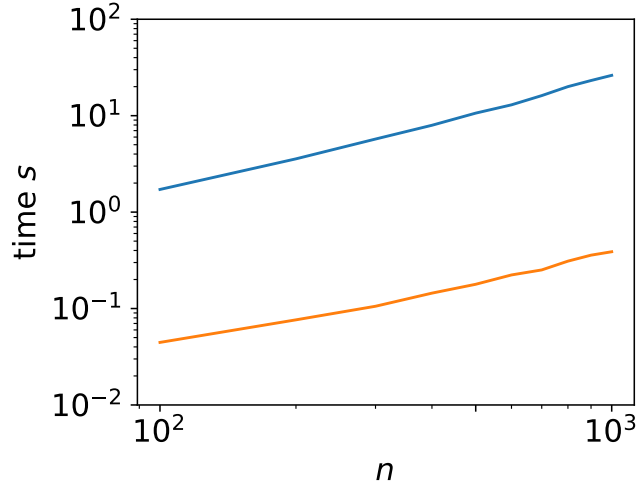
Figure E.2: Computation time for selecting one sequential sample using direct computation (E.12) (—) and the developed approach (—) for $n_{mc} = 2 * 10^5$ and varying n from 100 to 1000.

CPU. It is clear that the developed approach achieves a speedup of one and a half orders of magnitude.

E.4   Formula of Matern kernel

In this section, we introduce the Matern kernel for covariance function k, defined as

$$k(\mathbf{x}, \mathbf{x}') = \frac{\tau^2}{\Gamma(\nu)2^{\nu-1}} (\sqrt{2\nu}\, \text{dist}(\mathbf{x}, \mathbf{x}'))^\nu K_\nu\big(\sqrt{2\nu}\, \text{dist}(\mathbf{x}, \mathbf{x}')\big). \qquad (E.14)$$

The dist function in (E.14) is computed by:

$$\text{dist}(\mathbf{x}, \mathbf{x}') = ((\mathbf{x} - \mathbf{x}')^T \Lambda^{-1} (\mathbf{x} - \mathbf{x}'))^{\frac{1}{2}}, \qquad (E.15)$$

where τ and diagonal matrix Λ are hyperparameters representing the characteristic amplitude and length scales. $K_\nu(\cdot)$ is a modified Bessel function, and $\Gamma(\cdot)$ is the gamma function. ν is a pre-defined parameter controlling the continuity of the re-

alizations where a smaller value indicates a less smooth function. As $\nu \to \infty$, the Matern kernel becomes equivalent to the RBF kernel (infinitely differentiable) while $\nu = 1.5$ and 2.5 respectively indicate once and twice differentiable functions.

E.5 Results for Matern functions

In this section, we collect the results for two and three-dimensional Matern functions mentioned in §7.3.2. Figures E.3, E.4, and E.5 respectively correspond to figures 7.7, 7.8, and 7.10, but with the response functions (and GPR) generated by the Matern kernel (E.14). Conclusions from these cases with Matern functions are very similar to what we reach in §7.3.2 for the RBF functions. The only comment needed is that for the 3D Matern functions, the global optimal of α and t is achieved at $\alpha \approx 4$ and $t \approx 1$ instead of $\alpha \approx 3$ and $t \approx 1$ as in the RBF cases. However, the latter still provides a near-optimal performance for the Matern cases.

Figure E.3: Results for two-dimensional Matern functions. Error ϵ as function of number of samples for (a) $\alpha = 0$ and varying $t = 0.6$ (——), 0.8(——), 1(——), 1.2(——), 1.4(——), (b) $t = 1$ and varying $\alpha = 0$(——), 1 (——), 2(——), 3(——), 4(——), 6(——); (c) contour plot of $\log_{10} \epsilon$ at 146 sequential samples for varying t and α.
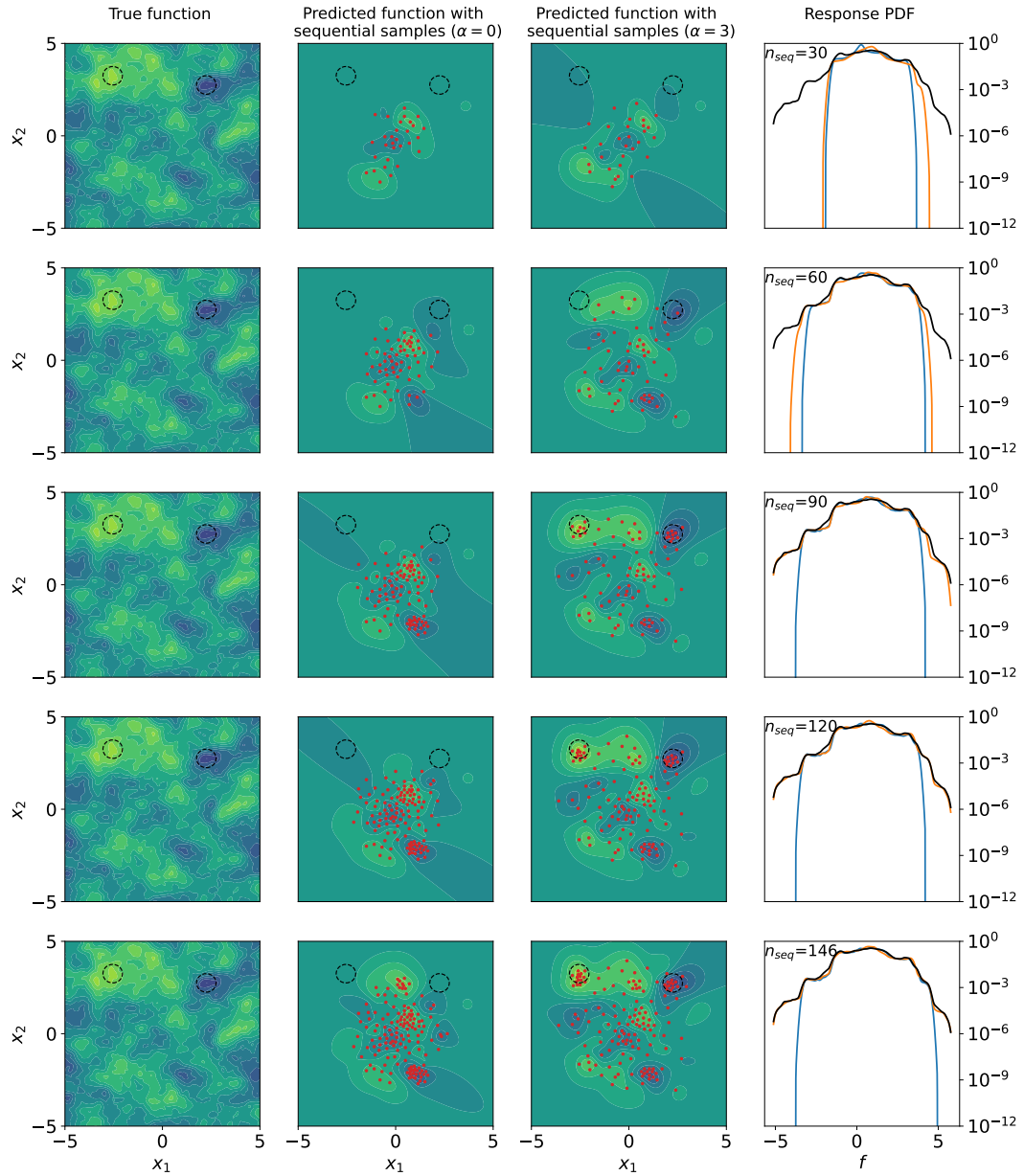
Figure E.4: First column: true response Matern function as a reference; second column: sequential samples (•) with $\alpha = 0$ on the predicted response function; third column: sequential samples (•) with $\alpha = 3$ on the predicted response function; fourth column: predicted PDF $p_{\hat{f}}(f)$ with $\alpha = 0$ (—) and $\alpha = 3$ (—) compared with the true PDF $p_f(f)$ (—). The top-to-bottom rows correspond to situations with number of sequential samples $n_{seq} = [30, 60, 90, 120, 146]$. The black circles shown in columns 1-3 mark the rare-event regions around $(-2.6, 3.2)$ and $(2.3, 2.8)$ that are missed by sequential samples with $\alpha = 0$ but captured with $\alpha = 3$.

(a)

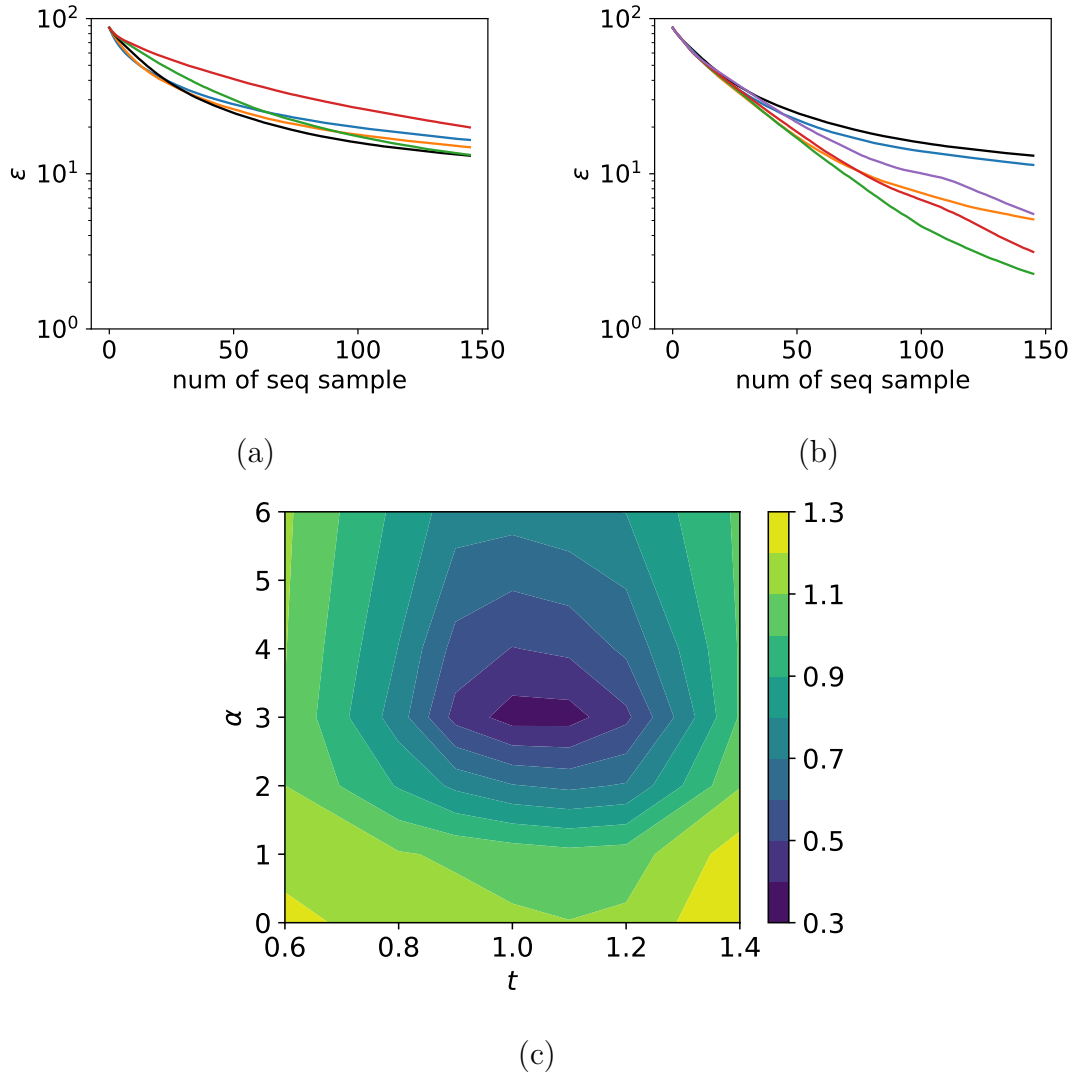(b)

(c)
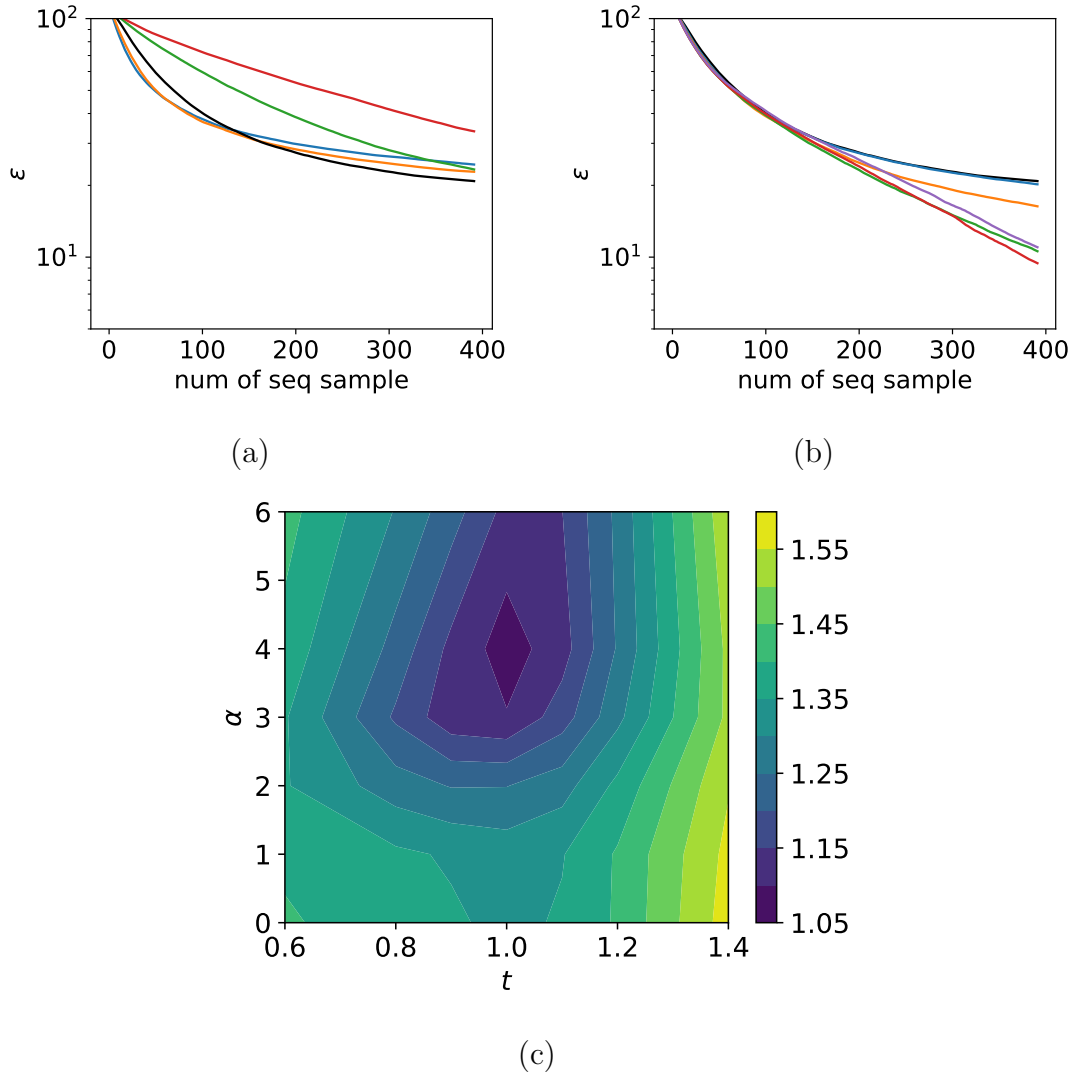
Figure E.5: Results for three-dimensional Matern functions. Error ϵ as function of number of samples for (a) $\alpha = 0$ and varying $t = 0.6$ (——), 0.8(——), 1(——), 1.2(——), 1.4(——), (b) $t = 1$ and varying $\alpha = 0$(——), 1 (——), 2(——), 3(——), 4(——), 6(——); (c) contour plot of $\log_{10} \epsilon$ at 392 sequential samples for varying t and α.

BIBLIOGRAPHY

[1] Panayiotis A Anastopoulos and Kostas J Spyrou. Ship dynamic stability assessment based on realistic wave group excitations. *Ocean Engineering*, 120:256–263, 2016.

[2] Panayiotis A Anastopoulos and Kostas J Spyrou. Evaluation of the critical wave groups method in calculating the probability of ship capsize in beam seas. *Ocean Engineering*, 187:106213, 2019.

[3] Panayiotis A Anastopoulos, Kostas J Spyrou, Christopher C Bassler, and Vadim Belenky. Towards an improved critical wave groups method for the probabilistic assessment of large ship motions in irregular seas. *Probabilistic Engineering Mechanics*, 44:18–27, 2016.

[4] Roy M Anderson and Robert M May. Population biology of infectious diseases: Part i. *Nature*, 280(5721):361–367, 1979.

[5] Christopher C Bassler, Vadim Belenky, and MJ Dipper. Application of wave groups to assess ship response in irregular seas. In *Proc. 11th int. ship stability workshop*, 2010.

[6] Christopher C Bassler, Vadim Belenky, and Martin J Dipper Jr. Characteristics of wave groups for the evaluation of ship response in irregular seas. In *International Conference on Offshore Mechanics and Arctic Engineering*, volume 49101, pages 227–237, 2010.

[7] Christopher C Bassler, Martin J Dipper, and Mark Melendez. Experimental ship dynamic stability assessment using wave groups. In *Contemporary Ideas on Ship Stability*, pages 507–520. Springer, 2019.

[8] V Belenky, K Weems, V Pipiras, D Glotzer, and T Sapsis. Tail structure of roll and metric of capsizing in irregular waves. In *Proc. 32nd Symp. Naval Hydrodynamics, Hamburg, Germany*, 2018.

[9] Vadim Belenky, Kenneth Weems, and Woei-Min Lin. Split-time method for estimation of probability of capsizing caused by pure loss of stability. *Ocean Engineering*, 122:333–343, 2016.

[10] Vadim Belenky, Kenneth M Weems, Christopher C Bassler, Martin J Dipper, Bradley L Campbell, and Kostas J Spyrou. Approaches to rare events in stochastic dynamics of ships. *Probabilistic Engineering Mechanics*, 28:30–38, 2012.

[11] Vadim L Belenky, Alexander B Degtyarev, and Alexander V Boukhanovsky. Probabilistic qualities of nonlinear stochastic rolling. *Ocean Engineering*, 25(1):1–25, 1998.

[12] Edin Berberović, Nils P. van Hinsberg, Suad Jakirlić, Ilia V. Roisman, and Cameron Tropea. Drop impact onto a liquid layer of finite thickness: Dynamics of the cavity evolution. *Phys. Rev. E*, 79:036306, 2009.

[13] Deniz A. Bezgin, Aaron B. Buhendwa, and Nikolaus A. Adams. Jax-fluids: A fully-differentiable high-order computational fluid dynamics solver for compressible two-phase flows. *Computer Physics Communications*, page 108527, 2022.

[14] Debby Bezzina and James Sayer. Safety pilot model deployment: Test conductor team report. *Report No. DOT HS*, 812(171):18, 2014.

[15] Barron J Bichon, Michael S Eldred, Laura Painton Swiler, Sandaran Mahadevan, and John M McFarland. Efficient global reliability analysis for nonlinear implicit performance functions. *AIAA journal*, 46(10):2459–2468, 2008.

[16] Mickael Binois and Nathan Wycoff. A survey on high-dimensional gaussian process modeling with application to bayesian optimization. *ACM Transactions on Evolutionary Learning and Optimization*, 2(2):1–26, 2022.

[17] Christopher M Bishop. *Pattern recognition and machine learning.* springer, 2006.

[18] Antoine Blanchard and Themistoklis Sapsis. Informative path planning for extreme anomaly detection in environment exploration and monitoring. *arXiv preprint arXiv:2005.10040*, 2020.

[19] Antoine Blanchard and Themistoklis Sapsis. Output-weighted optimal sampling for bayesian experimental design and uncertainty quantification. *arXiv e-prints*, pages arXiv–2006, 2020.

[20] Antoine Blanchard and Themistoklis Sapsis. Bayesian optimization with output-weighted optimal sampling. *Journal of Computational Physics*, 425:109901, 2021.

[21] Antoine Blanchard and Themistoklis Sapsis. Output-weighted optimal sampling for bayesian experimental design and uncertainty quantification. *SIAM/ASA Journal on Uncertainty Quantification*, 9(2):564–592, 2021.

[22] Bradley Campbell, Vadim Belenky, and Vladas Pipiras. Application of the envelope peaks over threshold (epot) method for probabilistic assessment of dynamic stability. *Ocean Engineering*, 120:298–304, 2016.

[23] Tianlong Chen, Xiaohan Chen, Wuyang Chen, Zhangyang Wang, Howard Heaton, Jialin Liu, and Wotao Yin. Learning to optimize: A primer and a benchmark. *The Journal of Machine Learning Research*, 23(1):8562–8620, 2022.

[24] Yutian Chen, Matthew W Hoffman, Sergio Gómez Colmenarejo, Misha Denil, Timothy P Lillicrap, Matt Botvinick, and Nando Freitas. Learning to learn without gradient descent by gradient descent. In *International Conference on Machine Learning*, pages 748–756. PMLR, 2017.

[25] Will Cousins, Miguel Onorato, Amin Chabchoub, and Themistoklis P Sapsis. Predicting ocean rogue waves from point measurements: An experimental study for unidirectional waves. *Physical Review E*, 99(3):032201, 2019.

[26] Will Cousins and Themistoklis P Sapsis. Reduced-order precursors of rare events in unidirectional nonlinear water waves. *Journal of Fluid Mechanics*, 790:368–388, 2016.

[27] Giovanni Dematteis, Tobias Grafke, and Eric Vanden-Eijnden. Extreme event quantification in dynamical systems with random components. *SIAM/ASA Journal on Uncertainty Quantification*, 7(3):1029–1059, 2019.

[28] Suraj S Deshpande, Lakshman Anumolu, and Mario F Trujillo. Evaluating the performance of the two-phase flow solver interFoam. *Computational Science & Discovery*, 5(1):014016, 2012.

[29] Douglas G Dommermuth and Dick KP Yue. A high-order spectral method for the study of nonlinear gravity waves. *Journal of Fluid Mechanics*, 184:267–288, 1987.

[30] Alexey Dosovitskiy, German Ros, Felipe Codevilla, Antonio Lopez, and Vladlen Koltun. CARLA: An open urban driving simulator. In *Proceedings of the 1st Annual Conference on Robot Learning*, pages 1–16, 2017.

[31] B Echard, N Gayton, and M Lemaire. Ak-mcs: an active learning reliability method combining kriging and monte carlo simulation. *Structural Safety*, 33(2):145–154, 2011.

[32] Mohammad Farazmand and Themistoklis P Sapsis. A variational approach to probing extreme events in turbulent dynamical systems. *Science advances*, 3(9):e1701533, 2017.

[33] Mohammad Farazmand and Themistoklis P Sapsis. Extreme events: Mechanisms and prediction. *Applied Mechanics Reviews*, 71(5), 2019.

[34] Shuo Feng, Yiheng Feng, Haowei Sun, Yi Zhang, and Henry X Liu. Testing scenario library generation for connected and automated vehicles: an adaptive framework. *IEEE Transactions on Intelligent Transportation Systems*, 2020.

[35] Shuo Feng, Yiheng Feng, Chunhui Yu, Yi Zhang, and Henry X Liu. Testing scenario library generation for connected and automated vehicles, part i: Methodology. *IEEE Transactions on Intelligent Transportation Systems*, 22(3):1573–1582, 2020.

[36] Shuo Feng, Haowei Sun, Xintao Yan, Haojie Zhu, Zhengxia Zou, Shengyin Shen, and Henry X Liu. Dense reinforcement learning for safety validation of autonomous vehicles. *Nature*, 615(7953):620–627, 2023.

[37] Shuo Feng, Xintao Yan, Haowei Sun, Yiheng Feng, and Henry X Liu. Intelligent driving intelligence test for autonomous vehicles with naturalistic and adversarial environment. *Nature communications*, 12(1):1–14, 2021.

[38] Alexander IJ Forrester, András Sóbester, and Andy J Keane. Multi-fidelity optimization via surrogate modelling. *Proceedings of the royal society a: mathematical, physical and engineering sciences*, 463(2088):3251–3269, 2007.

[39] Jacob Gardner, Geoff Pleiss, Kilian Q Weinberger, David Bindel, and Andrew G Wilson. Gpytorch: Blackbox matrix-matrix gaussian process inference with gpu acceleration. *Advances in neural information processing systems*, 31, 2018.

[40] M Ghil, P Yiou, Stéphane Hallegatte, BD Malamud, P Naveau, A Soloviev, P Friederichs, V Keilis-Borok, D Kondrashov, V Kossobokov, et al. Extreme events: dynamics, statistics and prediction. *Nonlinear Processes in Geophysics*, 18(3):295–350, 2011.

[41] Michael B Giles. Multilevel monte carlo methods. *Acta numerica*, 24:259–328, 2015.

[42] Xianliang Gong, Shuo Feng, and Yulin Pan. An adaptive multi-fidelity sampling framework for safety analysis of connected and automated vehicles. *IEEE Transactions on Intelligent Transportation Systems*, 2023.

[43] Xianliang Gong and Yulin Pan. Discussion:"bayesian optimal design of experiments for inferring the statistical expectation of expensive black-box functions"(pandita, p., bilionis, i., and panchal, j., 2019, asme j. mech. des., 141 (10), p. 101404). *Journal of Mechanical Design*, 144(5), 2022.

[44] Xianliang Gong and Yulin Pan. Effects of varying initial conditions of ship encountering wave groups in computing extreme motion statistics. *34th Symposium on Marine Hydrodynamics, Washington, DC*, 2022.

[45] Xianliang Gong and Yulin Pan. Quantification of extreme ship response probability in broadband wave fields. *38th International Workshop on Water Waves and Floating Bodies, Ann Arbor, USA*, 2022.

[46] Xianliang Gong and Yulin Pan. Sequential bayesian experimental design for estimation of extreme-event probability in stochastic input-to-response systems. *Computer Methods in Applied Mechanics and Engineering*, 395:114979, 2022.

[47] Xianliang Gong and Yulin Pan. A generalized likelihood-weighted optimal sampling algorithm for rare-event probability quantification. *arXiv preprint arXiv:2310.14457*, 2023.

[48] Xianliang Gong and Yulin Pan. Multi-fidelity bayesian experimental design to quantify extreme-event statistics. *SIAM Journal on Uncertainty Quantification*, In press, 2023.

[49] Xianliang Gong, Katerina Siavelis, Zhou Zhang, and Yulin Pan. Efficient computation of temporal exceeding probability of ship responses in a random wave field. *Applied Ocean Research*, 129:103405, 2022.

[50] Xianliang Gong, Zhou Zhang, Kevin J Maki, and Yulin Pan. Full resolution of extreme ship response statistics. *33rd symposium on naval hydrodynamics, Osaka, Japan*, 2020.

[51] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. MIT Press, 2016. http://www.deeplearningbook.org.

[52] John A Gubner. *Probability and random processes for electrical and computer engineers*. Cambridge University Press, 2006.

[53] Mengwu Guo and Jan S Hesthaven. Reduced order modeling for nonlinear structural analysis using Gaussian process regression. *Computer methods in applied mechanics and engineering*, 341:807–826, 2018.

[54] Stephen Guth, Alireza Mojahed, and Themistoklis P Sapsis. Evaluation of machine learning architectures on the quantification of epistemic and aleatoric uncertainties in complex dynamical systems. *arXiv preprint arXiv:2306.15159*, 2023.

[55] Achintya Haldar and Sankaran Mahadevan. Probability, reliability and statistical methods in engineering design. *(No Title)*, 2000.

[56] William V Harper and Sumant K Gupta. *Sensitivity/uncertainty analysis of a borehole scenario comparing Latin hypercube sampling and deterministic sensitivity approaches*. Office of Nuclear Waste Isolation, Battelle Memorial Institute Columbus, Ohio, 1983.

[57] Klaus Hasselmann. Stochastic climate models part i. theory. *tellus*, 28(6):473–485, 1976.

[58] Helge Holden, Bernt Øksendal, Jan Ubøe, and Tusheng Zhang. Stochastic partial differential equations. In *Stochastic partial differential equations*, pages 141–191. Springer, 1996.

[59] Zhen Hu and Sankaran Mahadevan. Global sensitivity analysis-enhanced surrogate (gsas) modeling for reliability analysis. *Structural and Multidisciplinary Optimization*, 53(3):501–521, 2016.

[60] Xun Huan and Youssef M Marzouk. Simulation-based optimal Bayesian experimental design for nonlinear systems. *Journal of Computational Physics*, 232(1):288–317, 2013.

[61] Zhiyuan Huang, Mansur Arief, Henry Lam, and Ding Zhao. Synthesis of different autonomous vehicles test approaches. In *2018 21st International Conference on Intelligent Transportation Systems (ITSC)*, pages 2000–2005. IEEE, 2018.

[62] Zhiyuan Huang, Henry Lam, and Ding Zhao. Towards affordable on-track testing for autonomous vehicle—a kriging-based statistical approach. In *2017 ieee 20th international conference on intelligent transportation systems (itsc)*, pages 1–6. IEEE, 2017.

[63] Hrvoje Jasak. OpenFOAM: Open source CFD in research and industry. *International Journal of Naval Architecture and Ocean Engineering*, 1(2):89–94, 2009.

[64] Hrvoje Jasak. Openfoam: open source cfd in research and industry. *International Journal of Naval Architecture and Ocean Engineering*, 1(2):89–94, 2009.

[65] Andre G Journel and Charles J Huijbregts. *Mining geostatistics*, volume 600. Academic press London, 1978.

[66] Nidhi Kalra and Susan M. Paddock. *Driving to Safety: How Many Miles of Driving Would It Take to Demonstrate Autonomous Vehicle Reliability?* RAND Corporation, Santa Monica, CA, 2016.

[67] Prabhjot Kaur, Samira Taghavi, Zhaofeng Tian, and Weisong Shi. A survey on simulators for testing self-driving cars. In *2021 Fourth International Conference on Connected and Autonomous Driving (MetroCAD)*, pages 62–70. IEEE, 2021.

[68] Marc C Kennedy and Anthony O'Hagan. Predicting the output from a complex computer code when fast approximations are available. *Biometrika*, 87(1):1–13, 2000.

[69] William Ogilvy Kermack and Anderson G McKendrick. A contribution to the mathematical theory of epidemics. *Proceedings of the royal society of london. Series A, Containing papers of a mathematical and physical character*, 115(772):700–721, 1927.

[70] Akira Kimura. Statistical properties of random wave groups. In *Coastal Engineering 1980*, pages 2955–2973. 1980.

[71] Solomon Kullback and Richard A Leibler. On information and sufficiency. *The annals of mathematical statistics*, 22(1):79–86, 1951.

[72] Miguel Lázaro-Gredilla and Michalis K Titsias. Variational heteroscedastic Gaussian process regression. In *ICML*, 2011.

[73] Seungjoon Lee, Felix Dietrich, George E Karniadakis, and Ioannis G Kevrekidis. Linking gaussian process regression with data-driven manifold embeddings for nonlinear data fusion. *Interface focus*, 9(3):20180083, 2019.

[74] Jing Li, Jinglai Li, and Dongbin Xiu. An efficient surrogate-based method for computing rare failure probability. *Journal of Computational Physics*, 230(24):8683–8697, 2011.

[75] Tony Lindeberg. Feature detection with automatic scale selection. *International journal of computer vision*, 30(2):79–116, 1998.

[76] Haitao Liu, Yew-Soon Ong, and Jianfei Cai. Large-scale heteroscedastic regression via Gaussian process. *IEEE transactions on neural networks and learning systems*, 32(2):708–721, 2020.

[77] Michael Selwyn Longuet-Higgins. Statistical properties of wave groups in a random sea state. *Philosophical Transactions of the Royal Society of London. Series A, Mathematical and Physical Sciences*, 312(1521):219–250, 1984.

[78] Pablo Alvarez Lopez, Michael Behrisch, Laura Bieker-Walz, Jakob Erdmann, Yun-Pang Flötteröd, Robert Hilbrich, Leonhard Lücken, Johannes Rummel, Peter Wagner, and Evamarie Wießner. Microscopic traffic simulation using sumo. In *The 21st IEEE International Conference on Intelligent Transportation Systems*. IEEE, 2018.

[79] Hui Lu, Shiqi Wang, Yuxuan Zhang, Jingru Lyu, and Kefei Mao. Adaptive sampling considering input and output space characteristics for the evaluation of autonomous systems. *Available at SSRN 4341960*.

[80] Zhan Ma and Wenxiao Pan. Data-driven nonintrusive reduced order modeling for dynamical systems with moving boundaries using Gaussian process regression. *Computer Methods in Applied Mechanics and Engineering*, 373:113495, 2021.

[81] Giovanni Malara, Pol D Spanos, and Felice Arena. Maximum roll angle estimation of a ship in confused sea waves via a quasi-deterministic approach. *Probabilistic Engineering Mechanics*, 35:75–81, 2014.

[82] Alonso Marco, Felix Berkenkamp, Philipp Hennig, Angela P Schoellig, Andreas Krause, Stefan Schaal, and Sebastian Trimpe. Virtual vs. real: Trading off simulations and physical experiments in reinforcement learning with bayesian optimization. In *2017 IEEE International Conference on Robotics and Automation (ICRA)*, pages 1557–1563. IEEE, 2017.

[83] Stefano Marelli and Bruno Sudret. An active-learning algorithm that combines sparse polynomial chaos expansions and bootstrap for structural reliability analysis. *Structural Safety*, 75:67–74, 2018.

[84] Kyle E Marlantes and Kevin J Maki. A hybrid data-driven model of ship roll. *arXiv preprint arXiv:2310.08651*, 2023.

[85] Alexandre N Marques, Remi R Lam, and Karen E Willcox. Contour location via entropy reduction leveraging multiple information sources. *arXiv preprint arXiv:1805.07489*, 2018.

[86] Joaquim RRA Martins and Andrew Ning. *Engineering design optimization*. Cambridge University Press, 2021.

[87] Romit Maulik, Themistoklis Botsas, Nesar Ramachandra, Lachlan R Mason, and Indranil Pan. Latent-space time evolution of non-intrusive reduced-order models using Gaussian process emulation. *Physica D: Nonlinear Phenomena*, 416:132797, 2021.

[88] Andrew McHutchon. Differentiating gaussian processes. *Cambridge (ed.)*, 2013.

[89] Michael D McKay, Richard J Beckman, and William J Conover. A comparison of three methods for selecting values of input variables in the analysis of output from a computer code. *Technometrics*, 42(1):55–61, 2000.

[90] Kevin A McTaggart. Ship capsize risk in a seaway using fitted distributions to roll maxima. *Journal of Offshore Mechanics and Arctic Engineering*, 122(2):141–146, 2000.

[91] Xuhui Meng, Hessam Babaee, and George Em Karniadakis. Multi-fidelity bayesian neural networks: Algorithms and applications. *Journal of Computational Physics*, 438:110361, 2021.

[92] Xuhui Meng and George Em Karniadakis. A composite neural network that learns from multi-fidelity data: Application to function approximation and inverse pde problems. *Journal of Computational Physics*, 401:109020, 2020.

[93] S. Mirjalili, S. S. Jain, and M. Dodd. Interface-capturing methods for two-phase flows: An overview and recent developments. *Annual Research Briefs*, page 117–135, 2017.

[94] Mustafa A Mohamad. *Direct and adaptive quantification schemes for extreme event statistics in complex dynamical systems*. PhD thesis, Massachusetts Institute of Technology, 2017.

[95] Mustafa A Mohamad and Themistoklis P Sapsis. Sequential sampling strategy for extreme event statistics in nonlinear dynamical systems. *Proceedings of the National Academy of Sciences*, 115(44):11138–11143, 2018.

[96] Galen E Mullins, Paul G Stankiewicz, R Chad Hawthorne, and Satyandra K Gupta. Adaptive generation of challenging scenarios for testing and evaluation of autonomous vehicles. *Journal of Systems and Software*, 137:197–215, 2018.

[97] Demin Nalic, Tomislav Mihalj, Maximilian Bäumler, Matthias Lehmann, Arno Eichberger, and Stefan Bernsteiner. Scenario based testing of automated driving systems: A literature survey. In *FISITA web Congress*, 2020.

[98] Ali Hasan Nayfeh, Dean T Mook, and P Holmes. Nonlinear oscillations. 1980.

[99] Nathan M. Newmark. A method of computation for structural dynamics. *Journal of the Engineering Mechanics Division*, 85:67–94, 1959.

[100] Jorge Nocedal. Updating quasi-newton matrices with limited storage. *Mathematics of computation*, 35(151):773–782, 1980.

[101] NVIDIA. Nvidia drive constellation: Virtual reality autonomous vehicle simulator, 2017.

[102] Michel K Ochi. *Ocean waves*. 1998.

[103] Miguel Onorato, L Cavaleri, S Fouques, O Gramstad, Peter AEM Janssen, Jaak Monbaliu, Alfred Richard Osborne, C Packozdi, Marina Serio, CT Stansberg, et al. Statistical properties of mechanically generated surface gravity waves: a laboratory experiment in a three-dimensional wave basin. *Journal of Fluid Mechanics*, 627:235–257, 2009.

[104] Art B. Owen. *Monte Carlo theory, methods and examples*. 2013.

[105] Piyush Pandita, Ilias Bilionis, and Jitesh Panchal. Bayesian optimal design of experiments for inferring the statistical expectation of expensive black-box functions. *Journal of Mechanical Design*, 141(10), 2019.

[106] Jeongbin Park, Bradford G Knight, Yingqian Liao, Marco Mangano, Bernardo Pacini, Kevin J Maki, Joaquim RRA Martins, Jing Sun, and Yulin Pan. Cfd-based design optimization of ducted hydrokinetic turbines. *Scientific Reports*, 13(1):17968, 2023.

[107] Paris Perdikaris, Maziar Raissi, Andreas Damianou, Neil D Lawrence, and George Em Karniadakis. Nonlinear information fusion algorithms for data-efficient multi-fidelity modelling. *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 473(2198):20160751, 2017.

[108] Paris Perdikaris, Daniele Venturi, Johannes O Royset, and George Em Karniadakis. Multi-fidelity modelling via recursive co-kriging and Gaussian–Markov random fields. *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 471(2179):20150018, 2015.

[109] Ethan Pickering, Stephen Guth, George Em Karniadakis, and Themistoklis P Sapsis. Discovering and forecasting extreme events via active learning in neural operators. *Nature Computational Science*, 2(12):823–833, 2022.

[110] Riccardo Poli, James Kennedy, and Tim Blackwell. Particle swarm optimization. *Swarm intelligence*, 1(1):33–57, 2007.

[111] Stefan Rahmstorf and Dim Coumou. Increase of extreme events in a warming world. *Proceedings of the National Academy of Sciences*, 108(44):17905–17909, 2011.

[112] C. E. Rasmussen and C. K. I. Williams. *Gaussian Process for Machine Learning*. MIT Press, 2006.

[113] Carl Edward Rasmussen. Gaussian processes in machine learning. In *Summer School on Machine Learning*, pages 63–71. Springer, 2003.

[114] Stefan Riedmaier, Thomas Ponn, Dieter Ludwig, Bernhard Schick, and Frank Diermeyer. Survey on scenario-based safety assessment of automated vehicles. *IEEE access*, 8:87456–87477, 2020.

[115] Jin Woo Ro, Partha S Roop, Avinash Malik, and Prakash Ranjitkar. A formal approach for modeling and simulation of human car-following behavior. *IEEE transactions on intelligent transportation systems*, 19(2):639–648, 2017.

[116] Samuel H Rudy and Themistoklis P Sapsis. Output-weighted and relative entropy loss functions for deep learning precursors of extreme events. *Physica D: Nonlinear Phenomena*, 443:133570, 2023.

[117] Themistoklis P Sapsis. Output-weighted optimal sampling for bayesian regression and rare event statistics using few samples. *Proceedings of the Royal Society A*, 476(2234):20190834, 2020.

[118] Themistoklis P Sapsis and Antoine Blanchard. Optimal criteria and their asymptotic form for data selection in data-driven reduced-order modelling with gaussian process regression. *Philosophical Transactions of the Royal Society A*, 380(2229):20210197, 2022.

[119] Soumalya Sarkar, Sudeepta Mondal, Michael Joly, Matthew E Lynch, Shaunak D Bopardikar, Ranadip Acharya, and Paris Perdikaris. Multifidelity and multiscale bayesian framework for high-dimensional engineering design and calibration. *Journal of Mechanical Design*, 141(12), 2019.

[120] Simo Särkkä. *Bayesian filtering and smoothing*. Number 3. Cambridge University Press, 2013.

[121] Roland Schobi, Bruno Sudret, and Joe Wiart. Polynomial-chaos-based kriging. *International Journal for Uncertainty Quantification*, 5(2), 2015.

[122] A Serani, R Pellegrini, J Wackers, C-E Jeanson, P Queutey, Michel Visonneau, and M Diez. Adaptive multi-fidelity sampling for cfd-based optimisation via radial basis function metamodels. *International Journal of Computational Fluid Dynamics*, 33(6-7):237–255, 2019.

[123] Shital Shah, Debadeepta Dey, Chris Lovett, and Ashish Kapoor. Airsim: High-fidelity visual and physical simulation for autonomous vehicles. In *Field and Service Robotics: Results of the 11th International Conference*, pages 621–635. Springer, 2018.

[124] Wanggang Shen and Xun Huan. Bayesian sequential optimal experimental design for nonlinear models using policy gradient reinforcement learning. *Computer Methods in Applied Mechanics and Engineering*, 416:116304, 2023.

[125] V. Shigunov. Probabilistic direct stability assessment. In *Proc. 15th Int. Ship Stability Workshop, 13–15 June, Stockholm, Sweden*, 2016.

[126] Vladimir Shigunov, Nikos Themelis, and Kostas J Spyrou. Critical wave groups versus direct monte-carlo simulations for typical stability failure modes of a container ship. In *Contemporary Ideas on Ship Stability*, pages 407–421. Springer, 2019.

[127] KT Shum and W Kendall Melville. Estimates of the joint statistics of amplitudes and periods of ocean waves using an integral transform technique. *Journal of Geophysical Research: Oceans*, 89(C4):6467–6476, 1984.

[128] Kevin M Silva and Kevin J Maki. Towards a computational fluid dynamics implementation of the critical wave groups method. *Ocean Engineering*, 235:109451, 2021.

[129] H Söding and E Tonguc. Computing capsizing frequencies in a seaway. In *Proc. 3rd Int. Conf. on Stability of Ships and Ocean Vehicles, Gdansk, Poland*, 1986.

[130] Mohamed S Soliman and JMT Thompson. Transient and steady state analysis of capsize phenomena. *Applied Ocean Research*, 13(2):82–92, 1991.

[131] Jialin Song, Yuxin Chen, and Yisong Yue. A general framework for multi-fidelity bayesian optimization with gaussian processes. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 3158–3167. PMLR, 2019.

[132] Kostas J Spyrou, Ioannis Tigkas, G Scanferla, and N Gavriilidis. Problems and capabilities in the assessment of parametric rolling. In *Proc 10th Int Ship Stab Workshop. Daejeon, Korea*, pages 47–55, 2008.

[133] Gilbert Strang. *Computational science and engineering*, volume 791. Wellesley-Cambridge Press Wellesley, 2007.

[134] Jian Sun, Huajun Zhou, Haochen Xi, He Zhang, and Ye Tian. Adaptive design of experiments for safety evaluation of automated vehicles. *IEEE Transactions on Intelligent Transportation Systems*, 2021.

[135] Zhili Sun, Jian Wang, Rui Li, and Cao Tong. Lif: A new kriging based learning function and its application to structural reliability analysis. *Reliability Engineering & System Safety*, 157:152–165, 2017.

[136] Rui Teixeira, Maria Nogal, and Alan O'Connor. Adaptive approaches in metamodel-based reliability analysis: A review. *Structural Safety*, 89:102019, 2021.

[137] Nikos Themelis and Kostas J Spyrou. Probabilistic assessment of ship stability. volume 115, pages 181–206, 2007.

[138] Michalis Titsias. Variational learning of inducing variables in sparse gaussian processes. In *Artificial intelligence and statistics*, pages 567–574. PMLR, 2009.

[139] Shanyin Tong, Anirudh Subramanyam, and Vishwas Rao. Optimization under rare chance constraints. *SIAM Journal on Optimization*, 32(2):930–958, 2022.

[140] Shanyin Tong, Eric Vanden-Eijnden, and Georg Stadler. Extreme event probability estimation using pde-constrained optimization and large deviation theory, with application to tsunamis. *Communications in Applied Mathematics and Computational Science*, 16(2):181–225, 2021.

[141] Simon Ulbrich, Till Menzel, Andreas Reschka, Fabian Schuldt, and Markus Maurer. Defining and substantiating the terms scene, situation, and scenario for automated driving. In *2015 IEEE 18th international conference on intelligent transportation systems*, pages 982–988. IEEE, 2015.

[142] Naoya Umeda, Hirotada Hashimoto, Dracos Vassalos, Shinichi Urano, and Kenji Okou. Nonlinear dynamics on parametric roll resonance with realistic numerical modelling. *International shipbuilding progress*, 51(2):205–220, 2004.

[143] Jeroen Wackers, Michel Visonneau, Simone Ficini, Riccardo Pellegrini, Andrea Serani, and Matteo Diez. Adaptive n-fidelity metamodels for noisy cfd data. In *AIAA AVIATION 2020 FORUM*, page 3161, 2020.

[144] Jeroen Wackers, Michel Visonneau, A Serani, R Pellegrini, R Broglia, and M Diez. Multi-fidelity machine learning from adaptive-and multi-grid rans simulations. In *33rd Symposium on Naval Hydrodynamics*, 2020.

[145] Eric A Wan and Rudolph Van Der Merwe. The unscented kalman filter for nonlinear estimation. In *Proceedings of the IEEE 2000 Adaptive Systems for Signal Processing, Communications, and Control Symposium (Cat. No. 00EX373)*, pages 153–158. Ieee, 2000.

[146] Zhong Yi Wan and Themistoklis P Sapsis. Reduced-space Gaussian process regression for data-driven probabilistic forecast of chaotic dynamical systems. *Physica D: Nonlinear Phenomena*, 345:40–55, 2017.

[147] Zhong Yi Wan, Pantelis Vlachas, Petros Koumoutsakos, and Themistoklis Sapsis. Data-assisted reduced-order modeling of extreme events in complex dynamical systems. *PloS one*, 13(5):e0197704, 2018.

[148] Guangyao Wang and Yulin Pan. Phase-resolved ocean wave forecast with ensemble-based data assimilation. *Journal of Fluid Mechanics*, 918, 2021.

[149] Hongqiao Wang, Guang Lin, and Jinglai Li. Gaussian process surrogates for failure detection: A bayesian experimental design approach. *Journal of Computational Physics*, 313:247–259, 2016.

[150] Kenneth Weems, Vadim Belenky, Bradley Campbell, Vladas Pipiras, and Themis Sapsis. Envelope peaks over threshold (epot) application and verification. In *Proc. 17th Intl. Ship Stability Workshop, Helsinki, Finland*, pages 71–79, 2019.

[151] Bruce J West, Keith A Brueckner, Ralph S Janda, D Michael Milder, and Robert L Milton. A new numerical method for surface hydrodynamics. *Journal of Geophysical Research: Oceans*, 92(C11):11803–11824, 1987.

[152] David C Wilcox et al. *Turbulence modeling for CFD*, volume 2. DCW industries La Canada, CA, 1998.

[153] Christopher Williams and Matthias Seeger. Using the nyström method to speed up kernel machines. *Advances in neural information processing systems*, 13, 2000.

[154] Zeping Wu, Donghui Wang, Wenjie Wang, Kun Zhao, Houcun Zhou, and Weihua Zhang. Hybrid metamodel of radial basis function and polynomial chaos expansions with orthogonal constraints for global sensitivity analysis. *Structural and Multidisciplinary Optimization*, 62:597–617, 2020.

[155] Wenting Xiao, Yuming Liu, Guangyu Wu, and Dick KP Yue. Rogue wave occurrence and dynamics by direct simulations of nonlinear wave-field evolution. *Journal of Fluid Mechanics*, 720:357–392, 2013.

[156] Jingxuan Yang, Haowei Sun, Honglin He, Yi Zhang, Henry X Liu, and Shuo Feng. Adaptive safety evaluation for connected and automated vehicles with sparse control variates. *IEEE Transactions on Intelligent Transportation Systems*, 2023.

[157] Yibo Yang, Antoine Blanchard, Themistoklis Sapsis, and Paris Perdikaris. Output-weighted sampling for multi-armed bandits with extreme payoffs. *Proceedings of the Royal Society A*, 478(2260):20210781, 2022.

[158] Jiaxiang Yi, Fangliang Wu, Qi Zhou, Yuansheng Cheng, Hao Ling, and Jun Liu. An active-learning method based on multi-fidelity kriging model for structural reliability analysis. *Structural and Multidisciplinary Optimization*, 63(1):173–195, 2021.

[159] Jiaqi Zhang, Louis Cammarata, Chandler Squires, Themistoklis P Sapsis, and Caroline Uhler. Active learning for optimal intervention design in causal models. *Nature Machine Intelligence*, pages 1–10, 2023.

[160] Ding Zhao, Xianan Huang, Huei Peng, Henry Lam, and David J LeBlanc. Accelerated evaluation of automated vehicles in car-following maneuvers. *IEEE Transactions on Intelligent Transportation Systems*, 19(3):733–744, 2017.

[161] Ding Zhao, Henry Lam, Huei Peng, Shan Bao, David J LeBlanc, Kazutoshi Nobukawa, and Christopher S Pan. Accelerated evaluation of automated vehicles safety in lane-change scenarios based on importance sampling techniques. *IEEE transactions on intelligent transportation systems*, 18(3):595–607, 2016.

[162] Hongyu Zheng, Fangfang Xie, Tingwei Ji, Zaoxu Zhu, and Yao Zheng. Multi-fidelity kinematic parameter optimization of a flapping airfoil. *Physical Review E*, 101(1):013107, 2020.

[163] Zhifu Zhu and Xiaoping Du. Reliability analysis with monte carlo simulation and dependent kriging predictions. *Journal of Mechanical Design*, 138(12), 2016.