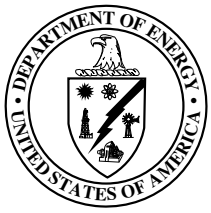
SAND2021-6951R



Preprint

Sandia National Laboratories

Surrogate Modeling for Efficiently, Accurately and Conservatively Estimating Measures of Risk

J. D. Jakeman[1], D. P. Kouri[1], and J. G. Huerta[2]

[1]Optimization and Uncertainty Quantification, Sandia National Laboratories, Albuquerque, NM, 87123
[2]Statistical Sciences, Sandia National Laboratories, Albuquerque, NM, 87123

Abstract

We present a surrogate modeling framework for conservatively estimating measures of risk from limited realizations of an expensive physical experiment or computational simulation. We adopt a probabilistic description of risk that assigns probabilities to consequences associated with an event and use risk measures, which combine objective evidence with the subjective values of decision makers, to quantify anticipated outcomes. Given a set of samples, we construct a surrogate model that produces estimates of risk measures that are always greater than their empirical estimates obtained from the training data. These surrogate models not only limit over-confidence in reliability and safety assessments, but produce estimates of risk measures that converge much faster to the true value than purely sample-based estimates. We first detail the construction of conservative surrogate models that can be tailored to the specific risk preferences of the stakeholder and then present an approach, based upon stochastic orders, for constructing surrogate models that are conservative with respect to families of risk measures. The surrogate models introduce a bias that allows them to conservatively estimate the target risk measures. We provide theoretical results that show that this bias decays at the same rate as the L^2 error in the surrogate model. Our numerical examples confirm that risk-aware surrogate models do indeed over-estimate the target risk measures while converging at the expected rate.

Article Info

Correspondence:
J. D. Jakeman
jdjakem@sandia.gov

Keywords:
Uncertainty quantification, risk, average value-at-risk, stochastic dominance, decision making, surrogate models, reliability

1 Introduction

Model-based decision making is subject to various sources of uncertainty. Risk assessment is needed to quantify the effect of these uncertainties on the severity of predicted outcomes. Conceptually, the term risk typically refers to the possibility of events with undesirable consequences. Following [49], we define risk formally, by a set of scenarios that could occur, consequences or outcomes of these scenarios, and the probability of these outcomes. For example, consider risk assessment of the truss structure depicted in Figure 1. Here, the set of possible scenarios are anticipated loads $P_1, \ldots, P_6$ and possible material properties A, E of the truss, which are often not exactly known. Each of these scenarios lead to different outcomes that

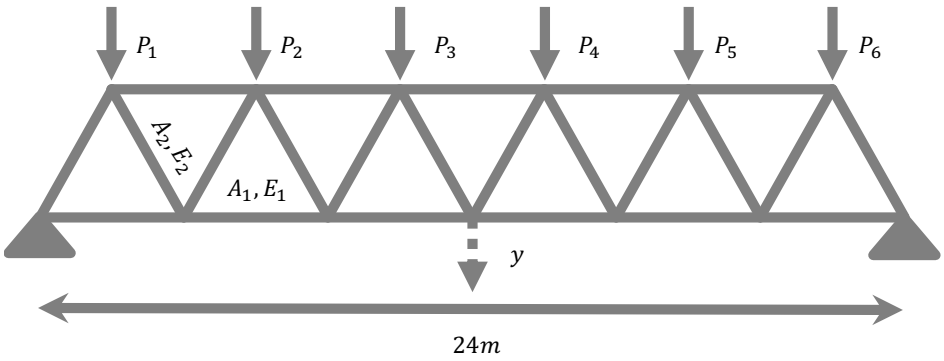


Figure 1: Two-dimensional truss structure with 23 bars and 13 nodes with uncertain loads $P_1, \ldots, P_6$ and material properties, specifically the cross section A_i and Young's modulus E_i of the horizontal and diagonal bars, $i = 1$ and $i = 2$ respectively. Our goal is to predict risk measures associated with the vertical displacement y at the mid-span.

we characterize by the vertical displacement y at the mid-span. Consequently, the mid-span displacement is a random variable Y.

The definition of risk that we use is objective and is simply the statement of the likelihood and consequences of outcomes. However, rational decision making requires quantitative metrics, known as *risk measures*, that combine objective evidence with subjective values. For our truss example, assume that the decision maker associates vertical displacement with decreased safety and specifies a constant C as the maximum displacement acceptable. Risk measures provide a quantitative answer to the question is $Y \leq C$, when uncertainty may occasionally lead to $Y > C$. If a risk measure is selected according to the preference of the stakeholder, then the single number that it produces can be easily interpreted.

Numerous risk measures have been studied in the literature (cf. [43] and the references therein). For example, the expectation, $\mathbb{E}[Y]$, is neutral to risk and can be used to certify a truss when the decision maker does not place any value judgment on deviations from average behavior. Using this risk measure, a system is deemed safe if $\mathbb{E}[Y] \leq C$. The mean-risk measure is inappropriate for risk-averse stakeholders, as it does not capture any notion of variability from the expected behavior. The probability of failure, $\text{prob}(Y > C)$, is commonly used for risk assessment and is often used when all failures are equally important to the decision maker, e.g. the cost of repairing a truss when it fails is always the same. However, when the stakeholder is interested in the magnitude of failure, e.g. larger truss deflections have greater repair costs, a risk measure such as average value-at-risk (AVaR) may be more appropriate. See [24, 25, 26] for recent works using risk measures to quantify and assess risk in engineering applications.

Estimating risk measures that characterize the nature of rare and undesirable events can be challenging. Monte Carlo (MC) sampling can be used to approximate risk measures. However, this typically requires a large number of experiments or simulations, which can make accurately estimating risk intractable when data generation is expensive. The cost of estimating tail statistics using MC sampling can be reduced using importance sampling (IS) [20, 30, 41] or subset simulation (SS) [3, 33]. However, similarly to MC, the estimates of rare events obtained using these methods converge slowly as the number of samples is increased. Replacing an expensive experiment or simulation with a cheap surrogate model is a popular approach for reducing the cost of sample-based approaches like MC, IS and SS methods [7, 14, 27]. Various types of surrogate models can be used for estimating risk measures, including polynomials [29, 31, 46], Gaussian processes (also known as kriging models) [5, 16, 44, 45, 47], neural networks [9, 32], and reduced order models [19, 50, 51]. However, all of these approaches introduce an error, in estimated measures of risk, that is often not accounted for.

Surrogate models are constructed using a finite number of data points. As a consequence, the resulting estimates are themselves random since they depend on the samples used. Without imposing constraints on surrogate construction, certain draws of samples can lead to an over-estimate of risk and other draws to an under-estimate. In this paper, we present a framework to produce biased surrogate models that

conservatively estimate target risk measures. Specifically, letting the outcomes obtained from a surrogate model be $\hat{Y}$, we will construct approximations of the form

$$\hat{Y} \approx Y$$

that *conservatively* approximate a risk measure $\mathcal{R}(Y)$, in the sense that

$$\mathcal{R}(\hat{Y}) \geq \mathcal{R}(Y).$$

The resulting surrogate models produce estimates of the target risk measures $\mathcal{R}$ that are always greater than the empirical estimate obtained from the training data. These surrogate models not only limit overconfidence in model reliability, but produce estimates of risk that converge much faster to the true risk than purely sample-based estimates from MC, IS, or SS.

We developed two approaches for constructing conservative surrogate models. The first approach uses the risk quadrangle [39] to translate a stakeholders risk aversion preferences, encoded by a utility or regret (disutility function), into a risk measure that is statistically consistent with the defined regret. The decision maker's notion of regret is then used to formulate an error measure that can be used to train the surrogate model, from data pairs of scenarios and associated outcomes, in a manner that is guaranteed to conservatively estimate the risk measure elicited from the stakeholder. The second approach uses stochastic dominance constraints [12] to construct surrogate models that are conservative with respect to a set of risk measures. This approach is useful when regret preferences are hard to elicit or when multiple stakeholders have competing subjective values. We construct surrogate models using both first-order and second-order dominance constraints. The surrogate-based $\hat{Y}$ output distribution has first-order dominance over the true Y distribution, if for any outcome y, the probability of exceeding y satisfies

$$\text{prob}(\hat{Y} > y) \geq \text{prob}(Y > y).$$

By enforcing first-order stochastic dominance, we ensure that we do not underestimate risk for almost any reasonable risk measure including the popular probability of failure (PoF). Similarly, second-order dominance is a weaker condition that we use to effectively enforce the conservativeness of the surrogate model when first-order dominance is too risk averse.

The subsequent sections are organized as follows. In Section 2, we introduce various risk measures and stochastic dominance, and discuss the challenges of using surrogate models to estimate risk measures. We then present a mathematical framework that uses the risk quadrangle to construct conservative surrogate models tailored to the specific risk preferences of the stakeholder in Section 3. Section 4 presents an alternative approach that uses stochastic dominance to construct surrogate models that are conservative with respect to broad classes of risk measures. Finally, we provide numerical examples that demonstrate the utility and efficacy of risk-adapted surrogate models in Section 5 and provide conclusions in Section 6.

2 Risk Assessment for Decision Making

Risk assessment is typically undertaken to aid a decision-making process. For example, it can be used compare the risk of two different truss structures of differing price. In this section, we discuss popular risk measures used to quantify risk based upon the subjective values of the decision maker. We then introduce stochastic dominance and discuss how it can be used when regret preferences are hard to elicit or when multiple stakeholders have competing subjective values.

2.1 Risk Measures

A risk measure is any real-valued function, $\mathcal{R}$, acting on a space of random variables, that satisfies $\mathcal{R}(C) = C$ for all constant values $C \in \mathbb{R}$. This requirement ensures that deterministic quantities are without risk. Practically speaking, risk measures provide a deterministic quantification of the distribution of possible outcomes associated with a risk and therefore are assumed to have the same units as the outcomes. There are numerous desirable axioms that a risk measure $\mathcal{R}$ should satisfy. For random variables Y and Y', four popular conditions from the literature are

(R1) **Convexity:** $\mathcal{R}(tY + (1-t)Y') \leq t\mathcal{R}(Y) + (1-t)\mathcal{R}(Y')$ for all $t \in [0,1]$;

(R2) **Monotonicity:** If $Y \leq Y'$ almost surely, then $\mathcal{R}(Y) \leq \mathcal{R}(Y')$;

(R3) **Translation Equivariance:** $\mathcal{R}(Y + C) = \mathcal{R}(Y) + C$ for all $C \in \mathbb{R}$;

(R4) **Positive homogeneity:** $\mathcal{R}(tY) = t\mathcal{R}(Y)$ for all $t \geq 0$;

(R5) **Law Invariance:** If $F_Y(t) = F_{Y'}(t)$ for all $t \in \mathbb{R}$, then $\mathcal{R}(Y) = \mathcal{R}(Y')$.

A risk measure that satisfies (R1), (R2) and (R3) is called a convex risk measure [15]. If a convex risk measure additionally satisfies (R4), then it is said to be coherent [2]. The monotonicity condition (R2) ensures that any consequence Y that is always less than a different consequence Y' should have less risk when measured by $\mathcal{R}$, while the translation equivariance properties (R3) ensures that the deterministic quantity C is without risk. Positive homogeneity (R4) of $\mathcal{R}$ ensures that if we change the units of Y, then the units of $\mathcal{R}(Y)$ also change in a consistent fashion. Finally, if Y and Y' are law invariant (R5), then its risk is not affected by rearranging the scenarios associated with Y or Y'. Below, we list various popular law-invariant risk measures.

- **Mean**. $\mathcal{R}(Y) = \mathbb{E}[Y]$. This coherent risk measure is often referred to as risk neutral and should not be used in high-consequence applications when extreme behavior of Y is important.

- **Mean-plus-standard-deviation**. $\mathcal{R}(Y) = \mathbb{E}[Y] + \lambda \mathbb{V}[Y]^{\frac{1}{2}}$, where $\lambda > 0$ is a user specified weight. This risk measure is often an improvement over mean-based risk as it quantifies the deviation of the outcomes from their average. However, this risk measure assigns the same weight to both positive and negative deviations, and therefore fails to account for skewness in the distribution. This risk measure does not satisfy the monotonicity condition (R2) and therefore is neither a coherent nor convex risk measure.

- **Worst-case**. $\mathcal{R}(Y) = \sup Y$. This coherent risk measure is extremely conservative and is typically difficult to estimate. It can often misinform decisions based on a single bad outcome because it does not consider how unlikely that outcome is and how "good" the alternative outcomes are.

- **Upper quantile**. $\mathcal{R}(Y) = q_p[Y] := \inf\{y \in \mathbb{R} \mid F_Y(y) \geq p\}$ for $p \in (0,1)$, where $F_Y(t) := \text{prob}(Y \leq t)$ denotes the cumulative distribution function (CDF) of a random variable Y. The p^{th} upper quantile (often called the value-at-risk or VaR in financial applications) provides a reasonable measure of conservativeness. In words, Y exceeds $q_p[Y]$ with probability $1-p$. Roughly speaking, when $p = 1$, $q_p[Y]$ is the largest value of Y and when $p = 0$, $q_p[Y]$ is the smallest value of Y. The upper quantile is closely related to the probability of failure. In particular, we have the equivalence

$$q_p(Y) \leq C \quad \iff \quad \text{prob}(X > C) \leq 1 - p. \tag{1}$$

  Unfortunately, the upper quantile does not characterize the weight of the distribution tail. Moreover, the upper quantile does not satisfy (R1) and therefore is neither a coherent nor convex risk measure.

- **Average value-at-risk**. $\mathcal{R}(Y) = \text{AVaR}_p[Y] := \frac{1}{1-p}\int_p^1 q_\alpha[Y]\,\mathrm{d}\alpha$ for $p \in (0,1)$. AVaR measures the average of the tail in excess of $q_p[Y]$ and therefore provides a conservative alternative to the upper quantile (i.e., $\text{AVaR}_p[Y] \geq q_p[Y]$ for all $p \in (0,1)$ and for all Y). When $p = 0$ AVaR returns the mean risk measure and when $p \to 1$ AVaR returns the worst-case risk measure. As shown in [38], one can compute AVaR by solving the one-dimensional convex optimization problem

$$\text{AVaR}_p[Y] = \inf_{t \in \mathbb{R}} \left\{ t + \frac{1}{1-p}\mathbb{E}[\max\{0, Y - t\}] \right\}.$$

  Moreover, AVaR is closely related to the concept of buffered probability of failure [35]. In particular, the buffered probability that Y exceeds C is defined as $\bar{p}_C(Y) = 1 - p$ where p is chosen so that $\text{AVaR}_p[Y] = C$. Analogous to the relation between the upper quantile and the probability of failure, we have the equivalence

$$\text{AVaR}_p[Y] \leq C \quad \iff \quad \bar{p}_C(Y) < 1 - p.$$

  AVaR is a coherent risk measure.

Figure 2 compares the probability density functions (PDFs) of two random variables with lognormal distributions. The preferred loss distribution should depend on the stakeholder's risk preference, which is encoded in the risk measure $\mathcal{R}$. If one focuses on mean risk, then Y_1 is preferred since $\mathbb{E}[Y_1] = 1.13 < 1.30 = \mathbb{E}[Y_2]$. In contrast, Y_2 produces a smaller mean-plus-standard-deviation risk with $\lambda = 1$, i.e.,

$$\mathbb{E}[Y_2] + \mathbb{V}[Y_2]^{\frac{1}{2}} = 0.42 < 0.60 = \mathbb{E}[Y_1] + \mathbb{V}[Y_1]^{\frac{1}{2}}.$$

Finally, the upper quantile with $p = 0.75$ cannot distinguish between these distributions as they produce the same value, whereas AVaR with $p = 0.75$ prefers Y_2 since $\text{AVaR}_{0.75}[Y_2] = 1.52 < 1.95 = \text{AVaR}_{0.75}[Y_1]$. In particular, the tail-weight of Y_2 is smaller than that of Y_1. The quantiles are represented by the left end points of the shaded regions in Figure 2 and the values of AVAR are the area of these shaded region.

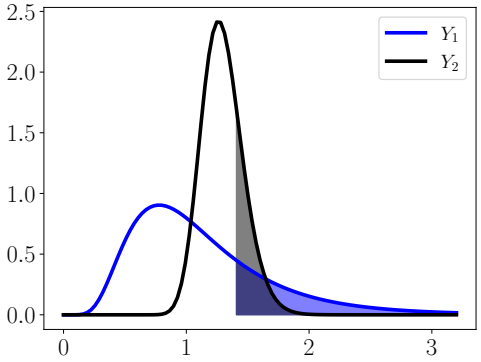


Figure 2: The preferred loss PDF depends on the risk measure. Y_1 is log-normally distributed with parameters $\mu_1 = 0$ and $\sigma_1 = 0.5$. Y_2 is log-normally distributed with parameters $\mu_2 = 0.25$ and $\sigma_2 = \sigma_1 - (4\sqrt{2}\text{erf}^{-1}(0.5))^{-1}$, where erf^{-1} is the inverse error function.

2.2 Eliciting Risk Preference

There are various intuitive ways to mathematically model a decision maker's aversion to risk. Perhaps the most intuitive approach is through so-called acceptance sets. In this approach, the decision maker defines a set of acceptable risks $\mathcal{A}$ [15]. For example, consider the truss structure depicted in Figure 1 and let Y denote the vertical mid-plane displacement beyond a reference value C. Using their acceptance set, the decision maker would deem Y to be acceptable if $Y \in \mathcal{A}$. Associated with $\mathcal{A}$ is the risk measure

$$\mathcal{R}_{\mathcal{A}}(Y) := \inf\{d \in \mathbb{R} \,|\, Y - d \in \mathcal{A}\}.$$

In the context of the truss structure, $\mathcal{R}_{\mathcal{A}}(Y)$ (units of displacement) is the smallest amount of certain deformation required to produce an acceptable mid-plane displacement. The minimal assumption on $\mathcal{A}$ is that if $Y \in \mathcal{A}$ and $Y' \leq Y$ almost surely, then $Y' \in \mathcal{A}$. Under this assumption, $\mathcal{R}_{\mathcal{A}}$ satisfies (R2) and (R3). If $\mathcal{A}$ is convex, then $\mathcal{R}_{\mathcal{A}}$ satisfies (R1) and if $\mathcal{A}$ is a cone, then $\mathcal{R}_{\mathcal{A}}$ satisfies (R4). In fact, there is a one-to-one correspondence between acceptance sets and risk measures. In particular, any risk measure $\mathcal{R}$ that satisfies (R2) and (R3) defines the acceptance set

$$\mathcal{A}_{\mathcal{R}} := \{Y \,|\, \mathcal{R}(Y) \leq 0\}.$$

For reliability applications, one may say that an output Y is acceptable if the probability that Y exceeds zero (i.e., the probability of failure) is less than $1 - p$ for some probability level $p \in (0, 1)$. Encoding this condition into the definition of the acceptance set $\mathcal{A}$ produces

$$\mathcal{A} = \{Y \,|\, \text{prob}(Y > 0) \leq 1 - p\}.$$

Given the equivalence (1), we can rewrite $\mathcal{A}$ as

$$\mathcal{A} = \{Y \,|\, q_p(Y) \leq 0\}.$$

Consequently, the risk measure associated with $\mathcal{A}$ is the upper quantile $\mathcal{R}_{\mathcal{A}}(Y) = q_p(Y)$.

An alternate approach to modeling risk aversion is based on utility theory. A utility measure $\mathcal{U}$ is an increasing function that quantifies the decision maker's happiness associated with a risk. A utility measure is called *risk averse* if it is concave and normalized if $\mathcal{U}(0) = 0$. A common approach to defining $\mathcal{U}$ is as the expected utility $\mathcal{U}(Y) = \mathbb{E}[u(Y)]$ where $u : \mathbb{R} \to \mathbb{R}$ is an increasing function. Figure 3 demonstrates different types of utility functions u. A common utility function in economic applications is the exponential utility $u(t) = 1 - \exp(-t)$. Other utility functions include the iso-elastic and hyperbolic absolute risk aversion utility functions. In this work, we consider the case for which large values of the output Y are undesirable. For this purpose, we will use the notion of regret or disutility. A regret measure $\mathcal{V}$ is defined using a utility function

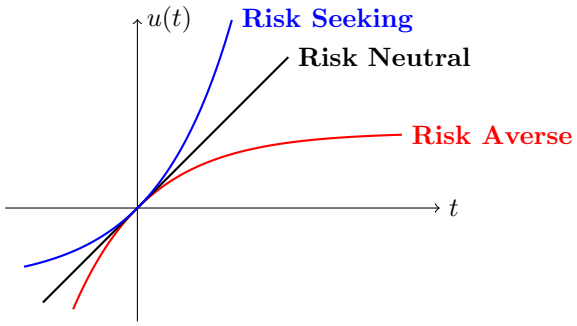


Figure 3: Utility function selection based on risk preference. If $u(t)$ is concave, then the decision maker is averse to risk. If $u(t) = t$, then they are neutral to risk. Finally, if $u(t)$ is convex, then they are risk seeking.

$\mathcal{U}$ by the relationship $\mathcal{V}(Y) = -\mathcal{U}(-Y)$. Relating the concept of regret to the concept of an acceptance set, a common approach in economic applications is to set

$$\mathcal{A} = \{Y \,|\, \mathcal{V}(Y) \leq 0\}.$$

In particular, the decision maker is willing to accept the risk Y if its regret is negative. On the other hand, the decision maker is not willing to accept Y if its regret is positive and they are indifferent to Y if the regret is zero. The risk associated with $\mathcal{A}$ is then

$$\mathcal{R}_{\mathcal{A}}(Y) = \inf\{d \in \mathbb{R} \,|\, \mathcal{V}(Y - d) \leq 0\},$$

which is often called the *u-mean* [8] and represents the certain value at which the decision maker is indifferent to the outcome. In the context of the truss example, $\mathcal{R}_{\mathcal{A}}(Y)$ is the mid-plane displacement beyond the reference displacement C that a decision maker must be willing to accept. One downside to this choice of risk measure is that it is often overly conservative. For example, consider the regret measure $\mathcal{V}(Y) = \frac{1}{1-p}\mathbb{E}[\max\{0, Y\}]$. For this regret measure, $\mathcal{V}(Y - d) \leq 0$ if and only if $d \geq \sup Y$. Consequently, $\mathcal{R}_{\mathcal{A}}(Y) = \sup Y$ is the worst-case scenario risk measure.

Related to the u-mean risk measures are the *optimized certainty equivalent* (OCE) risk measures [4] given by

$$\mathcal{R}(Y) = \inf_{d \in \mathbb{R}} \{d + \mathcal{V}(Y - d)\}.$$

To interpret the OCE risk measure, we again consider the truss example. For this example, we interpret the regret $\mathcal{V}(Y)$ as the anticipated displacement beyond the reference value C. If we can preload the truss structure to achieve a certain mid-plane displacement d, then the anticipated displacement after preloading is $d + \mathcal{V}(Y - d)$. Consequently, $\mathcal{R}(Y)$ represents the anticipated displacement that would result from optimal preloading. The OCE risk measures provide a relaxation of the u-mean risk measures in the following sense. By Lagrangian duality, the u-mean and OCE risk measures satisfy

$$\mathcal{R}_{\mathcal{A}}(Y) = \inf_{d \in \mathbb{R}} \sup_{\lambda \geq 0} \{d + \lambda \mathcal{V}(Y - d)\} = \sup_{\lambda \geq 0} \inf_{d \in \mathbb{R}} \{d + \lambda \mathcal{V}(Y - d)\} \geq \mathcal{R}(Y).$$

It is worth noting that if the regret measure is generated from the common exponential utility measure, i.e., $\mathcal{U}(Y) = \mathbb{E}[1 - \exp(-Y)]$, then the OCE and u-mean risk measures are the same $\mathcal{R}_{\mathcal{A}} = \mathcal{R}$. The OCE risk measure is closely related to generalized measures of deviation and error. This relationship is formalized by the risk quadrangle presented in [39].

2.3 Stochastic Dominance

The choice of risk measure should be tailored to the stakeholder's risk preference. Unfortunately, risk preferences are often difficult to elicit. In such situations, it may be more appropriate to assess risk with respect to a class of risk measures or even entire distributions. Stochastic dominance, also called stochastic

orders, can be used in such cases. For this work, we restrict our attention to the first and second stochastic orders.

The random variable Y is said to dominate another random variable Y' with respect to the first stochastic order, denoted $Y \succeq_{(1)} Y'$, if

$$F_Y(t) \leq F_{Y'}(t) \quad \forall\, t \in \mathbb{R}. \tag{2}$$

The first-order stochastic dominance condition (2) is equivalent to

$$\mathbb{E}[u(Y)] \leq \mathbb{E}[u(Y')]$$

for all nondecreasing functions $u : \mathbb{R} \to \mathbb{R}$ for which the expectations exist. On the other hand, Y dominates Y' with respect to the second stochastic order, denoted $Y \succeq_{(2)} Y'$, if

$$\begin{aligned} &\int_{-\infty}^{t} F_Y(\eta)\,\mathrm{d}\eta \leq \int_{-\infty}^{t} F_{Y'}(\eta)\,\mathrm{d}\eta \quad \forall\, t \in \mathbb{R} \\ \iff \quad &\mathbb{E}[\max\{0, t - Y\}] \leq \mathbb{E}[\max\{0, t - Y'\}] \quad \forall\, t \in \mathbb{R}. \end{aligned}$$

Second-order stochastic dominance (SSD) focuses on properties of the left tail, i.e., small values of Y. For engineering and science applications, we are typically interested in undesirable large values of Y. For this reason, we employ the so-called increasing convex order defined by

$$Y' \succeq_{\text{icx}} Y \quad \iff \quad \mathbb{E}[u(Y)] \leq \mathbb{E}[u(Y')] \tag{3}$$

for all nondecreasing convex functions $u : \mathbb{R} \to \mathbb{R}$ for which the expectations exist. In words, if $Y' \succeq_{\text{icx}} Y$, then the expected utility of Y is always smaller than the expected utility of Y'. The increasing convex order is related to SSD through the following equivalence

$$Y' \succeq_{\text{icx}} Y \quad \iff \quad -Y \succeq_{(2)} -Y'.$$

In particular, we have that $Y' \succeq_{\text{icx}} Y$ if and only if

$$\mathbb{E}[\max\{0, Y - t\}] \leq \mathbb{E}[\max\{0, Y' - t\}] \quad \forall\, t \in \mathbb{R}.$$

As such, we will refer to dominance with respect to the increasing convex order simply as SSD. It is clear from the utility representation of FSD and SSD that if $Y \succeq_{(1)} Y'$, then $-Y \succeq_{(2)} -Y'$. However, the converse is not true in general.

Figure 4 illustrates the difference between first-order (FSD) and second-order dominance. The FSD relation $Y_2 \succeq_{(1)} Y_1$ ensures that the CDF of Y_2 lies to the right of the CDF Y_1 on the plot (left) plot. This relation means that $Y_2 \succeq_{(1)} Y_1$ implies that the probability of failure of Y_2 will be greater than that computed using Y_1 for any threshold C. We will use this property to construct surrogates that conservatively estimate the probability of failure in Section 4.

Despite being a less risk-averse condition than FSD, the SSD relationship $-Y_3 \succeq_{(2)} -Y_2$ is more risk-averse than requiring

$$\mathcal{R}(Y_3) \geq \mathcal{R}(Y_2) \tag{4}$$

for many risk measures $\mathcal{R}$. In fact, the SSD relation $-Y_3 \succeq_{(2)} -Y_2$ ensures that $\text{AVaR}_p[Y_3]$ is larger than $\text{AVaR}_p[Y_2]$ for any confidence level $p \in (0, 1)$. Consequently, if the underlying probability space is nonatomic, then (4) holds for all risk measures that are proper, lower semicontinuous and satisfy (R1), (R2), (R3), and (R5) [43, Th. 6.51]. Note the area of the shaded regions in the middle plot equals the value of the conditional expectation for Y_2 and Y_3 for $t = 3.5$, which are shown as dots on the right plot.

2.4 Surrogate Modeling for Risk Assessment

To develop our surrogate modeling technique, we consider the data collection process

$$Y = f(X),$$

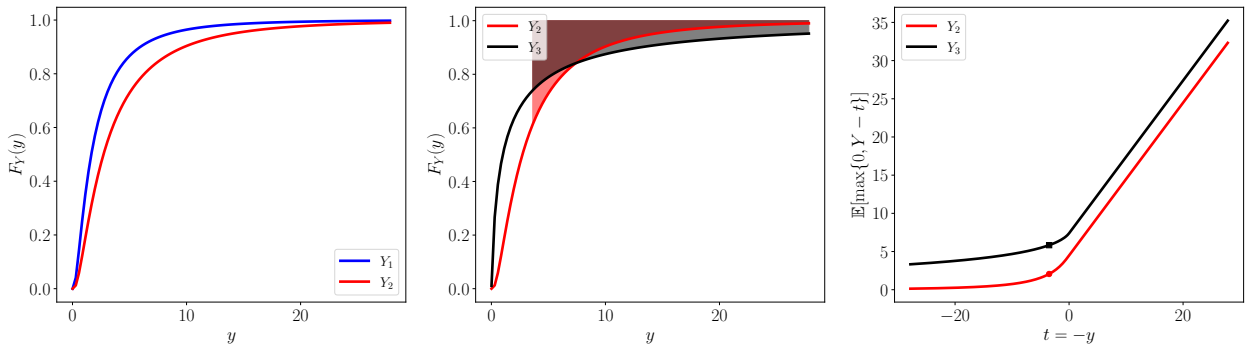


Figure 4: Comparison of first- and second-order stochastic dominance. (Left) $Y_2 \succeq_{(1)} Y_1$, (Middle) There is no first-order dominance but $-Y_3 \succeq_{(2)} -Y_2$. (Right) The values $\mathbb{E}[\max\{0, t-Y\}]$ confirming second-order dominance.

where $f(\cdot)$ represents a consequence or loss associated with input or environmental variables $X = (X_1, \ldots, X_D) \in \Omega_X \subseteq \mathbb{R}^D$ and $Y \in \Omega_Y \subseteq \mathbb{R}$ is the observed consequence. The random variable Y could be the output of an expensive physical experiment or numerical simulation. If we assign probability laws P_X and P_Y to describe the probability of the scenarios X and the associated outcomes Y, we can use sampling methods, such as MC, IS, and SS, to estimate $\mathcal{R}(Y)$ using sample averages. For example, denoting samples of X by $x^{(m)}$ and the evaluations of f at $x^{(m)}$ by $y^{(m)} = f(x^{(m)})$, for $m = 1, \ldots, M$, we can estimate the probability that the outcomes Y exceeds some threshold $t \in \mathbb{R}$ by

$$\text{prob}(Y > t) = \int_{\Omega_Y} \mathbb{I}_t(y)\, P_Y(\mathrm{d}y) \approx \sum_{m=1}^{M} \pi^{(m)} \mathbb{1}_{[t,\infty)}(y^{(m)}), \qquad \mathbb{1}_{[a,b]}(y) = \begin{cases} 1 & y \in [a,b] \\ 0 & \text{otherwise} \end{cases},$$

where $\pi^{(m)} \in [0,1]$ is the probability of drawing the sample $x^{(m)}$ for $m = 1, \ldots, M$ e.g., $\pi^{(m)} = \frac{1}{M}$ for MC methods. Unfortunately, the error in such sample-average estimates decays slowly as the number of samples is increased. For example, the error decays at a rate of $O(M^{-\frac{1}{2}})$ for MC methods. Consequently, the accurate sample average approximation of $\mathcal{R}(Y)$ is often impractical when the evaluations $f(x^{(m)})$ are expensive to generate. To circumvent this issue, we replace the expensive function $f(x)$ with a computationally tractable surrogate model $\theta_0 + g(x, \theta)$, where $\theta_0 \in \mathbb{R}$ and $\theta \in \Theta$ are parameters that must be estimated. To facilitate our results, we explicitly maintain the shift parameter θ_0. Throughout, we assume that Θ is a closed and nonempty subset of an Euclidean space. However, more general parameter sets can be incorporated with little difficulty.

Once θ_0 and θ have been determined, we can approximate $\mathcal{R}(Y)$ using $\mathcal{R}(\theta_0 + g(X, \theta))$. For example, given the samples of the surrogate model $\widehat{y}^{(n)} = \theta_0 + g(\widehat{x}^{(n)}, \theta)$, evaluated at the N input samples $\{\widehat{x}^{(1)}, \ldots, \widehat{x}^{(N)}\}$ with associated probabilities $\{\widehat{\pi}^{(1)}, \ldots, \widehat{\pi}^{(N)}\}$, we can estimate the probability that Y exceeds t by

$$\text{prob}(Y > t) \approx \sum_{n=1}^{N} \widehat{\pi}^{(n)} \mathbb{1}_{[t,\infty)}(\theta_0 + g(\widehat{x}^{(n)}, \theta)).$$

The error in the surrogate-based estimate depends on both the accuracy of the surrogate model and the number of evaluations of the surrogate model, N. Once constructed, the surrogate model can be evaluated at negligible cost and so the sampling error can be made arbitrarily small. The error in the estimate is therefore dominated by the accuracy of the surrogate model, which in turn depends on the number of evaluations M of the true function f.

It is well known that MC-based estimators are unbiased, but their error decays slowly with the sample size M. In contrast, surrogate-based estimates of risk are biased, but this bias decays quickly with the number of samples M. Surrogate-based estimates are often used as input to decisions regarding the control or response to high-consequence events, yet their bias is often not explicitly quantified. While the surrogate bias decays quickly, the surrogate model may underestimate the quantity $\mathcal{R}(Y)$ for any realization of the

training data. Again, consider Figure 2 and suppose Y_1 is the distribution of the model outcomes using the true data-generating process, while Y_2 is the distribution obtained from the surrogate model. By setting our risk measure to be AVaR, we observe that the surrogate model significantly underestimates the risk measure. This can lead to over-confidence in our estimate, which can lead to the system under investigation being incorrectly deemed safe or reliable. This incorrect assessment can have significant consequences, including severe financial losses and even loss of life.

A risk assessment can attempt to account for error by using a risk measure that is more risk-averse than necessary. For example, if one is interested in evaluating $\text{prob}(Y > t)$, they can instead compute $\text{prob}(\hat{Y} > t - \delta)$ using the surrogate model and some perturbation δ. Similarly, one could use $\text{AVaR}_{p+\delta}(\hat{Y})$ instead of $\text{AVaR}_p(Y)$. However, such an approach is heuristic and easily fallible. The appropriate size of δ, which ensures the risk measure is not under-estimated, is dependent on the error in the surrogate, yet the exact relationship is unclear. Even if an appropriate δ can be found, bounds on the error between the true risk and the approximated risk are not easily determined.

In the upcoming sections, we describe a mathematical framework producing surrogate models that are guaranteed to conservatively quantify risk with an accuracy that decreases as fast as the mean-squared error of the surrogate as the number of training data is increased.

3 Risk-Adapted Surrogate Modeling

The *risk quadrangle* provides a rigorous connection between stochastic optimization and statistical estimation that facilitates the translation of stakeholder preferences to mathematical definitions of measures of risk, deviation, error and regret. In this section, we leverage these connections to elicit risk measures tailored to the subjective values of decision makers and then to develop a strategy for constructing surrogates that accurately and conservatively estimate the chosen risk measure. We conclude the section by detailing error bounds on the estimates of risk obtained using our surrogate models and presenting numerical procedures for constructing such surrogate models.

3.1 Eliciting Risk Measures Using the Risk Quadrangle

Following [36], we discuss an approach for constructing risk measures based upon the decision maker's expected utility function using the risk quadrangle. Depicted in Figure 5, the risk quadrangle provides fundamental relationships that connect measures of risk, regret, deviation, and error. Table 1 details the risk, regret, deviation, and error measures of three common risk quadrangles.

Regret (or disutility) is intimately related to utility. Whereas utility functions quantify pleasure or gain from outcomes and typically seek high values, regret functions quantify displeasure or losses associated with high-values, which in risk assessment are typically associated with undesirable outcomes. Formally, a regret measure is a functional $\mathcal{V}$ that assigns to each random outcome a scalar value that quantifies the displeasure for uncertainty associated with possible realizations of the outcomes Y and satisfies $\mathcal{V}(Y) = -\mathcal{U}(-Y)$, where $\mathcal{U}$ is a measure of expected utility.

	Risk	$\mathcal{R}$	$\longleftrightarrow$	$\mathcal{D}$	Deviation	
Optimization		$\uparrow\downarrow$	$\mathcal{S}$	$\downarrow\uparrow$		Estimation
	Regret	$\mathcal{V}$	$\longleftrightarrow$	$\mathcal{E}$	Error	

$$\mathcal{R}(Y) = \mathbb{E}[Y] + \mathcal{D}(Y) \qquad \mathcal{D}(Y) = \mathcal{R}(Y) - \mathbb{E}[Y]$$

$$\mathcal{V}(Y) = \mathbb{E}[Y] + \mathcal{E}(Y) \qquad \mathcal{E}(Y) = \mathcal{V}(Y) - \mathbb{E}[Y]$$

$$\mathcal{R}(Y) = \inf_{t\in\mathbb{R}}\{t + \mathcal{V}(Y - t)\} \qquad \mathcal{D}(Y) = \inf_{t\in\mathbb{R}} \mathcal{E}(Y - t)$$

$$\mathcal{S}(Y) = \arg\min_{t\in\mathbb{R}}\{t + \mathcal{V}(Y - t)\} = \arg\min_{t\in\mathbb{R}} \mathcal{E}(Y - t)$$

Figure 5: The risk quadrangle provides rigorous mathematical connections between stochastic optimization and statistical estimation through these definitions of measures of risk, deviation, regret and error.

Table 1: Examples of different risk quadrangles

Safety Margins Risk Quadrangle	**p-th Quantile Risk Quadrangle**	**Entropic Risk Quadrangle**
$\mathcal{V}(Y) = \mathbb{E}[Y] + \lambda\|Y\|_2$	$\mathcal{V}(Y) = \frac{1}{1-p}\mathbb{E}[\max\{0, Y\}]$	$\mathcal{V}(Y) = \mathbb{E}[\exp(Y) - 1]$
$\mathcal{R}(Y) = \mathbb{E}[Y] + \lambda\mathbb{V}[Y]^{\frac{1}{2}}$	$\mathcal{R}(Y) = \mathrm{AVaR}_p[Y]$	$\mathcal{R}(Y) = \log\mathbb{E}[\exp(Y)]$
$\mathcal{E}(Y) = \lambda\|Y\|_2$	$\mathcal{E}(Y) = \mathbb{E}[\frac{p}{1-p}\max\{0, Y\} + \max\{0, -Y\}]$	$\mathcal{E}(Y) = \mathbb{E}[\exp(Y) - Y - 1]$
$\mathcal{D}(Y) = \lambda\mathbb{V}[Y]^{\frac{1}{2}}$	$\mathcal{D}(Y) = \mathrm{AVaR}_p[Y - \mathbb{E}[Y]]$	$\mathcal{D}(Y) = \log\mathbb{E}[\exp(Y - \mathbb{E}[Y])]$
$\mathcal{S}(Y) = \mathbb{E}[Y]$	$\mathcal{S}(Y) = q_p[Y]$	$\mathcal{S}(Y) = \log\mathbb{E}[Y]$

Many regret measures can be formulated as an expectation of a scalar regret function. That is, $\mathcal{V}(Y) = \mathbb{E}[v(Y)]$ where $v : \mathbb{R} \to \mathbb{R}$ is a regret function, i.e., $u(y) = -v(-y)$ is a utility function. The regret functions listed in Table 1 are depicted in Figure 6[1]. The quantile risk quadrangle uses the regret function $v(y) = \frac{1}{1-p}\max\{0, y\}$, which does not place any utility on negative values, but places linearly increasing displeasure on positive values. The rate of increase is controlled by the parameter p. Risk neutrality is recovered when $p = 0$. In particular, there is no difference in pleasure or displeasure between any values. The value $p = 1$ leads to the worst-case scenario risk measure, which places infinite displeasure on any positive value. In comparison, the entropic regret function $v(y) = \exp(y) - 1$ assigns some utility to negative values and exponentially increasing disutility to large positive values. For the entropic regret, we have that $u(y) = -v(-y)$ is the common exponential utility function.

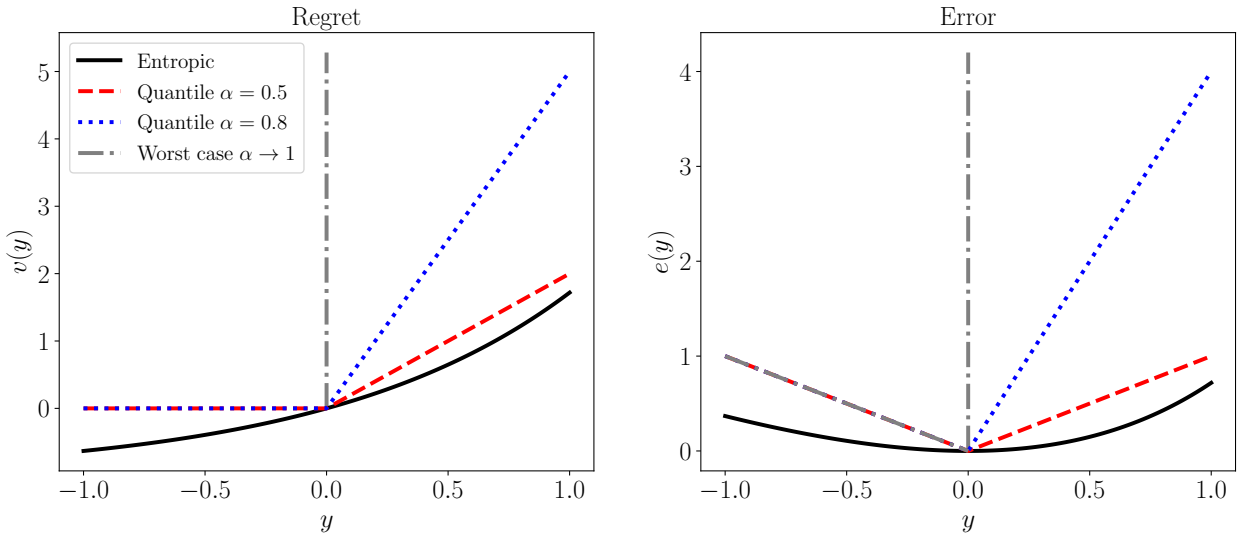


Figure 6: (left) Popular regret functions $v(y)$. (Right) Associated error functions $e(y)$ codified by the risk quadrangle.

3.2 Constructing Risk-Adapted Surrogate Models

The risk quadrangle can be used to formulate regression problems tailored to the notions of regret elicited from decision makers by exploiting the connections between regret and error shown in Figure 5 [37, 40]. Given an error measure that was generated from a regret measure, through the risk quadrangle, we can

[1]The regret measure of the mean risk quadrangle cannot be expressed as $\mathcal{V}(Y) = \mathbb{E}[v(y)]$ and so its regret function $v(y)$ is not shown in Figure 6.

calibrate the surrogate model parameters by solving the regression problem

$$\min_{\theta_0 \in \mathbb{R},\, \theta \in \Theta} \ \mathcal{E}(Y - \theta_0 - g(X, \theta)). \tag{5}$$

The L^2-norm, used with least-squares regression, is the error measure most commonly used to construct surrogates in engineering. However, alternative error measures can be used. Indeed, it is often advantageous to tailor the error measure to the risk preference of the stakeholder. When Y contains undesirable outcomes, producing surrogates that under-estimate the data Y, that is $\theta_0 + g(X, \theta) < Y$, is typically more concerning than producing surrogates that over-estimate the data.

As seen in Figure 5, the risk quadrangle provides a mechanism to formulate error measures tuned to the subjective values of the decision maker. Specifically, error measures use the stakeholder's notion of regret to measure how non-zero a random variable is through the relationship $\mathcal{E}(Y) = \mathcal{V}(Y) - \mathbb{E}[Y]$. The right panel of Figure 6 depicts the error functions $e(y) = v(y) - y$ associated with the regret functions in the left plot of the same figure. It is clear that different risk preferences lead to different error measures and thus penalize the residuals $Y - \theta_0 - g(X, \theta)$ differently, according to magnitude and sign. The worst-case regret is associated with an error function that assigns an infinitely large penalty to any positive value. Noting that we define residuals as the signed difference $Y - \theta_0 - g(X, \theta)$, we see that the worst-case error function attempts to produce a surrogate that is always larger than the data. The other error functions depicted also assign larger penalties to positive values, relative to negative values, but to differing degrees.

The deviation measure $\mathcal{D}$ associated with a regret measure $\mathcal{V}$ can also be used to construct surrogate models. The deviation $\mathcal{D}(Y)$ is a generalization of the standard deviation and measures the variability of a random variable Y. Like error measures, it too accounts for the subjective values of the decision maker and is often not symmetric, that is $\mathcal{D}(Y) \neq \mathcal{D}(-Y)$. Formally, the deviation is the smallest $\mathcal{E}$-weighted distance of Y to a constant. The value of t that attains this minimum is referred to as the statistic $\mathcal{S}(Y)$ of the risk quadrangle and can be thought of as the best scalar representative of the output Y according to the subjective beliefs of the decision maker.

The deviation, associated with an error measure, allows us to rewrite the regression problem (5) in the equivalent form

$$\theta \in \arg\min_{\theta' \in \Theta} \ \mathcal{D}(Y - g(X, \theta')) \quad \text{such that} \quad 0 \in \mathcal{S}(Y - \theta_0 - g(X, \theta)). \tag{6}$$

This form provides insight that is not immediately evident from (5). Specifically, the optimal model parameters θ minimize the deviation or uncertainty associated with the model. Moreover, the statistic of the residual must be zero. For least-squares regression, the optimal θ minimize the standard deviation of the model and the associated statistic is the average of the residual $\mathbb{E}[Y - \theta_0 - g(X, \theta)]$. Consequently, least-squares regression enforces that the residual is zero on average and thus is inappropriate when certain outcomes are highly undesirable. In contrast, the statistic of the quantile risk quadrangle is the p-th upper quantile of the residual, which can be more representative of rare undesirable events.

3.3 Conservatively Estimating Risk Using Surrogate Models

To construct risk-adapted surrogate models that conservatively estimate risk, in the sense that they try to avoid under-estimation, we use the following result, which we prove in Appendix A.

Proposition 3.1. *Suppose $\mathcal{R}$ is positively homogeneous and that there exists $\bar{\theta} \in \Theta$ for which $g(x, \bar{\theta}) = 0$ for all $x \in \Omega_X$. If $\theta^\star \in \Theta$ solves the deviation minimization problem*

$$\min_{\theta \in \Theta} \ \mathcal{D}(Y - g(X, \theta)), \tag{7}$$

and we set

$$\theta_0^\star := \mathcal{R}(Y - g(X, \theta^\star)), \tag{8}$$

then the following bounds hold

$$\theta_0^\star + \mathbb{E}[g(X, \theta^\star)] \leq \mathcal{R}(Y) \leq \theta_0^\star + \mathcal{R}(g(X, \theta^\star)) = \theta_0^\star + \mathbb{E}[g(X, \theta^\star)] + \mathcal{D}(g(X, \theta^\star)). \tag{9}$$

Remark 3.1 (Deviation Error Bounds). *An important consequence of* (9) *in Proposition 3.1 is the following error bound*

$$0 \leq \mathcal{R}(Y) - (\theta_0^\star + \mathbb{E}[g(X,\theta^\star)]) \leq \mathcal{D}(g(X,\theta^\star)), \tag{10}$$

which follows from the risk quadrangle relationships depicted in Figure 5. In other words, the average of the surrogate model $\theta_0^\star + \mathbb{E}[g(X,\theta^\star)]$ *under estimates the risk by at most the deviation* $\mathcal{D}(g(X,\theta^\star))$.

Remark 3.2 (Choice of Shift Parameter). *Although the deviation minimization problem* (7) *is intuitive—we seek parameters* $\theta^\star$ *that minimize the uncertainty associated with the residual—the numerical solution of* (7) *is often more challenging than the error minimization problem* (5). *Fortunately, the risk quadrangle relationships depicted in Figure 5 ensure that the optimal* $\theta^\star$ *in Proposition 3.1 also solves the error minimization problem* (5). *As a consequence, we can solve regression problem* (5) *and still ensure that the resulting surrogate model conservatively estimates* $\mathcal{R}(Y)$ *by computing the shift* (8) *from Proposition 3.1.*[2]

Proposition 3.1 and Remark 3.2 demonstrate that one can employ the following two-step procedure to conservatively estimate risk measures.

1. Construct a surrogate model, adapted to the decision maker's notion of regret, using (7) or (5).
2. Ensure that the surrogate model conservatively estimates $\mathcal{R}(Y)$ by introducing the shift (bias) $\theta_0^\star := \mathcal{R}(Y - g(\theta^\star))$.

In Section 3.4, we derive error estimates that show that the estimates of $\mathcal{R}(Y)$ obtained using this two-step procedure rapidly converge. Moreover, in some cases, the convergence rate is exponential.

3.4 Orthonormal Surrogate Models Using the Risk Quadrangle

The risk-quadrangle framework of Proposition 3.1 can be applied to most popular approximation strategies including polynomial chaos expansions (PCE) [48, 23], low-rank tensor decompositions [13, 17], and neural networks [18, 34]. In this section, we derive error bounds for estimates of $\mathcal{R}(Y)$ using surrogate models based on orthonormal basis functions, such as PCEs. For this, we assume that the probability spaces defined on Ω_X and Ω_Y are non-atomic. We investigate surrogate models with the form

$$f(x) \approx \theta_0 + \sum_{k=1}^{K} \psi_k(x)\theta_k := \theta_0 + g_K(x,\theta), \tag{11}$$

where K denotes the truncation order, $\Theta = \Theta_K := \mathbb{R}^K$, $\theta = [\theta_1, \ldots, \theta_K]^\top$, and $\{\psi_k\}_{k=1}^\infty$ is an orthonormal system. In particular, $\{\psi_k\}_{k=1}^\infty$ satisfies $\mathbb{E}[\psi_k(X)] = 0$ for all $k = 1, 2, \ldots$ and

$$\mathbb{E}[\psi_k(X)\psi_\ell(X)] = \begin{cases} 1 & \text{if } k = \ell \\ 0 & \text{otherwise} \end{cases} \qquad \text{for all } k,\, \ell = 1, 2, \ldots.$$

The Fourier coefficients associated with this basis are $\overline{\theta}_0 := \mathbb{E}[Y]$ and $\overline{\theta}_k := \mathbb{E}[\psi_k Y]$ for all $k = 1, 2, \ldots$. Provided Y has finite variance, the surrogate model (11) with Fourier coefficients $(\theta_0, \theta) = (\overline{\theta}_0, \overline{\theta})$, where $\overline{\theta} = [\overline{\theta}_1, \ldots, \overline{\theta}_K]$, satisfies

$$\int_{\Omega_X} |f(x) - \overline{\theta}_0 - g_K(x,\overline{\theta})|^2 \, P_X(\mathrm{d}x) \to 0 \qquad \text{as} \qquad K \to \infty.$$

In the following, we derive an upper bound on the error in the estimates of $\mathcal{R}(Y)$ obtained using orthonormal risk-adapted surrogate models and show that the bound decays as quickly as the truncation error of g_K, i.e.,

$$\mathbb{E}[(Y - \overline{\theta}_0 - g_K(X,\overline{\theta}))^2] = \tau_K := \sum_{k=K+1}^{\infty} \overline{\theta}_k^2,$$

[2]The optimal shift θ_0' from (5) satisfies $\theta_0' \in \mathcal{S}(Y - g(\theta^\star))$, but need not coincide with $\theta_0^\star$ unless $\mathcal{S}(\cdot) = \mathcal{R}(\cdot)$. This is the case, for example, with the entropic risk quadrangle for which $\mathcal{R}$ is the entropic risk measure, $\mathcal{R}(Z) = \ln \mathbb{E}[\exp(Z)]$, which is not positively homogeneous.

where τ_K is the order K tail energy associated with the approximation. With this goal, we first bound risk in terms of the symmetric deviation, which is useful for obtaining convergence rates of the error in estimates of risk measures. Specifically, we define the symmetric deviation measure generated from $\mathcal{D}$ by

$$\widehat{\mathcal{D}}(Y) := \tfrac{1}{2}(\mathcal{D}(Y) + \mathcal{D}(-Y)). \tag{12}$$

and the associated symmetric error measure projected from $\widehat{\mathcal{D}}$ by

$$\widehat{\mathcal{E}}(Y) := \tfrac{1}{2} \inf_{t\in\mathbb{R}} \{\mathcal{E}(Y - t) + \mathcal{E}(t - Y)\}.$$

Using these definitions, we have the following result, which is proved in Appendix A.

Proposition 3.2. *Let $\mathcal{R}$ be positively homogeneous. If $\theta^\star \in \Theta_K$ solves the symmetric deviation minimization problem*

$$\min_{\theta\in\Theta} \widehat{\mathcal{D}}(Y - g_K(X,\theta)) = \tfrac{1}{2} \min_{\theta\in\Theta} \{\mathcal{D}(Y - g_K(X,\theta)) + \mathcal{D}(g_K(X,\theta) - Y)\}, \tag{13}$$

and $\theta_0^\star = \mathcal{R}(Y - g_K(X,\theta^\star))$, then we have the bounds

$$0 \le \theta_0^\star + \mathcal{R}(g_K(X,\theta^\star)) - \mathcal{R}(Y) \le 2\widehat{\mathcal{D}}(Y - g_K(X,\theta^\star)). \tag{14}$$

The benefit of increasing the truncation order K to estimate the risk measure follows from (14) and is stated in the following proposition. See Appendix A for the proof of this proposition.

Proposition 3.3. *Suppose the error measure $\mathcal{E}$ is a positively homogeneous and that there exists $C > 0$ and $\alpha > 0$ such that*

$$\mathcal{E}(X) \le C\mathbb{E}[X^2]^\alpha. \tag{15}$$

Then, the surrogate model $g_K(X,\theta^\star)$ with coefficients $\theta^\star \in \Theta_K$ solve the symmetric deviation problem (13) satisfies the following bound

$$\widehat{\mathcal{D}}(Y - g_K(X,\theta^\star)) \le C\tau_K^\alpha.$$

When $\mathcal{E}$ is positively homogeneous, Propositions 3.2 and 3.3 ensure that the error in the risk evaluation computed using minimum symmetric deviation estimation satisfies

$$0 \le \theta_0^\star + \mathcal{R}(g_K(\theta^\star)) - \mathcal{R}(Y) \le 2C\tau_K^\alpha,$$

which decays to zero as K increases. In particular, in the setting of Proposition 3.3, we recover the spectral convergence of the L^2 estimator. However, the assumption that $\mathcal{E}$ satisfies (15) may not hold in general.

3.5 Computing Risk-Aware Surrogate Models

We now describe the numerical procedures used to construct surrogate models tailored to the popular risk measures mean-plus-standard-deviation and AVaR, which are associated with the safety margins and quantile risk quadrangles in Table 1. Note for both these risk quadrangles, the assumptions of Proposition 3.3 hold with $\alpha = \frac{1}{2}$ and $C = 1$ for mean-plus-standard-deviation risk, and $\alpha = \frac{1}{2}$ and $C = \frac{p}{1-p}$ for AVaR.

3.5.1 Estimating Mean-Plus-Standard-Deviation Risk Using Least-Squares Regression

According to the safety margins risk quadrangle in Table 1, we estimate the mean-plus-standard-deviation risk measure using least-squares regression

$$\min_{\theta_0\in\mathbb{R},\,\theta\in\Theta} \sum_{m=1}^{M} \pi^{(m)} \left[y^{(m)} - \theta_0 - g(x^{(m)},\theta)\right]^2. \tag{16}$$

For models that are linear in θ, such as PCE used in the numerical examples, this optimization problem can be easily solved using linear algebra (see, for example, the QR factorization). There are also numerous algorithms for solving unconstrained nonlinear least-squares problems when g is nonlinear in θ (see, for example, the Gauss-Newton and Levenberg-Marquardt methods).

After solving the least-squares problem (16), we compute $\theta_0^\star$ as in Proposition 3.1. Specifically, we compute the residuals

$$r^{(m)} = y^{(m)} - g(x^{(m)}, \theta^\star) \quad \text{for} \quad m = 1, \dots, M$$

and set

$$\theta_0^\star = \sum_{m=1}^{M} \pi^{(m)} r^{(m)} + \lambda \left[\sum_{m=1}^{M} \pi^{(m)} \left(r^{(m)} - \sum_{m=1}^{M} \pi^{(m)} r^{(m)} \right)^2 \right]^{\frac{1}{2}}.$$

Since $\mathcal{E}$ is positively homogeneous, the bound (9) ensures that the estimates of the risk measure obtained from the surrogate are conservative in the sense that

$$\theta_0^\star + \mathbb{E}[g(\theta^\star)] \le \mathbb{E}[Y] + \lambda \mathbb{V}[Y]^{\frac{1}{2}} \le \theta_0^\star + \mathbb{E}[g(\theta^\star)] + \lambda \mathbb{V}[g(\theta^\star)]^{\frac{1}{2}}.$$

3.5.2 Estimating AVaR Using Quantile Regression

To estimate AVaR, we use the quantile risk quadrangle in Table 1. The associated error measure can be minimized by solving the quantile regression problem

$$\min_{\theta_0 \in \mathbb{R},\, \theta \in \Theta_k} \sum_{m=1}^{M} \pi^{(m)} \gamma_p \left[y^{(m)} - \theta_0 - g(x^{(m)}, \theta) \right]$$
$$\text{where} \quad \gamma_p[u] = p \max(0, u) + (1 - p) \max(0, -u).$$

When g is linear in θ and $\Theta = \mathbb{R}^K$, the quantile regression problem can be efficiently solved as a linear program by splitting the residuals into positive and negative parts, i.e.,

$$r^{(m)} = y^{(m)} - \theta_0 - \sum_{k=1}^{K} \psi_k(x^{(m)}) \theta_k = u_m - v_m$$

Here, $u_m \ge 0$ is the positive part of $r^{(m)}$ and $v_m \ge 0$ is the corresponding negative part. This splitting yields the linear program

$$\begin{aligned} \min_{z,\, u,\, v} \quad & \begin{bmatrix} 0_K^\top & p\pi^\top & (1-p)\pi^\top \end{bmatrix} z \\ \text{subject to} \quad & Az = Y_M \\ & u \ge 0 \\ & v \ge 0 \end{aligned}$$

where $\pi = [\pi^{(1)}, \dots, \pi^{(M)}]^\top \in [0, 1]^M$, $Y_M := [y^{(1)}, \dots, y^{(M)}]^\top \in \mathbb{R}^M$,

$$A = \begin{bmatrix} \Psi & I_M & -I_M \end{bmatrix} \in \mathbb{R}^{M \times (K+2M)} \quad \text{and} \quad z = \begin{bmatrix} \theta^\top & u^\top & v^\top \end{bmatrix}^\top \in \mathbb{R}^{(K+2M)}.$$

Here, the entries of the matrix $\Psi \in \mathbb{R}^{M \times (K+1)}$ are given by $\Psi_{mk} = \psi_k(x^{(m)})$, for $k = 0, \dots, K$ and $m = 1, \dots, M$, where $\{\psi_k\}$ denote the factors defining g. This computational procedure is well known and is implemented in numerous software packages. Here, we use the implementation from PyApprox [22].

After solving the quantile regression problem, we compute the shift

$$\theta_0^\star = \mathrm{AVaR}_p \left[Y - g(X, \theta^\star) \right].$$

Specifically, assuming the residuals $R = [r^{(1)}, \dots, r^{(M)}]$ are sorted from smallest to largest then

$$\theta_0^\star = q_p(R) + \frac{1}{1-p} \sum_{m=k+1}^{M} \pi^{(m)} (r^{(m)} - q_p(R)),$$

where $q_p(R)$ is the p-th upper quantile and k is the index of the residual with a value greater than or equal to $q_p(R)$. Using this shift and noting that $\mathcal{E}$ is again positively homogeneous, we can use (9) to confirm that estimates of the risk measure obtained from the surrogate are conservative in the sense that

$$\theta_0^\star + \mathbb{E}[g(X, \theta^\star)] \le \mathrm{AVaR}_p [Y] \le \theta_0^\star + \mathrm{AVaR}_p [g(X, \theta^\star)].$$

4 Risk-Averse Surrogates Using Stochastic Dominance

The choice of risk measure should be tailored to the stakeholder's risk preference. Unfortunately, risk preferences are often difficult to elicit. In such situations, it may be more appropriate to construct approximations that enforce conservativeness with respect to a class of risk measures or even entire distributions. Stochastic orders can be used to achieve this goal. For this discussion, we restrict our attention to the first and second stochastic orders. However, one could consider higher stochastic orders, or more generally integral stochastic orders [43, Ch. 6.3.7], within the same framework.

With the goal of constructing accurate surrogate models that conservatively estimate the empirical CDF generated by the training data, we consider the estimation problem

$$\min_{\theta_0 \in \mathbb{R},\, \theta \in \Theta} \ \frac{1}{2}\mathbb{E}[|Y - \theta_0 - g(X,\theta)|^2] \tag{17a}$$

$$\text{subject to} \quad \theta_0 + g(\theta) \succeq_{(1)} Y. \tag{17b}$$

Given $\theta_0^\star \in \mathbb{R}$ and $\theta^\star \in \Theta$ that solve (17), we have the ordering

$$\text{prob}(\theta_0^\star + g(X,\theta^\star) > t) \geq \text{prob}(Y > t) \quad \forall\, t \in \mathbb{R}$$

and therefore $\theta_0^\star + g(X,\theta^\star)$ will produce conservative estimates of the probability that Y exceeds a threshold t. In addition, Theorem 6.50 in [43] ensures that

$$\mathcal{R}(\theta_0^\star + g(X,\theta^\star)) \geq \mathcal{R}(Y) \tag{18}$$

for all law-invariant (i.e., condition (R5)) risk measures $\mathcal{R}$ that satisfy the monotonicity condition (R2).

In a similar fashion, we can construct surrogate models that are second-order dominant using the estimation problem

$$\min_{\theta_0 \in \mathbb{R},\, \theta \in \Theta} \ \frac{1}{2}\mathbb{E}[|Y - \theta_0 - g(X,\theta)|^2] \tag{19a}$$

$$\text{subject to} \quad -Y \succeq_{(2)} -(\theta_0 + g(X,\theta)). \tag{19b}$$

See [12] for more information on stochastic-order constrained optimization problems. As discussed in Section 2.3, the disutility form of stochastic dominance, i.e., $-Y \succeq_{(2)} -(\theta_0 + g(X,\theta))$ (or equivalently $(\theta_0 + g(X,\theta) \succeq_{\text{icx}} Y$), is typically better suited for engineering applications, where large values are undesirable. This argument can be formalized by noting that if $\theta_0^\star \in \mathbb{R}$ and $\theta^\star \in \Theta$ solve (19), then

$$t + \frac{1}{1-p}\mathbb{E}[\max\{0, Y - t\}] \leq t + \frac{1}{1-p}\mathbb{E}[\max\{0, \theta_0^\star + g(X,\theta^\star) - t\}] \quad \forall\, t \in \mathbb{R},\ p \in (0,1),$$

which implies that

$$\text{AVaR}_p[Y] \leq \theta_0^\star + \text{AVaR}_p[g(X,\theta^\star)] \quad \forall\, p \in (0,1). \tag{20}$$

Therefore, (19) will produce the best approximation of Y with larger AVaR for all confidence levels p. This property is quite useful when considering law-invariant, coherent risk measures (i.e., risk measures that satisfy (R1), (R2), (R3), (R4), and (R5)). Note that if Y is exactly represented by the model $\theta_0 + g(X,\theta)$, then both (17) and (19) will produce parameter estimates that exactly reproduce Y. That is, the least-squares objective value will be zero and the constraints will be trivially satisfied since $Y = \theta_0^\star + g(X,\theta^\star)$.

4.1 Constructing Surrogate Models that Satisfy Stochastic Order Constraints

In this section, we discuss the optimization problems that must be solved to construct surrogate models that satisfy stochastic order constraints. Again, assuming that we only have access to a finite set of data Y of size M, we replace the first and second-order stochastic dominance constraints in (17) and (19) by

$$\sum_{m=1}^{M} \pi^{(m)} \mathbb{1}_{(-\infty,0]}(g(x^{(m)},\theta) - g(x^{(i)},\theta)) \leq \sum_{m=1}^{M} \pi^{(m)} \mathbb{1}_{(-\infty,0]}(y^{(m)} - \theta_0 - g(x^{(i)},\theta)) \tag{21}$$

and

$$\sum_{m=1}^{M} \pi^{(m)} \max\{0, g(x^{(m)},\theta) - g(x^{(i)},\theta)\} \leq \sum_{m=1}^{M} \pi^{(m)} \max\{0, y^{(m)} - \theta_0 - g(x^{(i)},\theta)\}$$

for $i = 1,\dots,M$, respectively. Note that since $y^{(m)}$ and $g(x^{(m)},\theta)$ only assume a discrete set of values, it is sufficient to enforce the dominance constraints on the discrete set of values $t \in \{\theta_0 + g(x^{(m)},\theta)\}_{m=1}^{M}$.

4.1.1 First-Order Stochastic Dominance

The numerical solution of the first-order stochastic dominance (FSD) constrained least-squares problem (17) can be obtained by solving

$$\min_{\theta_0\in\mathbb{R},\,\theta\in\Theta} \ \frac{1}{2}\sum_{m=1}^{M} \pi^{(m)}(y^{(m)} - \theta_0 - g^{(m)}(\theta))^2 \qquad \text{subject to} \qquad (21). \tag{FSD$_M$}$$

In practice, solving this optimization problem is challenging due to the discontinuity introduced by the Heaviside function $\mathbb{1}_{(-\infty,0]}$ in the first-order dominance constraints.

In the numerical examples that follow, we smooth the discontinuous constraints (21) (cf. [11]). Ideally, to ensure that (21) holds, we should replace $\mathbb{1}_{(-\infty,0]}$ with two smooth approximations h_1, $h_2 : \mathbb{R} \to [0,1]$ and solve

$$\min_{\theta_0\in\mathbb{R},\,\theta\in\Theta} \ \frac{1}{2}\sum_{m=1}^{M} \pi^{(m)}(y^{(m)} - \theta_0 - g(x^{(m)},\theta))^2 \tag{22a}$$

subject to

$$\sum_{m=1}^{M} \pi^{(m)} h_1(g^{(m)}(\theta) - g^{(i)}(\theta)) \leq \sum_{m=1}^{M} \pi^{(m)} h_2(y^{(m)} - \theta_0 - g^{(i)}(\theta)) \tag{22b}$$

$i = 1,\dots,M.$

If h_1 and h_2 satisfy

$$\mathbb{1}_{(-\infty,0]}(t) \leq h_1(t) \qquad \text{and} \qquad h_2(t) \leq \mathbb{1}_{(-\infty,0]}(t) \quad \forall\, t \in \mathbb{R}$$

and if $\theta_0^\star \in \mathbb{R}$ and $\theta^\star \in \Theta$ solve (22), then the original constraints (21) are also satisfied. Therefore, the resulting surrogate model of Y will be conservative with respect to the data $\{y^{(m)}\}$. Unfortunately, such h_1 and h_2 often leads to an inconsistent estimation problem. For example, suppose Y is exactly represented by the model $\theta_0 + g(X,\theta)$, then the smoothed constraint in (22) may not be satisfied at the optimal values of θ_0 and θ, i.e., the feasible set for (22) is smaller than the feasible set defined by (21). For our results, we set $h_1 = h_2$. This can theoretically produce a less conservative surrogate model. However, we found that the impact on the optimal solution is minimal.

Any number of smooth approximations of the Heaviside function can be used to solve (22). In this paper, we approximate the Heaviside function using the first derivative of the smooth max function

$$d(x;\delta,\epsilon) = x + \epsilon \log\left(-\frac{x+\delta}{\epsilon}\right). \tag{23}$$

The corresponding approximate right Heaviside function is given by

$$h_r(x;\delta_r,\epsilon) = \frac{\partial}{\partial x} d(x;\delta_r,\epsilon) = \frac{1}{1+\exp\left(-\frac{x+\delta_r}{\epsilon}\right)} \tag{24}$$

Here, δ_r and ϵ are tunable parameters. As $\epsilon \to 0$ the right Heaviside function converges to $\mathbb{1}_{[0,\infty)}(\cdot)$, so we use the left Heaviside function $h_l(x;\delta_l,\epsilon) = h_r(-x;-\delta_l,\epsilon)$. When $\delta_l = 0$, we have $h_l(0) = \frac{1}{2}\ \forall\epsilon$. This can lead to under-prediction of rare events, we found that setting $\delta_l = -\epsilon$, which results in $h_l(0) = \frac{1}{1+e^{-1}} \approx 0.73$ produced more conservative surrogates. This approximation of the Heaviside function is never exactly equal to the Heaviside function anywhere for non-zero ϵ. We investigated the performance of alternative smooth

approximations which are only inaccurate on a local interval of size ϵ, however we found solving (22) using these smoothers very challenging. The Jacobian of the constraints often has a low-rank because the gradient of the local Heaviside approximation are zero outside of the ϵ sized interval. This prevents the optimizer from choosing steps that will lead to optimal solutions satisfying the constraints. We also explored reformulating (FSD_M) as a mixed integer nonlinear program (see Appendix C.1), but had difficulty finding reasonable solutions.

Although the smoother we use in this paper performed better than the alternatives we investigated, using this smoother is still challenging especially as ϵ is decreased. Future work is needed to develop an more robust formulation for enforcing first-order stochastic dominance constraints, which is beyond the scope of this paper.

4.1.2 Second-Order Stochastic Dominance

Similar to the methods describe for FSD constraints, we solve a smooth approximation of (19) to ensure that AVaR is conservatively estimated for all values of $p \in (0, 1)$. In particular, we solve

$$\min_{\theta_0 \in \mathbb{R},\, \theta \in \Theta} \ \frac{1}{2} \sum_{m=1}^{M} \pi^{(m)} (y^{(m)} - \theta_0 - g^{(m)}(\theta))^2 \tag{25a}$$

subject to

$$\sum_{m=1}^{M} \pi^{(m)} d(g^{(m)}(\theta) - g^{(i)}(\theta)) \geq \sum_{m=1}^{M} \pi^{(m)} d(y^{(m)} - \theta_0 - g^{(i)}(\theta)) \tag{25b}$$

$i = 1, \ldots, M.$

Here, d is the smooth max function defined in (23).

See [12] for a discussion of sample-based approximation of optimization problems with SSD constraints. We note that (25) differs from the approximation problems considered in [12] due to the presence of the minus sign in (19b). Additionally, we note that (25) is not convex as a result of the constraint (25b), even if the model m is linear in θ. Again we investigated alternative formulations for enforcing SSD constraints (see Appendix C.1.2), but found that the smooth approximation presented here was the most effective for our examples.

4.2 Orthonormal Surrogate Models Using Stochastic Order Constraints

In this section, we compute an upper bound on the risk estimation error obtained from orthonormal surrogate models calibrated with stochastic dominance constrained least-squares estimation. Specifically, the following proposition bounds the surrogate model error when the stochastic dominance constraints are enforced on a compact subset of $\mathbb{R}$, $t \in S' \subset \mathbb{R}$. The proof of the proposition can be found in Appendix B. For this discussion, we employ the surrogate model g_K as described in Section 3.4.

Proposition 4.1. *Let $S' \subset \mathbb{R}$ be a compact set and $f : \mathbb{R}^D \to \mathbb{R}$ be a continuous function. Moreover, suppose that the unconstrained least-squares approximation $\bar{\theta}_0 + g_K(\cdot, \bar{\theta})$ satisfies the error bound*

$$\sup_{x \in f^{-1}(S')} |f(x) - (\bar{\theta}_0 + g_K(x, \bar{\theta}))| \leq \delta_K$$

with $\delta_K \downarrow 0$ as $K \to \infty$. Here, $f^{-1}(S') := \{x \in \mathbb{R}^D \mid f(x) \in S'\}$. Then, the stochastic dominance constrained least-squares estimators $(\theta_0^\star, \theta^\star) \in \mathbb{R} \times \Theta_K$ satisfy

$$\mathbb{E}[(Y - \theta_0^\star - g_K(X, \theta^\star))^2] \leq \tau_K + \delta_K^2$$

and

$$\|(\theta_0^\star, \theta^\star) - (\bar{\theta}_0, \bar{\theta})\|_2 \leq \delta_K.$$

Here, $\|\cdot\|_2$ denotes the Euclidean norm.

5 Numerical Examples

To confirm our theoretical results, we now present multiple numerical examples that highlight the properties and utility of risk-aware surrogates. In all examples, unless otherwise stated, we construct total-degree polynomial chaos expansions consisting of tensor product basis functions. Specifically, for each D-dimensional random vector X, we construct a univariate polynomial basis $\{\psi_{k,d}\}$, which is orthonormal with respect to the PDF of the variable and set

$$\psi_\lambda(x) = \prod_{d=1}^{D} \psi_{\lambda_d,d}(x_d),$$

where $\lambda = [\lambda_1, \ldots, \lambda_D]$ is a multi-index specifying the degree of the basis in each dimension. A total-degree polynomial expansion of degree Q is then given by

$$\theta_0 + \sum_{\lambda\in\Lambda} \psi_\lambda(x)\theta_\lambda, \qquad \Lambda = \left\{\lambda \mid 1 \le \sum_{d=1}^{D} \lambda_d \le Q\right\}, \qquad \text{with} \quad K = |\Lambda| = \binom{D+Q}{D} - 1, \tag{26}$$

where there is a one-to-one correspondence between the basis functions $\{\psi_\lambda\}_{\lambda\in\Lambda}$ and the basis functions $\{\psi_k\}_{k=1}^K$ used in (11).

For canonical random variables (e.g., normal or uniform), analytical expressions exist for the recursion coefficients needed to construct the univariate orthonormal basis functions. For other variables (e.g., log-normal or Gumbel), we compute the recursion coefficients using the numerical procedure presented in [28].

5.1 Risk Estimation for a Log-Normal Output

Let X be a multivariate normal random vector with mean and covariance μ and Σ, respectively. The PDF of X is given by

$$f_X(x) = \frac{1}{\sqrt{2\pi|\Sigma|^D}} \exp\left(-\frac{1}{2}(x-\mu)^\top\Sigma^{-1}(x-\mu)\right), \quad x \in \mathbb{R}^D.$$

We are interested in computing risk measures from outcomes generated by the model

$$Y = f(x) = \exp(a^\top x), \tag{27}$$

where $a = [a_1, \ldots, a_D]^\top$ and $a_d > 0$ for $d = 1, \ldots, D$. In this setting, the random outcome Y is a log-normal random variable for which we can compute various risk measures and dominance constraints analytically (see Appendix C). We construct PCE surrogates using training samples $\{x^{(m)}\}_{m=1}^M$, drawn randomly according to f_X, and set $\pi^{(m)} = \frac{1}{M}$. We the compute the PCE-based estimates of risk measures using 10^6 evaluations of the surrogate.

5.1.1 First-Order Stochastic Dominance

In Figure 7, we set $D = 1$, $a_1 = 1$ and use FSD to build a quadratic ($Q = 2$) PCE approximation (26) using 30 and 100 samples. The left plots compare the resulting approximation with the exact function f. There is obviously a large bias caused by enforcing the dominance constraints. However, the resulting surrogate is conservative with respect to the training data. The right plots show that the CDFs, computed using the FSD surrogate models, evaluated at the training data (labeled "FSD Train"), always lie to the right of the empirical estimate obtained using the same data (labeled "Train"). This means that the FSD CDF always over-estimates the probability in the tail as desired.

The right plots of Figure 7 also depict the exact CDF of Y (see Appendix C for its derivation) and the CDF computed using 10,000 Quasi Monte Carlo (QMC) samples of the FSD surrogate (labeled "FSD Validation"). The FSD CDF evaluated at the validation samples is still conservative with respect to the empirical estimate, but it is not conservative with respect to the exact CDF over the entire support of Y. Stochastic dominance, indeed any constraints, can only be enforced on the training data. When using 30 training samples, the FSD CDF does not conservatively predict the tail for CDF values greater than approximately 0.96. This can be improved by adding more samples. When using 100 training samples the CDF is conservative for CDF values up to 0.99.

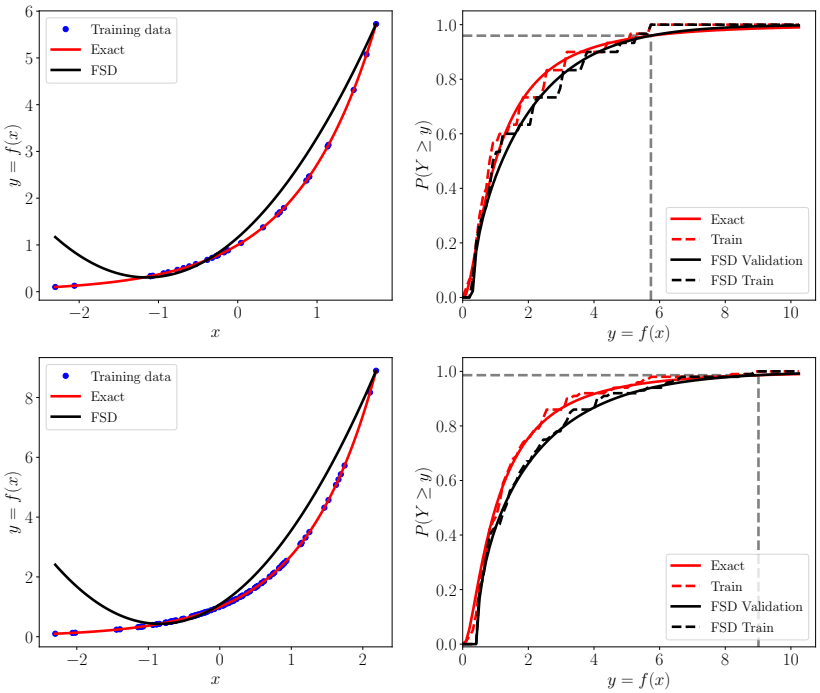


Figure 7: (Left column) The FSD approximation of f. (Right column) Comparison of the CDFs obtained analytically (Exact), empirically using the training data (Train), using the FSD approximation evaluated at the training data (FSD Train), and using the FSD approximation evaluated at the validation data (FSD Validation). (Top row) 30 training samples. (Bottom row) 100 training samples. The CDF estimated by evaluating the FSD surrogate at the validation sample is no longer conservative to the right of the vertical dashed gray lines. This occurs at (top) $y = 5.73$ and (bottom) $y = 9.01$, which correspond to CDF values of 0.96 and 0.99, respectively.

Figure 8 compares the accuracy of the unconstrained least squares and FSD surrogate-based estimates of probability of failure using 50 training samples with $D = 3$ and $a_1 = a_2 = a_3 = 1$. We define failure to occur when $Y > 8$ and set $\epsilon = 10^{-2}$ for the smooth max function in (23). The box plots depict the variation (from 100 realizations of the training data[3]) in the signed difference between the surrogate estimate of PoF and the empirical PoF. For this example, the least-squares surrogate model always underestimates the empirical PoF computed using the training data, whereas the FSD surrogate model always over-estimates the PoF. The FSD model is also conservative with respect to the exact PoF.

5.1.2 Second-Order Stochastic Dominance

SSD constraints are weaker than FSD constraints. In particular, FSD implies SSD, but the converse statement is not true. Thus, using SSD constraints do not guarantee that the resulting surrogate model will conservatively estimate the probability of failure. However, they are still more conservative than least-squares surrogate models and the risk-adapted surrogate models described in Section 3.3, which leverage the risk quadrangle.

In Figure 9, we set $D = 1$, $a_1 = 1$ and use SSD to build a Q-degree approximation with $Q = 2$ using 30 samples. The left plot compares the resulting approximation with the exact function f. Again, there is a

[3]93 of the 100 FSD estimation problems converged using a fixed set of optimization hyper-parameters. The failing problems can be solved by hand tuning the solver hyper-parameters. We list the failed runs here to demonstrate that solving the FSD constrained least squares is numerically challenging and future work is needed to develop optimization algorithms that can more robustly solve this problem without the need to fine tune optimization solver hyper-parameters. We note that robustness increases as ϵ increases.

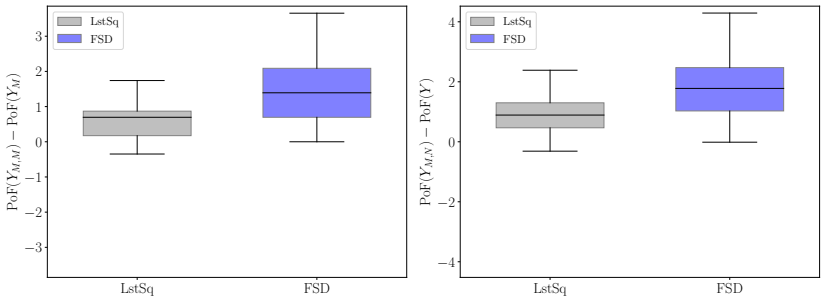


Figure 8: Comparison of the accuracy of least squares and FSD surrogate models when used to compute probability of failure (PoF) prob($Y > 8$). (Left) The signed difference between the surrogate estimate of PoF and the empirical PoF computed using the training data. (Right) The signed difference between the surrogate estimate of PoF and the empirical PoF computed using the validation data. In the y-axis labels Y_M denotes the output of the surrogate model and Y denotes the true output of the model. Here $\mathrm{PoF}(Y)$ denotes the exact PoF computed analytically, $\mathrm{PoF}(Y_M)$ the PoF computed using the M training samples, $\mathrm{PoF}(Y_{M,M})$ the PoF estimated using M evaluations of the surrogate at the training samples and $\mathrm{PoF}(Y_{M,N})$ the PoF estimated using $N = 10^6$ random evaluations of the surrogate.

large bias caused by enforcing the dominance constraints, which is needed to ensure that the surrogate model is conservative with respect to the training data. However, because the SSD constraints are weaker than FSD constraints, the SSD surrogate model has a better L^2 fit than the FSD surrogate model. Consequently, the SSD surrogate model becomes negative for some values of x unlike the FSD surrogate model (see Figure 7).

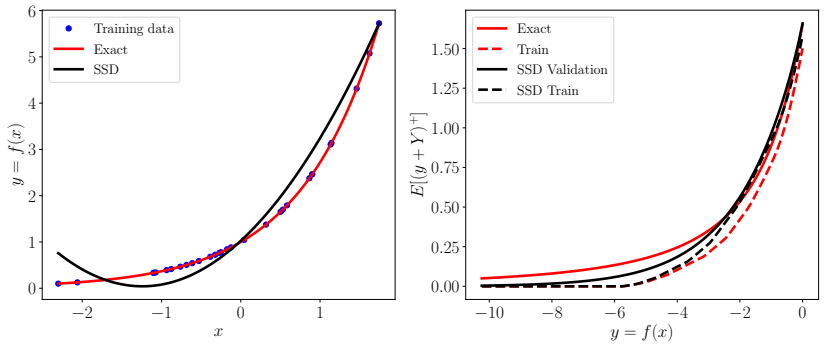


Figure 9: (Left) The SSD approximation of f. (Right) comparison of $\mathbb{E}[\max\{0, y+Y\}]$ obtained analytically (Exact), empirically using the training data (Train), using the SSD approximation evaluated at the training data (SSD Train), and using the SSD approximation evaluated at the validation data. (SSD Validation).

The right of Figure 9 shows that the values of $\mathbb{E}[\max\{0, y + Y\}]$, computed using the SSD surrogate model evaluated at the training data, always lie above the empirical estimate obtained using the same data. The estimates of $\mathbb{E}[\max\{0, y + Y\}]$ are also closer than the empirical estimate to the exact values when the surrogate is evaluated at the validation samples. The 30 samples do not contain enough extreme values to allow us to be conservative with respect to the exact function. Similar to FSD surrogate model, the conservativeness of the SSD surrogate model can be improved by using more data and increasing the polynomial degree.

Figure 10 compares the accuracy of least-squares and SSD surrogate-based estimates of $\mathrm{AVaR}_{0.8}(Y)$ using 100 training samples when $D = 2$ and $a_1 = a_2 = 1$. We set $\epsilon = 10^{-2}$ for the smooth positive-part

function in (23). The box plots depict the variation (from 100 realizations of the training data[4]) in the signed difference between the surrogate estimate of AVaR and the empirical AVaR. Again, least-squares regression does not conservatively estimate the empirical AVaR computed using the training data, but the SSD surrogate model does.

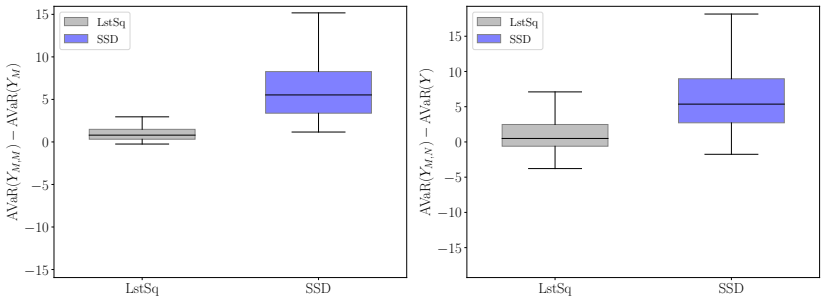


Figure 10: Comparison of the accuracy of least squares and SSD surrogates when used to compute the average value at risk (AVaR) with $p = 0.8$. (Left) The signed difference between the surrogate estimate of AVaR and the empirical AVaR computed using the training data. (Right) The signed difference between the surrogate estimate of AVaR and the empirical estimate computed using the validation data. In the y-axis labels Y_M denotes the output of the surrogate model and Y denotes the true output of the model. Here $\text{AVaR}(Y)$ denotes the exact AVaR computed analytically, $\text{AVaR}(Y_M)$ the AVaR computed using the M training samples, $\text{AVaR}(Y_{M,M})$ the AVaR estimated using M evaluations of the surrogate at the training samples and $\text{AVaR}(Y_{M,N})$ the AVaR estimated using $N = 10^6$ random evaluations of the surrogate.

5.1.3 Quantile Regression

FSD and SSD surrogate models provide conservative estimates of AVaR for a wide range of p. In this example, we compute a surrogate model that is conservative with respect to a single value of p using quantile regression. Here, we set $D = 1$, $a_1 = 1$. We use quantile regression to build a Q-degree approximation with $Q = 1$ using 30 training samples and compare the resulting approximation to other risk-adapted surrogate models (see the left image of Figure 11). The mean-squared error (MSE) evaluated using the training data increases with increasing conservatism. Specifically, the least-squares surrogate model has the smallest MSE followed by the quantile, SSD and FSD surrogate models in increasing order. Note that it is not just the coefficient θ_0 of the constant term in the polynomial expansion that is affected by the constraints. Higher-order coefficients are influenced as well. This is evident by the different slopes of the linear approximations.

The right plot of Figure 11 compares the accuracy of least-squares, quantile, SSD and FSD surrogate-based estimates of $\text{AVaR}_p(Y)$ for $p = 0.8$ using 30 training samples with $D = 4$. Again, we set $a_i = 1$ for $i = 1 \ldots, D$ and $\epsilon = 10^{-2}$. The box plots depict the variation (from 100 realizations of the training data) in the signed difference between the surrogate estimate of AVaR and the empirical AVaR.

Quantile regression performs as our theory suggests. It conservatively estimates AVaR when compared with the empirical AVaR, which uses the training data, at the value of $p = 0.8$ used by the quantile regression. The SSD and FSD surrogate models also conservatively estimate AVaR, but the bias they introduce is larger than that introduced by quantile regression. This highlights that if a stakeholder prefers a specific risk measure then the surrogate model should explicitly target that risk measure rather than simply using the more conservative surrogate methods such as FSD. Moreover, the regression algorithms based on the risk measures of the risk quadrangle are often computationally easier to solve than their stochastic dominance counterparts. For example, the quantile regression used here can be solved as a simple linear program (see Section 3.5).

[4]97 of the 100 SSD estimation problems converged. Similarly to FSD, solving the SSD constrained least squares estimation problem is challenging and future work is still needed to develop optimization algorithms that can more robustly solve this problem. Robustness increases as ϵ increases.

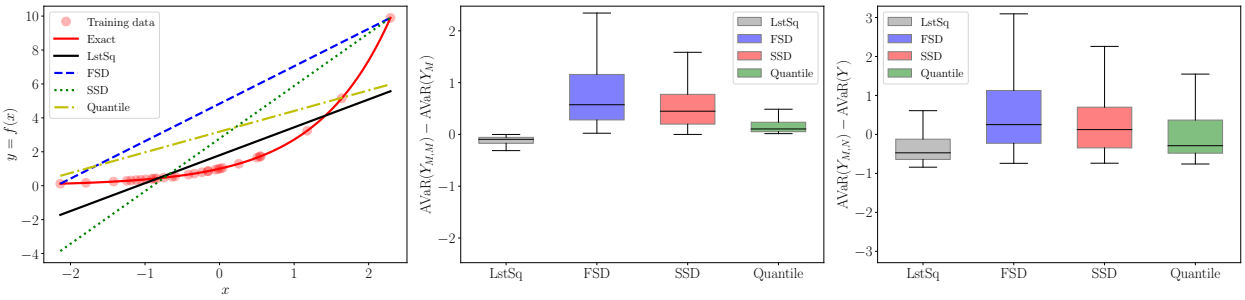


Figure 11: (Left) Linear polynomial approximations of f computed using various conservative constraints. (Middle) The signed difference between the surrogate estimates of AVaR and the empirical AVaR computed using the training data for $p = 0.8$. (Right) The signed difference between the surrogates estimate of AVaR and the empirical estimate computed using the validation data. Here $\text{AVaR}(Y)$ denotes the exact average value-at-risk computed analytically, $\text{AVaR}(Y_M)$ the AVaR computed using the M training samples, $\text{AVaR}(Y_{M,M})$ the AVaR estimated using M evaluations of the surrogate at the training samples and $\text{AVaR}(Y_{M,N})$ the AVaR estimated using $N = 10^6$ random evaluations of the surrogate.

5.1.4 Convergence

In the previous section, we demonstrated that, for a fixed polynomial degree, increased conservatism leads to a bias in the surrogate, e.g., a reduced ability to fit the training data. In this section, we will provide numerical evidence to support the theory presented in Sections 3.4 and 4.2, which states that error in estimates of the target risk measure decay at the same rate as the L^2 error decreases. Specifically, we show that as the degree of the PCE increases the estimates of the risk measure, associated with the analytical test function (27), decay exponentially fast.

In Figure 12, we set $a_i = \exp(-2(i-1))$ for $i = 1, 2$ and plot the error in estimates of $\text{AVaR}_{0.9}(Y)$ and $\text{prob}(Y > 3.6445)$ obtained using different types of risk-aware surrogate models as the polynomial degree increases. The number of samples was set to increase with degree. Specifically, the number of samples was set to 5 times the number of terms in the PCE. The threshold value used when computing the probability was chosen so that the $\text{prob}(Y > 3.6445) = 0.1$ and again we set the smoothing parameter of the FSD and SSD constrained least-squares to $\epsilon = 10^{-2}$. For both quantities, each surrogate model produces an estimate that converges exponentially fast. However, as shown in the previous examples the least-squares surrogate models often do not conservatively estimate the risk measure.

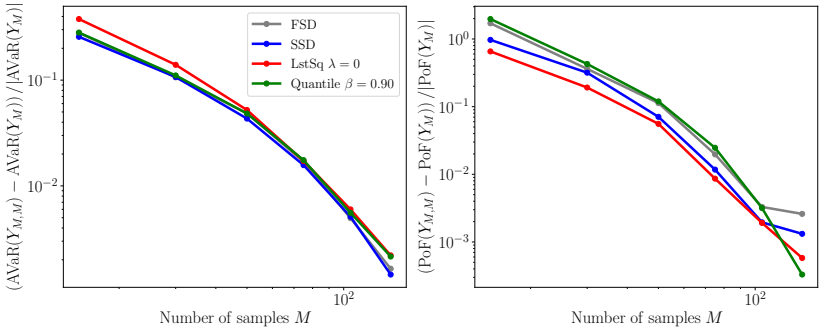


Figure 12: Median error (over 100 trials) in the estimate of (left) $\text{AVaR}_{0.9}$ and (right) $P[Y \geq 3.6445]$, obtained using accurate quadrature applied to a surrogate, as a function of polynomial degree. Here $\text{PoF}(Y)/\text{AVaR}(Y)$ denotes the exact risk measures computed analytically, and $\text{PoF}(Y_{M,N})/\text{AVaR}(Y_{M,N})$ the risk measures estimated using $N = 10^6$ random evaluations of the surrogate.

5.2 Risk Estimation for a Truss Structure

In this section, we investigate the utility of risk-aware surrogate models for a more realistic engineering benchmark. Specifically, we consider the two-dimensional truss structure depicted in Figure 1. The model of this structure is parameterized by 10 independent random variables representing uncertain loads and material properties. The Young's modulus parameters E_1 and E_2 (Pa) are log-normally distributed with mean 2.1×10^{11} and standard deviation 2.1×10^{10}. The cross-section of the horizontal truss A_1 (m^2) is log-normally distributed with mean 2.0×10^{-3} and standard deviation 2.0×10^{-4} and the cross-section of the diagonal truss A_2 (m^2) is log-normally distributed with mean 10^{-3} and standard deviation 10^{-4}. The loads $P_1, \ldots, P_6$ (N) satisfy a Gumbel distribution with mean 5.0×10^4 and standard deviation 7.5×10^3. This problem was first used in [6] and has been used to benchmark methods for estimating rare events in [29, 42]. The authors of these works provided a data-set containing the vertical displacement y at the mid-span for 10,000 random samples of the uncertain parameters. We used this data for training and validation of risk-aware surrogate models.

In the following, we approximate the model output with a quadratic PCE with 60 terms using 80 training samples. We use this surrogate model to estimate the probability of failure $\text{prob}(Y > 0.0995)$. We use $9,000$ validation samples, independent of the training samples, to estimate the reference failure probability $\text{prob}(Y > 0.0995) = 0.05$ and average value-at-risk $\text{AVaR}(Y) = 0.1055$.

The size of the PCE basis grows quickly with degree in 10 dimensions. Consequently, we use the adaptive basis selection algorithm from [21] to identify the best PCE truncation. This algorithm incrementally adds higher-degree terms to an initial basis. At each iteration, we solve a Least Absolute Shrinkage and Selection (LASSO) to identify a sparse basis. The procedure terminates when 10-fold cross validation errors increase. For this example, a basis with 60 terms was selected. This basis was used to construct different risk-aware surrogate models.

Table 2 displays the estimates of AVaR and the probability of failure for 5 different estimation methods. Entries annotated with a star indicate that the surrogate model conservatively estimates the target risk measure with respect to the training data. Entries annotated with a dagger indicate that the surrogate model conservatively estimates the risk measure with respect to the validation data. All methods conservatively estimate the probability of failure with respect to the training data, but only FSD over-estimates probability of failure with respect to the validation data. All methods except least-squares, over-estimate AVaR with respect to the training data. No surrogate over-estimates the reference value but the FSD surrogate produces the most accurate value.

Table 2: Estimates of risk for the truss structure. Estimates of risk greater than equal to the reference value computed using 9000 samples are marked by a dagger †. Estimates of risk greater than the empirical value computed using the training data are marked by an asterisk ⋆. The most accurate estimates of the reference risk are highlighted red.

Method	PoF (train)	PoF (validation)	AVaR (train)	AVaR (validation)
Empirical	$0.0625^\star$	0.0500	$0.1042^\star$	0.1055
FSD	$0.0625^\star$	$0.0501^\dagger$	$0.1043^\star$	0.1052
SSD	$0.0625^\star$	0.0471	$0.1042^\star$	0.1047
Quantile	$0.0625^\star$	0.0462	$0.1042^\star$	0.1046
LstSq	$0.0625^\star$	0.0463	0.1041	0.1046

5.3 Wing-weight model

In this section we show that our framework for constructing surrogates that conservatively estimate risk can be applied to approximation strategies other than traditional polynomial regression used in the previous section. Specifically we show that it can be coupled with dimension reduction to significantly reduce the cost of risk estimation in high-dimensions. With this goal consider the model of wing-weight first used in [1]

$$f(x) = 0.036\, S_w^{0.758} W_{fw}^{0.0035} \left(\frac{A}{\cos^2(\Lambda)}\right)^{0.6} q^{0.006} \lambda^{0.04} \left(\frac{100 t_c}{\cos(\Lambda)}\right)^{-0.3} (N_z W_{dg})^{0.49} + S_w W_p, \tag{28}$$

where $x = [S_w, W_{fw}, A, \Lambda, q, \lambda, t_c, N_z, W_{dg}, W_p]^\top$. The distribution of the random parameters are provided in Table 3. In the following we compute risk measures by finding important directions of the input space using active subspaces [10]. We compute the important directions by computing the eigenvalue decomposition of the average outer-products of gradients of the model f at a set of M samples, that is

$$W\Lambda W^\top = C = \frac{1}{M}\sum_{m=1}^{M} \nabla_x f(x^{(m)}) \nabla_x f(x^{(m)})^\top$$

Figure 13 plots the projection of $M = 30$ evaluations of f onto the dominant one-dimensional subspace, defined by the first eigenvector of W. It is evident that the wing-weight model can be reasonably approximated by a 1D function. However no 1D approximation can ever be exact due to small variations in the model response in the inactive directions. However we can still accurately and conservatively approximate risk using the framework presented in this paper. In Figure 13 we compare FSD, SSD and quantile ($p = 0.9$) surrogates computed using the 30 training samples depicted. Each surrogate conservatively and accurately approximates AVaR for $p = 0.9$. Specifically, using $N = 10^6$ evaluations of the true model Y and surrogate Y_M, the signed relative difference $(\text{AVaR}_{0.9}(Y_{M,N}) - \text{AVaR}_{0.9}(Y))/\text{AVaR}_{0.9}(Y)$ is 0.029, 0.008, and 0.01 respectively.

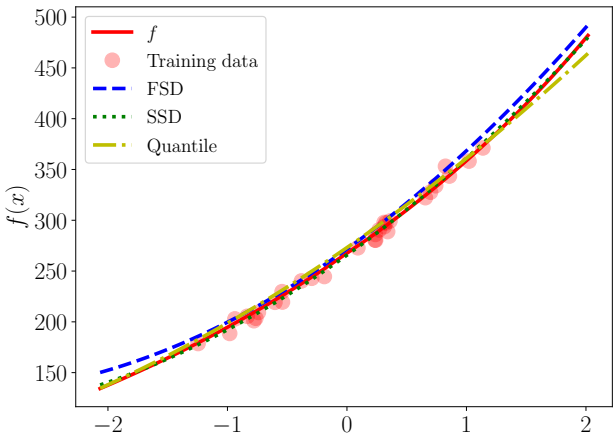


Figure 13: Various surrogate approximations of the wing-weight model (28).

Table 3: Random variables of the wing weight model (28)

Variable	Symbol	Distribution
Wing area	S_w	$\mathcal{U}(150, 200)$
Weight of fuel in the wing	W_{fw}	$\mathcal{U}(220, 300)$
Aspect ratio	A	$\mathcal{U}(6, 10)$
Quarter-chord sweep	Λ	$\mathcal{U}(-5\pi/9, 5\pi/9)$
Dynamic pressure at cruise	q	$\mathcal{U}(16, 45)$
Taper ratio	λ	$\mathcal{U}(0.5, 1)$
Aerofoil thickness to chord ratio	t_c	$\mathcal{U}(0.08, 0.18)$
Ultimate load factor	N_z	$\mathcal{U}(2.5, 6)$
Flight design gross weight	W_{dg}	$\mathcal{U}(1700, 2500)$
Paint weight	W_p	$\mathcal{U}(0.025, 0.08)$

6 Conclusions

We developed techniques for computing risk measures using a fixed data set. Such situations arise when archived data is available and there is no recourse to further integrate the data source. We introduced several estimation approaches that build surrogate models that conservatively (over-estimate) risk. Most uncertainty quantification literature focuses on constructing surrogate models that efficiently estimate the mean of a random variable Y. This approach is the least conservative of all methods considered in this paper. Least squares surrogates often underestimates risk measures such as the average value-at-risk as well as the probability of failure.

To overcome the limitations of least squares regression, we used the risk quadrangle to construct surrogates that conservatively estimate a risk measure associated with the subjective utility/regret preferences of a stakeholder. Stochastic dominance was used to enforce conservativeness with respect to a set of risk measures when a single utility preference cannot be elicited or agreed upon. For both approaches, we derived theoretical upper bounds on the error in estimates of risk measures obtained from orthonormal surrogate models. The error was shown to converge at the same rate as the L^2 error in the surrogate model when using polynomial chaos expansions. These results were supported by numerical examples that showed that risk-aware surrogate models do indeed over-estimate risk, while converging at the expected rate.

Risk-aware surrogate models are most useful in the limited data regime. In this setting, cross validation is challenging and the estimates from approximations, such as Gaussian processes, that are endowed with probabilistic estimates of error are heavily influenced by the prior assumptions of uncertainty in the approximation.

Although not considered in this paper, the approximation strategy presented can also be used within an adaptive sampling strategy to generate samples that can improve a surrogate models estimate of risk. Existing adaptive strategies build a surrogate model from available training data and use properties of the surrogate model to select the next set of data. Mean-squared loss functions that target accuracy in estimates of mean behavior are typically employed, but as we showed these loss functions can easily be replaced by other error functions that target the risk measure of interest to the stakeholder. We will explore such adaptive strategies in future work.

Acknowledgments

This work was sponsored by the Sandia National Laboratories Laboratory Directed Research Development (LDRD) program and the US Department of Energy, Office of Advanced Scientific Computing Research. Sandia National Laboratories is a multimission laboratory managed and operated by National Technology and Engineering Solutions of Sandia, LLC., a wholly owned subsidiary of Honeywell International, Inc., for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-NA0003525. This paper describes objective technical results and analysis. Any subjective views or opinions that might be expressed in the paper do not necessarily represent the views of the U.S. Department of Energy or the United States Government.

Appendix A Proof of Propositions 3.1 and 3.2

In this section, we provide proofs of Proposition 3.1 and 3.2 and provide some additional remarks regarding how some of the assumptions of this proposition can be relaxed.

Proof[Proof of Proposition 3.1] As a consequence of the relationships depicted in Figure 5, any risk measure $\mathcal{R}$ generated by the risk quadrangle is translation equivariant in the sense that

$$\mathcal{R}(Z + C) = \mathcal{R}(Z) + C \quad \forall\, C \in \mathbb{R} \tag{R1}$$

for all random variables Z. The translation equivariance of $\mathcal{R}$ ensures that $\mathcal{R}(\theta_0 + g(X, \theta)) = \theta_0 + \mathcal{R}(g(X, \theta))$ since the shift parameter θ_0 is deterministic. If $\mathcal{R}$ is positively homogeneous, i.e.,

$$\mathcal{R}(CZ) = C\mathcal{R}(Z) \quad \forall\, C \geq 0 \tag{R2}$$

for all random variables Z, then so are $\mathcal{D}$, $\mathcal{V}$ and $\mathcal{E}$. Moreover, positive homogeneity combined with convexity ensures that $\mathcal{R}$ is in fact subadditive and consequently,

$$\begin{aligned} \mathcal{R}(Y) &= \mathcal{R}(\theta_0 + g(X, \theta) + (Y - \theta_0 - g(X, \theta))) \\ &\leq \mathcal{R}(\theta_0 + g(X, \theta)) + \mathcal{R}(Y - \theta_0 - g(X, \theta)). \end{aligned} \tag{29}$$

By choosing $\theta_0 = \mathcal{R}(Y - g(X, \theta))$, a straightforward consequence of (29) and the translation equivariance of $\mathcal{R}$ is that $\mathcal{R}(\theta_0 + g(X, \theta)) \geq \mathcal{R}(Y)$. Moreover, this choice of θ_0 ensures that the expectation of the model satisfies

$$\mathbb{E}[\theta_0 + g(X, \theta)] = \mathbb{E}[Y] + \mathcal{D}(Y - g(X, \theta)). \tag{30}$$

The upper bound in (9) follows from (29) and our choice of $\theta_0^\star$. Specifically, we have that

$$\mathcal{R}(Y) \leq \mathcal{R}(\theta_0^\star + g(X, \theta^\star)) + \mathcal{R}(Y - \theta_0^\star - g(X, \theta^\star) = \theta_0^\star + \mathcal{R}(g(X, \theta^\star)),$$

where we use the facts that $\mathcal{R}(\theta_0^\star + g(X, \theta^\star)) = \theta_0^\star + \mathcal{R}(g(X, \theta^\star))$ and $\mathcal{R}(Y - \theta_0^\star - g(X, \theta^\star)) = 0$, which hold due to the translation equivariance of $\mathcal{R}$. On the other hand, the assumption that there exists $\bar{\theta}$ for which $g(x, \bar{\theta}) = 0$ for all $x \in \Omega_X$ and the optimality of $\theta^\star$ ensure that

$$\mathcal{R}(Y) = \mathbb{E}[Y] + \mathcal{D}(Y) = \mathbb{E}[Y] + \mathcal{D}(Y - g(X, \bar{\theta})) \geq \mathbb{E}[Y] + \mathcal{D}(Y - g(X, \theta^\star)).$$

Then (30) and the translation equivariance of $\mathcal{R}$ yield the lower bound in (9). □

The proof of Proposition 3.2 uses similar arguments as those used in the proof of proving Proposition 3.1.

Proof[Proof of Proposition 3.2] The lower bound in (14) follows from (29) and our choice of $\theta_0^\star$. Specifically, (29) implies

$$-\mathcal{R}(Y) \geq -\mathcal{R}(\theta_0^\star + g_K(X,\theta)) - \mathcal{R}(Y - \theta_0^\star - g_K(X,\theta)),$$

thus

$$\begin{aligned}\theta_0^\star + &\mathcal{R}(g_K(X,\theta^\star)) - \mathcal{R}(Y) \\ &\geq \theta_0^\star + \mathcal{R}(g_K(X,\theta^\star)) - \mathcal{R}(\theta_0^\star + g_K(X,\theta^\star)) - \mathcal{R}(Y - \theta_0^\star - g_K(X,\theta^\star)) \\ &= 0,\end{aligned}$$

where again we have used the fact that $\mathcal{R}(\theta_0^\star + g_K(X,\theta^\star)) = \theta_0^\star + \mathcal{R}(g_K(X,\theta^\star))$ and $\mathcal{R}(Y - \theta_0^\star - g_K(X,\theta^\star)) = 0$. On the other hand, by the definition of $\theta_0^\star$, we have that

$$\theta_0^\star + \mathcal{R}(g_K(X,\theta^\star)) - \mathcal{R}(Y) = \mathcal{D}(Y - g_K(X,\theta^\star)) + \mathcal{D}(g_K(X,\theta^\star)) - \mathcal{D}(Y).$$

Since $\mathcal{D}$ is convex and positively homogeneous, it is subadditive and we have that

$$\mathcal{D}(g_K(X,\theta^\star)) = \mathcal{D}((g_K(X,\theta^\star) - Y) + Y) \leq \mathcal{D}(g_K(X,\theta^\star) - Y) + \mathcal{D}(Y)$$

and the upper bound in (14) follows from the definition of symmetric deviation (12). □

Remark Appendix A.1 (Positive Homogeneity)**.** *If $\mathcal{R}$ is not positively homogeneous, we can modify* (29) *as*

$$\mathcal{R}(Y) \leq \mathcal{R}_\lambda(\theta_0 + g(X,\theta)) + \mathcal{R}_{1-\lambda}(Y - \theta_0 - g(X,\theta))$$

for any $\lambda \in (0,1)$, where $\mathcal{R}_t(X) := t\mathcal{R}(X/t)$ for $t > 0$. Note that $\mathcal{R}_t$ is associated with a risk quadrangle with components $\mathcal{E}_t$, $\mathcal{D}_t$ and $\mathcal{V}_t$ defined analogously. In this setting, the minimum deviation upper bound in (9) *holds with $\mathcal{R}$ replaced by $\mathcal{R}_{1-\lambda}$. However, the lower bound in* (9) *may no longer hold. To circumvent this, one may consider replacing the error measure used in* (5) *with either*

$$\mathcal{E}^1(X) := \inf_{0 \leq t \leq 1} \mathcal{E}_t(X) \quad \text{or} \quad \mathcal{E}^2(X) := \inf_{t \geq 0} \mathcal{E}_t(X).$$

Unfortunately, one can show that the first error measure is equal to $\mathcal{E}$ and the second is identically zero for many practical $\mathcal{E}$. To elaborate, for $\mathcal{E}^1$, one can show that its Fenchel conjugate is

$$(\mathcal{E}^1)^*(\theta) = \max\{0, \mathcal{E}^*(\theta)\}.$$

It follows from the Fenchel-Young inequality and the fact that $\mathcal{E}(0) = 0$ that $\mathcal{E}^*(\theta) \geq 0$ for all $\theta \in \mathcal{X}^*$. Therefore, $(\mathcal{E}^1)^* = \mathcal{E}^*$. Since $\mathcal{E} \in \Gamma_0(\mathcal{X})$, we have that

$$\mathcal{E}(X) = \mathcal{E}^{**}(X) = (\mathcal{E}^1)^{**}(X) \leq \mathcal{E}^1(X) \leq \mathcal{E}(X)$$

and hence $\mathcal{E}^1 = \mathcal{E}$. As for $\mathcal{E}^2$, its second Fenchel conjugate is the indicator function of the set

$$\{\theta \in \mathcal{X}^* \,|\, \mathcal{E}^*(\theta) \leq 0\} = \{\theta \in \mathcal{X}^* \,|\, \mathbb{E}[\theta X] \leq \mathcal{E}(X) \; \forall X \in \mathcal{X}\} = \partial\mathcal{E}(0).$$

This provides the general bounds

$$\sup_{\theta \in \partial\mathcal{E}(0)} \mathbb{E}[\theta X] = (\mathcal{E}^2)^{**}(X) \leq \mathcal{E}^2(X) \leq \mathcal{E}(X) \quad \forall X \in \mathcal{X}.$$

Here, the first inequality holds with equality if and only if $\mathcal{E}^2$ is lower semicontinuous. In this case, $\mathcal{E}^2$ is a regular measure of error as long as $\sup_{\theta \in \partial\mathcal{E}(0)} \mathbb{E}[\theta X] > 0$ for all $X \in \mathcal{X} \setminus \{0\}$. This is the case if $\partial\mathcal{E}(0) \neq \{0\}$. Note that $0 \in \partial\mathcal{E}(0)$ since $\mathcal{E}(0) = 0$ and $\mathcal{E}(X) > 0$ for all $X \neq 0$. As a simple example, consider the mean squared error $\mathcal{E}(X) = \frac{1}{2}\mathbb{E}[X^2]$. It follows that

$$\mathcal{E}_t(X) = \tfrac{1}{2t}\mathbb{E}[X^2],$$

which is strictly decreasing with respect to t and hence $\mathcal{E}^1 = \mathcal{E}$ and $\mathcal{E}^2 \equiv 0$.

Remark Appendix A.2 (Exact Reproducibility)**.** *In the case that Y is exactly represented by the model $\theta_0 + g(X, \theta)$, the error minimization problem (symmetric or otherwise) will return $\theta_0' \in \mathbb{R}$ and $\theta^\star \in \Theta$ that produce zero error, i.e., $Y = \theta_0' + g(X, \theta^\star)$. Likewise, $\theta^\star$ will produce zero deviation $\mathcal{D}(Y - g(X, \theta^\star)) = 0$. Therefore,*

$$\theta_0^\star + \mathbb{E}[g(X, \theta^\star)] = \mathbb{E}[Y] = \theta_0' + \mathbb{E}[g(X, \theta^\star)] \quad \Longrightarrow \quad \theta_0^\star = \theta_0'$$

and hence the above estimation procedures produce $\theta_0^\star$ and $\theta^\star$ that exactly reproduce Y.

This brings us to the proof of Proposition 3.3.

Proof[Proof of Proposition 3.3] We denote the first K Fourier coefficients by $\bar{\theta}_{1:K}$. The optimality of $\theta^\star$ ensures that

$$\widehat{\mathcal{D}}(Y - g_K(X, \theta^\star)) \leq \widehat{\mathcal{D}}(Y - g_K(X, \bar{\theta}_{1:K})) \leq \widehat{\mathcal{E}}(Y - \bar{\theta}_0 - g_K(X, \bar{\theta}_{1:K})).$$

By the definition of the symmetric error, we have that

$$\begin{aligned}&\widehat{\mathcal{E}}(Y - \bar{\theta}_0 - g_K(X, \bar{\theta}_{1:K})) \\ &\qquad = \tfrac{1}{2} \min_{t \in \mathbb{R}} \{\mathcal{E}((Y - \bar{\theta}_0 - g_K(X, \bar{\theta}_{1:K})) - t) + \mathcal{E}(t - (Y - \bar{\theta}_0 - g_K(X, \bar{\theta}_{1:K})))\}\end{aligned}$$

and therefore, (15) ensures that

$$\widehat{\mathcal{E}}(Y - \bar{\theta}_0 - g_K(X, \bar{\theta}_{1:K})) \leq C \min_{t \in \mathbb{R}} \mathbb{E}[((Y - \bar{\theta}_0 - g_K(X, \bar{\theta}_{1:K})) - t)^2]^\alpha.$$

The minimizing t is given by $t^\star = \mathbb{E}[Y - \bar{\theta}_0 - g_K(X, \bar{\theta}_{1:K})] = \mathbb{E}[Y] - \bar{\theta}_0 = 0$. Substituting $t^\star = 0$ in to the upper bound gives

$$\widehat{\mathcal{E}}(Y - \bar{\theta}_0 - g_K(X, \bar{\theta}_{1:K})) \leq C\mathbb{E}[(Y - \mathbb{E}[Y] - g_K(X, \bar{\theta}_{1:K}))^2]^\alpha \leq C\tau_k^\alpha$$

as was to be proved. □

Appendix B Proof of Proposition 4.1

The proof of Proposition 4.1 proceeds as follows.

Proof[Proof of Proposition 4.1] Note that we have the following inequalities:

$$\begin{aligned}\tau_K = \mathbb{E}[(Y - \bar{\theta}_0 - g_K(X, \bar{\theta}_{1:K}))^2] &\leq \mathbb{E}[(Y - \theta_0^\star - g_K(X, \theta^\star))^2] \\ &\leq \mathbb{E}[(Y - \theta_0 - g_K(X, \theta))^2],\end{aligned} \tag{31}$$

for all $(\theta_0, \theta) \in \mathbb{R} \times \Theta_K$ satisfying the stochastic dominance constraints for $t \in S'$. We now find a $\theta_0 \in \mathbb{R}$ such that $(\theta_0, \bar{\theta}_{1:K})$ is a feasible point. To that end, we recall that

$$\mathbb{E}[\max\{0, Y - t\}] = \mathbb{E}[\max\{0, \bar{\theta}_0 + g_K(X, \bar{\theta}_{1:K}) + r_K(X) - t\}],$$

where r_K denotes the remainder associated with the K^{th} approximation, i.e., $r_K(x) = f(x) - \bar{\theta}_0 - g_K(x, \bar{\theta}_{1:K})$. Therefore, by choosing

$$\theta_0 = \bar{\theta}_0 + \sup_{x \in f^{-1}(S')} r_K(x),$$

we have that $(\theta_0, \bar{\theta}_{1:K})$ satisfies the dominance constraints and thus (31) ensures that

$$\begin{aligned}\tau_k \leq \mathbb{E}[(Y - \theta_0^\star - g_K(X, \theta^\star))^2] &\leq \mathbb{E}[(Y - \theta_0 - g_K(X, \bar{\theta}_{1:K}))^2] \\ &= \tau_k + \left(\sup_{x \in f^{-1}(S')} r_K(x)\right)^2 \leq \tau_K + \delta_K^2.\end{aligned}$$

We note here that since f is continuous and S' is compact, then $f^{-1}(S')$ is compact and the supremum of r_k over $f^{-1}(S')$ exists and is finite. Expanding the quadratic objective function and utilizing the orthonormality of ψ_ℓ yields the desired result. □

Appendix C Tail statistics of log-normal variables

In this section, we derive the exact risk measures and dominance quantities of a log-normally distributed random variable, which were used in the numerical examples. With this goal, let Z be an uncorrelated multivariate normally distributed random variable with mean zero and unit variance. Then given a vector $a = [a_1, \ldots, a_D]^\top$, $a_d > 0$ for $d = 1, \ldots, D$ the variable $X = a^t Z$ is a univariate normally distributed random variable with mean $\mu = 0$ and variance $\sigma = \left(\sum_{d=1}^D a_d^2\right)^{\frac{1}{2}}$. The PDF of X is

$$f_X(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right), \quad x \in (\infty, \infty)$$

and the output of the function,

$$Y = f(X) = \exp(X),$$

is a log-normally distributed random variable. The log-normal distribution is well known and can be used to compute the following quantities.

Value-at-risk. The VaR for a log-normal random variable is

$$\mathrm{VaR}_p(Y) = F_Y^{-1}(p) = \exp(F_X^{-1}(p)) = \exp(\Phi_{\mu,\sigma^2}^{-1}(p)) = \exp(\mu + \sigma\sqrt{2}\mathrm{Erf}^{-1}(2p-1)), \tag{32}$$

for $p \in (0,1)$, where $\Phi_{\mu,\sigma^2}^{-1}(p) = \mu + \sigma\sqrt{2}\mathrm{Erf}^{-1}(2p-1)$ is the inverse CDF of a Gaussian variable with mean and variance μ and σ^2.

Average value-at-risk. The AVaR for a log-normal random variable is

$$\mathrm{AVaR}_p(Y) = \frac{1}{1-p}\exp\left(\mu + \frac{\sigma^2}{2}\right)\Phi\left(\frac{\mu + \sigma^2 - \log(\mathrm{VaR}_p(Y))}{\sigma}\right) \tag{33}$$

Second-order stochastic dominance. The upper bound for the SSD constraint with $-Y$ is

$$\begin{aligned}&\mathbb{E}\left[\max\{0, t+Y\}\right]\\ &= \left(1 - \Phi\left(\frac{\log(-t)-\mu}{\sigma}\right)\right)\left(t + \frac{1}{1-\Phi\left(\frac{\log(-t)-\mu}{\sigma}\right)}\exp(\mu + \frac{\sigma^2}{2})\Phi\left(\frac{\mu+\sigma^2-\log(-t)}{\sigma}\right)\right)\end{aligned} \tag{34}$$

Appendix C.1 Computational alternatives for enforcing stochastic dominance

In this section we detail alternative optimization problems for enforcing FSD and SSD constraints.

Appendix C.1.1 First-order stochastic dominance

The estimation problem (FSD_M) can be equivalently reformulated as the mixed integer nonlinear program (MINLP)

$$\min_{\substack{\theta_0\in\mathbb{R},\,\theta\in\Theta\\ b_{im}\in\mathbb{R},\,c_{im}\in\mathbb{R}}} \quad \frac{1}{2}\sum_{m=1}^M \pi^{(m)}(y^{(m)} - \theta_0 - g_K(x^{(m)};\theta))^2 \tag{35a}$$

$$\text{subject to} \quad \sum_{m=1}^M \pi^{(m)} b_{im} \le \sum_{m=1}^M \pi^{(m)} c_{im} \qquad i = 1, \ldots, M \tag{35b}$$

$$(1 - b_{im})(g_K(x^{(m)};\theta) - g_K(x^{(i)};\theta)) \ge 0 \qquad i,\, m = 1, \ldots, M \tag{35c}$$

$$c_{im}(y^{(m)} - \theta_0 - g_K(x^{(i)};\theta)) \le 0 \qquad i,\, m = 1, \ldots, M \tag{35d}$$

$$b_{im} \in \{0,1\}, \;\; c_{im} \in \{0,1\} \qquad i,\, m = 1, \ldots, M. \tag{35e}$$

We note that for any $b_{im} \in \{0,1\}$ and $c_{im} \in \{0,1\}$ for $i,\, m = 1,\dots,M$ at which (35c) and (35d) are satisfied, we have that

$$\sum_{m=1}^{M} \pi^{(m)} \mathbb{1}_{(-\infty,0]}(g_K(x^{(m)};\theta) - g_K(x^{(i)};\theta)) \leq \sum_{m=1}^{M} \pi^{(m)} b_{im}$$
$$\sum_{m=1}^{M} \pi^{(m)} \mathbb{1}_{(-\infty,0]}(y^{(m)} - \theta_0 - g_K(x^{(i)};\theta)) \geq \sum_{m=1}^{M} \pi^{(m)} c_{im}$$

for all $i = 1,\dots,M$ and thus the first-order dominance constraint (21) will be satisfied at a solution to (35). Since MINLPs can be difficult to solve in practice, one can relax (35) by replacing the binary constraints $b_{im} \in \{0,1\}$ and $c_{im} \in \{0,1\}$ with $b_{im} \in [0,1]$ and $c_{im} \in [0,1]$, respectively. This results in an "optimal" smoothing of the nonsmooth constraints (21).

Appendix C.1.2 Second-order stochastic dominance

An alternative approximate solution to the estimation problem (19), can be found by recalling that since $x \mapsto \max\{0,x\}$ is convex and continuous, and therefore is equal to its second Fenchel conjugate. That is,

$$\max\{0,x\} = \sup_{0 \leq t \leq 1} tx.$$

Using this fact, we can approximate the estimation problem (19) by the following optimization problem

$$\min_{\substack{\theta_0 \in \mathbb{R},\, \theta \in \Theta \\ s_{im} \in \mathbb{R},\, t_{im} \in \mathbb{R}}} \quad \frac{1}{2} \sum_{m=1}^{M} \pi^{(m)} (Y_m - \theta_0 - g_K(x^{(m)};\theta))^2 \tag{36a}$$

$$\text{subject to} \quad \sum_{m=1}^{M} \pi^{(m)} s_{im} (g_K(x^{(m)};\theta) - g_K(x^{(i)};\theta)) \geq \sum_{m=1}^{M} \pi^{(m)} t_{im} \qquad i = 1,\dots,M \tag{36b}$$

$$t_{im} + \theta_0 + g_K(x^{(i)};\theta) \geq Y_m \qquad i,\, m = 1,\dots,M \tag{36c}$$

$$0 \leq s_{im} \leq 1, \quad 0 \leq t_{im} \qquad i,\, m = 1,\dots,M. \tag{36d}$$

References

[1] F. A., A. Sobester, and A. Keane, *Engineering Design via Surrogate Modelling: A Practical Guide*, John Wiley & Sons, Ltd, 2008.

[2] P. Artzner, F. Delbaen, J.-M. Eber, and D. Heath, *Coherent measures of risk*, Mathematical Finance, 9 (1999), pp. 203–228.

[3] S.-K. Au and J. L. Beck, *Estimation of small failure probabilities in high dimensions by subset simulation*, Probabilistic Engineering Mechanics, 16 (2001), pp. 263 – 277.

[4] A. Ben-Tal and M. Teboulle, *An old-new concept of convex risk measures: The optimized certainty equivalent*, Mathematical Finance, 17 (2007), pp. 449–476.

[5] B. Bichon, M. Eldred, L. Swiler, S. Mahadevan, and J. McFarland, *Efficient global reliability analysis for nonlinear implicit performance functions*, AIAA Journal, 46 (2008), pp. 2459–2468.

[6] G. Blatman and B. Sudret, *Sparse polynomial chaos expansions and adaptive stochastic finite elements using a regression approach*, Comptes Rendus Mécanique, 336 (2008), pp. 518–523.

[7] C. Bucher and U. Bourgund, *A fast and efficient response surface approach for structural reliability problems*, Structural Safety, 7 (1990), pp. 57 – 66.

[8] H. Bühlmann, *Mathematical methods in risk theory*, vol. 172, Springer Science & Business Media, 2007.

[9] S. Chatterjee, S. Sarkar, S. Hore, N. Dey, A. S. Ashour, and V. E. Balas, *Particle swarm optimization trained neural network for structural failure prediction of multistoried rc buildings*, Neural Computing and Applications, 28 (2017), pp. 2005–2016.

[10] P. G. Constantine, E. Dow, and Q. Wang, *Active subspace methods in theory and practice: Applications to kriging surfaces*, SIAM Journal on Scientific Computing, 36 (2014), pp. A1500–A1524.

[11] S. Conti, M. Rumpf, R. Schultz, and S. Tölkes, *Stochastic dominance constraints in elastic shape optimization*, SIAM Journal on Control and Optimization, 56 (2018), pp. 3021–3034.

[12] D. Dentcheva and A. Ruszczyński, *Optimization with stochastic dominance constraints*, SIAM Journal on Optimization, 14 (2003), pp. 548–566.

[13] A. Doostan, A. Validi, and G. Iaccarino, *Non-intrusive low-rank separated approximation of high-dimensional stochastic models*, Computer Methods in Applied Mechanics and Engineering, 263 (2013), pp. 42–55.

[14] L. Faravelli, *Response surface approach for reliability analysis*, Journal of Engineering Mechanics, 115 (1989), pp. 2763–2781.

[15] H. Föllmer and A. Schied, *Convex measures of risk and trading constraints*, Finance Stoch., 6 (2002), pp. 429–447.

[16] B. Gaspar, A. Teixeira, and C. Guedes Soares, *Adaptive surrogate model with active refinement combining kriging and a trust region method*, Reliability Engineering and System Safety, 165 (2017), pp. 277–291. cited By 83.

[17] A. A. Gorodetsky and J. D. Jakeman, *Gradient-based optimization for regression in the functional tensor-train format*, Journal of Computational Physics, 374 (2018), pp. 1219 – 1238.

[18] E. Haber and L. Ruthotto, *Stable architectures for deep neural networks*, Inverse Problems, 34 (2017), p. 014004.

[19] M. Heinkenschloss, B. Kramer, T. Takhtaganov, and K. Willcox, *Conditional-value-at-risk estimation via reduced-order models*, SIAM/ASA Journal on Uncertainty Quantification, 6 (2018), pp. 1395–1423.

[20] L. J. Hong, Z. Hu, and G. Liu, *Monte carlo methods for value-at-risk and conditional value-at-risk: A review*, ACM Trans. Model. Comput. Simul., 24 (2014), pp. 22:1–22:37.

[21] J. Jakeman, M. Eldred, and K. Sargsyan, *Enhancing ℓ_1-minimization estimates of polynomial chaos expansions using basis selection*, Journal of Computational Physics, 289 (2015), pp. 18 – 34.

[22] J. D. Jakeman, *Pyapprox: Approximation and probabilistic analysis of data.* https://sandialabs.github.io/pyapprox/index.html, 2019.

[23] J. D. Jakeman, F. Franzelin, A. Narayan, M. Eldred, and D. Plfüger, *Polynomial chaos expansions for dependent random variables*, Computer Methods in Applied Mechanics and Engineering, 351 (2019), pp. 643 – 666.

[24] D. P. Kouri and A. Shapiro, *Optimization of PDEs with Uncertain Inputs*, Springer New York, New York, NY, 2018, pp. 41–81.

[25] D. P. Kouri and T. M. Surowiec, *Existence and optimality conditions for risk-averse PDE-constrained optimization*, SIAM/ASA Journal on Uncertainty Quantification, (2016). submitted.

[26] ——, *Risk-averse PDE-constrained optimization using the conditional value-at-risk*, SIAM Journal on Optimization, 26 (2016), pp. 365–396.

[27] J. Li, J. Li, and D. Xiu, *An efficient surrogate-based method for computing rare failure probability*, Journal of Computational Physics, 230 (2011), pp. 8683 – 8697.

[28] Z. Liu and A. Narayan, *On the computation of recurrence coefficients for univariate orthogonal polynomials*, 2021.

[29] S. Marelli and B. Sudret, *An active-learning algorithm that combines sparse polynomial chaos expansions and bootstrap for structural reliability analysis*, Structural Safety, 75 (2018), pp. 67 – 74.

[30] R. M. Neal, *Annealed importance sampling*, Statistics and Computing, 11 (2001), pp. 125–139.

[31] M. Paffrath and U. Wever, *Adapted polynomial chaos expansion for failure detection*, Journal of Computational Physics, 226 (2007), pp. 263 – 281.

[32] V. Papadopoulos, D. G. Giovanis, N. D. Lagaros, and M. Papadrakakis, *Accelerated subset simulation with neural networks for reliability analysis*, Computer Methods in Applied Mechanics and Engineering, 223-224 (2012), pp. 70 – 80.

[33] I. Papaioannou, W. Betz, K. Zwirglmaier, and D. Straub, *Mcmc algorithms for subset simulation*, Probabilistic Engineering Mechanics, 41 (2015), pp. 89 – 103.

[34] T. Qin, Z. Chen, J. D. Jakeman, and D. Xiu, *Deep learning of parameterized equations with applications to uncertainty quantification*, International Journal for Uncertainty Quantification, 11 (2021), pp. 63–82.

[35] R. T. Rockafellar and J. O. Royset, *On buffered failure probability in design and optimization of structures*, Reliability Engineering & System Safety, 95 (2010), pp. 499 – 510.

[36] R. T. Rockafellar and J. O. Royset, *Engineering decisions under risk averseness*, ASCE-ASME Journal of Risk and Uncertainty in Engineering Systems, Part A: Civil Engineering, 1 (2015), p. 04015003.

[37] R. T. Rockafellar and J. O. Royset, *Measures of residual risk with connections to regression, risk tracking, surrogate models, and ambiguity*, SIAM Journal on Optimization, 25 (2015), pp. 1179–1208.

[38] R. T. Rockafellar and S. Uryasev, *Conditional value-at-risk for general loss distributions*, Journal of Banking & Finance, 26 (2002), pp. 1443 – 1471.

[39] ——, *The fundamental risk quadrangle in risk management, optimization and statistical estimation*, Surveys in Operations Research and Management Science, 18 (2013), pp. 33 – 53.

[40] R. T. Rockafellar, S. Uryasev, and M. Zabarankin, *Risk tuning with generalized linear regression*, Mathematics of Operations Research, 33 (2008), pp. 712–729.

[41] G. Schuëller, H. Pradlwarter, and P. Koutsourelakis, *A critical appraisal of reliability estimation procedures for high dimensions*, Probabilistic Engineering Mechanics, 19 (2004), pp. 463 – 474.

[42] R. Schöbi, B. Sudret, and S. Marelli, *Rare event estimation using polynomial-chaos kriging*, ASCE-ASME Journal of Risk and Uncertainty in Engineering Systems, Part A: Civil Engineering, 3 (2017), p. D4016002.

[43] A. Shapiro, D. Dentcheva, and A. Ruszczynski, *Lectures on Stochastic Programming: Modeling and Theory, Second Edition*, MOS-SIAM Series on Optimization, Society for Industrial and Applied Mathematics, Philadelphia, 2014.

[44] Z. Sun, J. Wang, R. Li, and C. Tong, *Lif: A new kriging based learning function and its application to structural reliability analysis*, Reliability Engineering and System Safety, 157 (2017), pp. 152–165. cited By 120.

[45] V. S. Sundar and M. D. Shields, *Reliability analysis using adaptive kriging surrogates with multimodel inference*, ASCE-ASME Journal of Risk and Uncertainty in Engineering Systems, Part A: Civil Engineering, 5 (2019), p. 04019004.

[46] A. J. Torii, R. H. Lopez, and L. F. F. Miguel, *Probability of failure sensitivity analysis using polynomial expansion*, Probabilistic Engineering Mechanics, 48 (2017), pp. 76 – 84.

[47] S. Xiao, S. Oladyshkin, and W. Nowak, *Reliability analysis with stratified importance sampling based on adaptive kriging*, Reliability Engineering & System Safety, 197 (2020), p. 106852.

[48] D. Xiu, *Numerical Methods for Stochastic Computations: A Spectral Method Approach*, Princeton University Press, July 2010.

[49] E. Zio, *The future of risk assessment*, Reliability Engineering & System Safety, 177 (2018), pp. 176–190.

[50] Z. Zou, D. Kouri, and W. Aquino, *An adaptive local reduced basis method for solving pdes with uncertain inputs and evaluating risk*, Computer Methods in Applied Mechanics and Engineering, 345 (2019), pp. 302–322.

[51] Z. Zou, D. P. Kouri, and W. Aquino, *An adaptive sampling approach for solving pdes with uncertain inputs and evaluating risk*, in 19th AIAA Non-Deterministic Approaches Conference, 2017, p. 1325.