How Biased Is Your Model? Concentration Inequalities, Information and Model Bias

Konstantinos Gourgoulias, Markos A. Katsoulakis, Luc Rey-Bellet, and Jie Wang

***Abstract*— We derive tight and computable bounds on the bias of statistical estimators, or more generally of quantities of interest, when evaluated on a baseline model P rather than on the typically unknown true model Q. Our proposed method combines the scalable information inequality derived by P. Dupuis, K.Chowdhary, the authors and their collaborators together with classical concentration inequalities (such as Bennett's and Hoeffding-Azuma inequalities). Our bounds are expressed in terms of the Kullback-Leibler divergence $R(Q\|P)$ of model Q with respect to P and the moment generating function for the statistical estimator under P. Furthermore, concentration inequalities, i.e. bounds on moment generating functions, provide tight and computationally inexpensive model bias bounds for quantities of interest. Finally, they allow us to derive rigorous confidence bands for statistical estimators that account for model bias and are valid for an arbitrary amount of data.**

***Index Terms*— Uncertainty quantification, information theory, information bounds, model bias, model uncertainty, goal-oriented divergence, concentration inequalities, Kullback-Leibler divergence, statistical estimators.**

I. Introduction

AN ESSENTIAL ingredient of predictive modeling is the reliable calculation of specific statistics/quantities of interest of the predictive distribution. Such statistics are typically tied to the application domain, for instance, moments, covariance, failure probabilities, extreme events, arrival times, average velocity, energy and so on. Predictive models can involve (a) statistical aspects or data collection, and (b) physical/mathematical mechanisms with choices in complexity/resolution, some of them potentially computationally intractable. Therefore, to improve the predictive capabilities of models we face fundamental trade-offs between model complexity, amount of available data, computational efficiency, and model bias.

The main focus of the paper is the understanding and control of model bias which often inevitably occurs in model building and which is itself a measure of reliable predictions. Our primary tool are information-theoretic Uncertainty Quantification methods. Uncertainty Quantification (UQ) methods address questions related to model selection, model sensitivity, model reduction and misspecification, [1]–[3]. Sources of uncertainty are broadly classified in two categories: aleatoric, due to the inherent stochasticity of probabilistic models and the limited availability of data, and epistemic, stemming from the inability to accurately model all aspects of a complex system, [2], [4], [5]. Model bias is closely related to epistemic uncertainty, and probability metrics (Wasserstein, total variation) and divergences (Kullback-Leibler, Rényi, χ^2) [6] are important tools to quantify uncertainty by comparing models. Among the divergences, the Kullback-Leibler (KL) divergence (also known as relative entropy) is widely used because of its computational tractability. Specifically, KL-based methods have been used successfully in variational inference and expectation propagation [7], model selection [8], model reduction (coarse-graining) [9]–[12], optimal experiment design, [13], and UQ [14]–[17].

Information-theoretic methods for model building will typically induce bias for the various statistics and the quantities of interest (QoIs) of the predictive distribution compared to the "true" model–if known–or the available data. Managing the corresponding trade-offs between a range of less biased but more computationally expensive models naturally leads to the following main question for the paper:

Can we provide performance guarantees for model bias in models built via KL-based approximate inference, model misspecification, or model selection methods?

In this paper we ultimately seek to understand how a decrease in KL-divergence–associated with an increase in modeling and/or computational effort–can guarantee a model bias tolerance; and in addition, we seek the tightest possible control of model bias. Note that bounds on the model bias of a QoI between two distributions P and Q can be obtained, for example, in terms of their KL or χ^2 divergences using the classical Pinsker or Chapman-Robbins inequalities

Manuscript received June 29, 2017; revised January 31, 2020; accepted February 5, 2020. Date of publication February 28, 2020; date of current version April 21, 2020. The work of Konstantinos Gourgoulias was supported by the Office of Advanced Scientific Computing Research, U.S. Department of Energy, under Contract DE-SC0010723. The work of Markos A. Katsoulakis and Luc Rey-Bellet was supported in part by the National Science Foundation (NSF) under Grant DMS-1515712 and in part by the Air Force Office of Scientific Research (AFOSR) under Grant FA-9550-18-1-0214. The work of Jie Wang was supported by the Defense Advanced Research Projects Agency (DARPA) EQUiPS Program under Grant W911NF1520122. *(Corresponding author: Jie Wang.)*

Konstantinos Gourgoulias was with the Department of Mathematics and Statistics, University of Massachusetts Amherst, Amherst, MA 01003 USA. He is now with Babylon Health, London SW3 3DD, U.K. (e-mail: kostis.gourgoulias@babylonhealth.com).

Markos A. Katsoulakis, Luc Rey-Bellet, and Jie Wang are with the Department of Mathematics and Statistics, University of Massachusetts Amherst, Amherst, MA 01003 USA (e-mail: markos@math.umass.edu; luc@math.umass.edu; wang@math.umass.edu).

Communicated by A. Gohari, Associate Editor At Large.

Color versions of one or more of the figures in this article are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TIT.2020.2977067

respectively, [6], [18]. Clearly a decrease in divergence will improve bounds on the model bias. However, these classical inequalities are typically *non-tight* and *non-discriminating*, in the sense that they scale poorly with the size of data sets, with the number of variables in high-dimensional models (e.g. molecular systems), or with time in the context of stochastic processes; we refer to Sections 2.2–2.3 in [19] for a complete discussion, see also the example in Remark 21.

To tackle these challenges a class of new information inequalities have been introduced in [4] and further developed in [19], [20] by the authors and their collaborators (see also [21]–[23]). The resulting bounds on model bias bounds involve (a) the KL divergence $R(Q\|P)$[1] between a baseline model P and an alternative models Q, and (b) the moment generating function (MGF) for the QoI under the baseline model P. This inequality inherits the asymmetry of $R(Q\|P)$, which in turn allows us to exchange the roles of P and Q, depending on the context and/or availability of data from the true model Q or samples from the baseline model P. Considering a neighborhood of models around the baseline P, defined by the KL divergence $R(Q\|P)$, can be associated with a specified error tolerance and is non-parametric in nature. The crucial mathematical ingredient behind the inequality is the Donsker-Varadhan variational principle [24, Appendix C.] for the KL divergence, also known as the Gibbs variational formula [25]. This variational representation actually implies that the new inequalities are *tight*, i.e. they become an equality for a suitable model Q within a KL divergence neighborhood of the baseline model P. Furthermore, the dependence on the MGF renders the bounds scalable and *discriminating* for high-dimensional data sets and models, e.g. Markov Random Fields, long-time dynamics of stochastic processes and molecular models, as demonstrated recently in [19]. Also, broadly related methods in model misspecification and sensitivity analysis in financial risk measurement and queuing theory, using a robust optimization perspective, were proposed recently in [23] and [22], we also refer to references therein for other related work in operations research, finance and macroeconomics. Finally we also mention a related information theoretic approach developed in [26], [27] which provides uncertainty quantification bounds for rare events using a variational representation for the Rényi divergence which generalizes the Gibbs variational principle.

The primary goal of this paper is to use these new theoretical advances to develop practical tools to estimate and control model bias, and this raises new theoretical questions and implementation challenges. In particular evaluating or estimating MGFs can be very costly due to high variance of the estimators thus requiring either a large amount of data, see also Table 1, or multi-level/sequential Monte Carlo methods [7], [28]–[30]. In this paper we rather pursue the use of a variety of QoI-dependent *concentration inequalities* [31]–[33] to bypass the evaluation or estimation of the MGF and this leads to computable, tight bounds for model bias. Concentration inequalities are a fundamental mathematical tool in the study of rare events [34], model selection methods [35], statistical mechanics [33, Section 8.4] and random matrix theory [36]. Usually concentration inequalities are used to bound tail events, i.e. to provide bounds on the probability that a random variable deviates from typical behavior. In this paper we use concentration inequalities for the purpose of uncertainty quantification, specifically to control model bias, by efficiently implementing the new information inequalities developed in [4], [19], [20], while at the same time maintaining and expanding their theoretical advantages.

The new inequalities proved in this paper— which we call *concentration/information* inequalities—combine concentration inequalities with the variational principles underlying the bounds and lead to model bias bounds with the following key features:

(a) Easily computable bounds in terms of simple properties of the QoIs such as their mean, upper and lower bounds, suitable bounds on their variance, and so on; that is, without requiring the costly computation of MGFs.
(b) Scalability for QoIs that depend on large numbers of data such as statistical estimators, or for high dimensional probabilistic models.
(c) Derivation of rigorous confidence bands for statistical estimators that account for model bias and are valid for an arbitrary amount of data.
(d) Applicability to families of QoIs satisfying a concentration inequality, and not to just a single QoI.
(e) Tightness of the model bias bounds in the sense that the bounds are always attained within a prescribed KL-divergence *and* the class of QoIs in (d).

We also refer to the recent paper [37] where we applied the concentration inequalities ideas developed here to obtain robust uncertainty bounds for QoIs of random partial differential equations in subsurface flow problems, trained from sparse data. Furthermore, in the preprint [38] ideas in the present paper are developed further to provide uncertainty quantification bounds for path space QoIs for Markov process where concentration inequalities are intimately related to functional inequalities such as Poincaré and log-Sobolev.

The structure of the paper is as follows. In Section II we set-up the mathematical framework for the paper and discuss the information inequalities for QoIs of [4], [19], [20]. In Section III we use concentration inequalities to derive new concentration/information inequalities on model bias that are typically straightforward to implement. In Section IV, we discuss the tightness properties of the new concentration/information bounds. Finally in Section V we study the bias of statistical estimators, noting that such QoIs will require results that scale properly with the amount of available data. We also illustrate the bounds in a variety of examples. In Section VI we consider two elementary examples with bounded or unbounded QoIs. Two examples of systems with epistemic uncertainty are discussed in Section VII; the first one deals with failure probabilities for batteries and the second with Markov Random Fields such as Ising systems.

II. Tight Model Bias Bounds Using KL Divergence

In coarse-graining, model reduction, model selection, or variational inference, as well as in other uncertainty

[1]Also often denoted by $D(Q\|P)$ in the literature.

quantification and approximate inference problems, a baseline model P is compared to a "true" or simply a different model Q. In this case the notion of *risk or mean square error* plays a key role in assessing the quality of the corresponding estimators. Namely, if $\hat{f}$ is an unbiased estimator of the quantity of interest f for the true model Q (but not of the baseline model P) then the risk of the estimator is the mean squared error

$$\text{RISK} := \mathbb{E}_Q[(\hat{f} - \mathbb{E}_P[f])^2] = \underbrace{\text{Var}_Q[\hat{f}]}_{\text{Variance}} + \underbrace{|\mathbb{E}_P[f] - \mathbb{E}_Q[f]|^2}_{\text{ModelBias}}, \tag{1}$$

where we assume that first and second moments of f with respect to P, Q exist.

The main goal of this work is to understand how to transfer quantitative results on information metrics, specifically the KL divergence $R(Q\|P)$ (also known as relative entropy), to *bounds on the bias for quantities of interest* f. We formulate the corresponding mathematical problem next.

Mathematical Formulation. Let us consider a baseline model given by the probability measure P on the state space $\mathcal{X}$ which we assume to be a Polish (i.e. complete separable metric) space and we consider a QoI f, that is a measurable function $f : \mathcal{X} \to \mathbb{R}$. We specify next a family of alternative probability distributions in terms of the Kullback-Leibler (KL) divergence (or relative entropy) $R(Q\|P)$, which is defined as

$$R(Q\|P) = \mathbb{E}_Q[\log \frac{dQ}{dP}], \tag{2}$$

if Q is absolutely continuous with respect to P (and $+\infty$ otherwise). Note that $R(Q\|P)$ has the properties of a divergence that is $R(Q\|P) \geq 0$ for all Q and $R(Q\|P) = 0$ if and only if $Q = P$, see e.g. [18].

We fix a positive number η which we interpret as a level of *model misspecification*, quantified in terms KL divergence or, alternatively, as an information loss tolerance level between the baseline model P and alternative models described Q. We then define the set of alternative models as

$$\mathcal{Q}_\eta = \{Q : R(Q\|P) \leq \eta^2\}. \tag{3}$$

and any $Q \in \mathcal{Q}_\eta$ is referred to as an η-*admissible model*. We remark that our approach is *non-parametric*, i.e. it does not rely on any parametric form of the probability distributions considered. The relative entropy $R(Q\|P)$ is convex and lower-semicontinuous in (Q, P). In general the set $\mathcal{Q}_\eta$ is infinitely dimensional, but it is compact with respect to the weak topology, [24]. The fact that the KL divergence is not symmetric in its arguments can be advantageous in some situations. For example, in variational inference, it naturally imposes a constraint on the support of the possible approximations Q of a target model P [7].

Our primary mathematical challenge in this work lies in quantifying the model bias in (1) if we use an η-admissible model in Q_η rather than the baseline model P. That is we need to

$$\begin{aligned} &\text{Compute (or estimate)} \ \sup_{Q\in\mathcal{Q}_\eta} \{\mathbb{E}_Q[f] - \mathbb{E}_P[f]\} \\ &\text{and} \ \inf_{Q\in\mathcal{Q}_\eta} \{\mathbb{E}_Q[f] - \mathbb{E}_P[f]\}. \end{aligned} \tag{4}$$

Note that this approach is intrinsically goal-oriented since it includes not only a family of alternative models Q but also a specific choice of QoI f. In this context, for a fixed f the sup and the inf in (4) are attained and can be explicitly computed, see Theorem 2.

Goal-oriented divergence. We now define a divergence which incorporates the QoI f and hence is called goal-oriented; it was first introduced in the current form in [20] based on earlier work in [4]. Consider a QoI f and the moment-generating function (MGF)

$$M_P(c; \tilde{f}) := \mathbb{E}_P[e^{c\tilde{f}}] \tag{5}$$

of the centered QoI $\tilde{f}$,

$$\tilde{f}(x) := f(x) - \mathbb{E}_P[f]. \tag{6}$$

In general (see e.g. [34] for details) the MGF $M_P(c; \tilde{f})$ is finite for c in some interval I and equal to $+\infty$ otherwise. Throughout this paper we will make the standing assumption that $M_P(c; \tilde{f})$ is finite in the interval $I = (d_-, d_+)$ with $d_- < 0 < d_+$, then under this assumption [34], $M_P(c; \tilde{f})$ is C^∞ in I and strictly convex in I (unless f is constant) and f has finite moments of any order. We next define the goal-oriented (GO) divergence as

$$\Xi(Q\|P; f) = \inf_{c>0} \left\{ \frac{1}{c} \log M_P(c; \tilde{f}) + \frac{1}{c} R(Q\|P) \right\}. \tag{7}$$

for P, Q with $R(Q\|p) < \infty$. Note that if d_+ is finite then the infimum can be taken on $(0, d_+)$ and note also that if $R(Q\|P) = \infty$ then the goal oriented divergence can naturally be then set equal to $+\infty$. We note that for distributions such as Cauchy or lognormal, the MGF does not exist for $c \neq 0$, and the presented approach cannot be applied. For such cases new ideas become necessary, possibly in the same general spirit.

In [4], [20], [22] the following bound on the model bias was proved, along with certain mathematical properties. In the sequel the O notation signifies a quantity $O = O(x)$ such that $|O(x)| \leq C|x|$ for some positive constant C. Unless otherwise stated, $M_P(c; \tilde{f})$ will be finite in a neighborhood of the origin and set to be infinite everywhere else.

Theorem 1: Let P be a probability measure and let f be such that its MGF $M_P(c; \tilde{f})$ is finite in a neighborhood of the origin. Then for any Q with $R(Q\|P) < \infty$ we have

$$-\Xi(Q\|P; -f) \leq \mathbb{E}_Q[f] - \mathbb{E}_P[f] \leq \Xi(Q\|P; f). \tag{8}$$

The GO divergence $\Xi(Q\|P; f)$ has the following properties

1) *Divergence:* $\Xi(Q \mid\mid P; f) \geq 0$ and $\Xi(Q \mid\mid P; f) = 0$ if and only if either $Q = P$ or f is constant P-a.s.
2) *Asymptotics of* $\Xi(Q\|P; f)$ *in* $R(Q\|P)$:

$$\begin{aligned} \Xi(Q \mid\mid P; \pm f) &= \sqrt{\text{var}_P[f]}\sqrt{2R(Q \mid\mid P)} \\ &\pm \frac{1}{3}\gamma(f)\sqrt{\text{var}_P[f]}R(Q\|P) + O\Big((R(Q\|P))^{3/2}\Big) \end{aligned} \tag{9}$$

where $\gamma(f) = \mathbb{E}_P\big[(f-\mathbb{E}_P[f])^3\big]/\mathrm{var}_P[f]^{3/2}$ is the skewness of f. In particular

$$|\,\mathbb{E}_Q(f)-\mathbb{E}_P(f)\,| \le \sqrt{\mathrm{var}_P[f]}\sqrt{2R(Q\,||\,P)} + O(R(Q\,||\,P)). \tag{10}$$

Proof: For completeness the proof of properties 1 and 2 in Theorem 1 is given in the Appendix. We explain here how to obtain the bound (8) which plays a central role in the paper. The starting point is the *Gibbs variational principle* (see e.g [24] for a proof) which relates MGF and KL divergence via convex duality: provided $\mathbb{E}_p[e^f]$ is finite we have

$$\log \mathbb{E}_P[e^f] = \sup_{Q \ll P} \{\mathbb{E}_Q[f] - R(Q\|P)\} \tag{11}$$

where the sup is taken over the measures Q that are absolutely continuous with respect to P. From this we obtain that $\mathbb{E}_Q[f] \le \log \mathbb{E}_P[e^f] + R(Q\|P)$ for any Q with $R(Q\|P) < \infty$ and replacing f by $\pm c(f-\mathbb{E}_P[f])$ with $c>0$ we obtain

$$\pm c(\mathbb{E}_Q[f]-\mathbb{E}_P[f]) \le \log M_P(\pm c;\tilde{f}) + R(Q\|P); \tag{12}$$

optimizing over c gives the bounds (8). □

Note that it is often useful to translate the QoI, $\tilde{f}$, by some constant a. For such a translation, the goal-oriented divergence satisfies:

$$\Xi(Q\|P;f) = \inf_{c>0}\left\{\frac{1}{c}\log M_P(c;\tilde{f}-a) + \frac{1}{c}R(Q\|P)\right\} + a. \tag{13}$$

Re-centering the QoI can help to avoid numerical issues when estimating $\log M_P(c;\tilde{f})$ from samples $x_1,\dots,x_n \sim P$. This is often referred to as the "log-sum-exp trick" in the literature and a common choice for a is $\max\{f(x_1),\dots,f(x_n)\}$.

Tightness of goal-oriented divergence. Our next result complements Theorem 1 and demonstrates the tightness of the GO divergence bounds (8) for the bias of a QoI f; for the proof, we refer to Appendix. See [4] and [20] for earlier versions of that tightness result. To state our result we introduce the exponential family P^c given by

$$\frac{dP^c}{dP} = e^{cf-\log M_P(c;f)} = \frac{e^{cf}}{\mathbb{E}_P[e^{cf}]}, \tag{14}$$

which is well-defined for c in the interval $I=(d_-,d_+)$ where $M_P(c;f)$ is finite.

Theorem 2: Let P be a probability measure and let f be such that the MGF $M_P(c;\tilde{f})$ is finite in a neighborhood of the origin. Let $\mathcal{Q}_\eta = \{Q : R(Q\|P) \le \eta^2\}$ be the set all probability measures Q within a KL tolerance η^2 of P. Then there exist $\eta_\pm$, $0<\eta_\pm\le\infty$, such that for any $\eta \le \eta_\pm$ there are probability measures $Q^\pm = Q^\pm(\eta)$ that satisfy:

$$\Xi(Q^+\|P;f) = \mathbb{E}_{Q^+}[f]-\mathbb{E}_P[f] = \max_{Q\in\mathcal{Q}_\eta} \mathbb{E}_Q[f]-\mathbb{E}_P[f], \tag{15}$$

$$-\Xi(Q^-\|P;-f) = \mathbb{E}_{Q^-}[f]-\mathbb{E}_P[f] = \min_{Q\in\mathcal{Q}_\eta} \mathbb{E}_Q[f]-\mathbb{E}_P[f]. \tag{16}$$

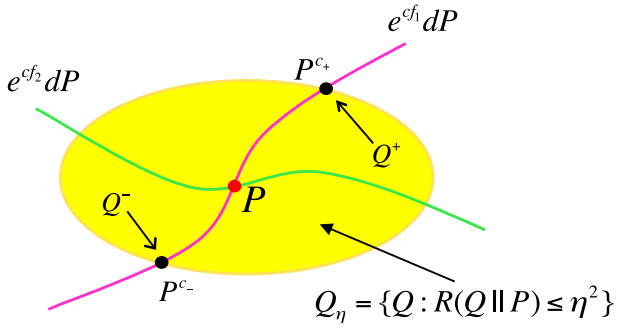

Fig. 1. The schematic depiction of Theorem 2 for the Quantities of Interest (QoIs) f_1, f_2 with tolerance η^2. The solid lines depict the one-parameter tilted probability distributions P^c (14) corresponding to the QoIs. The theorem implies that the upper and lower bounds in the family $\mathcal{Q}_\eta = \{Q : R(Q\|P) \le \eta^2\}$ are attained at the probability measures $Q^\pm = P^{c_\pm}$ for the QoI f_1.

The measures $Q_\pm$ are given by the elements $P^{c_\pm}$ of the exponential family (14) where $c_\pm$ are the unique solutions of

$$R(P^{c_\pm}\|P) = \eta^2. \tag{17}$$

Theorem 2 provides performance guarantees in the sense that, for all $Q\in\mathcal{Q}_\eta$, $\mathbb{E}_Q[f]$ belongs to the interval

$$-\Xi(Q^-\|P;-f)+\mathbb{E}_P[f] \le \mathbb{E}_Q[f] \le \mathbb{E}_P[f]+\Xi(Q^+\|P;f) \tag{18}$$

and the bounds are tight in $\mathcal{Q}_\eta$, in the sense that inequalities become equalities for $Q=Q^\mp$ respectively. This tightness property is crucial for our discussion because it implies that the GO divergence bounds in (8) are the best possible in the sense that they have attainable worst-case model scenarios $Q^\pm$ among all probability distributions Q within a KL tolerance $\eta^2>0$, see the schematic in Figure 1. The constants $\eta^\pm$ in Theorem 2 are often equal to $+\infty$ and may be finite only in special cases. The interested reader will find a detailed discussion of those cases, as well as a complete proof of Theorem 2 in the Appendix.

Remark 3: The tightness property (18) is a *non-parametric* result: the family $\mathcal{Q}_\eta$ of all alternative models Q cannot be parametrized in general and is only characterized by the property $R(Q\|P) \le \eta^2$. In spite of this non-parametric framework, we showed in Theorem 2 that the extremal models $Q^\pm$ that yield the tight bounds (18) belong to the parametrized family (14), see also Figure 1.

Remark 4 (Parametric vs. Non-Parametric UQ): The proposed bounds (18) can be pessimistic when considering uncertainty/sensitivity questions for models confined to a particular parametric family $\mathcal{Q}^{\text{PAR}} = \{Q^\theta : \theta\in\Theta\}$ over a parameter set Θ. Indeed, since the bounds (18) proposed above are based on the KL divergence, they are necessarily non-parametric and thus the resulting family of distributions $\mathcal{Q}_\eta$ allows for densities that may not be attainable within the particular parametric family. For example, if we already know that the set of models $\mathcal{Q}^{\text{PAR}}$ is a subset of a fixed parametric family, e.g. Q^θ's correspond to Gaussians, Poisson, or multinomial random variables, our non-parametric bounds (18) can be too wide since the family $\mathcal{Q}_\eta$ includes many other distributions outside

TABLE I
FOR THE ESTIMATION OF $\mathrm{var}_P[Y]$, WE ASSUME THAT $\mathbb{E}_P[Y]$ IS UNKNOWN AND THAT THE BIAS-ADJUSTED ESTIMATOR IS USED. FOR THE VARIANCE OF $\mathbb{E}_P[e^{cY}]$, A FIRST-ORDER APPROXIMATION IS USED (SEE [39]), ASSUMING THAT $\mathbb{E}_P[Y]$ IS SMALL

Quantity	Variance of estimator
$E_P[Y]$	$\mathrm{var}_P[Y]/n$
$\mathrm{var}_P[Y]$	$2(\mathrm{var}_P[Y])^2/(n-1)$
$M_P(c;Y)$	$c^2 e^{2c\mathbb{E}_P[Y]}\mathrm{var}_P[Y]/n$

the parametric family at hand, in the sense that the optimal $Q^{\pm}$ in Theorem 2 may not belong in $\mathcal{Q}^{\text{PAR}} = \{Q^\theta : \theta \in \Theta\}$.

On the other hand, if $Q^{\pm} \in \mathcal{Q}^{\text{PAR}}$ and $\mathcal{Q}^{\text{PAR}} \subset \mathcal{Q}_\eta$ then the bounds (18) are tight due to Theorem 2. An example in this direction are the exponential families, [7], where the QoIs f is any sufficient statistics. The exponential family is a very broad class of models, and the special case of Markov Random Fields and Ising models is discussed in Section VII-B; see also Section 4.1 in [20] for the linearized bounds in Theorem 1 in the case of the exponential family with sufficient statistics considered as QoIs.

Furthermore, in certain problems and due to the sparsity of available data–see for instance the battery failure probabilities in SectionVII-A–the family of alternative models $\mathcal{Q}_\eta$ is intrinsically non-parametric and is built around a baseline model P obtained, for example, through maximum likelihood or maximum a posteriori estimation. In that example the baseline model P is selected to be a Weibull distribution for the histogram of the battery failure times in Figure 7. But many alternative densities to P are possible, e.g. given by various choices of kernel density estimators of the histogram in Figure 7. Therefore considering the non-parametric family of models $\mathcal{Q} = \mathcal{Q}_\eta$ is a natural and, indeed, necessary choice.

The attractive properties of the GO bounds demonstrated in Theorem 1 and Theorem 2, come at a potentially significant cost since they require the knowledge or calculation of the MGF $M_P(c;\tilde{f})$ with respect to model P. If no simple formula for $M_P(c;\tilde{f})$ is known, this can be a data-intensive operation—compare the estimator variance of the MGF with that of other QoIs in Table I. Controlling the variance of an MGF estimator will require a large amount of data and/or the use of a multi-level Monte Carlo method, see also the discussion in Section I and Section VII. In the next section we introduce a new class of inequalities that share the aforementioned features of the GO divergence and satisfy Inequality (8), but they can bypass the estimation of an MGF by using the concept of concentration inequalities.

III. CONCENTRATION/INFORMATION INEQUALITIES FOR MODEL BIAS

To bypass the estimation or computation of the MGF in (7) we will use a QoI-dependent concentration bound for the MGF, i.e., a function $\Phi(c)$ taking values in $(0,\infty]$ such that

$$M_P(c;\tilde{f}) \le \Phi(c) \tag{19}$$

for all $c \ge 0$ and/or for $c \le 0$. Since the moment generating function $M_P(c;\tilde{f})$ can take the value $+\infty$ it is natural to allow the same for $\Phi(c)$.

Bounds of the form (19), for explicitly computable functions $\Phi(c)$, are called concentration inequalities and we discuss several such examples in Section III-A and Section III-C, as well as in Section V. Although we use only the simplest concentration inequalities here, the results are indicative to what can be accomplished using such information on f and P. In a continuation of this work, [38], using spectral gap estimate and log-Sobolev inequalities, concentration inequalities for path space observables for Markov processes are used to derive bounds for QoIs valid in the long-time regime. In upcoming work we will consider further applications for interacting particle systems with large number of degrees of freedom, arising in Kinetic Monte Carlo or molecular dynamics. Concentration inequalities is an important mathematical tool since they allow, via a Chernov bound, to control tail events, i.e. they provide explicit bounds on the probability that a random variable deviates from typical behavior. More specifically, such methods can address, among others, questions on rare events [34], model selection methods [35], statistical mechanics [33, Section 8.4] and random matrices [36]. Here we propose the use of concentration inequalities in tandem with the information inequalities (8) for uncertainty quantification and especially for providing model bias guarantees. In Theorem 5 we show how to construct new bounds for the model bias using a function Φ satisfying (19).

Theorem 5: Let P be a probability measure and let f be a QoI such that its MGF $M_P(c;\tilde{f})$ is finite in a neighborhood of the origin. Let $\Phi : \mathbb{R} \to (0,\infty]$ be a function with $\Phi(0) = 1$, $\Phi'(0) = 0$ and such that

$$M_P(c;\tilde{f}) \le \Phi(c) \tag{20}$$

for all $c \in \mathbb{R}$.

We define the set of admissible QoIs by

$$\mathcal{F}_P = \{g : M_P(c;\tilde{g}) \le \Phi(c)\}. \tag{21}$$

Then, $f \in \mathcal{F}_P$, and for every $Q \in \mathcal{Q}_\eta = \{Q : R(Q\|P) \le \eta^2\}$ and $g \in \mathcal{F}_P$ we have

$$-U_-(\eta;\mathcal{F}_P) \le \mathbb{E}_Q[g] - \mathbb{E}_P[g] \le U_+(\eta;\mathcal{F}_P), \tag{22}$$

where

$$U_\pm(\eta;\mathcal{F}_P) := \inf_{c>0}\left\{\frac{1}{c}\log\Phi(\pm c) + \frac{1}{c}\eta^2\right\}. \tag{23}$$

Proof: The proof follows immediately from the definition of GO divergence in (7), the bound (8) in Theorem 1, combined with the concentration inequality (20) and the definition of the admissible QoIs, $\mathcal{F}_P$. □

We discuss specific examples of inequalities of the type (20) and their corresponding admissible sets $\mathcal{F}_P$, in Sections III-A, III-B, and III-C below.

Remark 6 (Admissible set of QoIs): We note that the function Φ depends both on the QoI f and on P through (20) and therefore the set of admissible functions $\mathcal{F}_P$ also depends on the QoI f and on P. However, to keep notation simple, we suppress this dependence for both Φ and $\mathcal{F}_P$.

Remark 7 (Computing $U_{\pm}(\eta;\mathcal{F}_P)$): Some concentration bounds (20) such as the sub-Gaussian, sub-gamma and Hoeffding and Bernstein bounds discussed below provide explicit formulas for $U_{\pm}(\eta;\mathcal{F}_P)$, see for instance (32) and (38). However, in general—see the sharper Bennett bounds in (39) and (41)—we have an explicit formula for Φ but no explicit closed form solution of the optimization over c. However the elementary one-dimensional optimization in (23) can be easily carried out with standard solvers, e.g., Newton's method.

Divergence structure of $U_{\pm}(\eta;\mathcal{F}_P)$: The following properties of the bounds $U_{\pm}$ in (23) are analogous to the properties of the GO divergence (7) outlined in Theorem 1. One notable difference is that here the divergence structure defined by $U_{\pm}(\eta;\mathcal{F}_P)$ contains information about the entire family $\mathcal{F}_P$ in (21) and not just a single QoI f as was the case in the GO divergence (7).

Theorem 8: Under the assumptions of Theorem 5 and, in addition, if

$$\Phi(c) = M_{\bar{P}}(c;\tilde{h}), \tag{24}$$

i.e. $\Phi(c)$ is a MFG for some probability $\bar{P}$ and QoI h then $U_{\pm}(\eta;\mathcal{F}_P)$ satisfy:

1) *Divergence Properties:*
   a. $U_{\pm}(\eta;\mathcal{F}_P) \geq 0$, and
   b. $U_{\pm}(\eta;\mathcal{F}_P) = 0$ if and only if $\eta = 0$ or $\mathcal{F}_P$ is trivial, i.e. consists only of functions which are constant P-a.s.
2) *Linearization:* If $\Phi = \Phi(c)$ is twice differentiable in a neighborhood of $c = 0$, then we have the asymptotics $U_{\pm}(\eta;\mathcal{F}_P) = \sqrt{2\Phi''(0)}\eta + O(\eta^2)$ and thus,

$$|\mathbb{E}_Q[g] - \mathbb{E}_P[g]| \leq \sqrt{2\Phi''(0)}\eta + O(\eta^2) \quad \text{for all } g \in \mathcal{F}_P \text{ and all } Q \in \mathcal{Q}_\eta. \tag{25}$$

Proof: The proof follows from Theorem 1. Indeed since, by assumption, $\Phi(c) = M_{\bar{P}}(c;\tilde{h})$ we have

$$U_{\pm}(\eta;\mathcal{F}_P) = \Xi(Q\|\bar{P};\pm h) \tag{26}$$

for any probability Q such that $R(Q\|\bar{P}) = \eta^2$. Therefore, by Theorem 1, $U_{\pm}(\eta;\mathcal{F}_P) \geq 0$ and $U_{\pm}(\eta;\mathcal{F}_P) = 0$ if and only if $\eta = 0$ or h is constant $\bar{P}$ a.s. But if h is constant $\bar{P}$ a.s then $\Phi(c) = M_{\bar{P}}(c;\tilde{h}) = 1$ for all c and thus the set of admissible QoIs (21) becomes:

$$\mathcal{F}_P = \{g : M_P(c;\tilde{g}) \leq \Phi(c) = 1\}. \tag{27}$$

However for any $g \in \mathcal{F}_P$, by Jensen's inequality, $M_P(c;\tilde{g}) \geq 1$ since $\mathbb{E}_P[\tilde{g}] = 0$. Therefore the admissible set $\mathcal{F}_P$ consists only of constant functions thus g is constant P-a.s. Finally, the asymptotic result in (25) is proven in exactly the same way as for the GO divergence in Theorem 1, (see the proof in Appendix or in Section 3 of [20]). □

Theorem 5 and Theorem 8 motivate the following definition, in analogy to the goal oriented (GO) divergence (7) defined for a single QoI f:

Definition 9 (Concentration/Information Divergence): Given the notation and assumptions of Theorem 5 and Theorem 8, we define the concentration/information divergence between a baseline model P and the family of models $\mathcal{Q}_\eta$, satisfying (22) for all QoIs in $\mathcal{F}_P$:

$$U_{\pm}(\eta;\mathcal{F}_P) := \inf_{c>0}\left\{\frac{1}{c}\log\Phi(\pm c) + \frac{1}{c}\eta^2\right\}, \tag{28}$$

where $\mathcal{Q}_\eta$ and $\mathcal{F}_P$, are defined in (3) and (21) respectively.

Remark 10 (Features of Concentration/Information Inequalities): While the GO divergence bounds (8) are defined for a specific QoI f, key features of the new bounds in Theorem 5 include: (a) allow to consider whole families of admissible QoIs $\mathcal{F}_P$ defined in (21), and (b) they bypass the costly MGF calculations needed in the GO divergence (7). Finally, we next show that the new bounds (22) still share the advantages of the GO divergence bounds, namely: in Section IV we prove that, under suitable assumptions, (22) is, (c) tight in the family of models $\mathcal{Q}_\eta$, (3), and the family of QoIs $\mathcal{F}_P$, (21). in Section V we show that (22) is, (d) scalable to QoIs that depend on large numbers of data such as statistical estimators and to high dimensional probabilistic models.

Remark 11 (Why Concentration/Information Inequalities?): As shown in Lemma 2.11, Equation (2.28) of [20], the c^* that solves the optimization problem of the GO divergence bound in Equation (7) behaves like

$$c^* = c_1\eta + O(\eta^2), \tag{29}$$

for some explicit constant c_1 and $\eta^2 = R(Q\|P)$. Due to (29) and since estimator variance for the MGF increases exponentially with c, a larger uncertainty threshold η will quickly make the accurate estimation of $M_P(c^*;f)$ more demanding, as is readily clear from Table I and (29). This drawback becomes especially problematic when sampling from P is computationally expensive, e.g., requires MCMC sampling, P is multi-modal, etc. , see also the Markov Random Field example in Section VII-B, where sampling challenges can become more pronounced in higher dimensions. Even when P is simple to sample, as is the case with the baseline models in [40] and here, avoiding the estimation of $M_P(c^*;f)$ in the GO divergence can still save significant computational time, as Table I strongly suggests. For instance, the Bennett-(a,b) bound in (41) only requires (a) the bounds on the QoI, a, b, and (b) the expected value of the QoI with respect to P. Similarly, the Hoeffding bound (44) requires only the bounds on the QoI, a, b.

We will next discuss specific examples of the bound $\Phi(c)$ in the concentration bounds (20) and Theorem 5; furthermore, we also demonstrate how we can select such concentration bounds depending on the information we have regarding the distribution P. We divide our presentation into two cases, namely bounded and unbounded QoIs f.

A. Sub-Gaussian Bounds

For an unbounded QoI f and a probability distribution P, we can characterize the type of concentration by bounding either the tail probabilities $P(f(X) - \mathbb{E}_P[f] > a)$ for all a or $M_P(c;\tilde{f})$ for all c for which the MGF is finite. In this section, we discuss the (classical) sub-Gaussian concentration

bounds which are characterized by Gaussian decay of the tails. Sub-gamma bounds are discussed in Section III-B (see also Section VI-A); sub-Poissonian bounds could also be useful in various situations but we will not discuss them further here (see e.g. [33]).

Sub-Gaussian concentration bounds [31]: We say that $f = f(X)$ is a sub-Gaussian random variable if there exists a $\sigma_B > 0$ such that

$$M_P(c; \tilde{f}) \leq \Phi(c) := \exp(c^2\sigma_B^2/2) \text{ for all } c \in \mathbb{R}. \tag{30}$$

Now given a fixed σ_B, we can consider the family of QoIs defined in (21),

$$\mathcal{F}_P := \{g : M_P(c; \tilde{g}) \leq \Phi(c) = \exp(c^2\sigma_B^2/2)\}, \tag{31}$$

i.e. we consider all random variables with MGF bounded by the MGF of a normal random variable with variance σ_B^2. Furthermore, using (23) we can write an explicit formula for $U_\pm(\eta; \mathcal{F}_P) = \inf_{c>0}\{\frac{c\sigma_B}{2} + \frac{\eta^2}{c}\}$ as

$$U_\pm(\eta; \mathcal{F}_P) = \sigma_B\sqrt{2}\eta. \tag{32}$$

By expanding $M_P(c; \tilde{f})$ around $c = 0$, we can readily show that σ_B^2 is an upper bound of $\text{var}_P[f(X)]$. Relation (32) also implies that there is no η-admissible model $Q \in \mathcal{Q}_\eta$ for which the QoIs under consideration lie beyond the uncertainty region given by Theorem 5:

$$-\sigma_B\sqrt{2}\eta \leq \mathbb{E}_Q[g] - \mathbb{E}_P[g] \leq \sigma_B\sqrt{2}\eta \tag{33}$$

for all models $Q \in \mathcal{Q}_\eta$ and QoIs $g \in \mathcal{F}_P$. In Corollary 12, we consider the special case where P is a normal distribution which is compared against any models Q–possibly not normal–from $\mathcal{Q}_\eta$.

Corollary 12: Consider the QoI $f(x) = x$ where $P = N(\mu, \sigma^2)$. Also, let Q be any distribution such that $R(Q\|P) \leq \eta^2$. Then, if the coefficient of variation (also known as relative standard deviation) is $c_v := \sigma/|\mu|$, the relative model bias satisfies:

$$-c_v\sqrt{2}\eta \leq \frac{\mathbb{E}_Q[f] - \mathbb{E}_P[f]}{|\mathbb{E}_P[f]|} \leq c_v\sqrt{2}\eta. \tag{34}$$

In general, sub-Gaussianity is a strong assumption for an unbounded random variable. For example Poisson, gamma, and exponential random variables are not sub-Gaussian (see Section III-B). We also note that results like the McDiarmid's inequality, see Section V below, or the logarithmic Sobolev inequalities [32], [41], can provide values for the constant σ_B^2 for QoIs that satisfy specific properties, e.g., (60).

B. Sub-Gamma Bounds

We discuss here a bound which applies, in principle, to any QoI with a MGF finite in a neighborhood of the origin, that is to any QoI f which satisfy the conditions of Theorems 1 and Theorem 2. If $M_P(c; \tilde{f})$ is bounded for some $c > 0$ this implies at least exponential tails for the distribution of f and a typical example of random variables with (one-sided) exponential tails are the gamma random variables with parameters $a, b > 0$ which have the density $x^{a-1}e^{-x/b}/\Gamma(a)b^a$ ($\Gamma(a)$ is the Euler's Gamma function), mean ab, variance ab^2, and moment generating function $(1 - cb)^a$. Then for $f(X) = X$ and using the elementary inequality $-\log(1 - u) - u \leq u^2/2(1 - u)$ we have,

$$\log M_P(c; \tilde{f}) = -cab - a\log(1 - cb) \leq \frac{ab^2c^2}{2(1 - cb)}. \tag{35}$$

This calculation motivates the following definition.

Sub-gamma concentration bounds [31]: We say that $f = f(X)$ is a sub-gamma random variable if there exists constants $\sigma_B > 0$ and $b > 0$ such that

$$M_P(c; \tilde{f}) \leq \Phi(c) := \exp(c^2\sigma_B^2/2(1 - cb)) \tag{36}$$

or all $0 \leq c < b^{-1}$. Now given a fixed σ_B and $0 \leq c < b^{-1}$, we can consider the family of QoIs defined in (21),

$$\mathcal{F}_P := \{g : M_P(c; \tilde{g}) \leq \Phi(c) = \exp(c^2\sigma_B^2/2(1 - cb))\}. \tag{37}$$

The form of the bound is very convenient since, by straightforward calculation we obtain an explicit solution for the optimization problem (23): $U_\pm(\eta; \mathcal{F}_P) = \inf_{c>0}\{\frac{c\sigma_B}{2(1-cb)} + \frac{\eta^2}{c}\}$ as

$$U_\pm(\eta; \mathcal{F}_P) = \sigma_B\sqrt{2}\eta + b\eta^2 \tag{38}$$

Furthermore (see Sections 2.4 and 2.8 in [31]) one can show that any random variable with a finite moment generating function in a neighborhood of 0 is a sub-gamma random variable although it may not be easy in general to explicitly determine the constants σ_B^2 and b in (36).

C. Bennett, Hoeffding and Bernstein Bounds

Many quantities of interest are bounded such as failure probabilities or functions of random variables with bounded support. Bounded random variables are necessarily sub-Gaussian and and sub-gamma [31], but much sharper bounds for their MGFs, (20), can be derived and used to bound the worst-case bias through Theorem 5. In this direction, we next discuss some additional concentration bounds for bounded QoIs that we will also showcase in examples in this work. This list is not complete by any means and other concentration inequalities can be used here; see for instance [32] for other bounds. For each case below, the family of QoIs $\mathcal{F}_P$ is defined in terms of the concentration bound on the MGF, (21), as in Theorem 5.

Bennett concentration bound [34, Lemma 2.4.1]: Consider the random variable X where $X \sim P$ and the QoI $f = f(X)$ such that $f(X) \leq b$, for some $0 \leq b < \infty$. Setting $\mu := \mathbb{E}_P[f(X)]$, $\tilde{b} := b - \mu$, we have

$$M_P(c; \tilde{f}) \leq \Phi(c) := \frac{\tilde{b}^2}{\tilde{b}^2 + \sigma_B^2}\exp(-c\sigma_B^2/\tilde{b}) + \frac{\sigma_B^2}{\tilde{b}^2 + \sigma_B^2}\exp(c\tilde{b}), \tag{39}$$

for all $c \geq 0$ and where σ_B^2 is any upper bound of $\text{var}_P[f]$. Therefore, keeping in mind Remark 6, we define

$$\mathcal{F}_P = \{g : M_P(c; \tilde{g}) \leq \Phi(c)\}, \quad \text{where } \Phi \text{ is defined in (39).} \tag{40}$$

TABLE II

THE DIFFERENT MGF BOUNDS ALONG WITH THE CONDITIONS THEY IMPOSE ON P AND f AND THE QUANTITIES THEY DEPEND ON FOR THEIR IMPLEMENTATION IF WE ARE INTERESTED IN QUANTIFYING THE WORST-CASE BIAS. HOWEVER, BOUNDING THE WORST-CASE $\mathbb{E}_Q[f]$ **DOES NOT** REQUIRE $\mathbb{E}_P[f]$. GAUSSIAN DECAY OF THE TAILS OF THE DISTRIBUTION OF $f(X)$ IMPLIES THE SUB-GAUSSIAN MGF BOUND (SIMILAR ASSUMPTIONS ABOUT THE TAILS EXIST FOR THE REST OF THE BOUNDS). IN TERMS OF INFORMATION REQUIREMENTS, THE HOEFFDING BOUND REQUIRES THE LEAST AMOUNT, BUT IT IS ALSO THE LEAST TIGHT. AS AVAILABLE INFORMATION/ DATA FOR THE BOUNDS GROW, THE BOUNDS GET TIGHTER

Name	Conditions on f, P	$\Phi = \Phi(c)$ input
Hoeffding (43)	$a \le f(X) \le b$	a, b
Bennett-(a,b) (41)	$a \le f(X) \le b$	$\mathbb{E}_P[f], a, b$
Bennett (39)	$f(X) \le b$, $\mathrm{var}_P[f] \le \sigma_B^2$	$\mathbb{E}_P[f], b, \sigma_B$
Bernstein (46)	$f(X) \le b$, $\mathrm{var}_P[f] \le \sigma_B^2$	$\mathbb{E}_P[f], b, \sigma_B$
sub-Gaussian (30)	$M_P(c;\tilde{f}) \le \exp(\sigma_B^2 c^2/2)$	σ_B
sub-gamma (36)	$M_P(c;\tilde{f}) \le \exp(\sigma_B^2 c^2/2(1-c\tilde{b}))$	σ_B, b
GO bound (7)	$M_P(c;\tilde{f}) < \infty$	$\mathbb{E}_P[(f)^k]$ for all k

Bennett-(a,b) concentration bound [34, Corollary 2.4.5]: If the QoI f is such that $a \le f(X) \le b$, $X \sim P$, then $\sigma_B^2 = (\mu - a)(b - \mu)$ is a bound on the variance and from the Bennett bound we obtain, for all $c \in \mathbb{R}$, (with $\tilde{a} = a - \mu$)

$$M_P(c;\tilde{f}) \le \Phi(c) := \frac{\tilde{b}}{b-a}\exp(c\tilde{a}) - \frac{\tilde{a}}{b-a}\exp(c\tilde{b}). \quad (41)$$

The right-hand side of (41) is the MGF of a Bernoulli-distributed random variable with values $\{a, b\}$. Note that the Bernoulli is the distribution with the most "spread" around the mean value between all bounded random variables in $[a, b]$. Similarly to (40) we have,

$$\mathcal{F}_P = \{g : M_P(c;\tilde{g}) \le \Phi(c)\}, \quad \text{where } \Phi \text{ as in (41).} \quad (42)$$

Hoeffding concentration bound [34], [42]: When the QoI f is bounded as in the Bennett-(a,b) case, we can further bound the Bennett-(a,b) bound by a Gaussian MGF, giving rise to the (less tight) Hoeffding MGF bound,

$$M_P(c;\tilde{f}) \le \Phi(c) := \exp(c^2(b-a)^2/8) \text{ for all } c \in \mathbb{R}. \quad (43)$$

This bound can be obtained by Hoeffding's Lemma. Since $a \le f \le b$, we have $a - \mathbb{E}_P[f] \le \tilde{f} \le b - \mathbb{E}_P[f]$. By Hoeffding's Lemma applied on $\tilde{f}$, we obtain

$$\begin{aligned} M_P(c;\tilde{f}) &= E_P[e^{c\tilde{f}}] \le \exp\left(c^2 \frac{((b-\mathbb{E}_P[f]) - (a-\mathbb{E}_P[f]))^2}{8}\right) \\ &= \exp(c^2(b-a)^2/8). \end{aligned}$$

Unlike the Bennett bounds, the Hoeffding bound is independent of the location of the mean $\mu = \mathbb{E}_P[f]$ within the interval (a, b) and only depends on the length of the interval $[a, b]$. As such, it requires the least amount of information about f and P and is the least sharp of the bounds, as can be also seen in the example in Section VI-B. As in the sub-Gaussian case of Section III-A, we can now calculate $U_{\pm}(\eta;\mathcal{F}_P)$ explicitly:

$$U_{\pm}(\eta;\mathcal{F}_P) = (b-a)\sqrt{2}\eta, \quad (44)$$

where the set of QoIs is

$$\mathcal{F}_P = \{g : M_P(c;\tilde{g}) \le \Phi(c)\}, \quad \text{where } \Phi \text{ as in (43)} \quad (45)$$

Bernstein concentration bound [31]: As in the Bennett bound we assume that $f(X) \le b$ and $\mathrm{var}_P[f] \le \sigma_B^2$ and use the notation $\tilde{b} = b - \mu$. We have

$$M_P(c;\tilde{f}) \le \Phi(c) := \exp(c^2\sigma_B^2/2(1-\tilde{b}c)), c \in [0, 1/\tilde{b}) \quad (46)$$

which states that f is sub-gamma and as in the sub-gamma case of Section III-B we can compute $U_+(\eta;\mathcal{F}_P)$ explicitly:

$$U_+(\eta;\mathcal{F}_P) = \sqrt{2\sigma_B^2}\eta + \tilde{b}\eta^2. \quad (47)$$

where the set of QoIs is

$$\mathcal{F}_P = \{g : M_P(c;\tilde{g}) \le \Phi(c)\}, \quad \text{where } \Phi \text{ as in (46)} \quad (48)$$

A similar bound for $c < 0$ can be derived if $f > a$ is bounded below in which case we obtain,

$$U_-(\eta;\mathcal{F}_P) = \sqrt{2\sigma_B^2}\eta - \tilde{a}\eta^2, \quad (49)$$

where $\tilde{a} = a - \mu$.

Remark 13 (Hierarchy of Bounds): It is straightforward to demonstrate that we can order the bounds in terms of accuracy, noting that if the QoI f is bounded in $[a, b]$, then we always have the bound $\sigma_B^2 \le (\mathbb{E}_P[f] - a)(b - \mathbb{E}_P[f])$ in the Bennett bound (39). Therefore, we have the hierarchy of concentration bounds:

$$M_P(c;\tilde{f}) \le \text{Bennett} \le \text{Bennett-(a,b)} \le \text{Hoeffding}. \quad (50)$$

Unlike the two Bennett bounds, the Hoeffding bound is independent of the location of the mean μ within the interval $[a, b]$ and only depends on the length of the interval, $b-a$. As such, it requires the least amount of information about f and P and is the least sharp of the bounds, see Table II and the requirements for the QoI families $\mathcal{F}_P$, (40), (42) and (45). The Bernstein bound and Hoeffding bounds are not directly comparable: for small η, the Bernstein bound is better than Hoeffding, indeed it captures the exact asymptotic of the GO divergence, see (9) in Theorem 1, however for large η the Bernstein bound is worse than Hoeffding and for large η both those bounds are overly pessimistic and not informative for bounded QoIs. On the

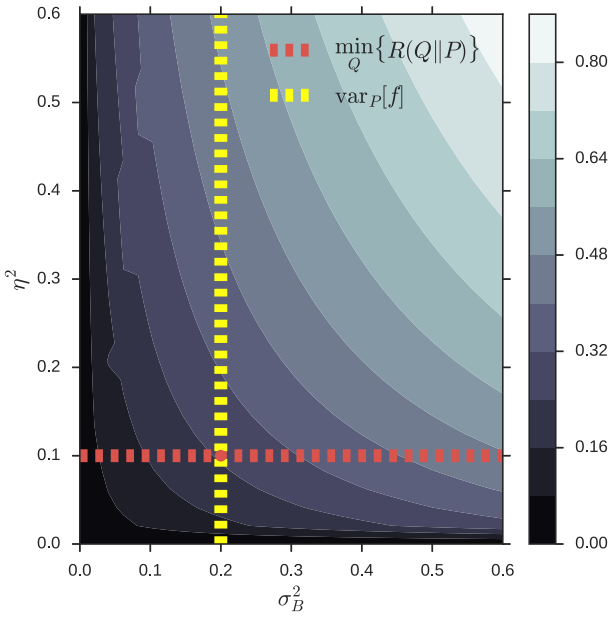

Fig. 2. Level curves of the upper model bias bound (22) with the Bennett bound (39) and assuming $b = 1$, $\mathrm{var}_P[f] \le \sigma_B^2$. Knowing $\eta^2 = R(Q\|P)$ for some model Q and an upper bound on the variance provides model-bias guarantees (through Theorem 5). Further reduction of the model bias bound requires a corresponding—and potentially expensive—decrease in KL and/or a tighter upper-bound for $\mathrm{var}_P[f]$, for example, by incorporating additional data. The tightest possible guarantee afforded by the Bennett bound is gained when $\sigma_B^2 = \mathrm{var}_P[f]$ and $\eta^2 = \min_Q R(Q\|P)$.

other end, the GO divergence bound—involving $M_P(c;\tilde{f})$—is the tightest, as we see in (50), but also the most expensive to implement, see Table I. We also refer to a demonstration of this hierarchy in the example in Section VI-B. Overall, as available information/data on the QoI f and and the baseline model P grows, concentration bounds and therefore model bias bounds become tighter. Finally, we refer to Figure 2, where we demonstrate the tightness of the model bias bounds (22), (23), in terms of both $\eta^2 = R(Q\|P)$ and σ_B, for the Bennett bounds (39).

Remark 14 (How Large Is the Class $\mathcal{F}_P$?): A plausible question is how rich is the set of admissible QoIs, $\mathcal{F}_P$, derived by the various concentration bounds on the MGF $M_P(c;\tilde{g})$ in (31), (40), (42) and (45). Here we address this question in the context of the Bennett bound, however the same argument also applies to the Bennett-(a,b) and Hoeffding bounds, as well as to the sub-Gaussian case in Section III-A. We can get a simple first insight in this direction based on (39). Indeed, based on the conditions for this inequality to hold, we readily have that

$$\mathcal{F}_P \supset \{g:\ g(X) \le b,\ \mathrm{var}_P[g] \le \sigma_B^2,\ \mathbb{E}_P g = \mu\}. \tag{51}$$

We also note that enforcing the condition on the mean, $\mathbb{E}_P g = \mu$, is trivial and involves only a translation of the QoI g.

IV. Tightness of the Concentration/Information Inequalities

In this section we show that, under suitable assumptions, for the concentration/information bounds derived in Section III the

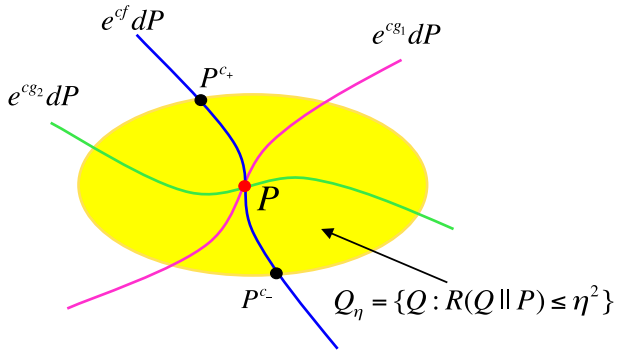

Fig. 3. The schematic depiction of Theorem 15 for a family of Quantities of Interest (QoIs) $\mathcal{F}_P$ and tolerance η^2. The solid lines depict the one-parameter tilted probability distributions P^c in (14) corresponding to the QoI $g_1, g_2, f \in \mathcal{F}_P$. The theorem implies that the upper and lower bounds in the family $\mathcal{Q}_\eta = \{Q : R(Q\|P) \le \eta^2\}$ are attained at the probability measures $Q^\pm = P^{c\pm}$.

divergence $U_\pm(\eta;\mathcal{F}_P)$ retains some of the tightness properties of the GO divergence $\Xi(Q\|P;f)$ established in Section II.

Theorem 15: Let P be a probability and $\mathcal{Q}_\eta = \{Q : R(Q\|P) \le \eta^2\}$. Assume $\Phi(c) = M_P(c;\tilde{f})$ is a MGF for some QoI f with respect to P and let

$$\mathcal{F}_P = \{g : M_P(c;\tilde{g}) \le \Phi(c) \ \text{ for all } c \in \mathbb{R}\}. \tag{52}$$

Then there exist $\eta_\pm$ such that $\eta \le \eta_\pm$ and probabilities $P^{c\pm} \in \mathcal{Q}_\eta$ (see (14)) that satisfy $R(P^{c\pm}\|P) = \eta^2$ as well as

$$\begin{aligned} U_+(\eta;\mathcal{F}_P) &= \mathbb{E}_{P^{c_+}}[f] - \mathbb{E}_P[f] \\ &= \max_{Q\in\mathcal{Q}_\eta, g\in\mathcal{F}_P} \mathbb{E}_Q[g] - \mathbb{E}_P[g], \end{aligned} \tag{53}$$

$$\begin{aligned} -U_-(\eta;\mathcal{F}_P) &= \mathbb{E}_{P^{c_-}}[f] - \mathbb{E}_P[f] \\ &= \min_{Q\in\mathcal{Q}_\eta, g\in\mathcal{F}_P} \mathbb{E}_Q[g] - \mathbb{E}_P[g], \end{aligned} \tag{54}$$

i.e., the maximum and minimum for model bias is attained within the family of models $\mathcal{Q}_\eta$ and the family of QoIs $\mathcal{F}_P$, see the schematic in Figure 3.

As a consequence we have the "confidence band" around the baseline model P,

$$\begin{aligned} -U_-(\eta;\mathcal{F}_P) + \mathbb{E}_P[g] \le \mathbb{E}_Q[g] \le \mathbb{E}_P[g] + U_+(\eta;\mathcal{F}_P) \\ \text{for all } Q \in \mathcal{Q}_\eta,\ g \in \mathcal{F}_P, \end{aligned} \tag{55}$$

with the two equalities holding if $Q = P^{c\mp}$ respectively and for $g = f \in \mathcal{F}_P$.

Proof: Since $f \in \mathcal{F}_P$, Theorem 2 implies that the probabilities $P^{c\pm}$ in (14), with $c_\pm$ chosen such that $R(P^{c\pm}\|P) = \eta^2$ satisfy

$$\Xi(P^{c\pm}\|P;\pm f) = U_\pm(\eta;\mathcal{F}_P). \tag{56}$$

Therefore, by Theorem 5, (22), for all $Q \in \mathcal{Q}_\eta, g \in \mathcal{F}_P$

$$-\Xi(P^{c-}||P;-f) \le \mathbb{E}_Q[g] - \mathbb{E}_P[g] \le \Xi(P^{c+}||P;f). \tag{57}$$

Finally, we apply (15) and (16) of Theorem 2 and use (56) to conclude the proof. □

Remark 16 (Connections to Mass Transport): Here we discuss one possible approach to verify the crucial assumption

of Theorem 15, namely that

$$\Phi(c) = M_P(c; \tilde{f}) \quad \text{for some QoI } f. \tag{58}$$

One natural way to ensure (58) holds is intimately related to mass transport methods, [43]. Instead of (58), we may assume the more easily checkable hypothesis that $\Phi(c) = M_{\bar{P}}(c; \tilde{h})$ for some h and some model $\bar{P}$; e.g. $h(x) = x$ and $\bar{P}$ a Gaussian distribution for the Hoeffding's bound, see also Example 18 below. To prove (58) one shows then that there exists a *transport map* between P and $\bar{P}$, namely a map T such that $\bar{P}(A) = P(T^{-1}(A))$ for any measurable set A [43]. If a transport map exists we have

$$\mathbb{E}_P[h \circ T] = \int h(Tx)P(dx) = \int h(y)\bar{P}(dy) = \mathbb{E}_{\bar{P}}[h].$$

and hence with $f = h \circ T$

$$\tilde{f} = f - \mathbb{E}_P[f] = h \circ T - \mathbb{E}_{\bar{P}}[h] = \tilde{h} \circ T.$$

This implies that

$$\begin{aligned}\Phi(c) =& M_{\bar{P}}(c; \tilde{h}) = \int e^{c\tilde{h}(y)}\bar{P}(dy) \\ &= \int e^{c\tilde{h}(Tx)}P(dx) = M_P(c; \tilde{f}),\end{aligned} \tag{59}$$

and thus the assumption (58) holds.

To ensure the existence of such a transport map T one needs some assumptions on P (and $\bar{P}$). For example, if P and $\bar{P}$ are non-atomic measures then a transport map always exists. If the measure P has a density then T can be constructed using the Knothe-Rosenblatt rearrangement or Brenier's L_2 optimal transport map; we refer to Chapter 1 [43], [44] for more details on these maps, and several other such transport maps and relevant conditions for their existence.

Next we demonstrate how to use Theorem 15 by interpreting $\Phi(c)$ as the MGF of a suitable QoI f with respect to the distribution P. In Example 17, we illustrate the tightness of the concentration bounds for the case of bounded random variables supported in $[-1, 1]$, while the arguments can be trivially generalized to any other bounded interval.

Example 17 (Bennett-(a,b) QoIs): Consider a distribution P such that there is an event $A \subset \mathbb{R}$ such that $P(A) = 1/2$; we consider the family of QoIs, $\mathcal{F}_P$, for which (41) is true with $a = -1$, $b = 1$, $\mathbb{E}_P[g] = 0$ for all $g \in \mathcal{F}_P$. The corresponding Bennett-(a,b) bound is

$$\Phi(c) = \frac{1}{2}e^c + \frac{1}{2}e^{-c}.$$

Then, if we choose $f(x) := 2 \cdot 1_A(x) - 1$, where 1_A is the characteristic function of the set A, we have $\Phi(c) = M_P(c; \tilde{f})$. Therefore Theorem 15 is immediately applicable.

The next example covers the case of sub-Gaussian QoIs which contains both bounded and unbounded random variables.

Example 18 (sub-Gaussian QoIs): Consider a probability measure P on $\mathbb{R}$ which has a density. For sub-gaussian QoIs (30) we have the bound $\Phi(c) = \exp(c^2\sigma_B^2/2)$, however, we can rewrite the bound as

$$\Phi(c) = M_{\bar{P}}(c; \tilde{h})$$

where $h(x) = x$ and $\bar{P} = N(0, \sigma_B^2)$ is a normal distribution. Since P has a density, we can use the measurable isomorphism, or any other applicable map discussed in Remark 16, to construct a transport map T between P and $\bar{P}$. Thus, we can show the existence of a QoI f that satisfies the condition (58) and we can readily apply Theorem 15 to show the tightness of the bounds given by (32).

V. Model Bias for Statistical Estimators

As discussed in Section II a key challenge is to control the risk involved in evaluating statistical estimator using the baseline model P rather than the true model Q. In addition it is important to control the bias of QoIs which are not necessarily expected values, for example the bias in the variance, i.e, $\mathrm{var}_P X - \mathrm{var}_Q X$, or other statistics such as correlation, skewness or quantiles; see [39]. Generally, given data $X_1, \ldots, X_n$, we aim to control the bias of statistical estimator $\psi = \psi(X_1, \ldots, X_n)$, for example the sample variance (69).

To obtain useful bounds on the bias of statistical estimators ψ, we need to exhibit and control the dependence of the inequalities in Sections III-C and III-A on the amount of data available, i.e. the dependence on n. We will exhibit a large and natural class of statistical estimators for which inequalities are asymptotically independent on n. As demonstrated in [19] the Concentration/Information inequalities of Sections II and III are the only known information equalities which scale properly with n.

The main tool we shall use is the key result used in the proof of the McDiarmid's inequality, see also the Hoeffding-Azuma bound, [34]. We refer to Chapter 2 of [32] or [45] for the proof.

Proposition 19: Let $X_1, \ldots, X_n$ be independent random variables with joint distribution $P^n = P_1 \times \cdots \times P_n$. Let $\psi(x_1, \ldots, x_n)$ satisfy the Lipschitz condition

$$\begin{aligned}\sup_{x_1,\ldots,x_n,x_k'} &|\psi(x_1, \ldots, x_k, \ldots, x_n) - \psi(x_1, \ldots, x_k', \ldots, x_n)| \\ &\leq d_k\end{aligned} \tag{60}$$

for some constants d_k, $k = 1, \ldots, n$. Then $\psi(X_1, \ldots, X_n)$ is a sub-Gaussian random variable and for all $c \in \mathbb{R}$ we have

$$\begin{aligned}M_{P^n}(c; \tilde{\psi}) &= \mathbb{E}_{P^n}[\exp(c(\psi - \mathbb{E}_{P^n}[\psi]))] \\ &\leq \exp\left(\frac{c^2}{8}\sum_{k=1}^n d_k^2\right).\end{aligned} \tag{61}$$

By combining the bound in (61) with the definition of $U_{\pm}(\eta; \mathcal{F}_P)$ in Theorem 5 for the sub-Gaussian case (30) we obtain immediately

Theorem 20: Consider two joint distributions $P^n = P_1 \times \cdots \times P_n$ and $Q^n = Q_1 \times \cdots \times Q_n$. For $X_1, \ldots, X_n$ and $\psi(x_1, \ldots, x_n)$ as in Proposition 19 we have

$$|\mathbb{E}_{P^n}[\psi(X_1, \ldots, X_n)] - \mathbb{E}_{Q^n}[\psi(X_1, \ldots, X_n)]| \tag{62}$$

$$\leq \left(\sum_{k=1}^n d_k^2\right)^{1/2} \sqrt{\frac{1}{2}\sum_{k=1}^n R(Q_i \| P_i)}. \tag{63}$$

If $X_1, \ldots, X_n$ are identically distributed with common distribution P and $Q_i = Q$, for $i = 1, 2, ..., n$, and if there exists a constant C such that

$$d_k \leq \frac{C}{n}, \quad k = 1, \ldots, n$$

then we have for any n

$$|\mathbb{E}_{P^n}[\psi(X_1, \ldots, X_n)] - \mathbb{E}_{Q^n}[\psi(X_1, \ldots, X_n)]| \leq C\sqrt{\frac{1}{2}R(Q\|P)}.$$

Proof: First, by independence it is easy to show that

$$R(Q^n\|P^n) = \sum_{i=1}^{n} R(Q_i\|P_i). \tag{64}$$

By Proposition 19, we have

$$E_{P^n}[\exp(c(\psi - \mathbb{E}_{P^n}[\psi]))] \leq \exp\left(\frac{c^2}{8}\sum_{k=1}^{n} d_k^2\right). \tag{65}$$

Consider the definition of $U_{\pm}(\eta; \mathcal{F}_P)$ in Theorem 5, where $\eta = \sqrt{R(Q^n\|P^n)} = \sqrt{\sum_{i=1}^{n} R(Q_i\|P_i)}$; then we readily obtain

$$\begin{aligned} U_{\pm}(\eta; \mathcal{F}_P) &= \inf_{c>0}\left\{\frac{1}{c}\log\Phi(\pm c) + \frac{1}{c}\eta^2\right\} \\ &= \inf_{c>0}\left\{\frac{c}{8}\sum_{k=1}^{n} d_k^2 + \frac{1}{c}R(Q^n\|P^n)\right\} \\ &= \frac{1}{\sqrt{2}}\sqrt{\sum_{k=1}^{n} d_k^2}\sqrt{\sum_{i=1}^{n} R(Q_i\|P_i)}. \end{aligned}$$

If $X_1, \ldots, X_n$ are identically distributed with common distribution P and if $d_k \leq \frac{C}{n}$ for some constant C then $\sum_{k=1}^{n} d_k^2 \leq \frac{C^2}{n}$ and $\sum_{i=1}^{n} R(Q_i\|P_i) = nR(Q\|P)$ and thus we obtain

$$\begin{aligned} U_{\pm}(\eta; \mathcal{F}_P) &= \frac{1}{\sqrt{2}}\sqrt{\sum_{k=1}^{n} d_k^2}\sqrt{\sum_{i=1}^{n} R(Q_i\|P_i)} \\ &\leq \frac{1}{\sqrt{2}}C\sqrt{R(Q\|P)}. \end{aligned}$$

□

Remark 21 (Poor Scalability of Certain Information Inequalities): A notable feature of the concentration/information inequalities is that they scale independently of the number of data/random variables n, at least for classes of QoIs that satisfy (60), as demonstrated in Theorem 20 and the subsequent examples. Furthermore, the bias bound (64) remains discriminating even if $n \to \infty$. The same scaling features are also shared with the GO divergence bounds (8), see [19]. On the other hand, classical information inequalities scale poorly with n. For example, in the case of the Pinsker inequality [6], [18], let us consider the QoI (estimator) (71) for the i.i.d. random variables $X_1, \ldots, X_n$,

$$\psi_n(X_1, \ldots, X_n) = \frac{1}{n}\sum_{i=1}^{n} f(X_i).$$

Then, the Pinsker inequality becomes

$$\begin{aligned} &|\mathbb{E}_{P^n}[\psi(X_1, \ldots, X_n)] - \mathbb{E}_{Q^n}[\psi(X_1, \ldots, X_n)]| \\ &\leq \|f\|_{\infty}\sqrt{2R(Q^n \,||\, P^n)} = O(\sqrt{n}), \end{aligned} \tag{66}$$

where we used that $\|\psi\|_{\infty} = \|f\|_{\infty}$, and $R(Q^n \,||\, P^n) = nR(Q \,||\, P)$. Therefore the Pinsker bound (66) blows up as $n \gg 1$, in contrast to the concentration/information inequality (64) that remains discriminating and informative for any n. Other model bias bounds based on the Renyi or χ^2 divergences (the latter known as the Chapman-Robbins inequality) or the Hellinger metric, also scale poorly with the size of data set and/or with the number of variables n; we refer to Sections 2.2–2.3 in [19] for a complete discussion.

Next, we apply these results towards obtaining model bias bounds for statistical estimators.

CDF estimator: If X is a real-valued random variable with cumulative distribution function (CDF) $F_P(x) = P\{X \leq x\}$, then given i.i.d. data $X_1, \ldots, X_n$,

$$\hat{F}_n(x) = \frac{1}{n}\sum_{k=1}^{n} I_{\{X_k \leq x\}}, \tag{67}$$

where I_A is the indicator function of the set A, $\hat{F}_n(x)$ is an estimator for the CDF, $F_P = F_P(x)$. It is easily verified that the conditions of Theorem 20 are satisfied with $C = 1$. Since the bound is uniform in x, and $\hat{F}_n(x)$ is an unbiased estimator of $F_P(x)$, we obtain

$$\begin{aligned} \sup_x |F_Q(x) - F_P(x)| &= \sup_x \left|\mathbb{E}_{Q^n}\left[\hat{F}_n(x)\right] - \mathbb{E}_{P^n}\left[\hat{F}_n(x)\right]\right| \\ &\leq \sqrt{2R(Q\|P)}, \end{aligned} \tag{68}$$

for any alternative model Q to the baseline P. As we also note in the sample variance example below, the estimator does not need to be unbiased.

Sample variance and general statistical estimators: McDiarmid's inequality and condition (60) can be used to control bias of QoIs which are not simply expected values, for example the sample variance

$$\begin{aligned} V_n(X_1, \ldots, X_n) &= \frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \frac{1}{n}\sum_{j=1}^{n} X_j\right)^2 \\ &= \frac{1}{2n(n-1)}\sum_{i,j=1}^{n}(X_i - X_j)^2. \end{aligned} \tag{69}$$

If we assume that $|X_i| \leq M$ for some $M > 0$ then we have

$$\sup_{\substack{|x_i|\leq M,\\ |x_k'|\leq M}} |V_n(x_1, \ldots, x_k, \ldots, x_n) - V_n(x_1, \ldots, x_k', \ldots, x_n)| \leq \frac{8M^2}{n-1}.$$

Then the sample variance satisfies (60) with $d_k = 8M^2/(n-1)$ for all k. Thus, we can bound the corresponding model bias by

$$\begin{aligned} |\mathrm{var}_P[X] - \mathrm{var}_Q[X]| &= |\mathbb{E}_{P^n}[V_n] - \mathbb{E}_{Q^n}[V_n]| \\ &\leq 8M^2\frac{n}{n-1}\sqrt{2R(Q\|P)}, \end{aligned} \tag{70}$$

which is valid for all $n > 1$. Note that if we take $n \to \infty$ we obtain the variance bound

$$|\mathrm{var}_P[X] - \mathrm{var}_Q[X]| \le 8M^2\sqrt{2R(Q\|P)}$$

which shows how the KL-divergence $R(Q\|P)$ controls the misspecification for QoIs beyond their expected values. The same analysis also applies to the (biased) plug-in estimator for the variance, namely

$$\tilde{V}_n(X_1,\dots,X_n) := \mathrm{var}_{\hat{F}_n}[X] = \frac{1}{n}\sum_{i=1}^n \left(X_i - \frac{1}{n}\sum_{j=1}^n X_j\right)^2.$$

Finally, we can easily generalize the sample variance calculation to more general QoIs and statistical estimators. The sample variance depends (up to a factor $\frac{n-1}{n}$) only on the two sample averages $\frac{1}{n}\sum_{i=1}^n X_i$ and $\frac{1}{n}\sum_{i=1}^n X_i^2$. It is not difficult to see that if $|X_i| \le M$ and the QoI has the form

$$\psi_n(X_1,\dots,X_n) = g\left(\frac{1}{n}\sum_{i=1}^n f_1(X_i),\dots,\frac{1}{n}\sum_{i=1}^n f_k(X_i)\right) \tag{71}$$

for some $f_1,\cdots f_k$ (say the the first k moments), and for some Lipschitz continuous function g, then one can apply Theorem 20 for a constant C which depends on M, the Lipschitz constant for g and $f_1,\cdots f_k$. One important example of the type (71) is the sample correlation, we refer to Example 2.16 in [46].

Confidence Bands and Model Bias To further illustrate our results we construct a non-parametric confidence band for the CDF $F_Q(x)$, in the same context as the setting in (68). We combine the bound (68) with the Dvoretzky-Kiefer-Wolfowitz (DKW) inequality [39], [46], i.e. the bound

$$P\left\{\sup_x |\hat{F}_n(x) - F_P(x)| \ge \epsilon\right\} \le 2e^{-2n\epsilon^2}, \tag{72}$$

which itself is obtained though concentration inequalities. For any n and $\alpha > 0$, we set $\epsilon_n = \sqrt{\log(2/\alpha)/2n}$ and

$$\begin{aligned} L_n(x;\eta) &= \max\{\hat{F}_n(x) - \sqrt{2}\eta - \epsilon_n,\, 0\} \\ U_n(x;\eta) &= \min\{\hat{F}_n(x) + \sqrt{2}\eta + \epsilon_n,\, 1\}. \end{aligned} \tag{73}$$

Since $\hat{F}_n(x)$ is an unbiased estimator for the baseline model P rather than for the (unknown) "true" model Q we obtain the α–confidence band for $F_Q(x)$:

$$\begin{aligned} &P\{L_n(x;\eta) \le F_Q(x) \le U_n(x;\eta) \text{ for all } x\} \ge 1-\alpha, \\ &\text{for all } \ Q \in \mathcal{Q}_\eta. \end{aligned} \tag{74}$$

Due to the fact that both our bound (68) and the DKW inequality (72) are valid for any data size n, the confidence band (74) does not require any asymptotic normality assumptions or a large data set $n \gg 1$.

Connections to the Vapnik-Chervonenkis inequality The DKW inequality is an effective tool for controlling deviations from the average for one dimensional distributions and their corresponding CDFs. However, the Vapnik-Chervonenkis (VC) theory [46] allows us to address the same issues in a more general setting that is applicable to higher-dimensional distributions, by considering the empirical probability distribution instead of the CDF. In particular, corresponding inequalities to (72), but for the empirical probability distribution, can be derived based on the VC theory, see for instance Theorem 2.41 and Theorem 2.43 in [46]. In turn the VC inequalities, along with our concentration information bounds (64) can allow us to obtain confidence intervals for higher dimensional distributions, similarly to (74).

VI. Elementary Examples

Prior to discussing applications involving more complex models in Section VII, here we demonstrate the concentration/information inequalities we developed earlier to two elementary examples that allow easy analytic and computational implementations.

A. Exponential Distribution

We first consider the model bias bounds using the GO divergence in Theorem 1, contrasted to the concentration/information divergence in Theorem 5. In our first example, the baseline model P is an exponential distribution. The models Q can be any distributions which are absolutely continuous with respect to P, hence $R(Q\|P) < \infty$. Let P be an exponential distribution with parameter $\lambda_P = 1$. The QoI is $f(X) = X$. The MGF of P is $M_P(c;X) = 1/(1-c)$ and thus it is finite in $(0,1)$, while otherwise it is infinite. Next, we let η be a model uncertainty threshold and Q any distribution, not necessarily exponential or in any parametric family, such that $R(Q\|P) \le \eta^2$. We note that the distribution P exhibits sub-exponential behavior, namely

$$\begin{aligned} M_P(c;f) &= 1 + c + \frac{c^2}{1-c} \le 1 + c + 2c^2 \\ &\le \exp(c + c^2/(2\sigma_B^2)) := \Phi(c), \quad c \in (-0.5, 0.5), \end{aligned} \tag{75}$$

where $\sigma_B = 1/2$ and the interval $(-0.5, 0.5)$ is selected so that the bounds remain finite. In general, if we have additional information on the location of λ_P, e.g., from data, then we can adjust the interval that c lies in accordingly. Here, the concentration/information bound (23) is then adjusted according to (75), using Theorem 5 and the general concentration bound (20). Although the MGF is known in this particular example, the use of the concentration bound (75) allows us to quantify the worst-case model bias for all QoIs $g \in \mathcal{F}_P$, where

$$\mathcal{F}_P = \{g : M_P(c;\tilde{g}) \le \Phi(c),\ c \in (-0.5, 0.5),\ \sigma_B^2 \le 1/4\}. \tag{76}$$

Figure 4 is a comparison of the GO-divergence and the concentration/information bound based on (75), along with the exact model bias for the case that Q is also an exponential distribution with $R(Q\|P) \le \eta^2$ and $\eta \in [0, 1.6]$.

Finally, we can consider other types of tail decay, and thus corresponding concentration inequalities, besides the sub-Gaussian and the sub-exponential cases discussed thus far. For example, we can also consider Poisson-type tail decay, see for instance [32, Section 3.3.5] and [47].

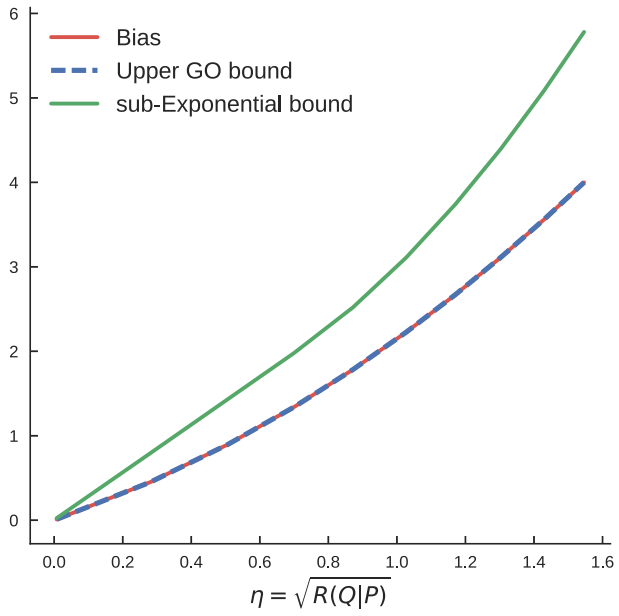

Fig. 4. Comparison of model bias bounds based on the GO divergence and the concentration/information (75), with the exact model bias $1 - \lambda_Q$, $\lambda_Q \in (1.01, 10)$, where Q is an exponential distribution with mean λ_Q. The sub-exponential bound (75) is less sharp as the KL divergence increases, since it captures the worst-case performance over the family of QoIs $\mathcal{F}_P$. Although $R(Q\|P)$ is computed with Q being an exponential distribution, the bounds to the model bias are valid for any Q that is absolutely continuous with respect to P and has $R(Q\|P)$ in the range of the figure.

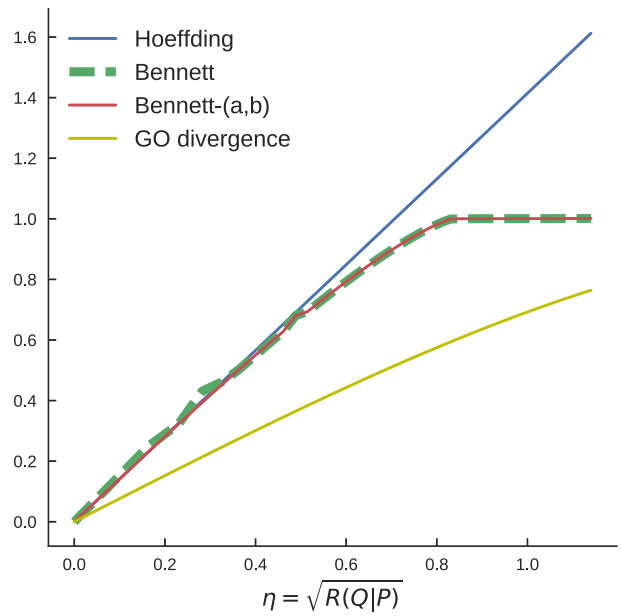

Fig. 5. Comparison of the different bounds for the bias in the truncated Normal example (see Section VI-B), assuming that the observable of interest is $f(X) = X$. This plot makes no assumptions on the form of Q except that $R(Q\|P) = \eta^2 \in (0.0, 4.0)$. As in Figure 4, here the concentration/information bounds capture the worst-case performance over the family of QoIs $\mathcal{F}_P$, hence perform worse than the GO divergence bounds which are suitable only for a single QoI, see also (50). Notice that Bennett and Bennett-(a, b) track better the bound of the GO divergence for large values of the KL whereas the Hoeffding is sufficient only for small values of the KL, i.e., at the linearized regime of the GO bounds. Only the upper bounds for the bias are shown here.

Remark 22: The bias is an unbounded function of the KL divergence in this example—a consequence of the QoI $f(X) = X$ being unbounded under P. Therefore, any decrease in KL divergence translates to an improvement in worst-case model bias, see Figure 4; this fact is in sharp contrast with the truncated Normal example in Section VI-B, where even large improvements to larger values of the KL divergence may not help much in reducing model bias, see Figure VI-B.

B. Truncated Normal

In this example the distributions we consider are bounded, allowing us to deploy the hierarchy of concentration/ information bounds (50) developed in Section III-C. We assume the random variable X follows the truncated Normal distribution, $P = TN(0, 1, -1, 1)$, where $[-1, 1]$ is the interval of support. Here we will bound the model bias, $\mathbb{E}_Q[f] - \mathbb{E}_P[f]$, for any Q such that $R(Q\|P) = \eta^2$, where $\eta \in [0.01, 1]$ and for any f in a suitable family of QoIs, $\mathcal{F}_P$. Apart from these, the bounds make no other assumptions on Q and f. Figure 5 contains a comparison of the different concentration/information bounds (50) from Section III-C.

As a general observation, we notice that for large values of $\eta = \sqrt{R(Q\|P)}$, small perturbations of η will not change the Bennett/GO (see Relations (7) and (39)) bounds significantly. Therefore, for some QoIs, e.g., $f(X) = X$, small improvements to large values of the KL will barely improve the worst-case bias (as captured by the bounds, see Figure 5). The existence of such QoIs is guaranteed by the sharpness of the bounds demonstrated in Section IV. Finally, we note that even for the tighter concentration/information bounds, i.e., the ones associated with the two Bennett bounds (39) and (41), there is some discrepancy with the GO divergence bound. This discrepancy is due to the fact that the GO bound is applied only for a specific QoI, while the concentration/information bounds are tight over the broad classes of QoIs defined in Section III-C, see also Remark 14.

VII. Epistemic Uncertainty Quantification via Concentration/Information Inequalities

We apply the concentration/information inequalities to control model bias between baseline and alternative models in two more complex examples. The type of model bias considered here arises in epistemic uncertainty quantification, where modelers are unsure if their baseline model included all necessary complexity or lacks sufficient data, [2], [5]. The KL divergence and in particular the GO divergence bounds provide a non-parametric framework to mathematically describe this type of epistemic uncertainties, as first shown in [4]. Here, we consider two such examples that illustrate different aspects of epistemic uncertainty, namely a data-driven model for the lifetime of lithium batteries, as well as a high-dimensional Markov Random Field model subject to various localized uncertainties such as local defects. A key aspect of our discussion in both examples is the necessity and the (ease of) implementation of concentration/information model bias bounds, see for instance Remark 11.

TABLE III
FAILURE TIMES OF TEST SAMPLES [48]

Specimen number	01	02	03	04	05	06	07	08	09	10	11	12
Failure time	1373	1470	1520	1427	892	814	777	637	927	688	857	866

A. Epistemic Uncertainty for Failure Probabilities

Here, we apply the bounds of Theorem 5, and in particular the inequalities in Section III-C, to the life-time analysis of lithium secondary batteries. Firstly, we introduce the Weibull distribution which is widely used in for analyzing life-time data, see [48] and references therein. The probability density function of a Weibull random variable is

$$f(t) = \frac{\beta}{\xi}\left(\frac{t}{\xi}\right)^{\beta-1} e^{-(\frac{t}{\xi})^{\beta}}, t > 0, \tag{77}$$

where $\beta > 0$ is called a shape parameter and $\xi > 0$ is a scale parameter of the distribution [49]. The shape parameter explains the types of failure and the scale parameter explains the characteristic life cycle of devices. The cumulative distribution function F can be expressed as:

$$F(T) = 1 - e^{-(\frac{T}{\xi})^{\beta}},$$

where T denotes the time of failure (or the lifetime) of the battery.

In Table III, experimental data based on life cycle tests are obtained from [48]. By fitting the data in Table III to the parameters of the Weibull distribution, we obtain the corresponding maximum likelihood estimator (MLE) for ξ and β are $\hat{\xi} = 1138$ and $\hat{\beta} = 3.55$, respectively. Now, we consider this MLE Weibull distribution as the baseline model P, which is a data-driven approximation to the unknown true model. Next we consider the family of alternative models within a fixed tolerance η^2, namely the non-parametric family of models $\mathcal{Q}_\eta$, see (3). This family accounts for unknown features not necessarily captured in the baseline model which was arbitrarily assumed to be Weibull. Furthermore, the family $\mathcal{Q}_\eta$ can account for perturbations in the baseline model—constructed based on the specific dataset in Table III—due to additional data that may become available or for any errors in the data used in the MLE step. Next, we assess the impact of model uncertainty within the family of models $\mathcal{Q}_\eta$ on two QoIs associated with lifetime probabilities of the batteries:

$$f_1(t) = 1_{\{0 \le t \le T\}}(t), t > 0, \tag{78}$$

$$f_2(t; w) = \frac{1}{1 + e^{w(t-T)}}, t > 0. \tag{79}$$

The function $f_2(t; w)$ is a commonly used smooth approximation to the indicator function $f_1(t)$ and is usually referred as the logistic function, see Section 39.1 of [50]). The parameter w, $w \ge 1$, controls the smoothness of the approximation. The QoI for the life-time probability is defined exactly as $F_P[T] := E_P[f_1(t)] = P(0 \le t \le T)$ or through the smooth approximation $\mathbb{E}_P[f_2]$.

Since the QoI $f_1(t)$ is bounded in $[0, 1]$, we can apply the Bennett (39), Bennett-(a,b) (41) and Hoeffding bounds (43)) to obtain the uncertainty region, where $a = 0$, $b = 1$, $\tilde{a_1} = -F_P(T)$, $\tilde{b_2} = 1 - F_P(T)$ and $\sigma_B^2 = Var_P[f_1(t)]$, the latter needed just in the Bennet bound. For f_2, we estimate $\mathbb{E}_P[f_2]$ by sampling from P, thus computing $\mu_2 = \mathbb{E}_P[f_2]$, needed in both Bennett bounds. Then, $\tilde{a}_2 = -\mu_2$ and $\tilde{b}_2 = 1 - \mu_2$. In Figure 6 we compare the lifetime probabilities given by f_1 and f_2, where for the latter we set $w = 5$. In this figure, we also observe that the logistic function f_2 gives a good approximation of the indicator function f_1 since lifetime probabilities based on them are almost the same. Moreover, we set $\eta^2 = 0.1$ and also plot the GO divergence bounds of Theorem 1 based on f_1 and Bennett-(a,b) bounds based on f_2. We notice that the bounds almost coincide. We also consider the Bennett-(a,b) bounds based on a smaller tolerance $\eta^2 = 0.01$. As we see in the figure, we obtain a significantly narrower model bias region.

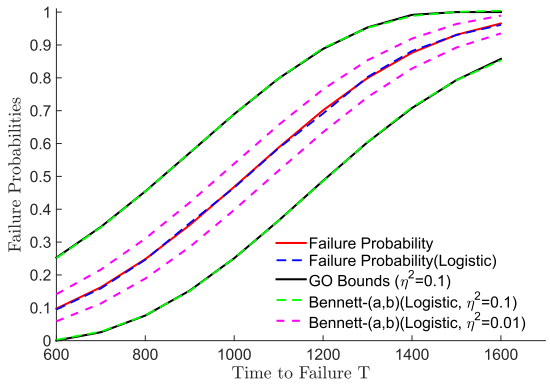

Fig. 6. The blue line is the failure probability based on the logistic function f_2; The red line is the failure probability based on the indicator function f_1; The black lines are the GO bounds based on f_1 with $\eta^2 = 0.1$; The green lines are the Bennett-(a,b) bounds based on f_2 with model uncertainty $\eta^2 = 0.1$. The magenta lines are the Bennett-(a,b) bounds based on f_2 with $\eta^2 = 0.01$.

The non-parametric setting using KL divergence is natural in this example where the available data is sparse, inducing significant epistemic uncertainty in the predictions of failure probabilities. In this general line of thought, instead of considering the MLE Weibull distribution as the baseline model and the non-parameteric family Q_η, we can also apply Bayesian estimation to calculate a posterior $P(\xi, \beta|data)$ for (ξ, β). Then we can consider the Weibull corresponding to the maximum a posteriori estimator (MAP) parameter as our baseline model P. In this case we can consider other alternative models $Q = Q^\theta$ where Q^θ is a parametric family of Weibull models and the parameters θ are distributed according to the posterior, $\theta \sim P(\xi, \beta|data)$. Here our setting is entirely parametric although one can also consider non-parametric Bayesian methods; we also refer to the discussion in Remark 4.

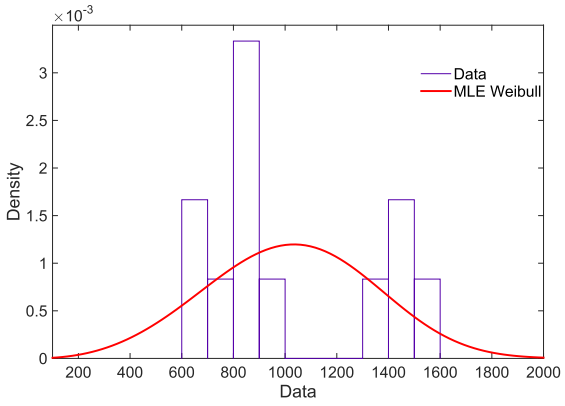

Fig. 7. The Histogram is represented for the data in Table III; The red line is the baseline model P, which is the MLE Weibull distribution.

How to choose η? In Figure 6, we set η at various arbitrary fixed values that correspond to model perturbations associated with *local* (η small) or *global* (η larger) sensitivity analysis, in a single unified framework for both. On the other hand, from a practical point of view η can be calculated as the KL distance of the baseline model P from the available data set from real model. To this end, if Q can be constructed from a data set coming from the real model by using the histogram or more generally with a kernel density estimator; then we can estimate $\eta = R(Q||P)$ directly. For example, we consider Q to be a histogram with density given by

$$q^{hist}(x) = \sum_{k=1}^{m} \frac{\nu_k}{nh} I_{B_k}(x). \tag{80}$$

Here $B_1, \ldots, B_m$ denote the histogram bins, h is the bin width, n is the number of observations and ν_k is the number of data in B_k, while I_{B_k} is the indicator function on the bin B_k. We plot the histogram of the data with $h = 100$ as well as the fitted Weibull distribution in Figure 7. The corresponding relative entropy between the histogram Q and the baseline model P is $\eta = R(Q||P) \approx 0.8$. In fact, as we also readily see in Figure 7, the Weibull is not a good fit for the data, and that is why we have a larger value of η here. In order to improve the performance of the predictive model, we can either acquire more data or choose a better family of parametric or nonparametric densities to fit the data in place of the Weibull, but at the same time avoid overfitting.

B. Uncertainty Quantification for Markov Random Fields

Here we consider the impact on QoIs of localized perturbations to statistical probability distributions of Markov Random Fields [50] such as Gibbs measures. Such distributions are inherently high-dimensional, allowing us to focus on this aspect of model bias bounds. In particular, we consider Gibbs measures for particle systems defined on a fixed finite subset Λ_N of the infinite dimensional lattice $\mathbb{Z}^d$. Specifically we consider $\Lambda_N = \{x \in \mathbb{Z}^d, |x_i| \leq n\}$ the square lattice with $N = (2n+1)^d$ lattice sites, where typically $n \gg 1$. Before we describe the model, we will specify some necessary notation: we let S be the configuration space of a single particle at a lattice site $x \in \mathbb{Z}^d$. For example in a lattice gas model $S = \{0, 1\}$, i.e. the lattice site can be empty or occupied, and in a Potts model $S = \{0, 1, ..., q\}$, i.e. the site is empty or occupied by particles of q different species. In Ising magnetization models studied below, we have that $S = \{-1, 1\}$, corresponding to down or up spins respectively. Then S^X is the configuration space for the particles in any subset $X \subset \mathbb{Z}^d$; we denote by $\sigma_X = \{\sigma_x\}_{x \in X}$ an element of S^X. Next, in order to define a Gibbs measure on Λ_N, we first specify the Hamiltonian $H_N(\sigma_{\Lambda_N})$ of a set of particles in the region Λ_N. An interaction $\Phi = \{\Phi_X : X \subset \mathbb{Z}^d, X \text{ finite}\}$ associates to any finite subset X a function $\Phi_X(\sigma_X)$ which depends only on the particle configuration in X and accounts for all particle interactions within X, see [25] for details. Given an interaction Φ we then define the Hamiltonian H_N^Φ (with free boundary conditions) by

$$H_N^\Phi(\sigma_{\Lambda_N}) = \sum_{X \subset \Lambda_N} \Phi_X(\sigma_X), \tag{81}$$

and Gibbs measure μ_N^Φ by

$$d\mu_N^\Phi(\sigma_{\Lambda_N}) = \frac{1}{Z_N^\Phi} e^{-H_N(\sigma_{\Lambda_N})} dP_N(\sigma_{\Lambda_N}), \tag{82}$$

where P_N is the counting measure on S^{Λ_N} and $Z_N^\Phi = \sum_{\sigma_{\Lambda_N}} e^{-H_N(\sigma_{\Lambda_N})}$ is the normalization constant, also known as the partition function, [25].

Here we consider classes of perturbed models with corresponding interaction Ψ that includes only local perturbations to the interaction Φ, e.g. *local defects* encoded in the interaction potential J, or localized perturbations to the external field h in the example of the Ising-type Hamiltonian (84). We also note that defects of finite temperature multi-scale probability distributions are a continuous source of interest in the computational materials science community, see, for instance, [51]; in fact, lattice probability distributions such as (82), constitute an important class of simplified prototype problems. In the case of localized perturbations to the interaction Φ in (81), the Hamiltonians scale as follows:

$$H_N^\Psi(\sigma_{\Lambda_N}) = H_N^\Phi(\sigma_{\Lambda_N}) + O(1).$$

Thus the corresponding relative entropy satisfies

$$\begin{aligned} R(\mu_N^\Psi \,||\, \mu_N^\Phi) &= \log E_{\mu_N^\Psi}(e^{\Delta H}) + E_{\mu_N^\Psi}(-\Delta H) \\ &= O(1), \end{aligned} \tag{83}$$

uniformly in the system size N, where we define $\Delta H = H_N^\Psi - H_N^\Phi$. However, in most cases, we do not know the exact local perturbation as well as the perturbed Gibbs measure μ_Ψ. Instead, based on (83) we can consider a family of perturbed models

$$\mathcal{Q}_\eta = \{\mu_\Psi : R(\mu_N^\Psi \| \mu_N^\Phi) \leq \eta^2\}.$$

This family will include any perturbation *within that tolerance* η^2, for example: defects located at different lattice sites, and of different magnitudes, as the scaling (83) demonstrates rigorously.

As a concrete example of a Hamiltonian (81), we consider μ_Φ to be a one-dimensional Ising model probability distributions on the one-dimensional lattice Λ_N, labeled successively

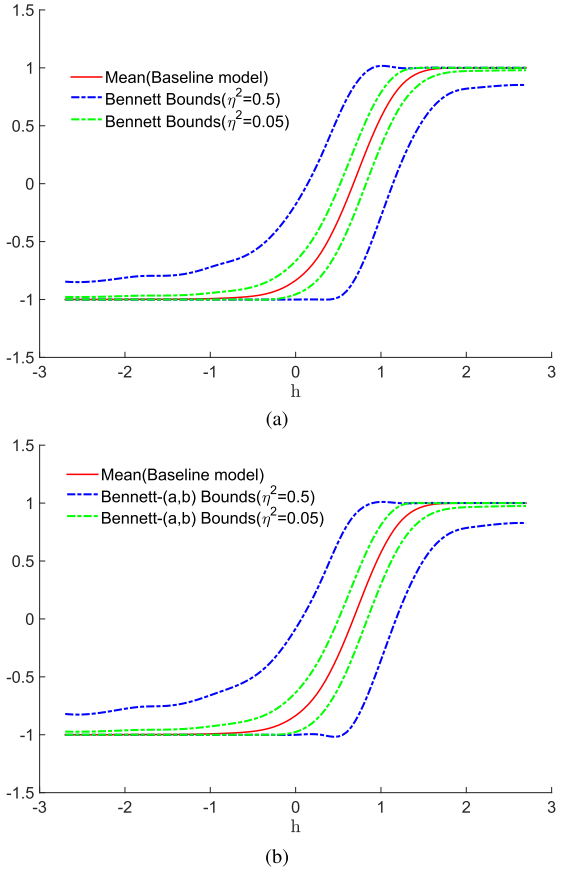

Fig. 8. (a) The red line is the mean of the QoI (85) with $m = 1$ for the baseline model (84) with $J = 1$ and $\beta = 1$. The green and blue lines are the Bennett bounds for $\eta^2 = 0.05$ and $\eta^2 = 0.5$, respectively. (b) The red line is the mean of the same QoI for the baseline model (84) with $J = 1$ and $\beta = 1$; The green and blue lines are the Bennett-(a,b) bounds for $\eta^2 = 0.05$ and $\eta^2 = 0.5$, respectively. In both figures the lattice size is N=100.

by $x = 1, 2, ..., N$. To each site corresponds a spin $\sigma(x)$, with two possible values: $+1$ or -1. The Hamiltonian is given by

$$H_N^{\Phi}(\sigma_{\Lambda_N}) = -\beta \sum_{x=1}^{N-1} J(x)\sigma(x)\sigma(x+1) - \beta h \sum_{x=1}^{N} \sigma(x). \tag{84}$$

Using the concentration/information inequalities developed in Section III, we can obtain model bias bounds for QoIs, such as the localized average around any lattice site x,

$$f(\sigma_{\Lambda_N}) = \frac{1}{2m+1} \sum_{\{y:|y-x|\leq m\}} \sigma(y), \tag{85}$$

for a fixed radius m. In the demonstration below we we pick $m = 1$ for concreteness. Since the QoI f (85) is bounded, $-1 \leq f \leq 1$, we can use the Bennett-(a,b) bound (41). Alternatively, we can use the Bennett bound (39)), which however requires estimating in addition to $E_{\mu_N^{\Phi}}[f]$, the variance $Var_{\mu_N^{\Phi}}[f]$ by sampling from μ_{Φ}, see also Section III-C. The latter is not unreasonable given that variance computations are necessary in many applications because they ensure suitable confidence intervals for the averaged QoIs. In Figure 8, we implement both Bennett and Bennett-(a,b) bounds by considering two different KL divergence tolerances, $\eta^2 = 0.5, 0.05$. A comparisons between Figure 8a and Figure 8b indicates that Bennett and Bennett-(a,b) bounds are fairly close for this example.

Notable computational advantages of these concentration/information inequalities over direct numerical simulation of alternative models $Q = \mu_{\Psi}$, as well as over the GO divergence bounds in Theorem 1 are the following: (1) when using Theorem 5 along with Bennett-type bounds (39) or (41), we can deploy computational resources to estimate $E_{\mu_N^{\Phi}}[f]$ or possibly $\text{Var}_{\mu_N^{\Phi}}[f]$—see also Table II—just for the baseline model $P = \mu_{\Phi}$, instead of simulating all alternative models $Q = \mu_{\Psi}$ models; (2) we do not need to use the full GO divergence bounds in Theorem 1, which require potentially expensive full MGF calculations, also recalling Remark 11.

VIII. Conclusion

In this paper we combined the uncertainty quantification information inequality of [4], [19], [20] together with classical concentration inequalities [31] to obtain easily implementable bounds for the model bias of quantities of interest (QoIs). The bounds control the model bias in terms of the relative entropy between different models and intrinsic statistical quantities associated to the QoIs in a baseline model, e.g. mean, variance, L^{∞} bound. Our results improve substantially on classical information bounds such as the Pinsker inequality. First, our bound scales correctly with the size of the data sets/number of degrees of freedom while classical inequalities do not, see Remark 21. This scaling property is illustrated in Section V where we discuss bias bounds for general statistical estimators. In addition, we demonstrate the tightness of our bounds in Sections II and IV: given suitable families of QoIs and a family of models whose Kullback-Leibler divergence with respect to a given baseline model is less than a tolerance η^2, there always exists a QoI and models which saturate the upper and lower bounds. This demonstrates rigorously the precise sense our model bias bound is optimal. This approach can be generalized in various ways. The recent preprint [38] proves UQ bounds for Markov process in the long-time regime by using concentration inequalities obtained via spectral gap estimates, Poincaré and log-Sobolev inequalities. In forthcoming work we will study further bias in phase diagrams of Gibbs-Markov random fields and for molecular dynamics, as well as model bias of coarse-grained models for equilibrium and non-equilibrium molecular dynamics built via variational inference methods, [9], [52].

Appendix
Proofs of Theorem 1 and Theorem 2

In the appendix we use the notation $\Lambda(c) = \log M_P(c; \tilde{f})$ so that the GO divergence is

$$\Xi(Q\|P; \pm f) = \inf_{c>0} \left\{ \frac{\Lambda(\pm c) + R(Q\|P)}{c} \right\}$$

Note that $\Lambda(c)$ is convex function which we assume to be finite on an interval (d_-, d_+) with $d_- < 0 < d+$. On that

interval $\Lambda(c)$ is infinitely differentiable and strictly convex. Since we centered the QoI we have $\Lambda(0) = \Lambda'(0) = 0$ and $\Lambda''(0) = \mathrm{var}_P[f]$.

Proof of Theorem 1: We first establish the property 1) of the GO divergence. To show that the GO divergence is non-negative we note that the KL divergence $R(Q\|P)$ is non-negative and that by Jensen's inequality,

$$\begin{aligned}\Lambda(c) = \log E_P[e^{c(f-E_P[f])}] &\geq \log e^{E_P[c(f-E_P[f])]} \\ &= E_P[f - E_P[f]] = 0.\end{aligned}$$

and thus $\Xi(Q\|P; f)$ is non-negative.

If $f = E_P[f]$ is constant then $\Lambda(c) = 0$ and since $R(Q\|P) \in [0, \infty)$,

$$\Xi(Q\|P; f) = \inf_{c>0} \frac{R(Q\|P)}{c} = 0.$$

If $Q = P$ then $R(Q\|P) = 0$ and

$$0 \leq \Xi(Q\|P; f) = \inf_{c>0} \frac{\Lambda(c)}{c} \leq \lim_{c\to 0} \frac{\Lambda(c)}{c} = \Lambda'(0) = 0.$$

Conversely assume that $\Xi(Q\|P; f) = 0$, then we may assume that $R(Q\|P) > 0$ otherwise we have $Q = P$. Then we must have $d_+ = \infty$ and the infimum must be obtained in the limit $c \to \infty$. Since $\Lambda(c)/c = \frac{1}{c}\int_0^c \Lambda'(s)ds$ and Λ' is non-decreasing then $\Lambda(c)/c$ is also non-decreasing. Since we have $\lim_{c\to 0} \frac{\Lambda(c)}{c} = 0$ and $\lim_{c\to\infty} \frac{\Lambda(c)}{c} = 0$ we must have $\Lambda(c) = 0$ for all $c \geq 0$. But then $\Lambda''(0) = \mathrm{var}_P[f] = 0$ which implies that f is constant P-a.s. This concludes the proof of Property 1) in Theorem 1.

We turn next to Property 2). We set $\eta = \sqrt{R(Q\|P)}$ so that

$$\Xi(Q||P; f) = \inf_{c\geq 0}\left\{\frac{\Lambda(c) + \eta^2}{c}\right\}.$$

Recall that we have $\Lambda(0) = \Lambda'(0) = 0$ and

$$\Lambda''(0) \equiv \kappa_2 = \mu_2 \quad \Lambda'''(0) \equiv \kappa_3 = \mu_3,$$

where $\kappa_k = \Lambda^{(k)}(0)$ is the k-th cumulant and $\mu_k = \mathbb{E}_P[(f - \mathbb{E}_P[f])^k]$ is k-th the centered moment of f. For $c^* = c^*(\eta)$ to be a minimum we must have

$$c^*\Lambda'(c^*) - \Lambda(c^*) - \eta^2 = 0$$

and then the minimum is equal to

$$\inf_{c\geq 0} \frac{\Lambda(c) + \eta^2}{c} = \Lambda'(c^*)$$

Since $c(0) = 0$ we expand c in powers of η

$$c(\eta) = c_1\eta + c_2\eta^2 + O(\eta^3)$$

and Λ in powers of c

$$\Lambda(c) = \frac{1}{2}\kappa_2\ c^2 + \frac{1}{6}\kappa_3\ c^3 + 0(c^4).$$

We have

$$c^2(\eta) = c_1^2\eta^2 + 2\ c_1c_2\eta^3 + O(\eta^4) \qquad c^3(\eta) = c_1^3\eta^3 + O(\eta^4)$$

and thus

$$\begin{aligned}\eta^2 &= c\Lambda'(c) - \Lambda(c) \\ &= \kappa_2\ c^2 + \frac{1}{2}\kappa_3\ c^3 - (\frac{1}{2}\kappa_2\ c^2 + \frac{1}{6}\kappa_3\ c^3) + O(c^4) \\ &= \frac{1}{2}\kappa_2 c^2 + \frac{1}{3}\kappa_3\ c^3 + O(c^4) \\ &= \frac{1}{2}\kappa_2 c_1^2\eta^2 + \left(\kappa_2\ c_1c_2 + \frac{1}{3}\kappa_3 c_1^3\right)\eta^3 + O(\eta^4)\end{aligned}$$

from which we obtain

$$\begin{aligned}1 &= \frac{1}{2}\kappa_2\ c_1^2 \implies c_1 = \sqrt{\frac{2}{\kappa_2}} \\ 0 &= \kappa_2\ c_1\ c_2 + \frac{1}{3}\kappa_3\ c_1^3 \Longrightarrow c_2 = -\frac{1}{3}\frac{\kappa_3}{\kappa_2}c_1^2 = -\frac{2}{3}\frac{\kappa_3}{\kappa_2^2}\end{aligned}$$

Finally we have

$$\begin{aligned}\Lambda'(c) &= \kappa_2\ c + \frac{1}{2}\kappa_3\ c^2 + 0(c^3) \\ &= \kappa_2\ c_1\eta + \kappa_2\ c_2\eta^2 + \frac{1}{2}\kappa_3\ c_1^2\eta^2 \\ &= \sqrt{2}\sqrt{\kappa_2}\eta + \frac{1}{3}\frac{\kappa_3}{\kappa_2}\eta^2\end{aligned}$$

If we use the skewness

$$\gamma(f) = \frac{\mathbb{E}_P\left[(f - \mathbb{E}_P[f])^3\right]}{\mathrm{var}_P[f]^{3/2}} = \frac{\kappa_3}{\kappa_2^{3/2}}$$

we have

$$\inf_{c\geq 0} \frac{\Lambda(c) + \eta^2}{c} = \sqrt{2\mathrm{var}_P[f]}\eta \pm \frac{1}{3}\sqrt{\mathrm{var}_P[f]}\gamma(f)\eta^2 + O(\eta^3).$$

By replacing f by $-f$ we obtain the optimization problem for $\Lambda(-c)$ and thus

$$\begin{aligned}\Xi(Q\ ||\ P; \pm f) = &\sqrt{\mathrm{var}_P[f]}\sqrt{2R(Q\ ||\ P)} \\ &+ \frac{1}{3}\gamma\sqrt{\mathrm{var}_P[f]}R(Q\|P) + O((R(Q\|P))^{\frac{3}{2}})\end{aligned}$$

and this proves Property 2).

Proof of Theorem 2: Theorem 2 is contained in parts 1) and 2) of the following Theorem which gives more details on the case where $\eta_\pm$ is finite. Most ingredients of the proof of Theorem 2 are already present in [4], [20], (see in particular [20][Theorem 2.9]). In its present form Theorem 23 is proved in [27][Proposition 3]. We note that the proof can also be carried out using Lagrange multipliers taking into account the KL inequality constraint, see for instance [23].

Theorem 23: Suppose (d_-, d_+) is the largest open set such that $\Lambda(c) = \log M_P(c; \tilde{f}) < \infty$ for all $c \in (d_-, d_+)$.

1) For any $M \geq 0$ the optimization problems

$$\inf_{c>0} \frac{\Lambda(\pm c) + M}{c}$$

have unique minimizers $c^\pm \in [0, \pm d_\pm]$. Let $M_\pm$ be defined by

$$M_\pm = \lim_{c\nearrow \pm d_\pm} \pm c\Lambda'(\pm c) - \Lambda(\pm c).$$

Then the minimizers $c_\pm = c_\pm(M)$ are finite for $M < M_\pm$ and $c_\pm(M) = \pm d_\pm$ if $M \geq M_\pm$.

2) If $c_\pm(M) < \pm d_\pm$ then

$$\frac{\Lambda(\pm c_\pm)+M}{c_\pm} = \inf_{c>0} \frac{\Lambda(\pm c)+M}{c} = \pm\Lambda'(\pm c_\pm) = \pm\Big(\mathbb{E}_{P_{\pm c_\pm}}[f] - \mathbb{E}_P[f]\Big), \tag{86}$$

where $c_\pm(M)$ is strictly increasing in M and is determined by the equation

$$\mathcal{R}\big(P_{\pm c_\pm} \,||\, P\big) = M. \tag{87}$$

3) $M_\pm$ is finite in two distinct cases.
   a) If $\pm d_\pm < \infty$ (in which case g must be unbounded above/below) $M_\pm$ is finite if $\lim_{c\to\pm d_\pm} \Lambda(\pm c) := \Lambda(d_\pm) < \infty$ and $\lim_{c\to\pm d_\pm} \pm\Lambda'(\pm c) := \pm\Lambda'(d_\pm) < \infty$, and for $M \geq M_\pm$ we have

$$\inf_{c>0} \frac{\Lambda(\pm c)+M}{c} = \frac{\Lambda(d_\pm)+M}{\pm d_\pm} = \pm\Big(\mathbb{E}_{P_{d_\pm}}[f] - \mathbb{E}_P[f]\Big) + \frac{M-M_\pm}{\pm d_\pm}. \tag{88}$$

   b) If $\pm d_\pm = \infty$ and $M_\pm$ is finite then f is P-a.s. bounded above/below and for $M \geq M_\pm$ we have

$$\inf_{c>0} \frac{\Lambda(\pm c)+M}{c} = \operatorname{ess\,sup}_{x\in\mathcal{X}}\{\pm(f(x) - \mathbb{E}_P[f])\}. \tag{89}$$

Proof of Theorem 23: First note that it is enough to prove the result for $\Lambda(c)$ since the result for $\Lambda(-c)$ is obtained by replacing f by $-f$. We also use the notation $\tilde{f}_+ = \operatorname{ess\,sup}\{f(x) - \mathbb{E}_P[f]\}$.

We first claim that automatically

$$\Lambda(d_+) = \lim_{c\nearrow d_+} \Lambda(c),$$

where $\Lambda(d_+)$ may be infinite. By monotone convergence

$$\mathbb{E}_P[1_{\{\tilde{f}\geq 0\}} e^{c\tilde{f}}] \nearrow \mathbb{E}_P[1_{\{\tilde{f}\geq 0\}} e^{d_+\tilde{f}}]$$

as $c \nearrow d_+$. By dominated convergence

$$\mathbb{E}_P[1_{\{\tilde{f}<0\}} e^{c\tilde{f}}] \searrow \mathbb{E}_P[1_{\{\tilde{f}<0\}} e^{d_+\tilde{f}}]$$

as $c \nearrow d_+$, and the claim follows. A very similar argument shows that $\Lambda'(c)$ also has a limit as $c \nearrow d_+$.

Let

$$B(c; M) = \frac{\Lambda(c)+M}{c}. \tag{90}$$

We divide into cases.

1) $\tilde{f}_+ < \infty$. In this case $\Lambda'(c) \nearrow \tilde{f}_+ < \infty$ as $c \to \infty$ and $\Lambda'(0) < \tilde{f}_+$. If $M = 0$ then the infimum is $\Lambda'(0)$ and attained at $c_+ = 0$ since $\Lambda(c)/c$ is an increasing function. If $M > 0$ then

$$B'(c; M) = \frac{c\Lambda'(c) - \Lambda(c) - M}{c^2}$$

for $c \geq 0$. The function $c\Lambda'(c) - \Lambda(c)$ strictly increases from 0 at $c = 0$ to some limit $M_+ > 0$ at $c = \infty$, and the minimizer is at the unique finite root of $c\Lambda'(c) - \Lambda(c) = M$ for $M < M_+$ and $c_+ = \infty$ for $M \geq M_+$.

2) $\tilde{f}_+ = \infty$. In this case there are two subcases.
   a) $d_+ = \infty$. In this case since $\tilde{f}_+ = \infty$ we have $\Lambda'(c) \nearrow \infty$ as $c \to \infty$ and $c\Lambda'(c) - \Lambda(c) \to \infty$ as $c \to \infty$. Since $0\Lambda'(0) - \Lambda(0) = 0$, in all cases of $M \geq 0$ there is a unique root to $c\Lambda'(c) - \Lambda(c) = M$ and hence a unique minimizer.
   b) $d_+ < \infty$. We know that $\Lambda'(c)$ converges as $c \nearrow d_+$ to a well defined left hand limit which we call $\Lambda'(d_+)$ (note that this value could be ∞). Thus we have that $c\Lambda'(c) - \Lambda(c)$ ranges from 0 at $c = 0$ to $M_+ = d_+\Lambda'(d_+) - \Lambda(d_+)$. For $M \in [0, M_+)$ there is a unique minimizer in $[0, d_+)$. For $M \geq M_+$ the unique minimizer is at $c_+ = d_+$.

To conclude the proof we note that if $c_+ < d_+$ then an easy computation shows that

$$c_+\Lambda'(c_+) - \Lambda(c_+) = \mathcal{R}(P_{c_+} \,||\, P) = M,$$

and thus

$$B(c_+, M) = \Lambda'(c_+) = \mathbb{E}_{P_{c_+}}[f] - \mathbb{E}_P[f]$$

which proves (86) and (87). Finally if $d_+ = \infty$ and g is P-a.s. bounded above then the infimum is equal to $\lim_{c\to\infty} \frac{\Lambda(c)}{c}$ and this establishes (89). If $d_+ < \infty$ and $M_+ < \infty$ then the bound takes the form (88). □

REFERENCES

[1] M. S. Eldred *et al.*, "Dakota, a multilevel parallel object-oriented framework for design optimization, parameter estimation, uncertainty quantification, and sensitivity analysis: Version 5.0 user's manual," Sandia Nat. Lab., Albuquerque, NM, USA, Tech. Rep. SAND2010-2183, Dec. 2009.

[2] R. C. Smith, *Uncertainty Quantification: Theory, Implementation, and Applications*. Philadelphia, PA, USA: Society for Industrial and Applied Mathematics, 2013.

[3] A. Saltelli *et al.*, *Global Sensitivity Analysis: The Primer*. Hoboken, NJ, USA: Wiley, 2008.

[4] K. Chowdhary and P. Dupuis, "Distinguishing and integrating aleatoric and epistemic variation in uncertainty quantification," *ESAIM, Math. Model. Numer. Anal.*, vol. 47, no. 3, pp. 635–662, Mar. 2013.

[5] T. J. Sullivan, *Introduction to Uncertainty Quantification*. Cham, Switzerland: Springer, 2015.

[6] A. B. Tsybakov, *Introduction to Nonparametric Estimation*. New York, NY, USA: Springer, 2008.

[7] C. M. Bishop, *Pattern Recognition and Machine Learning*. New York, NY, USA: Springer-Verlag, 2006.

[8] K. P. Burnham and D. R. Anderson, *Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach*. New York, NY, USA: Springer, 2003.

[9] A. Chaimovich and M. S. Shell, "Relative entropy as a universal metric for multiscale errors," *Phys. Rev. E, Stat. Phys. Plasmas Fluids Relat. Interdiscip. Top.*, vol. 81, no. 6, Jun. 2010, Art. no. 060104.

[10] M. A. Katsoulakis and P. Plecháč, "Information-theoretic tools for parametrized coarse-graining of non-equilibrium extended systems," *J. Chem. Phys.*, vol. 139, no. 7, Aug. 2013, Art. no. 074115.

[11] J. F. Rudzinski and W. G. Noid, "Coarse-graining entropy, forces, and structures," *J. Chem. Phys.*, vol. 135, no. 21, Dec. 2011, Art. no. 214101.

[12] A. J. Majda, R. V. Abramov, and M. J. Grote, *Information Theory and Stochastics for Multiscale Nonlinear Systems* (CRM Monograph Series). Providence, RI, USA: American Mathematical Society, 2005.

[13] A. Atkinson, A. Doney, and R. Tobias, *Optimum Experimental Designs, With SAS*. London, U.K.: Oxford Univ. Press, 2007.

[14] A. J. Majda and B. Gershgorin, "Quantifying uncertainty in climate change science through empirical information theory," *Proc. Nat. Acad. Sci. USA*, vol. 107, no. 34, pp. 14958–14963, Aug. 2010.

[15] A. J. Majda and B. Gershgorin, "Improving model fidelity and sensitivity for complex systems through empirical information theory," *Proc. Nat. Acad. Sci. USA*, vol. 108, no. 25, pp. 10044–10049, Jun. 2011.
[16] M. Komorowski, M. J. Costa, D. A. Rand, and M. P. H. Stumpf, "Sensitivity, robustness, and identifiability in stochastic chemical kinetics models," *Proc. Nat. Acad. Sci. USA*, vol. 108, no. 21, pp. 8645–8650, May 2011.
[17] Y. Pantazis and M. A. Katsoulakis, "A relative entropy rate method for path space sensitivity analysis of stationary complex stochastic dynamics," *J. Chem. Phys.*, vol. 138, no. 5, Feb. 2013, Art. no. 054115.
[18] T. M. Cover and J. A. Thomas, *Elements of Information Theory*. Hoboken, NJ, USA: Wiley, Jul. 2006.
[19] M. A. Katsoulakis, L. Rey-Bellet, and J. Wang, "Scalable information inequalities for uncertainty quantification," *J. Comput. Phys.*, vol. 336, pp. 513–545, May 2017.
[20] P. Dupuis, M. A. Katsoulakis, Y. Pantazis, and P. Plecháč, "Path-space information bounds for uncertainty quantification and sensitivity analysis of stochastic dynamics," *SIAM/ASA J. Uncertain. Quantif.*, vol. 4, no. 1, pp. 80–111, Jan. 2016.
[21] T. Breuer and I. Csiszár, "Systematic stress tests with entropic plausibility constraints," *J. Banking Finance*, vol. 37, no. 5, pp. 1552–1559, May 2013. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0378426612001112
[22] H. Lam, "Robust sensitivity analysis for stochastic systems," *Math. Oper. Res.*, vol. 41, no. 4, pp. 1248–1275, Nov. 2016.
[23] P. Glasserman and X. Xu, "Robust risk measurement and model risk," *Quant. Finance*, vol. 14, no. 1, pp. 29–58, Sep. 2013.
[24] P. G. Dupuis and R. S. Ellis, *A Weak Convergence Approach to the Theory of Large Deviations* (Wiley Series in Probability and Statistics). New York, NY, USA: Wiley, 1997.
[25] B. Simon, *The Statistical Mechanics of Lattice Gases*. Princeton, NJ, USA: Princeton Univ. Press, 2014.
[26] R. Atar, K. Chowdhary, and P. Dupuis, "Robust bounds on risk-sensitive functionals via Rényi divergence," *SIAM/ASA J. Uncertain. Quantif.*, vol. 3, no. 1, pp. 18–33, Jan. 2015.
[27] P. Dupuis, M. A. Katsoulakis, Y. Pantazis, and L. Rey-Bellet, "Sensitivity analysis for rare events based on Rényi divergence," *Ann. Appl. Probab.*, 2019. [Online]. Available: https://imstat.org/journals-and-publications/annals-of-applied-probability/annals-of-applied-probability-future-papers/
[28] J. S. Liu, *Monte Carlo Strategies in Scientific Computing* (Springer Series in Statistics). New York, NY, USA: Springer-Verlag, 2001.
[29] T. Lelièvre, G. Stoltz, and M. Rousset, *Free Energy Computations: A Mathematical Perspective*. Singapore: World Scientific, 2010.
[30] P. Del Moral, A. Doucet, and A. Jasra, "On adaptive resampling strategies for sequential Monte Carlo methods," *Bernoulli*, vol. 18, no. 1, pp. 252–278, Feb. 2012.
[31] S. Boucheron, G. Lugosi, P. Massart, and M. Ledoux, *Concentration Inequalities: A Nonasymptotic Theory of Independence*. Oxford, U.K.: Oxford Univ. Press, 2013.
[32] M. Raginsky and I. Sason, "Concentration of measure inequalities in information theory, communications, and coding," *Found. Trends Commun. Inf. Theory*, vol. 10, nos. 1–2, pp. 1–247, 2013.
[33] M. Ledoux, *The Concentration of Measure Phenomenon*. Providence, RI, USA: American Mathematical Society, 2005.
[34] A. Dembo and O. Zeitouni, *Large Deviations Techniques and Applications*, vol. 38. Berlin, Germany: Springer-Verlag, 2010.
[35] P. Massart, *Concentration Inequalities and Model Selection*. Berlin, Germany: Springer, 2007.
[36] J. A. Tropp, "An introduction to matrix concentration inequalities," *Found. Trends Mach. Learn.*, vol. 8, nos. 1–2, pp. 1–230, 2015.
[37] E. J. Hall and M. A. Katsoulakis, "Robust information divergences for model-form uncertainty arising from sparse data in random PDE," *SIAM/ASA J. Uncertain. Quantif.*, vol. 6, no. 4, pp. 1364–1394, Jan. 2018, doi: 10.1137/17M1143344.
[38] J. Birrell and L. Rey-Bellet, "Uncertainty quantification for Markov processes via variational principles and functional inequalities," 2018, *arXiv:1812.05174*. [Online]. Available: http://arxiv.org/abs/1812.05174
[39] L. Wasserman, *All of Statistics: A Concise Course in Statistical Inference*. New York, NY, USA: Springer-Verlag, 2004.
[40] J. Li and D. Xiu, "Computation of failure probability subject to epistemic uncertainty," *SIAM J. Sci. Comput.*, vol. 34, no. 6, pp. A2946–A2964, Jan. 2012.
[41] S. G. Bobkov and F. Götze, "Exponential integrability and transportation cost related to logarithmic sobolev inequalities," *J. Funct. Anal.*, vol. 163, no. 1, pp. 1–28, Apr. 1999.
[42] W. Hoeffding, "Probability inequalities for sums of bounded random variables," *J. Amer. Stat. Assoc.*, vol. 58, no. 301, pp. 13–30, 1963.
[43] C. Villani, *Optimal Transport: Old and New*. Berlin, Germany: Springer, 2008.
[44] G. Carlier, A. Galichon, and F. Santambrogio, "From Knothe's transport to Brenier's map and a continuation method for optimal transport," *SIAM J. Math. Anal.*, vol. 41, no. 6, pp. 2554–2576, Jan. 2010.
[45] C. McDiarmid, "On the method of bounded differences," *Surv. Combinatorics*, vol. 141, no. 1, pp. 148–188, 1989.
[46] L. Wasserman, *All of Nonparametric Statistics*. New York, NY, USA: Springer-Verlag, 2006.
[47] S. Boucheron, G. Lugosi, and P. Massart, "On concentration of self-bounding functions," *Electron. J. Probab.*, vol. 14, pp. 1884–1899, 2009.
[48] S.-W. Eom, M.-K. Kim, I.-J. Kim, S.-I. Moon, Y.-K. Sun, and H.-S. Kim, "Life prediction and reliability assessment of lithium secondary batteries," *J. Power Sources*, vol. 174, no. 2, pp. 954–958, Dec. 2007.
[49] A. Papoulis and S. U. Pillai, *Probability, Random Variables, and Stochastic Processes*. New York, NY, USA: McGraw-Hill, 2002.
[50] D. J. C. MacKay, *Information Theory, Inference and Learning Algorithms*. Cambridge, U.K.: Cambridge Univ. Press, 2003.
[51] J. Marian, G. Venturini, B. L. Hansen, J. Knap, M. Ortiz, and G. H. Campbell, "Finite-temperature extension of the quasicontinuum method using langevin dynamics: Entropy losses and analysis of errors," *Model. Simul. Mater. Sci. Eng.*, vol. 18, no. 1, Dec. 2009, Art. no. 015003.
[52] V. Harmandaris, E. Kalligiannaki, M. Katsoulakis, and P. Plecháč, "Path-space variational inference for non-equilibrium coarse-grained systems," *J. Comput. Phys.*, vol. 314, pp. 355–383, Jun. 2016.

Konstantinos Gourgoulias received the bachelor's degree in applied mathematics from the University of Crete in 2011, and the Ph.D. degree in mathematics from the University of Massachusetts Amherst in 2017. He is currently a Senior Research Scientist with the Artificial Intelligence Research Lab, Babylon Health, co-leading the modelling and inference research efforts of the lab. His research interests include the development of robust machine learning models under uncertainty, uncertainty quantification, applications of information theory, and causal reasoning.

Markos A. Katsoulakis received the bachelor's degree from the University of Athens, Greece, in 1987, and the Ph.D. degree in applied mathematics from Brown University in 1993. He is a Professor at the Department of Mathematics and Statistics, University of Massachusetts, Amherst, MA, USA, where he is currently the Director of the Center for Applied Mathematics and Computation. His research interests include probabilistic and multiscale predictive modeling, uncertainty quantification, and information theory. He has also worked extensively in interdisciplinary collaborations involving chemical engineering, materials science, computer science, and atmospheric and oceanic sciences. He is a member of the Editorial Boards of the *SIAM/ASA Journal on Uncertainty Quantification*, the *SIAM Mathematical Modeling and Computation Book Series,* and the *Communications in Mathematical Sciences*. He was a member of the Editorial Board of the *SIAM Journal in Mathematical Analysis* (2002–2014). He is a member of the Executive Committee of the UMass TRIPODS Institute for Theoretical Foundations of Data Science.

Luc Rey-Bellet received the bachelor's degree in physics from the Swiss Federal Institute of Technology, Zurich, Switzerland, in 1994, and the Ph.D. degree in mathematics from the University of Geneva, Switzerland, in 1998. He is currently a Professor with the Department of Mathematics and Statistics, University of Massachusetts, Amherst, MA, USA. His research interests include statistical mechanics, applied probability, and uncertainty quantification as well as Monte-Carlo methods and game theory. His research was supported by multiple grants from the NSF, DOE, and AFOSR. He was an invited speaker at the International Congress of Mathematical Physics in 2003, and was a member of the Editorial Board of the *Journal of Statistical Physics* from 2013 to 2018.

Jie Wang received the bachelor's degree from Henan Normal University in 2010, the master's degree from the University of Chinese Academy of Science in 2013, and the Ph.D. degree in mathematics from the University of Massachusetts Amherst in 2019. She is currently a Modeler in card marketing and acquisition risk at Discover Financial Services. Her research interests include information theory, uncertainty quantification, and machine learning.