Uncertainty Principles in Risk-Aware Statistical Estimation

Nikolas P. Koumpis and Dionysios S. Kalogerias

***Abstract*— We present a new uncertainty principle for risk-aware statistical estimation, effectively quantifying the inherent trade-off between mean squared error (mse) and risk, the latter measured by the associated average predictive squared error variance (sev), for every admissible estimator of choice. Our uncertainty principle has a familiar form and resembles fundamental and classical results arising in several other areas, such as the Heisenberg principle in statistical and quantum mechanics, and the Gabor limit (time-scale trade-offs) in harmonic analysis. In particular, we prove that, provided a joint generative model of states and observables, the product between mse and sev is bounded from below by a computable model-dependent constant, which is explicitly related to the Pareto frontier of a recently studied sev-constrained minimum mse (MMSE) estimation problem. Further, we show that the aforementioned constant is inherently connected to an intuitive new and rigorously topologically grounded statistical measure of distribution skewness in multiple dimensions, consistent with Pearson's moment coefficient of skewness for variables on the line. Our results are also illustrated via numerical simulations.**

I. Introduction

Designing decision rules aiming for least expected losses is a standard and commonly employed objective in statistical learning, estimation, and control. Still, achieving optimal performance on average is insufficient without safeguarding against less probable though statistically significant, i.e., *risky*, events, and this is especially pronounced in modern, critical applications. Examples appear naturally in many areas, including robotics [1], [2], wireless communications and networking [3], [4], edge computing [5], health [6], and finance [7], to name a few. Indeed, risk-neutral decision policies smoothen unexpected events by construction, thus exhibiting potentially large statistical performance volatility, since the latter remains uncontrolled. In such situations, risk-aware decision rules are highly desirable as they systematically guarantee robustness, in the form of various operational specifications, such as safety [8], [9], fairness [10], [11], distributional robustness [12], [13], and prediction error stability [14].

In the realm of Bayesian mean squared error (mse) statistical estimation, a risk-constrained reformulation of the standard minimum mse (MMSE) estimation problem was recently proposed in [15], where, given a generative model (i.e., distribution) of states and observables, risk is measured by the average predictive squared error variance (sev) associated with every feasible square integrable (i.e., admissible) estimator. Quite remarkably, such a constrained functional estimation problem admits a unique closed-form solution; as compared with classical risk-neutral MMSE estimation (i.e., conditional mean), the optimal risk-aware estimator nonlinearly interpolates between the risk-neutral MMSE estimator (i.e., conditional mean) and a new, maximally risk-aware statistical estimator, minimizing average errors while constraining risk under a designer-specified threshold.

From the analysis presented in [15], it becomes evident that low-risk estimators deteriorate performance on average and vice-versa. However, although mse and sev (i.e., risk) are shown to trade between each other within the class of optimal risk-aware estimators proposed in [15], a mathematical statement that expresses this fundamental interplay for *general* estimators is non-trivial and currently unknown. This paper is precisely on the discovery, quantification and analysis of this interplay. Our contributions are as follows.

–A New mse/sev Uncertainty Principle (Section III). We quantify the trade-off between mse and sev associated with any square integrable estimator of choice by bounding their product by a model-dependent, estimator-independent characteristic constant. This fundamental lower bound, which we call the *optimal trade-off*, is always attained within the class of optimal risk-aware estimators of [15], and provides a universal benchmark of the trade-off efficiency of every possible admissible estimator. Our uncertainty relation comes in the natural form of an *uncertainty principle*; similar relations are met in different contexts, e.g., in statistical mechanics (Heisenberg principle) [16] and harmonic analysis (Gabor limit) [17]. In essence, uncertainty principles are bounds on the concentration or spread of a quantity in two different domains. In our case, sev measures the squared error statistical spread, while mse measures the squared error average (expected) value. Our uncertainty principle states that, *in general*, both quantities cannot be simultaneously small, let alone minimized; in the latter case exceptions exist, herein called the class of *skew-symmetric* models. In fact, conditionally Gaussian models are a canonical example in this exceptional class.

–Hedgeable Risk Margins and Lower Bound Characterization (Sections IV-V). We present an intuitive geometric interpretation of the class of risk-aware estimators of [15], inherently related to the optimal trade-off involved in our uncertainty principle. We define a new quantity, called the *expected hedgeable risk margin* associated with the underlying generative model by projecting the stochastic parameterized curve induced by the class of risk-aware estimators of [15] onto the line that links the risk-neutral (i.e., MMSE) with the maximally risk-averse estimator. Intuitively, such a projection expresses the margin to potentially counteract against risk (as measured by the sev), on average relative to the distribution of the observables. Subsequently, we show that, under mild assumptions, the optimal trade-off is *order-equivalent* to the corresponding expected risk margin. We

The Authors are with the Department of ECE, Michigan State University, East Lansing, MI. email: {dionysis, koumpisn}@msu.edu.

do this by proving explicit and order-matching upper and lower bounds on the optimal trade-off that depend strictly proportionally to the expected risk margin. The importance of this result is that a large (small) risk margin implies a large (small) optimal trade-off, and vice versa.

–Topological/Statistical Interpretation of Risk Margins, and Skewness in High Dimensions (Section VI). The significance of the risk-margin functional is established by showing that it admits a dual topological and statistical interpretation, within a rigorous technical framework. First, we prove that the space of all generative models with a finite risk-margin becomes a topological space endowed with a (pseudo)metric, the latter induced by a certain risk-margin-related functional. This functional vanishes for all skew-symmetric models, and therefore is rigorously interpretable as a distance to all members of this exceptional class (via the (pseudo)metric). Simultaneously, the aforementioned risk-margin-related functional corresponds to an intuitive model statistic which can be regarded as a generalized measure of distribution skewness, consistent with the familiar Pearson's moment coefficient of skewness, to which it reduces exactly for totally unobservable variables on the line. Similarly, the induced (pseudo)metric may be regarded as a measure of the *relative skewness* between (filtered) distributions.

Lastly, our results are supported by indicative numerical examples, along with a relevant discussion (Section VII).

II. SEV-Constrained MMSE Estimation

On a probability space $(\Omega, \mathscr{F}, \mathcal{P})$, consider random elements $\boldsymbol{X} : \Omega \to \mathbb{R}^n$ and $\boldsymbol{Y} : \Omega \to \mathbb{R}^m$ following a joint Borel probability measure $\mathcal{P}_{(\boldsymbol{X},\boldsymbol{Y})} \equiv \mathcal{P}$. Intuitively, $\boldsymbol{X}$ may be thought of as a *hidden* random state of nature, and $\boldsymbol{Y}$ as the corresponding *observables*. Also, hereafter, let $\mathcal{L}_{2|\mathscr{Y}}$ be the space of square-integrable $\mathscr{Y} \triangleq \sigma\{\boldsymbol{Y}\}$-measurable estimators (i.e., deterministic functions of the observables). Provided a generative model $\mathcal{P}_{(\boldsymbol{X},\boldsymbol{Y})}$, we consider the *mean squared error* and *squared error variance* functionals $\mathrm{mse} : \mathcal{L}_{2|\mathscr{Y}} \to \mathbb{R}_+$ and $\mathrm{sev} : \mathcal{L}_{2|\mathscr{Y}} \to \mathbb{R}_+$ defined respectively as

$$\mathrm{mse}(\hat{\boldsymbol{X}}) \triangleq \mathbb{E}\{\|\boldsymbol{X} - \hat{\boldsymbol{X}}\|_2^2\}, \quad \text{and} \tag{1}$$

$$\begin{aligned}\mathrm{sev}(\hat{\boldsymbol{X}}) &\triangleq \mathbb{E}\{\mathbb{V}_{\boldsymbol{Y}}\{\|\boldsymbol{X} - \hat{\boldsymbol{X}}\|_2^2\}\} \\ &\equiv \mathbb{E}\{\mathbb{E}\{(\|\boldsymbol{X} - \hat{\boldsymbol{X}}\|_2^2 - \mathbb{E}\{\|\boldsymbol{X} - \hat{\boldsymbol{X}}\|_2^2|\boldsymbol{Y}\})^2|\boldsymbol{Y}\}\},\end{aligned} \tag{2}$$

where $\hat{\boldsymbol{X}} \in \mathcal{L}_{2|\mathscr{Y}}$. Note that both functionals mse and sev are *law invariant*, i.e., they depend exclusively on $\mathcal{P}_{(\boldsymbol{X},\boldsymbol{Y})}$ [18]. As such, they may be equivalently thought of as mappings whose domain is the space of Borel probability measures on the product space $\mathbb{R}^n \times \mathbb{R}^m$.

While mse quantifies the squared error incurred by a given estimator $\hat{\boldsymbol{X}}$ on average and is a *gold-standard* performance criterion in estimation and control [19], sev quantifies the *risk* of $\hat{\boldsymbol{X}}$, as measured by the average predictive variance of the associated instantaneous estimation error around its MMSE-sense prediction given the observable $\boldsymbol{Y}$. In other words, sev quantifies the *statistical variability* of $||\boldsymbol{X} - \hat{\boldsymbol{X}}||_2^2$ against the *predictable statistical benchmark* $\mathbb{E}\{||\boldsymbol{X} - \hat{\boldsymbol{X}}||_2^2|\boldsymbol{Y}\}$. Such statistical variability is left uncontrolled in standard MMSE estimation; in fact, this is a natural flaw of MMSE estimators (i.e., conditional means) by construction, resulting in statistically unstable prediction errors, especially in problems involving skewed and/or heavy-tailed distributions [15].

To counteract risk-neutrality of MMSE estimators, a *constrained* reformulation of the MMSE problem was recently introduced in [15], where the mse is minimized subject to an explicit constraint on the associated sev. The resulting risk-aware stochastic variational (i.e., functional) problem is

$$\begin{array}{ll}\underset{\hat{\boldsymbol{X}} \in \mathcal{L}_{2|\mathscr{Y}}}{\text{minimize}} & \mathrm{mse}(\hat{\boldsymbol{X}}) \\ \text{subject to} & \mathrm{sev}(\hat{\boldsymbol{X}}) \le \varepsilon\end{array}, \tag{3}$$

where $\varepsilon > 0$ is a user-prescribed tolerance. As problem (3) may be shown to be convex [15], [20], prominent role in the analysis of (3) plays its variational Lagrangian relaxation

$$\inf_{\hat{\boldsymbol{X}} \in \mathcal{L}_{2|\mathscr{Y}}} \mathrm{mse}(\hat{\boldsymbol{X}}) + \mu\, \mathrm{sev}(\hat{\boldsymbol{X}}), \tag{4}$$

for fixed $\mu \ge 0$, dependent of the particular ε of choice. By defining the third-order posterior statistic

$$\boldsymbol{R}(\boldsymbol{Y}) \triangleq \mathbb{E}\{||\boldsymbol{X}||_2^2 \boldsymbol{X}|\boldsymbol{Y}\} - \mathbb{E}\{||\boldsymbol{X}||_2^2|\boldsymbol{Y}\}\mathbb{E}\{\boldsymbol{X}|\boldsymbol{Y}\}, \tag{5}$$

and under the mild condition that $\mathbb{E}\{\|\boldsymbol{X}\|_2^3|\boldsymbol{Y}\} \in \mathcal{L}_{2|\mathscr{Y}}$ (also assumed hereafter), an essentially unique optimal solution to (4) may be expressed in closed form as

$$\hat{\boldsymbol{X}}^*_\mu(\boldsymbol{Y}) = \frac{\mathbb{E}\{\boldsymbol{X}|\boldsymbol{Y}\} + \mu \boldsymbol{R}(\boldsymbol{Y})}{\boldsymbol{I} + 2\mu \boldsymbol{\Sigma}_{\boldsymbol{X}|\boldsymbol{Y}}}, \tag{6}$$

for all $\mu \ge 0$, where $\boldsymbol{\Sigma}_{\boldsymbol{X}|\boldsymbol{Y}} \succeq 0$ denotes the conditional convariance of $\boldsymbol{X}$ relative to $\boldsymbol{Y}$. When $\mu \equiv \infty$, we also *define* the *maximally risk-averse estimator* (corresponding to the tightest choice of ε)

$$\hat{\boldsymbol{X}}^*_\infty(\boldsymbol{Y}) \triangleq \frac{1}{2}\boldsymbol{\Sigma}^\dagger_{\boldsymbol{X}|\boldsymbol{Y}} \boldsymbol{R}(\boldsymbol{Y}) + \boldsymbol{U}\begin{bmatrix}\boldsymbol{0}_r \\ [\boldsymbol{U}^\top \mathbb{E}\{\boldsymbol{X}|\boldsymbol{Y}\}]_n^{n-r}\end{bmatrix}, \tag{7}$$

where $\boldsymbol{\Sigma}^\dagger_{\boldsymbol{X}|\boldsymbol{Y}} \succeq 0$ denotes the Moore–Penrose pseudoinverse of $\boldsymbol{\Sigma}_{\boldsymbol{X}|\boldsymbol{Y}}$, the latter with spectral decomposition $\boldsymbol{\Sigma}_{\boldsymbol{X}|\boldsymbol{Y}} \equiv \boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{U}^\top$ and of rank r. It is then standard procedure to show that $\lim_{\mu \to \infty} \hat{\boldsymbol{X}}^*_\mu = \hat{\boldsymbol{X}}^*_\infty$, implying that the paratererization $\boldsymbol{X}^*_{(\cdot)}$ is continuous on $[0, \infty]$. Lastly, as also proved in [15], whenever $\mathcal{P}_{\boldsymbol{X}|\boldsymbol{Y}}$ satisfies the condition

$$\mathbb{E}\{(X_i - \mathbb{E}\{X_i \mid \boldsymbol{Y}\})^2 (\boldsymbol{X} - \mathbb{E}\{\boldsymbol{X} \mid \boldsymbol{Y}\}) \mid \boldsymbol{Y}\} \equiv \boldsymbol{0} \tag{8}$$

for all $i \in \mathbb{N}^+_n$, it follows that

$$\hat{\boldsymbol{X}}^*_\mu = \mathbb{E}\{\boldsymbol{X}|\boldsymbol{Y}\}, \quad \forall \mu \in [0, \infty]. \tag{9}$$

In particular, this is the case when $\mathcal{P}_{\boldsymbol{X}|\boldsymbol{Y}}$ is jointly Gaussian. Hereafter, every generative model $\mathcal{P}_{(\boldsymbol{X},\boldsymbol{Y})}$ satisfying (9) for almost all $\boldsymbol{Y}$ will be called *skew-symmetric*; this terminology is justified later in Section V.

III. Uncertainty Principles

Already from (6) we can see that there is an inherent trade-off between mse and sev for the family of optimal estimators $\{\hat{\boldsymbol{X}}^*_\mu\}_\mu$. Of course, the resulting mse and sev define the Pareto frontier of problem (4). In this section, we quantify the mse/sev trade-off for all admissible estimators. We do that by deriving a non-trivial lower bound on the product between mse and sev.

We start by stating two technical lemmata, useful in our development. To this end, let $\sigma_{\max}(\boldsymbol{Y})$ denote the maximum eigenvalue of $\boldsymbol{\Sigma}_{\boldsymbol{X}|\boldsymbol{Y}}$, and define $\widehat{\boldsymbol{\Delta X}} \triangleq \hat{\boldsymbol{X}}^*_0 - \hat{\boldsymbol{X}}^*_\infty$.

Lemma 1 (Monotonicity). *The functions* $\mathrm{mse}(\hat{\boldsymbol{X}}^*_{(\cdot)})$ *and* $\mathrm{sev}(\hat{\boldsymbol{X}}^*_{(\cdot)})$ *are increasing and decreasing on* $[0,\infty]$*, respectively.*

Lemma 2 (Continuity). *The same functions* $\mathrm{mse}(\hat{\boldsymbol{X}}^*_{(\cdot)})$ *and* $\mathrm{sev}(\hat{\boldsymbol{X}}^*_{(\cdot)})$ *are continuous on* $[0,\infty]$ *and Lipschitz continuous on* $[0,\infty)$ *with respective constants*

$$\begin{aligned}\mathcal{K}_{\mathrm{mse}} &= \mathbb{E}\{\sigma_{\max}(\boldsymbol{Y})\|\widehat{\boldsymbol{\Delta X}}\|_2^2\} \quad \textit{and}\\ \mathcal{K}_{\mathrm{sev}} &= \mathbb{E}\{\sigma_{\max}(\boldsymbol{Y})^2\|\widehat{\boldsymbol{\Delta X}}\|_2^2\}.\end{aligned} \tag{10}$$

Utilizing the lemmata above, we may now state the main result of the paper, which provides a new and useful characterization of the region of allowable mse-sev combinations *ever* possibly achievable by *any* square-integrable estimator, given a generative model. Essentially, our result, which now follows, quantifies that inherent trade-off between average estimation performance and risk.

Theorem 1 (Uncertainty Principles). *Every admissible estimator* $\hat{\boldsymbol{X}} \equiv \hat{\boldsymbol{X}}(\boldsymbol{Y}) \in \mathcal{L}_{2|\mathscr{Y}}$ *satisfies the lower bounds*

$$\mathrm{mse}(\hat{\boldsymbol{X}})\mathrm{sev}(\hat{\boldsymbol{X}}) \ge \mathfrak{h} \ge \mathrm{mse}(\hat{\boldsymbol{X}}^*_0)\mathrm{sev}(\hat{\boldsymbol{X}}^*_\infty), \tag{11}$$

where the characteristic number $\mathfrak{h}$ *is given by*

$$\mathfrak{h}(\mathcal{P}) \equiv \mathrm{mse}(\hat{\boldsymbol{X}}^*_{\mu^\star})\mathrm{sev}(\hat{\boldsymbol{X}}^*_{\mu^\star}), \tag{12}$$

for any $\mu^\star \in \mathrm{argmin}_{\mu\in[0,\infty]}\{\mathrm{mse}(\hat{\boldsymbol{X}}^*_\mu)\mathrm{sev}(\hat{\boldsymbol{X}}^*_\mu)\} \neq \emptyset$.

Proof of Theorem 1. We may examine the following three mutually exclusive cases:

Case 1: $\mathrm{sev}(\hat{\boldsymbol{X}}) \in (\mathrm{sev}(\hat{\boldsymbol{X}}^*_\infty), \mathrm{sev}(\hat{\boldsymbol{X}}^*_0)]$. Then, from the intermediate value theorem, it follows that there is $\mu_{\hat{\boldsymbol{X}}} \in [0,\infty)$ such that $\hat{\boldsymbol{X}}^*_{\mu_{\hat{\boldsymbol{X}}}}$ matches the performance of $\hat{\boldsymbol{X}}$, i.e.,

$$\mathrm{sev}(\hat{\boldsymbol{X}}^*_{\mu_{\hat{\boldsymbol{X}}}}) \equiv \mathrm{sev}(\hat{\boldsymbol{X}}). \tag{13}$$

This fact, together with optimality of $\hat{\boldsymbol{X}}^*_{\mu_{\hat{\boldsymbol{X}}}}$ for the Lagrangian relaxation (4), implies

$$\mathrm{mse}(\hat{\boldsymbol{X}})+\mu_{\hat{\boldsymbol{X}}}\mathrm{sev}(\hat{\boldsymbol{X}}) \ge \mathrm{mse}(\hat{\boldsymbol{X}}^*_{\mu_{\hat{\boldsymbol{X}}}})+\mu_{\hat{\boldsymbol{X}}}\mathrm{sev}(\hat{\boldsymbol{X}}^*_{\mu_{\hat{\boldsymbol{X}}}}), \tag{14}$$

which further gives

$$\mathrm{mse}(\hat{\boldsymbol{X}}) \ge \mathrm{mse}(\hat{\boldsymbol{X}}^*_{\mu_{\hat{\boldsymbol{X}}}}). \tag{15}$$

Therefore, it is true that

$$\begin{aligned}\mathrm{mse}(\hat{\boldsymbol{X}})\mathrm{sev}(\hat{\boldsymbol{X}}) &\ge \mathrm{mse}(\hat{\boldsymbol{X}}^*_{\mu_{\hat{\boldsymbol{X}}}})\mathrm{sev}(\hat{\boldsymbol{X}}^*_{\mu_{\hat{\boldsymbol{X}}}})\\ &\ge \inf_{\mu\in[0,\infty]}\mathrm{mse}(\hat{\boldsymbol{X}}^*_\mu)\mathrm{sev}(\hat{\boldsymbol{X}}^*_\mu),\end{aligned} \tag{16}$$

proving the claim of the theorem in this case.

Case 2: $\mathrm{sev}(\hat{\boldsymbol{X}}) \equiv \mathrm{sev}(\hat{\boldsymbol{X}}^*_\infty)$. Because $\mathrm{sev}(\hat{\boldsymbol{X}})$ is convex quadratic in $\hat{\boldsymbol{X}}$ and bounded below, it is fairly easy to show that $\mathrm{sev}(\hat{\boldsymbol{X}}^*_\infty) \equiv \inf_{\hat{\boldsymbol{X}}\in\mathcal{L}_{2|\mathscr{Y}}}\mathrm{sev}(\hat{\boldsymbol{X}})$. Now, it either holds that $\mathrm{mse}(\hat{\boldsymbol{X}}) \ge \mathrm{mse}(\hat{\boldsymbol{X}}^*_\infty)$ $(\equiv \lim_{\mu\uparrow\infty}\mathrm{mse}(\hat{\boldsymbol{X}}^*_\mu))$, giving

$$\mathrm{mse}(\hat{\boldsymbol{X}})\mathrm{sev}(\hat{\boldsymbol{X}}) \ge \mathrm{mse}(\hat{\boldsymbol{X}}^*_\infty)\mathrm{sev}(\hat{\boldsymbol{X}}^*_\infty), \tag{17}$$

or it must be true that $\mathrm{mse}(\hat{\boldsymbol{X}}) < \mathrm{mse}(\hat{\boldsymbol{X}}^*_\infty)$. In the latter case, the intermediate value property implies the existence of a multiplier $\mu_{\hat{\boldsymbol{X}}} \in [0,\infty)$ such that $\mathrm{mse}(\hat{\boldsymbol{X}}) \equiv \mathrm{mse}(\hat{\boldsymbol{X}}^*_{\mu_{\hat{\boldsymbol{X}}}})$. If $\mu_{\hat{\boldsymbol{X}}} > 0$, optimality of $\hat{\boldsymbol{X}}^*_{\mu_{\hat{\boldsymbol{X}}}}$ for (4) yields

$$\mathrm{mse}(\hat{\boldsymbol{X}})+\mu_{\hat{\boldsymbol{X}}}\mathrm{sev}(\hat{\boldsymbol{X}}) \ge \mathrm{mse}(\hat{\boldsymbol{X}}^*_{\mu_{\hat{\boldsymbol{X}}}})+\mu_{\hat{\boldsymbol{X}}}\mathrm{sev}(\hat{\boldsymbol{X}}^*_{\mu_{\hat{\boldsymbol{X}}}}), \tag{18}$$

or, equivalently, $\mathrm{sev}(\hat{\boldsymbol{X}}) \ge \mathrm{sev}(\hat{\boldsymbol{X}}^*_{\mu_{\hat{\boldsymbol{X}}}})$. Note that $\mathrm{sev}(\hat{\boldsymbol{X}}) \equiv \mathrm{sev}(\hat{\boldsymbol{X}}^*_\infty)$, and so this actually implies that $\mathrm{sev}(\hat{\boldsymbol{X}}^*_\infty) \equiv \mathrm{sev}(\hat{\boldsymbol{X}}^*_{\mu_{\hat{\boldsymbol{X}}}})$. Regardless, we obtain

$$\mathrm{mse}(\hat{\boldsymbol{X}})\mathrm{sev}(\hat{\boldsymbol{X}}) \ge \mathrm{mse}(\hat{\boldsymbol{X}}^*_{\mu_{\hat{\boldsymbol{X}}}})\mathrm{sev}(\hat{\boldsymbol{X}}^*_{\mu_{\hat{\boldsymbol{X}}}}). \tag{19}$$

If $\mu_{\hat{\boldsymbol{X}}} \equiv 0$, then $\hat{\boldsymbol{X}} \equiv \hat{\boldsymbol{X}}^*_0 \equiv \mathbb{E}\{\boldsymbol{X}|\boldsymbol{Y}\}$ almost everywhere, which implies that $\mathrm{sev}(\hat{\boldsymbol{X}}^*_\infty) \equiv \mathrm{sev}(\hat{\boldsymbol{X}}) \equiv \mathrm{sev}(\hat{\boldsymbol{X}}_0)$, and

$$\mathrm{mse}(\hat{\boldsymbol{X}})\mathrm{sev}(\hat{\boldsymbol{X}}) \equiv \mathrm{mse}(\hat{\boldsymbol{X}}^*_0)\mathrm{sev}(\hat{\boldsymbol{X}}^*_0). \tag{20}$$

From (17), (19) and (20), we readily see that

$$\mathrm{mse}(\hat{\boldsymbol{X}})\mathrm{sev}(\hat{\boldsymbol{X}}) \ge \inf_{\mu\in[0,\infty]}\mathrm{mse}(\hat{\boldsymbol{X}}^*_\mu)\mathrm{sev}(\hat{\boldsymbol{X}}^*_\mu), \tag{21}$$

whenever $\hat{\boldsymbol{X}}$ is such that $\mathrm{sev}(\hat{\boldsymbol{X}}) \equiv \mathrm{sev}(\hat{\boldsymbol{X}}^*_\infty)$.

Case 3: $\mathrm{sev}(\hat{\boldsymbol{X}}) \notin [\mathrm{sev}(\hat{\boldsymbol{X}}^*_\infty), \mathrm{sev}(\hat{\boldsymbol{X}}^*_0)]$. Then we must necessarily have

$$\mathrm{sev}(\hat{\boldsymbol{X}}) > \mathrm{sev}(\hat{\boldsymbol{X}}^*_\mu), \quad \forall \mu \in [0,\infty]. \tag{22}$$

In this case, either $\mathrm{mse}(\hat{\boldsymbol{X}}) \equiv \mathrm{mse}(\hat{\boldsymbol{X}}^*_{\mu_{\hat{\boldsymbol{X}}}})$ for some $\mu_{\hat{\boldsymbol{X}}} \in [0,\infty]$, implying that

$$\mathrm{mse}(\hat{\boldsymbol{X}})\mathrm{sev}(\hat{\boldsymbol{X}}) \ge \mathrm{mse}(\hat{\boldsymbol{X}}^*_{\mu_{\hat{\boldsymbol{X}}}})\mathrm{sev}(\hat{\boldsymbol{X}}^*_{\mu_{\hat{\boldsymbol{X}}}}), \tag{23}$$

or $\mathrm{mse}(\hat{\boldsymbol{X}}) > \mathrm{mse}(\hat{\boldsymbol{X}}^*_\mu)$ for all $\mu \in [0,\infty]$, which gives

$$\mathrm{mse}(\hat{\boldsymbol{X}})\mathrm{sev}(\hat{\boldsymbol{X}}) > \mathrm{mse}(\hat{\boldsymbol{X}}^*_\mu)\mathrm{sev}(\hat{\boldsymbol{X}}^*_\mu), \ \forall \mu \in [0,\infty]. \tag{24}$$

Again, it follows that

$$\mathrm{mse}(\hat{\boldsymbol{X}})\mathrm{sev}(\hat{\boldsymbol{X}}) \ge \inf_{\mu\in[0,\infty]}\mathrm{mse}(\hat{\boldsymbol{X}}^*_\mu)\mathrm{sev}(\hat{\boldsymbol{X}}^*_\mu), \tag{25}$$

and the proof is now complete. □

The practical aspects of Theorem 1 are summarized as follows: Provided an adequate threshold of mse(sev), the corresponding sev(mse) is always, at least, inversely proportional to that level. Except for its analogy to classical uncertainty principles from physics and analysis (see Section

I), our uncertainty relation resembles classical lower bounds in unconstrained/unbiased estimation, such as the Cramèr-Rao and Chapman-Robins bounds, in the sense that it provides a universal benchmark for any admissible estimator. Such an estimator might be chosen from the class of risk-aware estimators $\{\hat{\boldsymbol{X}}^*_\mu\}_\mu$, or even from many other estimator classes (possibly more computationally friendly), such as linear estimators, deep neural networks, adaptive estimators, convex combinations of $\hat{\boldsymbol{X}}^*_0$ and $\hat{\boldsymbol{X}}^*_\infty$, etc. However, under the setting of Theorem 1, any estimator outside the family of risk-aware estimators $\{\hat{\boldsymbol{X}}^*_\mu\}_\mu$ calls for Pareto improvement. Further, estimators achieving the lower bound $\mathfrak{h}(\mathcal{P})$ must be equivalent to $\hat{\boldsymbol{X}}^*_{\mu^\star}$ (note that $\mu^\star$ might not be unique).

IV. Hedgeable Risk Margins

As expected, the Bayesian lower bound $\mathfrak{h}(\mathcal{P})$ is achieved within the class of risk-aware estimators $\{\hat{\boldsymbol{X}}^*_\mu\}_\mu$. In this section, we are interested in answering the following question: Where are the estimators that achieve the lower bound with respect to the risk aversion parameter μ localized, and how does the width of such a localization area relate with the generative model $\mathcal{P}_{(\boldsymbol{X},\boldsymbol{Y})}$? To this end, we now introduce a function that measures the *total projection* onto $\widehat{\boldsymbol{\Delta X}}$ of the transformed risk-averse μ-parameterized stochastic curve generated by $\hat{\boldsymbol{X}}^*_\mu$ for all $\mu \geq 0$, defined as

$$\mathbb{C}(\boldsymbol{Y}) \triangleq \int_0^{+\infty} \left\langle \boldsymbol{\Sigma}^\dagger_{\boldsymbol{X}|\boldsymbol{Y}} \frac{d\hat{\boldsymbol{X}}^*_\tau(\boldsymbol{Y})}{d\tau}, \widehat{\boldsymbol{\Delta X}}(\boldsymbol{Y}) \right\rangle d\tau. \quad (26)$$

As we will see, $\mathbb{C}(\boldsymbol{Y})$ is actually nonnegative and expresses the *margin to potentially counteract or hedge against risk*, as the latter is quantified by the sev functional; hereafter, we suggestively refer to $\mathbb{C}(\boldsymbol{Y})$ as the *hedgeable risk margin* associated with observation $\boldsymbol{Y}$. By letting $\sigma_{\min}(\boldsymbol{Y})$ be the smallest *non-zero* eigenvalue of $\boldsymbol{\Sigma}_{\boldsymbol{X}|\boldsymbol{Y}}$, we have the following result.

Theorem 2 (Expected Hedgeable Risk Margin). *For fixed generative model $\mathcal{P}_{(\boldsymbol{X},\boldsymbol{Y})}$, $\mathbb{C}(\boldsymbol{Y})$ may be expressed as*

$$\mathbb{C}(\boldsymbol{Y}) = \|\widehat{\boldsymbol{\Delta X}}(\boldsymbol{Y})\|^2_{\boldsymbol{\Sigma}^\dagger_{\boldsymbol{X}|\boldsymbol{Y}}} \geq 0, \quad (27)$$

and its expected value satisfies the standard bounds

$$\mathbb{E}\{\mathcal{E}^2_L(\boldsymbol{Y})\} \leq \mathbb{E}\{\mathbb{C}(\boldsymbol{Y})\} \leq \mathbb{E}\{\mathcal{E}^2_U(\boldsymbol{Y})\}, \quad (28)$$

where

$$\mathcal{E}_{U(L)}(\boldsymbol{Y}) \triangleq \begin{cases} \dfrac{\|\widehat{\boldsymbol{\Delta X}}(\boldsymbol{Y})\|_2}{\sqrt{\sigma_{\min(\max)}(\boldsymbol{Y})}}, & \textit{if } \sigma_{\min}(\boldsymbol{Y}) > 0 \\ 0, & \textit{if not} \end{cases}. \quad (29)$$

Proof of Theorem 2. In case $\sigma_{\min}(\boldsymbol{Y}) = 0$ (which happens if and only if $\sigma_{\max}(\boldsymbol{Y}) = 0$), the situation is trivial and the result holds. Therefore, in what follows we may assume that $\sigma_{\min}(\boldsymbol{Y}) > 0$. In that case, to obtain an expression for (26) we have to differentiate (6) with respect to μ which for brevity may be written as $\hat{\boldsymbol{X}}^*_\mu = \boldsymbol{\zeta}(\mu)(\hat{\boldsymbol{X}}^*_0 + \mu\boldsymbol{R})$, where $\boldsymbol{\zeta}(\mu) \triangleq (\boldsymbol{I} + 2\mu\boldsymbol{\Sigma}_{\boldsymbol{X}|\boldsymbol{Y}})^{-1}$. By fixing an observation $\boldsymbol{Y}$, we obtain the linear μ-varying system

$$\frac{d\hat{\boldsymbol{X}}^*_\mu}{d\mu} = -2\boldsymbol{\zeta}(\mu)\boldsymbol{\Sigma}_{\boldsymbol{X}|\boldsymbol{Y}}\hat{\boldsymbol{X}}^*_\mu + \boldsymbol{\zeta}(\mu)\boldsymbol{R}. \quad (30)$$

Then, given the commutator $\left[\boldsymbol{\zeta}(\mu), \boldsymbol{\Sigma}_{\boldsymbol{X}|\boldsymbol{Y}}\right] = \boldsymbol{0}$, (30) can be written as

$$\frac{d\hat{\boldsymbol{X}}^*_\mu}{d\mu} = \boldsymbol{\zeta}(\mu)^2(\boldsymbol{R} - 2\boldsymbol{\Sigma}_{\boldsymbol{X}|\boldsymbol{Y}}\hat{\boldsymbol{X}}^*_0). \quad (31)$$

Now, from (7) we have

$$\hat{\boldsymbol{X}}^*_\infty = \boldsymbol{U}\boldsymbol{K}\boldsymbol{U}^\top\hat{\boldsymbol{X}}^*_0 + \frac{1}{2}\boldsymbol{U}\boldsymbol{D}^\dagger_{\boldsymbol{\Sigma}_{\boldsymbol{X}|\boldsymbol{Y}}}\boldsymbol{U}^\top\boldsymbol{R}, \quad (32)$$

where

$$\boldsymbol{D}^\dagger_{\boldsymbol{\Sigma}_{\boldsymbol{X}|\boldsymbol{Y}}} = \operatorname{diag}\left(\{(\sigma_i(\boldsymbol{Y}))^{-1}\}_{i\in\mathbb{N}^+_r}, \boldsymbol{0}\right) \quad (33)$$

and

$$\boldsymbol{K} = \operatorname{diag}\left(\{0\}_{i\in\mathbb{N}^+_r}, \boldsymbol{1}\right). \quad (34)$$

Thus,

$$\begin{aligned} \boldsymbol{U}^\top\boldsymbol{R} &= \boldsymbol{D}_{\boldsymbol{\Sigma}_{\boldsymbol{X}|\boldsymbol{Y}}}\left(2\boldsymbol{U}^\top\hat{\boldsymbol{X}}^*_\infty - 2\boldsymbol{K}\boldsymbol{U}^\top\hat{\boldsymbol{X}}^*_0\right) \\ &= 2\boldsymbol{D}_{\boldsymbol{\Sigma}_{\boldsymbol{X}|\boldsymbol{Y}}}\boldsymbol{U}^\top\hat{\boldsymbol{X}}^*_\infty. \end{aligned} \quad (35)$$

From (31), and (35), we have:

$$\frac{d\hat{\boldsymbol{X}}^*_\mu(\boldsymbol{Y})}{d\mu} = 2\boldsymbol{U}\boldsymbol{\Lambda}(\boldsymbol{Y})^2\boldsymbol{D}_{\boldsymbol{\Sigma}_{\boldsymbol{X}|\boldsymbol{Y}}}\boldsymbol{U}^\top\widehat{\boldsymbol{\Delta X}}, \quad (36)$$

where

$$\boldsymbol{\Lambda}(\boldsymbol{Y})^2 = \operatorname{diag}\left(\{(1 + 2\mu\sigma_i(\boldsymbol{Y}))^{-2}\}_{i\in\mathbb{N}^+_r}, \boldsymbol{1}\right). \quad (37)$$

Therefore, the integrand reads:

$$\begin{aligned} &\left\langle \boldsymbol{\Sigma}^\dagger_{\boldsymbol{X}|\boldsymbol{Y}}\frac{d\hat{\boldsymbol{X}}^*_\mu(\boldsymbol{Y})}{d\mu}, \widehat{\boldsymbol{\Delta X}} \right\rangle \\ &= 2\widehat{\boldsymbol{\Delta X}}^\top\boldsymbol{U}\boldsymbol{\Lambda}(\boldsymbol{Y})^2(\boldsymbol{D}^\dagger_{\boldsymbol{\Sigma}_{\boldsymbol{X}|\boldsymbol{Y}}})\boldsymbol{D}_{\boldsymbol{\Sigma}_{\boldsymbol{X}|\boldsymbol{Y}}}\boldsymbol{U}^\top\widehat{\boldsymbol{\Delta X}}, \end{aligned} \quad (38)$$

from which it follows that

$$\mathbb{C}(\boldsymbol{Y}) = [\boldsymbol{U}^\top\widehat{\boldsymbol{\Delta X}}]^\top\boldsymbol{D}^\dagger_{\boldsymbol{\Sigma}_{\boldsymbol{X}|\boldsymbol{Y}}}\boldsymbol{U}^\top\widehat{\boldsymbol{\Delta X}}. \quad (39)$$

Thus, provided the assumptions from [15] we obtain

$$\mathbb{E}\left\{\frac{\|\widehat{\boldsymbol{\Delta X}}\|^2_2}{\sigma_{\max}(\boldsymbol{Y})}\right\} \leq \mathbb{E}\{\mathbb{C}(\boldsymbol{Y})\} \leq \mathbb{E}\left\{\frac{\|\widehat{\boldsymbol{\Delta X}}\|^2_2}{\sigma_{\min}(\boldsymbol{Y})}\right\}, \quad (40)$$

and we are done. □

At this point it is worth attributing geometric meaning in the above result; by integrating (36) in $(0, \mu)$ we obtain:

$$\hat{\boldsymbol{X}}^*_\mu(\boldsymbol{Y}) = \hat{\boldsymbol{X}}^*_0(\boldsymbol{Y}) + \boldsymbol{U}\boldsymbol{G}(\mu)\boldsymbol{U}^\top\widehat{\boldsymbol{\Delta X}}, \quad (41)$$

where

$$\boldsymbol{G}(\mu) = \operatorname{diag}\left(\{2\mu\sigma_i(\boldsymbol{Y})(1 + 2\mu\sigma_i(\boldsymbol{Y}))^{-1}\}_{i\in\mathbb{N}^+_r}, \boldsymbol{0}\right). \quad (42)$$

We observe that the risk-aware estimator shifts the conditional mean estimator by the transformed difference $\widehat{\boldsymbol{\Delta X}}$. Thus, motivated by the one dimensional case we may interpret $\widehat{\boldsymbol{\Delta X}}$ as the direction of asymmetry of the posterior (for the given observation), and note the following: referring to (38), $[\boldsymbol{U}^\top\widehat{\boldsymbol{\Delta X}}]_i$ being large enough for most of $i \in \mathbb{N}^+_n$ implies that large estimation errors incurred by the conditional mean estimator are mostly due to the built-in riskiness of the posterior. In this case, the projection from (38) decreases with μ over a long width before fading-out.

To put it differently, large projections indicate enough margin with respect to μ to potentially hedge against risk, justifying the meaning ascribed in $\mathbb{C}(\boldsymbol{Y})$. Inequality (40) implies that, on average, the information regarding the *active* risk-aware estimates -and subsequently those that achieve the lower bound- is completely embodied to the limit points of the curve. Thus, recalling (9) and (8), we expect that ńearą skew symmetric generative model those risk-averse estimates which actively account for risk will be limited. Such a claim might also be justified from the fact that

$$\mathbb{E}\left\{\left|\frac{d\Pi(\mu)}{d\mu}\right|\right\} \le 2\frac{\mathbb{E}\{\mathbb{C}(\boldsymbol{Y})\}}{\mu^2}\ , \tag{43}$$

where $\Pi(\mu)$ is the integrand of (26). Thus, (43) implies that for every $\varepsilon > 0$ there exists $\tilde{\mu} = \sqrt{2\varepsilon^{-1}\mathbb{E}\{\mathbb{C}(\boldsymbol{Y})\}}$ such that the right-hand side of (43) is upper bounded by ε.

V. Lower Bound Characterization

As we saw earlier, provided a generative model, there exist risk-aware estimators that result in both good (even optimal) performance on average, and an adequate level of robustness; a standard example is the efficient frontier family $\{\hat{\boldsymbol{X}}^*_\mu\}_\mu$, and in particular for $\mu \equiv \mu^*$ (see Theorem 1). But still, how far can the trade-off incurred by any member of the efficient frontier class $\{\hat{\boldsymbol{X}}^*_\mu\}_\mu$ be from achieving the ultimate lower bound $\mathrm{mse}(\hat{\boldsymbol{X}}^*_0)\mathrm{sev}(\hat{\boldsymbol{X}}^*_\infty)$, and how is this distance related to the assumed generative model? We answer these questions by showing that the difference between the parameterization $\mathrm{mse}(\hat{\boldsymbol{X}}^*_\mu)\mathrm{sev}(\hat{\boldsymbol{X}}^*_\mu)$ and $\mathrm{mse}(\hat{\boldsymbol{X}}^*_0)\mathrm{sev}(\hat{\boldsymbol{X}}^*_\infty)$ is bounded from above *and* below by a function of another positive, risk margin-related, model-dependent functional.

Theorem 3 (Uncertainty Bound Characterization). *Suppose that there exists $\rho_{\max} \ge 0$, such that* $\operatorname{ess\,sup}\sigma_{\max}(\boldsymbol{Y}) \le \rho_{\max}$*. Then, the products $\mathrm{mse}(\hat{\boldsymbol{X}}^*_\mu)\mathrm{sev}(\hat{\boldsymbol{X}}^*_\mu)$, $\mu \in [0,\infty]$ and $\mathrm{mse}(\hat{\boldsymbol{X}}^*_0)\mathrm{sev}(\hat{\boldsymbol{X}}^*_\infty)$ satisfy the uniform upper bound*

$$\mathrm{mse}(\hat{\boldsymbol{X}}^*_\mu)\mathrm{sev}(\hat{\boldsymbol{X}}^*_\mu) - \mathrm{mse}(\hat{\boldsymbol{X}}^*_0)\mathrm{sev}(\hat{\boldsymbol{X}}^*_\infty) \le \mathbb{U}(\mathcal{P}), \tag{44}$$

where

$$\begin{aligned}\mathbb{U}(\mathcal{P}) = \big((\rho_{\max})^2\mathrm{mse}(\hat{\boldsymbol{X}}^*_0) + \rho_{\max}\mathrm{sev}(\hat{\boldsymbol{X}}^*_\infty)\big)d(\mathcal{P})^2 \\ + (\rho_{\max})^3 d(\mathcal{P})^4,\end{aligned} \tag{45}$$

and $d(\mathcal{P}) \triangleq 2\sqrt{\mathbb{E}\{\mathbb{C}(\boldsymbol{Y})\}}$. If, further, there exists $\rho_{\min} > 0$, such that $\operatorname{ess\,inf}\sigma_{\min}(\boldsymbol{Y}) \ge \rho_{\min}$*, then the same products satisfy the lower bound*

$$\mathbb{L}(\mathcal{P},\mu) \le \mathrm{mse}(\hat{\boldsymbol{X}}^*_\mu)\mathrm{sev}(\hat{\boldsymbol{X}}^*_\mu) - \mathrm{mse}(\hat{\boldsymbol{X}}^*_0)\mathrm{sev}(\hat{\boldsymbol{X}}^*_\infty) \tag{46}$$

where

$$\begin{aligned}\mathbb{L}(\mathcal{P},\mu) = \big(\alpha(\mu)\mathrm{mse}(\hat{\boldsymbol{X}}^*_0) + \rho_{\min}\mu^2\alpha(\mu)\mathrm{sev}(\hat{\boldsymbol{X}}^*_\infty)\big)d(\mathcal{P})^2 \\ + (\rho_{\min})\mu^2\alpha(\mu)^2 d(\mathcal{P})^4,\end{aligned} \tag{47}$$

and $\alpha(\mu) = (1/4)\rho_{\min}^2(1+2\mu\rho_{\max})^{-2}$.

Proof of Theorem 3. To begin with, under the setting of the theorem, let us integrate (36) in (μ, μ'), obtaining

$$\hat{\boldsymbol{X}}^*_\mu - \hat{\boldsymbol{X}}^*_{\mu'} = (\mu - \mu')\boldsymbol{U}\boldsymbol{H}(\mu,\mu')\boldsymbol{U}^\top\widehat{\boldsymbol{\Delta X}}, \tag{48}$$

where

$$\begin{aligned}&\boldsymbol{H}(\mu,\mu') \\ &= \mathrm{diag}\left(\left\{\frac{2\sigma_i(\boldsymbol{Y})}{\big(1+2\mu\sigma_i(\boldsymbol{Y})\big)\big(1+2\mu'\sigma_i(\boldsymbol{Y})\big)}\right\}_{i\in\mathbb{N}_r^+}, \boldsymbol{0}\right).\end{aligned} \tag{49}$$

Subsequently, consider the difference

$$\begin{aligned}&|\mathrm{mse}(\hat{\boldsymbol{X}}^*_\mu) - \mathrm{mse}(\hat{\boldsymbol{X}}^*_{\mu'})| \\ &= \big|\mathbb{E}\{(\hat{\boldsymbol{X}}^*_\mu - \hat{\boldsymbol{X}}^*_0 + \hat{\boldsymbol{X}}^*_{\mu'} - \hat{\boldsymbol{X}}^*_0)^\top(\hat{\boldsymbol{X}}^*_\mu - \hat{\boldsymbol{X}}^*_{\mu'})\}\big|.\end{aligned} \tag{50}$$

After substituting $\mu' = 0$ and subsequently applying (48) and *Lemma* 2, we get

$$\begin{aligned}\Lambda_{\mathrm{mse}}(\mu) &\triangleq \mathrm{mse}(\hat{\boldsymbol{X}}^*_\mu) - \mathrm{mse}(\hat{\boldsymbol{X}}_0) \\ &= \mathbb{E}\{(\hat{\boldsymbol{X}}^*_\mu - \hat{\boldsymbol{X}}^*_0)^\top(\hat{\boldsymbol{X}}^*_\mu - \hat{\boldsymbol{X}}^*_0)\}.\end{aligned} \tag{51}$$

Additionally, recalling the QCQP reformulation of the sev-constrained MMSE estimation problem in [15], we may write

$$\begin{aligned}&|\mathrm{sev}(\hat{\boldsymbol{X}}_\mu) - \mathrm{sev}(\hat{\boldsymbol{X}}_{\mu'})| \\ &= \big|\mathbb{E}\{(\hat{\boldsymbol{X}}^*_\mu - \hat{\boldsymbol{X}}^*_\infty + \hat{\boldsymbol{X}}^*_{\mu'} - \hat{\boldsymbol{X}}^*_\infty)^\top\boldsymbol{\Sigma}_{\boldsymbol{X}|\boldsymbol{Y}}(\hat{\boldsymbol{X}}^*_\mu - \hat{\boldsymbol{X}}^*_{\mu'})\}\big|.\end{aligned} \tag{52}$$

Thus, by substituting $\mu' = +\infty$, (52) yields

$$\begin{aligned}\Lambda_{\mathrm{sev}}(\mu) &\triangleq \mathrm{sev}(\hat{\boldsymbol{X}}^*_\mu) - \mathrm{sev}(\hat{\boldsymbol{X}}^*_\infty) \\ &= \mathbb{E}\{(\hat{\boldsymbol{X}}^*_\mu - \hat{\boldsymbol{X}}^*_\infty)^\top\boldsymbol{\Sigma}_{\boldsymbol{X}|\boldsymbol{Y}}(\hat{\boldsymbol{X}}^*_\mu - \hat{\boldsymbol{X}}^*_\infty)\}.\end{aligned} \tag{53}$$

From (48), Lemma 1 and Theorem 2, it is easy to show that

$$\Lambda_{\mathrm{mse}}(\mu) \le \frac{\rho_{\max}}{4}d(\mathcal{P})^2 \text{ and } \Lambda_{\mathrm{sev}}(\mu) \le \frac{\rho_{\max}^2}{4}d(\mathcal{P})^2, \tag{54}$$

Therefore, from (54), we may write

$$\begin{aligned}&\mathrm{mse}(\hat{\boldsymbol{X}}^*_\mu)\mathrm{sev}(\hat{\boldsymbol{X}}^*_\mu) - \mathrm{mse}(\hat{\boldsymbol{X}}^*_0)\mathrm{sev}(\hat{\boldsymbol{X}}^*_\infty) \\ &\quad = \Lambda_{\mathrm{sev}}(\mu)\mathrm{mse}(\hat{\boldsymbol{X}}^*_0) + \Lambda_{\mathrm{mse}}(\mu)\mathrm{sev}(\hat{\boldsymbol{X}}^*_\infty) \\ &\quad\quad + \Lambda_{\mathrm{mse}}(\mu)\Lambda_{\mathrm{sev}}(\mu) \\ &\quad \le \mathbb{U}(\mathcal{P}), \quad \forall \mu \in [0, +\infty].\end{aligned} \tag{55}$$

Lastly, when a $\rho_{\min}$ exists, again from (48) and Theorem 2 we may also fairly easily find that

$$\Lambda_{\mathrm{mse}}(\mu) \ge \rho_{\min}\mu^2\alpha(\mu)d(\mathcal{P})^2 \quad \text{and} \tag{56}$$

$$\Lambda_{\mathrm{sev}}(\mu) \ge \alpha(\mu)d(\mathcal{P})^2, \tag{57}$$

and thus in a similar manner obtain the lower bound $\mathbb{L}(\mathcal{P},\mu)$. Enough said. □

Theorem 3 implies that for sufficiently small $\varepsilon > 0$ for which $d(\mathcal{P}) < \varepsilon$, it is true that, *uniformly* over $\mu \in [0, +\infty]$,

$$\mathrm{mse}(\hat{\boldsymbol{X}}^*_\mu)\mathrm{sev}(\hat{\boldsymbol{X}}^*_\mu) \simeq \mathfrak{h}(\mathcal{P}) \simeq \mathrm{mse}(\hat{\boldsymbol{X}}^*_0)\mathrm{sev}(\hat{\boldsymbol{X}}^*_\infty). \tag{58}$$

In other words, when $d(\mathcal{P})$ is very small, we can select the risk aversion parameter μ almost freely and still achieve simultaneously both a good average performance and an adequate level of robustness; this is of course a feature of (near-)skew-symmetric models. On the contrary, highly skewed models displace the optimal trade-off $\mathfrak{h}(\mathcal{P})$ away from the ultimate lower bound, thus rendering the exchangeability between mse and sev highly nontrivial. Given fixed values of $\rho_{\min}$ and $\rho_{\max}$, Theorem 3 also implies that the optimal trade-off $\mathfrak{h}(\mathcal{P})$ is fully characterized by three numbers: $d(\mathcal{P})$, the minimum mse and the minimum sev. Next, we show that $d(\mathcal{P})$ admits simultaneously well-defined and intuitive topological *and* statistical interpretations, within a rigorous framework.

VI. Risk Margins as Complete Metrics and Measures of Skewness in High Dimensions

In what follows, denote the product of state and observable spaces as $S \triangleq \mathbb{R}^n \times \mathbb{R}^m$, and let $\mathsf{P}(S)$ be the set of all Borel probability measures on S. Also recall the risk margin-related functional $d : \mathsf{P}(S) \to \mathbb{R}_+$ defined in Theorem 3 as

$$\begin{aligned} d(\mathcal{P}) &\equiv 2\sqrt{\mathbb{E}_{\mathcal{P}_{\boldsymbol{Y}}}\{\mathbb{C}(\mathcal{P}_{\boldsymbol{X}|\boldsymbol{Y}})\}} \\ &= 2\sqrt{\mathbb{E}_{\mathcal{P}_{\boldsymbol{Y}}}\left\{\|\widehat{\boldsymbol{\Delta X}}(\mathcal{P}_{\boldsymbol{X}|\boldsymbol{Y}})\|^2_{\boldsymbol{\Sigma}^{\dagger}_{\boldsymbol{X}|\boldsymbol{Y}}}\right\}}, \end{aligned} \tag{59}$$

where we now explicitly highlight the dependence on the generative model $\mathcal{P} \equiv \mathcal{P}_{(\boldsymbol{X},\boldsymbol{Y})}$. Then, we consider the space

$$\mathsf{P}_{\mathbb{S}}(S) \triangleq \{\mathcal{P} \in \mathsf{P}(S) | d(\mathcal{P}) < \infty\}, \tag{60}$$

as well as the feasibility set

$$\mathcal{F} \triangleq \{\alpha \geq 0 | d(\mathcal{P}) = \alpha, \text{for some } \mathcal{P} \in \mathsf{P}_{\mathbb{S}}(S)\} \subseteq \mathbb{R}_+. \tag{61}$$

Our discussion will concentrate on endowing $\mathsf{P}_{\mathbb{S}}(S)$ with a topological structure based on appropriate handling of the functional d, resulting among other things in a meaningful and intuitive topological interpretation for the latter.

Indeed, for every number $\alpha \in \mathcal{F}$, take an arbitrary element $\mathcal{P}_\alpha \in \mathsf{P}_{\mathbb{S}}(S)$ such that $d(\mathcal{P}_\alpha) = \alpha$. Then, we may construct a measure-valued multifunction $\mathcal{C} : \mathcal{F} \rightrightarrows \mathsf{P}_{\mathbb{S}}(S)$ as

$$\mathcal{C}(\alpha) \triangleq \{\mathcal{P} \in \mathsf{P}(S) | \mathcal{P} \sim \mathcal{P}_\alpha\}, \tag{62}$$

which in turn defines an equivalence class of $\mathcal{P}_\alpha$ for each $\alpha \in \mathcal{F}$, as well as consider *any* selection of the multifunction $\mathcal{C}$, say $C : \mathcal{F} \to \mathsf{P}_{\mathbb{S}}(S)$, i.e., a measure-valued function such that $C(\cdot) \in \mathcal{C}(\cdot)$ on $\mathcal{F}$. Next, we define a set of *equivalence class representatives* as

$$\mathcal{R} = \mathrm{range}(C). \tag{63}$$

Based on our construction, one may always choose $C(\cdot) = \mathcal{P}_{(\cdot)}$ on $\mathcal{F}$, in which case $\mathcal{R} = \{\mathcal{P}_\alpha\}_{\alpha \in \mathcal{F}}$. There is a bijective mapping between $\mathcal{R}$ and the collection of equivalence classes $\{\mathcal{C}(\alpha)\}_{\alpha \in \mathcal{F}}$. Therefore, we may define the *canonical projection map*

$$\Pi(\mathcal{P}) = \arg\min_{\tilde{\mathcal{P}} \in \mathcal{R}} |d(\mathcal{P}) - d(\tilde{\mathcal{P}})| \in \mathcal{R}, \tag{64}$$

which maps every Borel measure $\mathcal{P}$ in $\mathsf{P}_{\mathbb{S}}(S)$ to its representative $\Pi(\mathcal{P})$ in $\mathcal{R}$ and equivalently, to its corresponding equivalence class. In other words, the canonical map Π *separates* or *partitions* $\mathsf{P}_{\mathbb{S}}(S)$ on the basis of the values of the risk margin statistic $d(\mathcal{P})$, for each $\mathcal{P} \in \mathsf{P}_{\mathbb{S}}(S)$.

Let us now define another related functional $d_{\mathbb{S}} : \mathsf{P}_{\mathbb{S}}(S) \times \mathsf{P}_{\mathbb{S}}(S) \to \mathbb{R}_+$ as

$$\begin{aligned} d_{\mathbb{S}}(\mathcal{P}, \mathcal{P}') &\triangleq \sqrt{\left|\mathbb{E}_{\mathcal{P}_{\boldsymbol{Y}}}\{\mathbb{C}(\mathcal{P}_{\boldsymbol{X}|\boldsymbol{Y}})\} - \mathbb{E}_{\mathcal{P}'_{\boldsymbol{Y}}}\{\mathbb{C}(\mathcal{P}'_{\boldsymbol{X}|\boldsymbol{Y}})\}\right|} \\ &= \sqrt{\left|(d(\mathcal{P}))^2 - (d(\mathcal{P}'))^2\right|}. \end{aligned} \tag{65}$$

It is easy to see that the pair $(\mathcal{R}, d_{\mathbb{S}})$ is a metric space. Then, $d_{\mathbb{S}}$ induces a topology on the representative set $\mathcal{R}$, which we suggestively call the *(hidden) skewed topology* on $\mathcal{R}$. Similarly, $(\mathsf{P}_{\mathbb{S}}(S), d_{\mathbb{S}})$ is a pseudometric space. In this case, we say that $d_{\mathbb{S}}$ induces the *(hidden) skewed pseudometric topology* on $\mathsf{P}_{\mathbb{S}}(S)$. In fact, we may prove more [21].

Theorem 4. *The metric space $(\mathcal{R}, d_{\mathbb{S}})$ is Polish, and the pseudometric space $(\mathsf{P}_{\mathbb{S}}(S), d_{\mathbb{S}})$ is pseudoPolish.*

Therefore, there is a standard topological structure induced by $d_{\mathbb{S}}$ (and thus by d) on $\mathsf{P}_{\mathbb{S}}(S)$ with a complete description and favorable properties in terms of separation, closeness and limit point behavior.

Under these structural considerations, it is then immediate to observe that for any given Borel measure $\mathcal{P} \in \mathsf{P}_{\mathbb{S}}(S)$ we have that $d(\mathcal{P}) = |(d(\Pi(\mathcal{P})))^2 - 0|^{1/2} = d_{\mathbb{S}}(\Pi(\mathcal{P}), \mathcal{P}_0)$, and therefore we may *interpret $d(\mathcal{P})$ as the distance of $\mathcal{P}$ relative to all equivalent to each other skew-symmetric Borel measures on S* (i.e., with $\|\widehat{\boldsymbol{\Delta X}}\|_2 = 0$ almost everywhere), which are precisely the measures for which risk-neutral and risk-aware estimators, $\hat{\boldsymbol{X}}_0$ and $\hat{\boldsymbol{X}}_\infty$ respectively, coincide and thus the corresponding mse and sev are simultaneously minimal. This fact is significant, not only because it provides a clear topological meaning for the (expected) risk margin analyzed earlier in Section IV (Theorem 2), but also because $d(\mathcal{P})$ induces a similar interpretation to the optimal trade-off $\mathfrak{h}(\mathcal{P})$ via Theorem 3, and consequently completely characterizes the general $\mathrm{mse}/\mathrm{sev}$ trade-off of the uncertainty principle of Theorem 1.

Simultaneously, both functionals d and $d_{\mathbb{S}}$ admit a convenient and intuitive *statistical interpretation*, as well. To see this, let us consider the simplest case of a totally hidden, real-valued state variable, say X. We have $\mathcal{P} \equiv \mathcal{P}_{X|\boldsymbol{0}} \equiv \mathcal{P}_X$, where $\boldsymbol{0}$ denotes a fictitious trivial observation. Then, denoting the mean and variance of X as μ and σ^2, respectively, $d(\mathcal{P})$ may be expressed as

$$\begin{aligned} d(\mathcal{P}) &= 2\frac{1}{\sigma}\left|\widehat{\boldsymbol{\Delta} X}(\mathcal{P}_X)\right| \\ &= 2\frac{1}{\sigma}\left|\mathbb{E}\{X\} - \frac{\mathbb{E}\{X^3\} - \mathbb{E}\{X^2\}\mathbb{E}\{X\}}{2\sigma^2}\right| \\ &= \left|\frac{2\sigma^2\mu - \mathbb{E}\{X^3\} + (\mu^2 + \sigma^2)\mu}{\sigma^3}\right| \\ &= \left|\frac{-\mathbb{E}\{X^3\} + \mu^3 + 3\sigma^2\mu}{\sigma^3}\right|, \end{aligned} \tag{66}$$

or, equivalently,

$$d(\mathcal{P}) = \left|\mathbb{E}\left\{\left(\frac{X - \mu}{\sigma}\right)^3\right\}\right|, \tag{67}$$

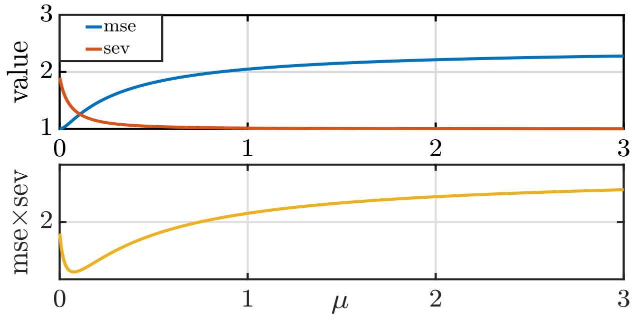

Fig. 1. Normalized (mse)/(sev) and their corresponding product in case of state-dependent noise.

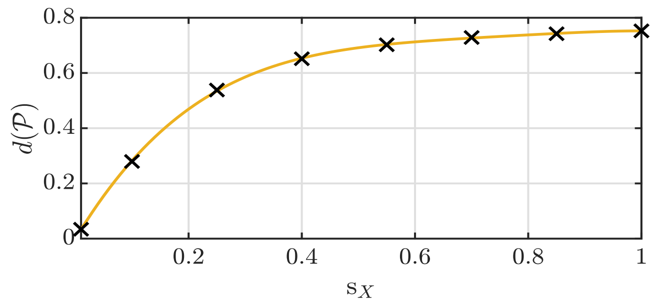

Fig. 2. Measure of skewness as a function of the parameter s_X.

which is nothing but the *absolute value of Pearson's moment coefficient of skewness* (i.e., excluding directionality). In other words, Pearson's moment cofficient of skewness may be interpreted itself as the *difference of a pair of optimal estimators; these are the mean of X (in the MMSE sense), and the maximally risk-averse estimator of X, optimally biased towards the tail of the distribution $\mathcal{P}_X$*. Further, via our topological interpretation of $d(\mathcal{P})$, Pearson's moment coefficient of skewness *expresses, in absolute value, the distance (in a topologically consistent sense) of the distribution of X relative to any non-skewed distribution on the real line*, with the most obvious representative being $\mathcal{N}(0, 1)$.

Consequently, the risk margin functional d (intuitively a scalar quantity) may be thought as a measure of skewness *magnitude* in multiple dimensions, corresponding to a consistent non-directional generalization of Pearson's moment skewness coefficient, also fully applicable the hidden state model setting, and tacitly exploiting statistical dependencies of both the conditional and marginal measures $\mathcal{P}_{\boldsymbol{X}|\boldsymbol{Y}}$ and $\mathcal{P}_{\boldsymbol{Y}}$. In the same fashion, the risk margin (pseudo)metric $d_{\mathbb{S}}$ may be conventiently thought as a measure of the *relative skewness between (filtered) distributions*.

VII. Numerical Simulations and Discussion

Our theoretical claims are now justified through indicative numerical illustrations, along with a relevant discussion. We justify our claims by presenting the following working examples: First, we consider the problem of inferring an exponentially distributed hidden state X, with $\mathbb{E}\{X\} = 2$ while observing $Y = X + v$ [15]. The random variable v expresses a state-dependent, zero-mean, normally distributed noise, whose (conditional) variance is given by $\mathbb{E}\{v^2|X\} = 9X^2$. Fig. 1 shows $\mathrm{mse}(\hat{\boldsymbol{X}}^*_\mu)$, $\mathrm{sev}(\hat{\boldsymbol{X}}^*_\mu)$, as well as their product $\mathrm{mse}(\hat{\boldsymbol{X}}^*_\mu)\mathrm{sev}(\hat{\boldsymbol{X}}^*_\mu)$, all with respect to the risk-aversion parameter μ. The former two have been normalized with respect to their corresponding minimum values while the product results after the aforementioned normalization step. From the figure, it is evident that the optimal trade-off (in the sense implied by Theorem 1) is attained close to the origin; note, though, that such an optimal μ^* does *not* correspond to the value of μ for which (normalized) mse and sev curves intersect.

Next, we consider the problem of estimating another real-valued hidden state X while observing $Y = X \times W$, with $(X, W) \sim \mathrm{Lognormal}(\boldsymbol{0}, \boldsymbol{S})$, $\boldsymbol{S} = \mathrm{diag}(s_X, 0.25)$. The variable $s_X > 0$ defines a parametric family of probability measures whose skewness increases with s_X. We would like to examine the impact of our theoretical results by varying the skewness of the aforementioned model. However, we are not aware that by increasing s_X the posterior skewness alters as well. In addition, even if skewness varies with s_X, the way it does so is not apparent. For these reasons, we employ our new distance/skewness measure to trial the model with respect to s_X. This experiment is shown in Fig. 2 where we verify that at least for the examined s_X-values, the average posterior skewness increases.

Fig. 3 illustrates how the profiles of $\mathrm{mse}(\hat{\boldsymbol{X}}^*_\mu)$, $\mathrm{sev}(\hat{\boldsymbol{X}}^*_\mu)$, and their product $\mathrm{mse}(\hat{\boldsymbol{X}}^*_\mu)\mathrm{sev}(\hat{\boldsymbol{X}}^*_\mu)$ scale with s_X. As above, we normalize $\mathrm{mse}(\hat{\boldsymbol{X}}^*_\mu)$ and $\mathrm{mse}(\hat{\boldsymbol{X}}^*_\mu)\mathrm{sev}(\hat{\boldsymbol{X}}^*_\mu)$ with respect to their minimum possible value, respectively, and $\mathrm{sev}(\hat{\boldsymbol{X}}^*_\mu)$ with respect to its maximum one. First, although the average performance deteriorates faster as the skewness increases (e.g., for the most skewed model, depicted in cyan), a 15% deterioration of mse corresponds to a 20% safety improvement, indicating that, there might be particular models allowing for an even more advantageous exchange.

Further, Fig. 3 shows that, for the smallest skewness level (blue), almost all risk-aware estimates achieve a near-optimal bound. As the skewness increases, the optimal -with respect to the product- estimators become strongly separated from each other within the class $\{\hat{\boldsymbol{X}}^*_\mu\}_\mu$. In this one-dimensional example, there is a unique optimal value for $\mu^\star$ with respect to the product; however, this might be only an exception to the rule, especially for higher-dimensional models. Note that a graphical representation of the product like the one depicted in Fig. 3 is all that we need to do to at least approximately determine the optimal value for μ (a single parameter).

Lastly, Fig. 4 presents the course of the upper bound $\mathbb{U}(\mathcal{P})$ with respect to the skewness parameter s_X. To clarify its behavior close to zero, we sample s_X additionally at 0.01 and 0.1. Expectedly, while $d(\mathcal{P})$ approaches zero, the bound approaches zero as well regardless of the chosen limit $\rho_{\max}$, and the values $\mathrm{mse}(\hat{\boldsymbol{X}}^*_0)$, and $\mathrm{sev}(\hat{\boldsymbol{X}}^*_\infty)$.

VIII. Conclusion

This work quantified the inherent trade-off between mse and sev by lower bounding the product between the two over all admissible estimators. Provided a level of performance (resp. risk), the introduced uncertainty relation reveals the

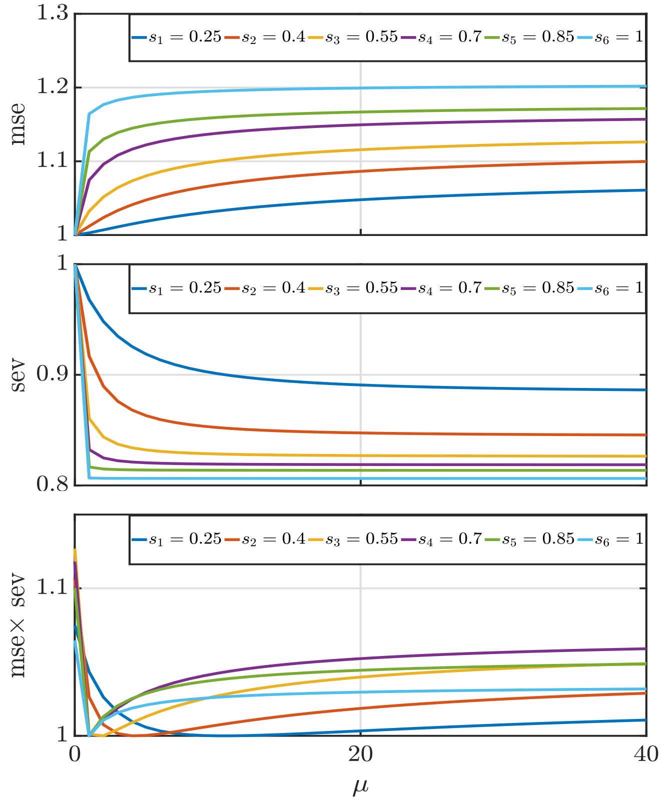

Fig. 3. Up (Center): mse (sev) percent increase (decrease) relative to risk-aversion parameter μ for different skewness levels. Down: Normalized trade-off relative to risk-aversion parameter μ for different skewness levels.

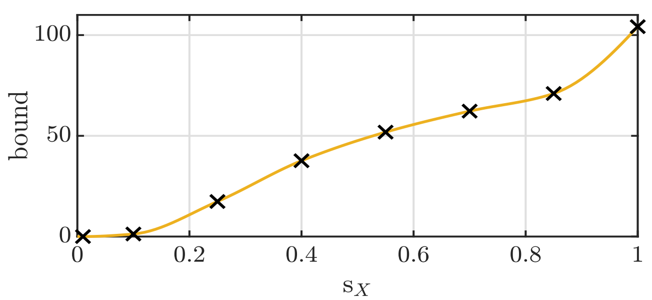

Fig. 4. The bound $\mathbb{U}(\mathcal{P})$ as a function of the parameter s_X. In this example $\rho_{\max} = 10$.

minimum risk (resp. performance) tolerance for the problem and assesses how effective any estimator is with respect to the optimal Bayesian trade-off. Projecting the risk-averse stochastic μ-parameterized curve on the link between the MMSE and the maximally risk-averse estimator, we defined as analyzed the so-called hedgeable risk margin of the model. Its significance stems from the fact that it admits both a rigorous topological and an intuitive statistical interpretations, fitting our risk-aware estimation setting. In particular, the risk margin functional induces a new measures of the skewness of the conditional evidence regarding the state provided the observables. Connecting the dots, we showed that the optimal trade-off is order-equivalent to this new measure of skewness, thus fully characterizing our uncertainty principle from a statistical perspective.

References

[1] S.-K. Kim, R. Thakker, and A.-A. Agha-Mohammadi, "Bi-directional value learning for risk-aware planning under uncertainty," *IEEE Robotics and Automation Letters*, vol. 4, no. 3, pp. 2493–2500, 2019.

[2] A. Wang, X. Huang, A. Jasour, and B. Williams, "Fast risk assessment for autonomous vehicles using learned models of agent futures," *arXiv preprint arXiv:2005.13458*, 2020.

[3] W.-J. Ma, C. Oh, Y. Liu, D. Dentcheva, and M. M. Zavlanos, "Risk-averse access point selection in wireless communication networks," *IEEE Transactions on Control of Network Systems*, vol. 6, no. 1, pp. 24–36, 2018.

[4] M. Bennis, M. Debbah, and H. V. Poor, "Ultrareliable and low-latency wireless communication: Tail, risk, and scale," *Proceedings of the IEEE*, vol. 106, no. 10, pp. 1834–1853, 2018.

[5] Y. Li, D. Guo, Y. Zhao, X. Cao, and H. Chen, "Efficient Risk-Averse Request Allocation for Multi-Access Edge Computing," *IEEE Communications Letters*, pp. 1–1, sep 2020.

[6] A. R. Cardoso and H. Xu, "Risk-Averse Stochastic Convex Bandit," in *International Conference on Artificial Intelligence and Statistics*, apr 2019, vol. 89, pp. 39–47.

[7] D. Dentcheva, S. Penev, and A. Ruszczyński, "Statistical estimation of composite risk functionals and risk optimization problems," *Annals of the Institute of Statistical Mathematics*, vol. 69, no. 4, pp. 737–760, 2017.

[8] M. P. Chapman, J. Lacotte, A. Tamar, D. Lee, K. M. Smith, V. Cheng, J. F. Fisac, S. Jha, M. Pavone, and C. J. Tomlin, "A Risk-Sensitive Finite-Time Reachability Approach for Safety of Stochastic Dynamic Systems," in *Proceedings of the American Control Conference*. jul 2019, vol. 2019-July, pp. 2958–2963, Institute of Electrical and Electronics Engineers Inc.

[9] S. Samuelson and I. Yang, "Safety-Aware Optimal Control of Stochastic Systems Using Conditional Value-at-Risk," in *Proceedings of the American Control Conference*. aug 2018, vol. 2018-June, pp. 6285–6290, Institute of Electrical and Electronics Engineers Inc.

[10] R. Williamson and A. Menon, "Fairness risk measures," in *International Conference on Machine Learning*. PMLR, 2019, pp. 6786–6797.

[11] J. Tao, L. Shao, Z. Guan, W. Ho, and S. Talluri, "Incorporating risk aversion and fairness considerations into procurement and distribution decisions in a supply chain," *International Journal of Production Research*, vol. 58, no. 7, pp. 1950–1967, 2020.

[12] M. Gürbüzbalaban, A. Ruszczyński, and L. Zhu, "A Stochastic Subgradient Method for Distributionally Robust Non-Convex Learning," *arXiv preprint, arXiv:2006.04873*, jun 2020.

[13] S. Curi, K. Levy, S. Jegelka, A. Krause, et al., "Adaptive sampling for stochastic risk-averse learning," *arXiv preprint arXiv:1910.12511*, 2019.

[14] D. S. Kalogerias, "Noisy Linear Convergence of Stochastic Gradient Descent for CV@ R Statistical Learning under Polyak-Łojasiewicz Conditions," *arXiv preprint arXiv:2012.07785*, 2020.

[15] D. S. Kalogerias, L. F. O. Chamon, G. J. Pappas, and A. Ribeiro, "Better Safe Than Sorry: Risk-Aware Nonlinear Bayesian Estimation," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 5480–5484.

[16] C. Cohen-Tannoudji, B. Diu, and F. Laloe, "Quantum mechanics. Vol. 2," 2008.

[17] D. Vakmann, *Sophisticated signals and the uncertainty principle in radar*, vol. 4, Springer Science & Business Media, 2012.

[18] A. Shapiro, D. Dentcheva, and A. Ruszczyński, *Lectures on Stochastic Programming: Modeling and Theory*, Society for Industrial and Applied Mathematics, 2nd edition, 2014.

[19] J. L. Speyer and W. H. Chung, *Stochastic Processes, Estimation, and Control*, vol. 17, Siam, 2008.

[20] D. S. Kalogerias, L. F. O. Chamon, G. J. Pappas, and A. Ribeiro, "Risk-Aware MMSE Estimation," *arXiv preprint, arXiv:1912.02933, to be submitted*, 2020.

[21] N. P. Koumpis and D. S. Kalogerias, "Uncertainty Principles in Risk-Aware Statistical Estimation," *arXiv preprint*, 2021.