Journal of Computational Physics 336 (2017) 513–545

Contents lists available at ScienceDirect

Journal of Computational Physics

www.elsevier.com/locate/jcp

Scalable information inequalities for uncertainty quantification

Markos A. Katsoulakis, Luc Rey-Bellet, Jie Wang *

Department of Mathematics and Statistics, University of Massachusetts Amherst, Amherst, MA 01003, USA

ARTICLE INFO

Article history:
Received 20 April 2016
Received in revised form 30 January 2017
Accepted 6 February 2017
Available online 10 February 2017

Keywords:
Kullback Leibler divergence
Information metrics
Uncertainty quantification
Statistical mechanics
High dimensional systems
Nonlinear response
Phase diagrams

ABSTRACT

In this paper we demonstrate the only available scalable information bounds for quantities of interest of high dimensional probabilistic models. Scalability of inequalities allows us to (a) obtain uncertainty quantification bounds for quantities of interest in the large degree of freedom limit and/or at long time regimes; (b) assess the impact of large model perturbations as in nonlinear response regimes in statistical mechanics; (c) address model-form uncertainty, i.e. compare different extended models and corresponding quantities of interest. We demonstrate some of these properties by deriving robust uncertainty quantification bounds for phase diagrams in statistical mechanics models.

1. Introduction

Information Theory provides both mathematical methods and practical computational tools to construct probabilistic models in a principled manner, as well as the means to assess their validity, [1]. One of the key mathematical objects of information theory is the concept of information metrics between probabilistic models. Such concepts of distance between models are not always metrics in the strict mathematical sense, in which case they are called divergences, and include the relative entropy, also known as the Kullback–Leibler (KL) divergence, the total variation and the Hellinger metrics, the χ^2 divergence, the F-divergence, and the Rényi divergence, [2]. For example, the relative entropy between two probability distributions $P = P(\sigma)$ and $Q = Q(\sigma)$ on $\mathbb{R}^N$ is defined as

$$R(Q \parallel P) = \int_{\mathbb{R}^N} \log\left(\frac{Q(\sigma)}{P(\sigma)}\right) Q(\sigma) d\sigma, \tag{1}$$

when the integral exists. The relative entropy is not a metric but it is a divergence, that is it satisfies the properties: (i) $R(Q \parallel P) \geq 0$, (ii) $R(Q \parallel P) = 0$ if and only if $P = Q$ a.e.

We may for example think of the model Q as an approximation, or a surrogate model for another complicated and possibly inaccessible model P; alternatively we may consider the model Q as a misspecification of the true model P. When measuring model discrepancy between the two models P and Q, tractability depends critically on the type of distance used between models. In that respect, the relative entropy has very convenient analytic and computational properties, in

* Corresponding author.
E-mail addresses: markos@math.umass.edu (M.A. Katsoulakis), luc@math.umass.edu (L. Rey-Bellet), wang@math.umass.edu (J. Wang).

http://dx.doi.org/10.1016/j.jcp.2017.02.020
0021-9991/

particular regarding the scaling properties of the system size N which could represent space and/or time. Obtaining bounds which are valid for high dimensional ($N \gg 1$) or spatially extended systems and/or long time regimes is the main topic of the paper and we will discuss these properties in depth in the upcoming sections.

Information metrics provide systematic and practical tools for building approximate statistical models of reduced complexity through variational inference methods [3–5] for machine learning [6,7,4] and coarse-graining of complex systems [8–16]. Variational inference relies on optimization problems such as

$$\min_{Q\in\mathcal{Q}} R(P \parallel Q) \quad \text{or} \quad \min_{Q\in\mathcal{Q}} R(Q \parallel P)\,, \tag{2}$$

where $\mathcal{Q}$ is a class of simpler, computationally more tractable probability models than P. Subsequently, the optimal solution Q^* of (2) replaces P for estimation, simulation and prediction purposes. The choice of order in P and Q in (2) can be significant and depends on implementation methods, availability of data and the specifics of each application, e.g. [3,4,14,5]. In the case of coarse-graining the class of coarse-grained models $\mathcal{Q}$ will also have fewer degrees of freedom than the model P, and an additional projection operator is needed in the variational principle (2), see for instance [8,16]. In addition, information metrics provide fidelity measures in model reduction, [17–23], sensitivity metrics for uncertainty quantification, [24–29] and discrimination criteria in model selection [30,31]. For instance, for the sensitivity analysis of parametrized probabilistic models $P^\theta = P^\theta(\sigma)$, $\theta \in \Theta$ the relative entropy $R(P^\theta \parallel P^{\theta+\epsilon})$ measures the loss of information due to an error in parameters in the direction of the vector $\epsilon \in \Theta$. Different directions in parameter space provide a ranking of the sensitivities. Furthermore, when $|\epsilon| \ll 1$ we can also consider the quadratic approximation $R(P^\theta \parallel P^{\theta+\epsilon}) = \epsilon \mathbf{F}(P^\theta)\epsilon^\top + O(|\epsilon|^3)$ where $\mathbf{F}(P^\theta)$ is the Fisher Information matrix, [27,26,28].

Based on this earlier discussion it is natural and useful to approximate, perform model selection and/or sensitivity analysis in terms of information theoretical metrics between probability distributions. However, one is often interested in assessing model approximation, fidelity or sensitivity on concrete quantities of interest and/or statistical estimators. More specifically, if P and Q are two probability measures and $f = f(\sigma)$ is some quantity of interest or statistical estimator, then we measure the discrepancy between models P and Q with respect to the Quantity of Interest (QoI) f by considering the model bias,

$$E_Q(f) - E_P(f)\,. \tag{3}$$

Indeed, in a statistics context, f could be an unbiased statistical estimator for model P which is either complicated to compute or possibly inaccessible and Q is a computationally tractable nominal or reference model, e.g., a surrogate model. Thus, (3) is the estimator bias when using model Q instead of P. Alternatively, P could be the nominal model, for instance a model obtained through a careful statistical inference method, e.g. some type of best-fit approach such as maximum likelihood or variational inference, or just simply our best guess. However, due to the uncertainty whether this is a suitable model for the QoI f, we would like to measure the performance on f over a family of alternative models, for example all models within a KL tolerance η, $\mathcal{Q} = \{Q : R(Q \| P) \le \eta\}$. In this case, a bound on $\max_{Q\in\mathcal{Q}} |E_Q(f) - E_P(f)|$ will provide a performance guarantee for the model P. Our main mathematical goal is to understand how to transfer quantitative results on information metrics into bounds for quantities of interest in (3). In this direction, information inequalities can provide a method to relate quantities of interest (3) and information metrics (1), a classic example being the Csiszar–Kullback–Pinsker (CKP) inequality, [2]:

$$|E_Q(f) - E_P(f)| \le \|f\|_\infty \sqrt{2R(Q \parallel P)}\,, \tag{4}$$

where $\|f\|_\infty = \sup_{\sigma\in\mathbb{R}^N} |f(\sigma)|$. In other words relative entropy controls how large the model discrepancy (3) can become for the quantity of interest f. More such inequalities involving other probability metrics such as Hellinger distance, χ^2 and Rényi divergences are discussed in the subsequent sections.

In view of (4) and other such inequalities, a natural question is whether these are sufficient to assess the fidelity of complex systems models. In particular complex systems such as molecular or multi-scale models are typically high dimensional in the degrees of freedom and/or often require controlled fidelity (in approximation, uncertainty quantification, etc.) at long time regimes; for instance, in building coarse-grained models for efficient and reliable molecular simulation. Such an example arises when we are comparing two statistical mechanics systems determined by Hamiltonians H_N and $\bar{H}_N$ describing say N particles with positions $X = (x_1, ..., x_N)$. The associated canonical Gibbs measures are given by

$$P_N(X)dX = Z_N^{-1} e^{-H_N(X)} dX \quad \text{and} \quad Q_N(X)dX = \bar{Z}_N^{-1} e^{-\bar{H}_N(X)} dX\,, \tag{5}$$

where Z_N and $\bar{Z}_N$ are normalizations (known as partition functions) that ensure the measures (5) are probabilities. Example (5) is a ubiquitous one, given the importance of Gibbs measures in disparate fields ranging from statistical mechanics and molecular simulation, pattern recognition and image analysis, to machine and statistical learning, [32,3,4]. In the case of (5), the relative entropy (1) readily yields,

$$R(Q_N \parallel P_N) = E_{Q_N}(H_N - \bar{H}_N) + \log Z_N - \log \bar{Z}_N\,. \tag{6}$$