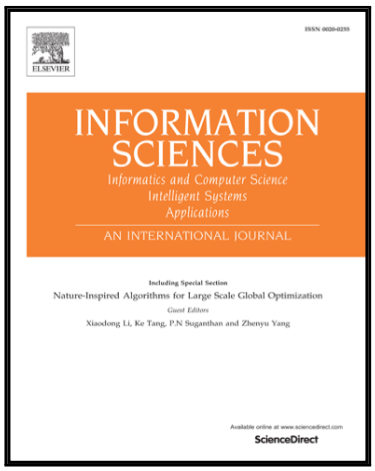

Fully Probabilistic Design of Hierarchical Bayesian Models

Anthony Quinn, Miroslav Kárný, Tatiana V. Guy

PII: S0020-0255(16)30516-3
DOI: 10.1016/j.ins.2016.07.035
Reference: INS 12361

To appear in: *Information Sciences*

Received date: 30 May 2015
Revised date: 8 July 2016
Accepted date: 13 July 2016

Please cite this article as: Anthony Quinn, Miroslav Kárný, Tatiana V. Guy, Fully Probabilistic Design of Hierarchical Bayesian Models, *Information Sciences* (2016), doi: 10.1016/j.ins.2016.07.035

Fully Probabilistic Design of Hierarchical Bayesian Models

Anthony Quinn[†1], **Miroslav Kárný**[*], **Tatiana V. Guy**[*]

[a]† *Trinity College Dublin, Dublin 2. Ireland.* aquinn@tcd.ie
∗*The Institute of Information Theory and Automation, The Czech Academy of Sciences. Czech Republic.* {karny,guy}@utia.cas.cz

Abstract

The minimum cross-entropy principle is an established technique for design of an unknown distribution, processing linear functional constraints on the distribution. More generally, fully probabilistic design (FPD) chooses the distribution—within the knowledge-constrained set of possible distributions—for which the Kullback-Leibler divergence to the designer's ideal distribution is minimized. These principles treat the unknown distribution deterministically. In this paper, fully probabilistic design is applied to hierarchical Bayesian models for thre first time, yielding optimal design of a (possibly nonparametric) stochastic model for the unknown distribution. This equips minimum cross-entropy and FPD distributional estimates with measures of uncertainty. It enables robust choice of the optimal model, as well as randomization of this choice. The ability to process non-linear functional constraints in the constructed distribution significantly extends the applicability of these principles. Currently available FPD procedures for a) merging of external knowledge, b) approximate learning and stabilized forgetting, c) decision strategy design, and d) local adaptive control design, are unified for the first time via the hierarchical FPD framework of this paper.

Keywords: Fully probabilistic design; ideal distribution; minimum cross-entropy principle; Bayesian conditioning; Kullback-Leibler divergence; Bayesian nonparametric modelling

[1] Corresponding author.

1. Introduction

A central concern in inductive science is to construct a model of an unknown quantity of interest, x, associated with the environment of an observer (also called a 'modeller' or 'designer'), $\mathcal{I}$, with which they[2] can only partially interact, and only partially understand. A fundamental task is for $\mathcal{I}$ optimally to construct their stochastic model for x, consistent with their knowledge and preferences. Knowledge, K, refers to anything that can serve towards model construction: theoretical or empirical facts, physical laws, values/observations/set-memberships for related variables and functions, expert opinions, *etc.* This knowledge rarely determines $\mathcal{I}$'s constructed model uniquely, and so they have freedom to select a model from among the possible knowledge-constrained models, while aiming to satisfy their preferences with respect to these models. In this paper, $\mathcal{I}$ expresses their preferences quantitatively via an ideal distribution. The key challenge is to provide a consistent, well-justified methodology with which $\mathcal{I}$ may choose their model uniquely and optimally, consistent with K, and taking account of these preferences.

The optimal choice of the model is undoubtedly the most important open question in the empirical sciences, known variously as inductive inference [11, 30], learning [31], knowledge and/or preference elicitation [5, 9, 22], *etc.* In this paper, we adopt the Bayesian philosophy, which builds the model, not of x itself, but of $\mathcal{I}$'s quantified beliefs about x. We adopt probability as an axiomatic framework [7] for such belief quantification, consistently addressing key inference tasks such as computation of conditional, total (marginal) and inverse beliefs [33].

If an explicit probability model conditioning $\mathcal{I}$'s knowledge, K, on possible values of x is available, then Bayes' rule is the consistent mechanism for processing K into a unique model for x itself. If—as is typically the case—this conditional model of K is not available, then the minimum cross-entropy[3] principle [38] may be applied. It extends the maximum entropy

[2]The gender-neutral pronoun—'they'—and gender-neutral possessive adjective—'their'—are used throughout.

[3]The Kullback-Leibler divergence (KLD) [28]—which we will recall in equation (5)—is also widely known in the literature as the *cross-entropy* [38]. In this paper, we will refer to the divergence itself as the KLD, and the design principle which minimizes it as the *minimum cross-entropy principle*, in reference to its first proposal in [38], or *fully*

principle [12], and provides a mechanism for processing K into an uncertainty (probability) model for x, when K includes the following elements[4]:

Remark 1 (K in the minimum cross-entropy (MXE) case).

(i) constraints on linear functionals of $\mathcal{I}$'s uncertainty model; and

(ii) an explicit model (i.e. choice or guess) of $\mathcal{I}$'s prior beliefs about x.

It has been shown [4] that the MXE principle and Bayesian conditioning are usually consistent—in the sense that they yield the same probability model (i.e. quantified belief) for x—in the case where they both process the same knowledge, K, as specified in Remark 1. Rare counter-examples are discussed in [4] . The necessity of these conditions for consistency of the MXE principle has not been explored in the literature, and this paper demonstrates that, indeed, further relaxations of the knowledge specification in Remark 1 are possible.

By relaxing the MXE-specified knowledge, K (Remark 1), the current paper extends its applicability to the probabilistic model design problem outlined above. To achieve this, we relax the ubiquitous constraint that $\mathcal{I}$'s uncertainty model for x conditioned on K—quantified via the distribution $\mathsf{A}(x|K)$—should be deterministic and should be constructed via deterministic optimization. For the first time, $\mathsf{A} \sim \mathsf{S}$ is treated also as an uncertain quantity (essentially an unknown parameter) in a *hierarchical* uncertainty model, M, for x and A, where S is an appropriately defined uncertainty model for A. Our purpose is to process knowledge and preferences relating to x and A. A consequence of this hierarchical setting, as we will see, is that it allows K to be expressed as *nonlinear* functionals of A (extending (i) in Remark 1). It is a concern, however, to ensure that the design remains consistent with Bayesian conditioning. We do this by proving that these more general constraints are consistent with K in Remark 1 *if* a hierarchical uncertainty model for A is adopted.

probabilistic design [23], depending on the context.

[4]The various knowledge specifications addressed in this paper will be summarized in remark environments such as this one. They differ, but we will use the same symbol, K, to refer to all cases. Confusion will be avoided by referring to the numbered remark where the specific knowledge is described.

Fully probabilistic design (FPD) [23, 25], which is the main focus of this paper, provides a generalization of the minimum cross-entropy principle for optimal model design. Firstly, it allows preferences about the model for x to be processed. Secondly, it does not require a prior for x (element (ii) in Remark 1) to be specified. In the hierarchical context, an explicit prior for x corresponds to a prior estimate or choice for the unknown model, A, indicative of the deterministic way in which the MXE principle handles unknown A. FPD does not require any such estimate or choice to be made.

FPD continues to be applied to practical problems in decision-making, control, machine learning and signal processing, as specified in the final item of the next paragraph. The extension of FPD to allow processing of nonlinear functional constraints—as developed in this paper for the first time—provides the unified and consistent theoretical setting for all these contexts.

In summary, the main original contributions of this paper are:

- The application of Fully Probabilistic Design (FPD) (i.e. the minimum cross-entropy (MXE) principle, and its special case, the maximum entropy (ME) principle) to the optimal design of hierarchical Bayesian models. This provides new theoretical insights into both.
- Hierarchical FPD equips the long-establised and widely applied designs of MXE and ME—crisp model choices without quantification of uncertainty—with measures of uncertainty. This also enables robust choice of the optimal model, as well as randomization of this choice.
- The extension of MXE and ME to allow processing of non-linear constraints in the constructed distribution significantly extends the applicability of these principles.
- Currently available FPD procedures for a) merging of external knowledge [20, 26], b) approximate recursive learning and stabilized forgetting [16, 17, 18], c) decision strategy design [2, 36], and d) local adaptive control design [24], are unified for the first time via the hierarchical FPD framework of this paper.

1.1. Layout of the Paper

Fully probabilistic design of unknown distributions and its relationship to the minimum cross-entropy principle and to Bayesian conditioning are

reviewed in Section 2. A complete hierarchical model for the unknown distribution is introduced, and the FPD-optimal design of this hierarchy is deduced under various specifications of the ideal distribution. The specialization to maximum-entropy-type designs—which arise when uniform ideals are adopted—is also explored. FPD-optimal designs for various functionally constrained sets of hierarchical distributions are derived in Section 3. These yield stochastic relaxations of the deterministic distributional estimates arising from conventional FPD and MXE. The various choices of functional constraints in the hierarchical context are fully explored, emphasizing the fact that nonlinear constraints can be accommodated. Specializations involving partially specified and noncommittal ideals are thoroughly explored. In Section 4, the requirement that the unknown distribution be finitely parameterized is relaxed, leading to the optimal design of nonparametric process distributions. The range of applications facilitated by FPD for hierarchical models is outlined in Section 5, key findings of the paper are discussed, and conclusions are drawn.

1.2. Notational Conventions

A lower-case math-italic symbol, e.g. x, denotes the realization of a (finite-dimensional) random variable. We do not specify its dimension or type, nor do we distinguish notationally between the random variable and its realization. The set of possible values of any quantity is denoted by the bold version of the symbol used for that quantity, e.g. $x \in \boldsymbol{x}$, $\mathsf{A} \in \mathbf{A}$. $|\boldsymbol{x}|$ denotes the counting measure of $\boldsymbol{x}$ when the latter has finite cardinality (i.e. when x is a discrete-valued random variable), and it denotes the Lebesgue measure of $\boldsymbol{x}$, otherwise. $\emptyset$ denotes the empty set, and $\times$—when operating on two sets—denotes their set product.

The specific probability distribution of x, conditional on some knowledge, K, is always denoted by an upper-case math-*sans-serif* letter, e.g. $x \sim \mathsf{A}(x|K)$, meaning that x, given K, is distributed as A. This notation refers interchangeably to the probability density function or probability mass function, depending on the type of x. The dominating measure with respect to which $\mathsf{A}(x)$ is defined is denoted by dx. The context will make clear which is meant in each case. The support of A, denoted $\mathrm{supp}(\mathsf{A})$, is the smallest subset of $\boldsymbol{x}$ having probability one under A.

Decorations are used to denote the specific way in which a distribution is chosen:

- the distributional symbol is decorated with superscript-o (for 'optimal'), e.g. $\mathsf{A}^o(x|K)$, to denote the FPD-optimal design (choice) of the distribution on $\boldsymbol{x}$, as explained in the text;
- the distributional symbol is decorated with subscript-I (for 'ideal'), e.g. $\mathsf{A}_\mathsf{I}(x|K)$, to denote a specific ideal choice of distribution on $\boldsymbol{x}$ in the FPD context;
- the distributional symbol is decorated with subscript-P (for 'prior'), e.g. $\mathsf{A}_\mathsf{P}(x|K)$, to denote a deterministic choice of prior distribution on $\boldsymbol{x}$ in the MXE context.

$\mathsf{M}(\cdot)$ is used exclusively to denote an unspecified distribution of its argument(s), while $\delta(x-x_0)$ denotes the distribution that is singular (degenerate) at $x_0 \in \boldsymbol{x}$ (typically, Dirac on continuous x and Kronecker on discrete x).

If $\mathsf{g}(x)$ is a finite-dimensional, real mapping from $\boldsymbol{x}$, then its expectation is denoted and defined as

$$\hat{\mathsf{g}} \equiv \mathsf{E}_{\mathsf{A}(x|K)}[\mathsf{g}] \equiv \int_{\boldsymbol{x}} \mathsf{g}(x)\, \mathsf{A}(x|K)\, dx.$$

Other symbols and decorations will be defined as they arise in the paper.

2. FPD and the MXE Principle

We consider an observer, $\mathcal{I}$, who interacts with their environment. $\mathcal{I}$ possesses knowledge and preferences in respect of their environment, and, specifically, with respect to an unknown quantity of interest, $x \in \boldsymbol{x}$. Within the adopted probabilistic framework, x is treated as a real, finite-dimensional random variable. Following [15], $\mathcal{I}$'s uncertainty model about x is expressed via the probability model, A, with support in $\boldsymbol{x}$. $\mathcal{I}$'s uncertainty extends to A itself, which they assume to belong to an appropriate set, $\mathbf{A}$, of possible A's. Therefore, $\mathcal{I}$ augments their probability model, quantifying their uncertainty about A via a chosen probability model, $\mathsf{S} \in \mathbf{S}$. In summary, $\mathcal{I}$'s probabilistic uncertainty model, $\mathsf{M} \in \mathbf{M}$, models their joint belief about

$$(x, \mathsf{A}) \in \boldsymbol{x} \times \mathbf{A}, \tag{1}$$

the extended set of the joint unknowns, x and A. A specific uncertainty model, $\mathsf{M} \in \mathbf{M}$, is prescribed by $\mathsf{S} \in \mathbf{S}$, where the set, $\mathbf{S}$, is consistent with

any knowledge K (sets $\boldsymbol{x}$, $\mathbf{A}$, $\mathbf{S}$, *etc.*) that informs $\mathcal{I}$ about x and A. This knowledge *does not* determine the model, $\mathsf{M} \in \mathbf{M}$, uniquely. The specific $\mathsf{S} \in \mathbf{S}$ is optimally chosen by $\mathcal{I}$ according to their joint *preferences* about x and A. Following the adopted FPD principle, these preferences are quantified by an ideal distribution, M_I, as follows:

Definition 1 (The Ideal Distribution and FPD). *The ideal distribution, M_I, also called the target or desired distribution, specifies $\mathcal{I}$'s preferred form for M. As such, it does not necessarily satisfy the constraints imposed by K (i.e. $\mathsf{M}_\mathsf{I} \notin \mathbf{M}$ in general). Within the FPD framework for optimal model design, M_I enters the KLD as the second argument. Therefore, it acts as the zero-KLD datum against which to rank all possible distributions consistent with knowledge, K. The FPD-optimal choice, $\mathsf{M}^o \in \mathbf{M}$, is the distribution closest to M_I in the minimum-KLD sense, while also being consistent with K. An axiomatic justification of FPD, and the role of the ideal distribution, is found in [25].*

Several settings of the problem formalisation are readily encountered:

(a) A is parametric, so that

$$x|\mathsf{A} \equiv x|\theta, \quad \theta \in \boldsymbol{\theta},$$

in the sense that x has the same distribution given either by A or θ. This asserts that $\mathcal{I}$'s model for x is determined by a finite number of unknown parameters, θ, implying the standard parametric hierarchical model [2],

$$\mathsf{M}(x, \theta|\mathsf{S}, K) \equiv \mathsf{A}(x|\theta, K)\mathsf{S}(\theta|K),$$

where θ fulfils the conventional roles of a hyperparameter, hidden field, missing data, *etc*, depending on the context, and S is a hyperparameter distribution.[5]

(b) x is discrete-valued, with $|\boldsymbol{x}| < \infty$. This gives a special case of (a), with $\theta = \mathsf{A}$ (being, here, a probability mass function on $\boldsymbol{x}$), and so $\boldsymbol{\theta} = \mathbf{A} \equiv \boldsymbol{\Delta}$, the $(|\boldsymbol{x}| - 1)$-dimensional open probability simplex. Hence, S is a distribution on the simplex.

[5] As already stated, FPD does not need to distinguish between prior and posterior knowledge. In the special case of sequential Bayesian inference, different parts of K condition the factors in the hierarchy. We will say more about Bayesian conditioning in Remark 7.

(c) A is an unknown (infinite-dimensional) distribution, and so we model[6] $\mathsf{A} \sim \mathcal{S}$ as a nonparametric process. This implies the following hierarchical model :

$$\begin{array}{rcll} x|\mathsf{A}, K & \sim & \mathsf{A}(x|K), & x \in \boldsymbol{x}, \quad \mathsf{A} \in \mathbf{A}, \\ \mathsf{A}|\mathcal{S}, K & \sim & \mathcal{S}(\mathsf{A}|K), & \mathcal{S} \in \boldsymbol{\mathcal{S}}. \end{array} \tag{2}$$

Hence, the distribution of x is a nonparametric process mixture, with the nonparametric process model, $\mathcal{S}$, taking the role of the mixing distribution [32]. $\mathcal{I}$'s hierarchical uncertainty model, M, for x and A (1) is specified by (2) in this case.

Whichever context is adopted, the hierarchical model is truncated at $\mathsf{S} \in \mathbf{S}$ or $\mathcal{S} \in \boldsymbol{\mathcal{S}}$, so that K conditions M on a specific choice of $\mathsf{S} \in \mathbf{S}$ or $\mathcal{S} \in \boldsymbol{\mathcal{S}}$. The flexibility of the nonparametric process in (c) ensures that any further relaxation of $\mathcal{S}$—via a hierarchy of models on $\mathcal{S} \in \boldsymbol{\mathcal{S}}$—does not extend the set, $\mathbf{M}$ of the hierarchical models, M (2). Hence, (2) is unrestricted [1]. This is not true of the truncated hierarchies in (a) and (b). We will return to the nonparametric setting in Section 4. For the present, to avoid technicalities, we assume that x is the finite-state image of y (a continuous random variable) under a specified, finite, measurable partition of $\boldsymbol{y}$, and that x is the modelled quantity. Then, we arrive at case (b) above, and, in particular, can express the hierarchical probability model via the standard chain rule of probability.

Definition 2 (Hierarchical Model of x). *$\mathcal{I}$'s uncertainty model for x is defined hierarchically, via the following joint distribution:*

$$\begin{aligned} \mathsf{M}(x, \mathsf{A}|\mathsf{S}, K) &\equiv \mathsf{M}(x|\mathsf{A}, \mathsf{S}, K)\mathsf{M}(\mathsf{A}|\mathsf{S}, K) \\ &\equiv \mathsf{A}(x|K)\mathsf{S}(\mathsf{A}|K), \quad \mathsf{A} \in \mathbf{A}, \ \mathsf{S} \in \mathbf{S}. \end{aligned} \tag{3}$$

As stated in Section 1, a conditional probability model, $\mathsf{M}(K|x)$, for knowledge, K, is assumed to be unavailable, and so the model, $\mathsf{M}(x, \mathsf{A}|\mathsf{S}, K)$, cannot be computed via Bayes' rule. *The main purpose of this paper is to extend the minimum cross-entropy principle [38] and fully probabilistic design [23] to the hierarchical model context, for the purpose of optimally processing K in this case.*

[6] Math-calligraphic $\mathcal{S}$ denotes the distribution of a *nonparametric* process, A.

$\mathcal{I}$'s objective is to optimize their model, $\mathsf{M}(x, \mathsf{A}|\mathsf{S}, K)$, *deterministically*, by tuning the only free factor in the model, namely S. This is a model design (decision) problem. It was shown axiomatically in [38]—though not in this hierarchical context—that its solution involves minimization of the KLD from M to M_P in the case where K involves specification of a prior, M_P, along with linear, finite-dimensional functional constraints on M. This *minimum cross-entropy principle* was shown in [4] usually to give the same result as Bayesian conditioning on this particular knowledge, K. However, $\mathcal{I}$'s knowledge, K, may not specify a prior model, M_P, at all, nor the imposition of functional constraints, as specified by K in Remark 1. Instead, S is specified only in terms of its role as a hypermodel in Definition 2. In this case, $\mathcal{I}$'s processed knowledge, K, includes only the following constraints:

Remark 2 (K in the Hierarchical FPD case).

(i) the unknown parameter, A*, on the left-hand-side of (2) is specialized to a probability distribution on* x*;*

(ii) x *is conditionally independent of* S*, given* A*.*

This K induces the joint model asserted in (2). Effectively, $\mathcal{I}$ seeks to process the very flexibly defined knowledge in Remark 2 in place of the knowledge in Remark 1, which is a specialization of the knowledge in Remark 2. In the decision-making context, $\mathcal{I}$ aims to select S so as to optimize their probability model, $\mathsf{M}(x, \mathsf{A}|\mathsf{S}, K)$. Processing of K in Remark 2 does not, of course, imply a unique optimal design. In order to achieve uniqueness in the design, $\mathcal{I}$ specifies their joint preferences in respect of x and A, quantified as their *ideal distribution*, and denoted by $\mathsf{M}_\mathsf{I}(x, \mathsf{A}|\mathsf{S}, K)$. In this context, FPD [23] dictates the following optimal choice for the free factor, S, in the hierarchical model (3)[7]:

$$\mathsf{S}^o(A|K) \equiv \arg\min_{\mathsf{S}\in\mathbf{S}} \mathcal{D}(\mathsf{M}||\mathsf{M}_\mathsf{I}). \tag{4}$$

[7] Any S equal to S^o almost everywhere (a.e.) [35] yields the same KLD minimum as in (4), and S^o is to be understood as equating to any member of the equivalence class of (4).

$\mathcal{D}(\mathsf{M}||\mathsf{M}_\mathsf{I})$ is the Kullback-Leibler divergence (KLD) [28] from M to M_I:

$$\begin{aligned}\mathcal{D}(\mathsf{M}||\mathsf{M}_\mathsf{I}) &\equiv \mathsf{E}_\mathsf{M}\left[\ln\left(\frac{\mathsf{M}}{\mathsf{M}_\mathsf{I}}\right)\right] \\ &\equiv \int_{\boldsymbol{x}\times\mathbf{A}} \ln\left(\frac{\mathsf{M}(x,\mathsf{A}|\mathsf{S},K)}{\mathsf{M}_\mathsf{I}(x,\mathsf{A}|\mathsf{S},K)}\right)\mathsf{M}(x,\mathsf{A}|\mathsf{S},K)\;\mathrm{d}x\;\mathrm{d}\mathsf{A}. \qquad (5)\end{aligned}$$

In this paper, it is assumed that the support of the ideal, M_I, is $\boldsymbol{x}\times\mathbf{A}$, so that the KLD is finite, $\forall\mathsf{M}\in\mathbf{M}$. In general, $\mathsf{M}_\mathsf{I}\notin\mathbf{M}$, i.e. the ideal may not be one of the possible joint models, M. Indeed, this is typically the case. However, the ideal should respect the knowledge, K, as shown in the notation.

Remark 3 (Fully Probabilistic Design (FPD)). *The unique Bayesian dissimilarity measure for ranking possible alternative distributions (e.g. an ideal, prior or approximate alternative) against an unreduced distribution was shown axiomatically in [3] to be the KLD from the unreduced distribution to the possible alternatives, as in (5)*[8]*. The optimal choice of an approximate alternative therefore requires minimization of this KLD with respect to its* second *argument.*

In contrast, the inferential design problem specified above requires the ranking of possible unreduced distributions against a specified—and therefore fixed—ideal. In this case, [25] axiomatically justify the same KLD (5). Therefore, the optimal choice of the unreduced distribution involves minimizing this KLD with respect to its first *argument, as in (4). This principle for optimally choosing the unreduced distribution in the context of a specified ideal is called* fully probabilistic design.

The following theorem provides the FPD-optimal model in the general case where no special constraints are imposed by K on $\mathsf{S}\in\mathbf{S}$ or $\mathsf{A}\in\mathbf{A}$.

Theorem 1 (Fully Probabilistic Design of the Hierarchical Model). *Let $\mathcal{I}$'s hierarchical probability model in the extended measurable space, $(\boldsymbol{x}\times\mathbf{A},\,\sigma(\boldsymbol{x}\times\mathbf{A}))$, be M, as given in Definition 2. Here, $\sigma(\cdot)$ denotes the σ-algebra of Borel sets in the extended set (1). We assume that $|\boldsymbol{x}|<\infty$. No*

[8] By unreduced distributions, we mean the distributions that use the full knowledge, K, of the observer, $\mathcal{I}$, for describing the modelled x and A.

special constraints are placed on the sets, $\mathbf{A} \neq \emptyset$, $\mathbf{S} \neq \emptyset$, *of* A *and* S, *respectively. Let* $\mathcal{I}$*'s joint ideal model for* x *and* A *be given by the same factored form:*

$$\mathsf{M}_{\mathsf{I}}(x, \mathsf{A}|\mathsf{S}, K) \equiv \mathsf{A}_{\mathsf{I}}(x|\mathsf{S}_{\mathsf{I}}, K)\mathsf{S}_{\mathsf{I}}(\mathsf{A}|K), \tag{6}$$

where A_{I} *and* S_{I} *are chosen by* $\mathcal{I}$ *to quantify their joint preferences about* x *and* A, *respectively. The following regularity condition is assumed:*

$$\mathcal{D}(\mathsf{A}||\mathsf{A}_{\mathsf{I}}) < \infty, \quad \forall \mathsf{A} \in \mathsf{supp}(\mathsf{S}_{\mathsf{I}}) \equiv \mathbf{A}. \tag{7}$$

Then, the FPD-optimal design of S, *i.e. the minimizer of (4), is*

$$\mathsf{S}^o(\mathsf{A}|K) \propto \mathsf{S}_{\mathsf{I}}(\mathsf{A}|K) \exp\left(-\mathcal{D}(\mathsf{A}||\mathsf{A}_{\mathsf{I}})\right), \tag{8}$$

where proportionality, $\propto$, *is discussed in Remark 4. The FPD-optimal hierarchical model of Definition 2 is therefore*

$$\mathsf{M}^o(x, \mathsf{A}|\mathsf{S}^o, K) = \mathsf{A}(x|K)\mathsf{S}^o(\mathsf{A}|K). \tag{9}$$

Proof: From (3) and (6):

$$\begin{aligned}
\mathcal{D}(\mathsf{M}||\mathsf{M}_{\mathsf{I}}) &= \int_{\boldsymbol{x}\times\mathbf{A}} \ln\left(\frac{\mathsf{A}(x|K)\mathsf{S}(A|K)}{\mathsf{A}_{\mathsf{I}}(x|K)\mathsf{S}_{\mathsf{I}}(A|K)}\right) \mathsf{A}(x|K)\mathsf{S}(\mathsf{A}|K)\ \mathrm{d}x\ \mathrm{d}\mathsf{A} \\
&= \int_{\boldsymbol{x}\times\mathbf{A}} \ln\left(\frac{\mathsf{A}(x|K)}{\mathsf{A}_{\mathsf{I}}(x|K)}\right) \mathsf{A}(x|K)\mathsf{S}(\mathsf{A}|K)\ \mathrm{d}x\ \mathrm{d}\mathsf{A} \\
&+ \int_{\boldsymbol{x}\times\mathbf{A}} \ln\left(\frac{\mathsf{S}(\mathsf{A}|K)}{\mathsf{S}_{\mathsf{I}}(\mathsf{A}|K)}\right) \mathsf{A}(x|K)\mathsf{S}(\mathsf{A}|K)\ \mathrm{d}x\ \mathrm{d}\mathsf{A} \\
&= \int_{\mathbf{A}} \mathcal{D}(\mathsf{A}||\mathsf{A}_{\mathsf{I}})\mathsf{S}(\mathsf{A}|K)\ \mathrm{d}\mathsf{A} + \int_{\mathbf{A}} \ln\left(\frac{\mathsf{S}(\mathsf{A}|K)}{\mathsf{S}_{\mathsf{I}}(\mathsf{A}|K)}\right) \mathsf{S}(\mathsf{A}|K)\ \mathrm{d}\mathsf{A},
\end{aligned} \tag{10}$$

using Fubini's theorem. The first term follows from the definition of the KLD (5), and the second from the fact that $\int_{\boldsymbol{x}} \mathsf{A}(x|K)\ \mathrm{d}x = 1$. Then:

$$\mathcal{D}(\mathsf{M}||\mathsf{M}_{\mathsf{I}}) = \int_{\mathbf{A}} \ln\left(\frac{\mathsf{S}(A|K)}{\frac{1}{c_{\mathsf{S}^o}}\mathsf{S}_{\mathsf{I}}(\mathsf{A}|K)\exp\left(-\mathcal{D}(\mathsf{A}||\mathsf{A}_{\mathsf{I}})\right)}\right) \mathsf{S}(\mathsf{A}|K)\ \mathrm{d}\mathsf{A} - \ln c_{\mathsf{S}^o}, \tag{11}$$

where

$$c_{\mathsf{S}^o} \equiv \int_{\mathbf{A}} \mathsf{S}_{\mathsf{I}}(\mathsf{A}|K)\exp\left(-\mathcal{D}(\mathsf{A}||\mathsf{A}_{\mathsf{I}})\right)\ \mathrm{d}\mathsf{A} \tag{12}$$

normalizes the distribution (8). Here, we have imposed the regularity condition (7), which guarantees that $\mathsf{S}^o(\mathsf{A}|K)$ is positive and finite, and therefore proper (i.e. $c_{\mathsf{S}^o} < \infty$ (12)). Also, we have used the fact that $\int_{\mathbf{A}} \mathsf{S}(\mathsf{A}|K)\ \mathrm{d}\mathsf{A} = 1$ in the second term of (11). The latter can be expressed as

$$\mathcal{D}(\mathsf{M}||\mathsf{M}_\mathsf{I}) = \mathcal{D}(\mathsf{S}||\mathsf{S}^o) - \ln(c_{\mathsf{S}^o}).$$

Since $\mathsf{S}^o(\mathsf{A}|K) \in \mathbf{S}$, it follows from the standard properties of $\mathcal{D}(\cdot||\cdot)$ that $\mathsf{S}^o(\mathsf{A}|K)$ (8) is the unique FPD minimizer, defined in (4). The FPD-optimal hierarchical model, M^o (9), follows immediately from Definition 2, by letting $\mathsf{S} = \mathsf{S}^o$. □

Remark 4 (Normalization of the FPD-optimal design). *In (8), the symbol, $\propto$, denotes equality up to the normalizing constant, $c_{\mathsf{S}^o} \equiv c_{\mathsf{S}^o}$ (12), of the expression on the right. Here, as throughout the paper, we adopt the following agreements: (a) finiteness of the integral (12) is guaranteed, since S_I is a distribution and $\mathsf{supp}(\mathsf{A}_\mathsf{I}) \equiv \boldsymbol{x}$ as assumed following (5) above; and (b) the appropriate normalizing constant is denoted, as here, by c subscripted by the distribution that it normalizes.*

The significance of Theorem 1 is that it furnishes an optimal mechanism for processing knowledge, K, and preferences into a probabilistic uncertainty model. In this sense, it replaces Bayes' rule with an optimal mechanism for construction of the posterior inference of the unknowns (x and A) in the hierarchical model, in those cases where an explicit observation model, $\mathsf{M}(K|x, \mathsf{A})$, is not available. We will comment on the consistency of this design with Bayesian conditioning in Remark 7.

Remark 5 (Conflicting ideals). *The factors, A_I and S_I, in (6) can quantify conflicting preferences in respect of x. This reflects the often inconsistent nature of $\mathcal{I}$'s preferences in practice. For instance, it may be the case that $\mathsf{E}_{\mathsf{S}_\mathsf{I}}[\mathsf{A}] \neq \mathsf{A}_\mathsf{I}$, reflecting the case where $\mathcal{I}$ specifies preferences about x and A that do* not *respect the hierarchical relationship between them (Definition 2). $\mathsf{S}^o(\mathsf{A}|K)$ (8) is the FPD-optimal compromise between these inconsistent ideals, A_I and S_I, about x and A, respectively, in the case where S is constrained to being the distribution of $\mathsf{A}(x|K)$. Essentially, $\mathsf{S}^o(\mathsf{A}|K)$ attempts to place maximal probability mass around A_I, but this design is modulated by S_I (see (6)). Ultimately, the extent to which S^o achieves a compromise between the specified ideals is dependent on how conflicted the specified ideals actually*

are. For a review of preference elicitation and its consistent quantification, see [22].

The induced FPD-optimal model for x follows by marginalizing over A in (9):

$$\begin{aligned}\mathsf{M}^o(x|\mathsf{S}^o,K) &\equiv \mathsf{A}^o(x|K)=\int_{\mathbf{A}}\mathsf{M}^o(x,\mathsf{A}|\mathsf{S}^o,K)\,\mathrm{d}\mathsf{A}=\int_{\mathbf{A}}\mathsf{A}(x|K)\mathsf{S}^o(\mathsf{A}|K)\,\mathrm{d}\mathsf{A}\\ &= \mathsf{E}_{\mathsf{S}^o}[\mathsf{A}(x|K)]. \qquad (13)\end{aligned}$$

$\mathsf{A}^o(x|K)$ optimally processes knowledge, K, and preferences for the purposes of constructing $\mathcal{I}$'s posterior inference about x. Again, the design (13) replaces Bayes' rule in those cases where an explicit model, $\mathsf{M}(K|x)$, is unavailable or unspecified.

$\mathsf{S}^o(\mathsf{A}|K)$ (8) is a fully probabilistic quantification of $\mathcal{I}$'s beliefs about unknown $\mathsf{A}(x|K)$, equipping the point estimate, $\mathsf{A}^o(x|K)$ (13), with quantified uncertainty.

In Theorem 1, both x and A have specified ideals (6), which act jointly as constraints in the FPD-optimal design of the uncertainty model, M. However, it may be that only one of the ideal factors is specified or available. The design of the other factor is then unconstrained (referred to as being 'left to its fate' in [21]). The possible specifications of the ideal, and the resulting FPD-optimal designs of the hierarchy, are addressed in the following corollary of Theorem 1.

Corollary 1 (of Theorem 1). *The FPD problem specified in Theorem 1 is solved under two incomplete specifications of the ideal model, respectively.*
(a) *If*

$$\mathsf{M}_{\mathsf{I}}(x,\mathsf{A}|\mathsf{S},K)\equiv\mathsf{A}_{\mathsf{I}}(x|\mathsf{S},K)\mathsf{S}(\mathsf{A}|K), \qquad (14)$$

then the FPD minimizer, defined by (4), is

$$\mathsf{S}^o(A|K)=\delta(\mathsf{A}-\mathsf{A}^o), \qquad (15)$$

where $\delta(\cdot)$ is the distribution that is singular at $\mathsf{A}=\mathsf{A}^o\in\mathbf{A}$, and where

$$\mathsf{A}^o(x|K)\equiv\arg\min_{\mathsf{A}\in\mathbf{A}}\mathcal{D}(\mathsf{A}||\mathsf{A}_{\mathsf{I}}) \qquad (16)$$

is assumed to exist. In this case, the FPD-optimal hierarchical model is

$$\mathsf{M}^o(x,\mathsf{A}|\mathsf{S}^o,K)=\mathsf{A}^o(x|\mathsf{S},K)\delta(\mathsf{A}-\mathsf{A}^o). \qquad (17)$$

(b) *If*

$$\mathsf{M}_{\mathsf{I}}(x, \mathsf{A}|K) \equiv \mathsf{A}(x|K)\mathsf{S}_{\mathsf{I}}(\mathsf{A}|K), \tag{18}$$

then the FPD minimizer, defined by (4), is

$$\mathsf{S}^o(A|K) = \arg\min_{\mathsf{S}\in\mathbf{S}} \mathcal{D}(\mathsf{S}||\mathsf{S}_{\mathsf{I}}), \tag{19}$$

which is assumed to exist. In this case, the FPD-optimal hierarchical model is

$$\mathsf{M}^o(x, \mathsf{A}|\mathsf{S}^o, K) = \mathsf{A}(x|K)\mathsf{S}^o(\mathsf{A}|K). \tag{20}$$

Proof:
(a) Letting $\mathsf{S}_{\mathsf{I}}(\mathsf{A}|K) = \mathsf{S}(\mathsf{A}|K)$ (consistent with (14)) in the second term on the right-hand-side of (10), then,

$$\mathcal{D}(\mathsf{M}||\mathsf{M}_{\mathsf{I}}) = \int_{\mathbf{A}} \mathcal{D}(\mathsf{A}||\mathsf{A}_{\mathsf{I}})\mathsf{S}(\mathsf{A}|K)\ \mathrm{d}\mathsf{A}. \tag{21}$$

Adopting $\mathsf{M} = \mathsf{M}^o$, as asserted in (17), which is equivalent to substituting $\mathsf{S}(\mathsf{A}|K) = \delta(\mathsf{A} - \mathsf{A}^o)$ (15) in the right-hand side of (21), then $\mathcal{D}(\mathsf{M}^o||\mathsf{M}_{\mathsf{I}}) = \mathcal{D}(\mathsf{A}^o||\mathsf{A}_{\mathsf{I}})$. Assuming that the minimizer, $\mathsf{A}^o \in \mathbf{A}$ (16), of $\mathcal{D}(\mathsf{A}||\mathsf{A}_{\mathsf{I}})$ exists, then (15) is the minimizer of (4), by definition, under the ideal specification (14), and the associated FPD-optimal hierarchical model is (17), from the definition of $\delta(\cdot)$.
(b) Letting $\mathsf{A}_{\mathsf{I}}(x|\mathsf{S}_{\mathsf{I}}, K) = \mathsf{A}(x|K)$ (consistent with (18)) in the first term on the right-hand-side of (10), then

$$\mathcal{D}(\mathsf{M}||\mathsf{M}_{\mathsf{I}}) = \int_{\mathbf{A}} \ln\left(\frac{\mathsf{S}(\mathsf{A}|K)}{\mathsf{S}_{\mathsf{I}}(\mathsf{A}|K)}\right)\mathsf{S}(\mathsf{A}|K)\ \mathrm{d}\mathsf{A} \equiv \mathcal{D}(\mathsf{S}||\mathsf{S}_{\mathsf{I}}) \tag{22}$$

Adopting $\mathsf{M} = \mathsf{M}^o$, as asserted in (20), which is equivalent to substituting $\mathsf{S}(\mathsf{A}|K) = \mathsf{S}^o(\mathsf{A}|K)$ (19) in the right-hand side of (22), then $\mathcal{D}(\mathsf{M}^o||\mathsf{M}_{\mathsf{I}}) = \mathcal{D}(\mathsf{S}^o||\mathsf{S}_{\mathsf{I}})$. Again, by definition, it follows that (19) is the minimizer of (4) under the ideal specification (18), and the associated FPD-optimal hierarchical model is given by (20) . □

Hence, a *deterministic design* for $\mathsf{S}(\mathsf{A}|K)$—one which chooses $\mathsf{A}^o(x|\mathsf{S}, K)$ with probability 1—arises in case **(a)**. A^o (16) can be interpreted as the optimal (in the minimum-KLD sense prescribed by FPD) projection of A_{I} into the allowed set, $\mathbf{A} \neq \emptyset$. In contrast, case **(b)** chooses the *randomized* design, S^o (19), as the model for A in this case, being the FPD-optimal

projection of S_I into the allowed set, $\mathbf{S} \neq \emptyset$. The FPD-optimal model for x in case **(b)** is given by (13), using (19).

In summary, Corollary 1 demonstrates that the FPD-optimal design is achieved by minimization of the marginal KLD—(16) or (19)—in the case of a partially specified ideal.

2.1. Fully Probabilistic Design with Uniform Ideals

It is important to distinguish between the incomplete ideal specifications explored in Corollary 1, in which either A_I or S_I is unspecified (i.e. the design of the respective distribution is *unconstrained*), and the explicit ideal specifications of ignorance or non-commitment in respect of one or both of the unknowns, x and/or A in (3). In the latter case, the FPD-optimal design seeks a compromise between the possibly conflicting informed and ignorance ideals specified for the factors in (6), as noted in Remark 5.

There is an extensive Bayesian literature on the elicitation of priors to express such states of ignorance or non-commitment [14, 41]. A particular choice can be adopted as a quantifier of non-committal preferences, and then processed by the FPD principle (Theorem 1 and Corollary 1). Here, we focus on the *uniform ideal*, $\mathsf{U}(\cdot)$, specializing (6) to $\mathsf{M}_\mathsf{I}(x, \mathsf{A}|\mathsf{S}, K) \equiv \mathsf{U}(x, \mathsf{A})$. Note that

$$\mathcal{D}(\mathsf{A}||\mathsf{U}) = -\mathcal{H}_\mathsf{A} + \ln|\boldsymbol{x}|, \tag{23}$$

where, by arrangement, $|\boldsymbol{x}| < \infty$, and where

$$\mathcal{H}_\mathsf{A} \equiv -\int_{\boldsymbol{x}} \ln(\mathsf{A}(x|K))\mathsf{A}(x|K)\ \mathrm{d}x. \tag{24}$$

(23) is the *differential entropy* of $x \sim \mathsf{A}(x|K)$ [6]. Here, we adopt the notation, $\mathcal{H}_\mathsf{A}$, to emphasize that this is a functional of unknown $\mathsf{A} \in \mathbf{A}$. Hence, the FPD-optimal design under a joint uniform ideal, U(x, A), is obtained by substituting (23) into (8):

$$\mathsf{S}^o(\mathsf{A}|K) \propto \exp\left(\mathcal{H}_\mathsf{A}\right). \tag{25}$$

It follows that the FPD-optimal maximum *a posteriori* estimate, $\hat{\mathsf{A}}_{\mathsf{MAP}}$, of A in this case is

$$\hat{\mathsf{A}}_{\mathsf{MAP}}(x|K) = \arg\max_{\mathsf{A}\in\mathbf{A}} \mathcal{H}_\mathsf{A}. \tag{26}$$

From (25), the FPD-optimal design of the log-distribution of A—when processing uniform ideals—is proportional to the entropy, $\mathcal{H}_\mathsf{A}$, of A. Entropy

(24) and its maximizer (26) are widely studied in statistical physics [13], and applied to problems such as parametric prior design [15] and image restoration [39]. The FPD principle provides a hierarchical Bayesian justification for these approaches.

For completeness, we specialize Corollary 1 to uniform ideals. In case **(a)**, we substitute (23) into (21), yielding

$$\mathcal{D}(\mathsf{M}||\mathsf{M}_\mathsf{I}) = -\int_{\mathbf{A}} \mathcal{H}_\mathsf{A} \mathsf{S}(\mathsf{A}|K)\ \mathrm{d}\mathsf{A} + \ln|\boldsymbol{x}|.$$

Since $\mathcal{H}_\mathsf{A} \geq 0$, the FPD-optimal design is the deterministic one, selecting a maximum entropy design for A (26):

$$\mathsf{S}^o(\mathsf{A}) = \delta(\mathsf{A} - \hat{\mathsf{A}}_{\mathsf{MAP}}), \tag{27}$$

In case **(b)**, $\mathcal{D}(\mathsf{M}||\mathsf{M}_\mathsf{I}) = -\mathcal{H}_\mathsf{S} + \ln|\mathbf{A}|$, from (22), implying a maximum entropy design, in this case for S.

3. Fully Probabilistic Design under Functional Constraints

The KLD, $\mathcal{D}(\cdot||\cdot)$, has an information-theoretic interpretation as cross-entropy. The axiomatic setting of the minimum cross-entropy principle [38] is a *deterministic* one: an optimal point estimate, $\mathsf{A}_{\mathsf{MXE}}(x|K)$, of $\mathsf{A}(x|K)$—the unknown uncertainty model for x—is constructed, where $\mathcal{I}$'s knowledge, K, is expressed as inequality constraints on specified linear functionals of A, thereby constraining the set[9], $\mathbf{A}_K \subset \mathbf{A}$, of A. $\mathcal{I}$ also quantifies their *prior beliefs* about x via a deterministic prior distribution, $\mathsf{A}_\mathsf{P}(x|K)$. Specifically, the MXE principle processes a knowledge structure, K, of the kind in Remark 1. We now specify this further.

Remark 6 (K in the MXE case: detailed specification).

(i) $\mathcal{I}$'s uncertainty model for x is quantified by the unknown distribution, $\mathsf{A}(x|K)$. This is not equipped with a probabilistic model, but treated deterministically.

[9] In this Section, the symbol for a set of functionally constrainted distributions will be subscripted by K, e.g. $\mathbf{A}_K$.

(ii) Linear functional constraints are imposed on A*:*

$$\mathbf{A}_K \equiv \left\{ \mathsf{A}(x|K) : \int_{\boldsymbol{x}} \mathsf{g}(x)\mathsf{A}(x|K)\ \mathrm{d}x \leq 0 \right\} \neq \emptyset. \tag{28}$$

$\mathsf{g}(x)$ *is a specified finite-dimensional real function of* $x \in \boldsymbol{x}$*, and the inequality is applied element-wise.*

(iii) $\mathcal{I}$ *makes a prior choice or estimation of their quantified beliefs about* x*, being* $\mathsf{A}_\mathsf{P}(x|K)$*.*

For completeness, we now recall the MXE-optimal design of A implied by K in Remark 6 [38].

Theorem 2 (MXE Design of $\mathsf{A}(x|K)$). *The MXE-optimal design of* A *is defined as*

$$\mathsf{A}_{\mathsf{MXE}}(x|K) \equiv \arg\min_{\mathsf{A}\in\mathbf{A}_K} \mathcal{D}(\mathsf{A}||\mathsf{A}_\mathsf{P}). \tag{29}$$

If the knowledge, K*, specified in Remark 6, is processed, then*

$$\mathsf{A}_{\mathsf{MXE}}(x|K) \propto \mathsf{A}_\mathsf{P}(x|K)\exp[-\mu^T\mathsf{g}(x)], \tag{30}$$

where μ *is a vector of non-negative constants, of dimension compatible with the scalar product,* $\mu^T\mathsf{g}(x)$*, and where* μ^T *denotes the transpose of* μ*.*

Proof: $\mathcal{D}(\mathsf{A}||\mathsf{A}_\mathsf{P})$ in (29) is a strictly convex functional of A on the non-empty convex set, $\mathbf{A}_K$ (28), and it is bounded from below. Therefore, the infimum exists. Moreover, assuming that $\mathcal{D}(\mathsf{A}||\mathsf{A}_\mathsf{P})$ is finite for some $\mathsf{A} \in \mathbf{A}_K$, then the infimum is the unique minimizer. This minimizer can be found as that of the corresponding Kuhn-Tucker functional [27], [29]. Operations of the kind specified in the proof of Theorem 1 yield the minimizer in (30). Inserting (30) into (29), μ satisfies

$$\int_{\boldsymbol{x}} \mathsf{g}(x)\mathsf{A}_{\mathsf{MXE}}(x|K)\ \mathrm{d}x \leq 0. \tag{31}$$

The elements of μ are either zero—for each functional inequality in (28) satisfied by $\mathsf{A}_\mathsf{P}(x|K)$—or else unique and positive, satisfying the respective strict equality constraint in (31). □

In the special case where the prior guess, $\mathsf{A}_\mathsf{P}(x|K)$, satisfies all the constraints in (28)—in which case $\mathsf{A}_\mathsf{P} \in \mathbf{A}_K$—then $\mu = 0$ and so, from (30),

$$\mathsf{A}_\mathsf{MXE}(x|K) = \mathsf{A}_\mathsf{P}(x|K).$$

Remark 7 (Interpreting the prior, $\mathsf{A}_\mathsf{P}(x|K)$). *The MXE principle for design of the unknown distribution, $\mathsf{A}(x|K)$, neglects the fully Bayesian hierarchical modelling of Definition 2 in favour of deterministic optimization of $\mathsf{A} \in \mathbf{A}_K$ (28). From the hierarchical FPD perspective, this prior, A_P, in MXE can be interpreted as an initial distributional estimate of A, based on knowledge, K_P, available to $\mathcal{I}$* prior to *acquiring further knowledge, K_δ. The latter is formulated as the constraint on the allowed set of A (28), which is element (ii) of K (Remark 6). This implies a sequential processing of $K \equiv K_\delta \cup K_\mathsf{P}$. Indeed, denoting by $\mathsf{F}(K_\delta|x)$ a direct model of K_δ, then Bayes' rule provides the uniquely consistent deductive mechanism for combining this with A_P, to yield $\mathsf{A}(x|K)$. It has been proved in [4] that $\mathsf{A}(x|K) = \mathsf{A}_\mathsf{MXE}(x|K)$ (30) in this case. This means that the MXE principle, if processing knowledge, K, in Remark 6, is consistent with Bayesian conditioning (rare counterexamples are discussed in [4]). However, it remains the case in the standard setting of MXE (Theorem 2) that a point estimate or choice of A must be specified* a priori*, which is then relaxed in the subsequent variational optimization (29) over the set, $\mathbf{A}_K$.*

We emphasize that the deterministic design of A prescribed by Theorem 2 is replaced in this paper by an optimal design of a stochastic model for A, via hierarchical FPD. For this reason, a prior guess of A, being A_P, is typically not processed in our hierarchical FPD context.

For completeness, we note that *FPD-optimal* design of unknown but *unmodelled* A, in the presence of $\mathcal{I}$'s *ideal*, A_I, about x, and processing the functional constraints in (28), is defined by (29) with A_P replaced by A_I. It follows that

$$\mathsf{A}^o(x|K) \propto \mathsf{A}_\mathsf{I}(x|K) \exp[-\mu^T \mathsf{g}(x)]. \tag{32}$$

Interpreted formally in the context of the hierarchical model (3), this MXE-type FPD design specifies $\mathsf{S}^o(\mathsf{A}|K) = \delta(\mathsf{A} - \hat{\mathsf{A}})$, the degenerate choice at some unknown $\hat{\mathsf{A}} \in \mathbf{A}_K$. FPD yields the deterministic FPD design, $\hat{\mathsf{A}} = \mathsf{A}^o$ (32), in this case. Since, typically, $\mathsf{A}_\mathsf{I} \notin \mathbf{A}_K$ (28), as noted in Section 2, then at least some of the constraints in (28) are active, giving $\mu \neq 0$ and $\mathsf{A}^o \neq \mathsf{A}_\mathsf{I}$.

3.1. Hierarchical FPD under Functional Constraints

The hierarchical Bayesian modelling of A via $\mathsf{S}(\mathsf{A}|K)$ in Definition 2 has not been considered in the axiomatic setting of FPD (nor of MXE, as noted above) before. Nevertheless, the principle can be applied without further theoretical development—as shown in the next theorem—to the optimal design of the Bayesian hierarchical model in the parametric case ($|\boldsymbol{x}| < \infty$), i.e. where $\mathsf{A} \in \mathbf{A}_K \subset \mathbf{A} = \boldsymbol{\Delta}$, as in case (ii) of Section 2. The knowledge, K, processed by FPD in this case is as follows:

Remark 8 (K in hierarchical FPD case; functional constraints on S).

(i)

$$\mathsf{S}(\mathsf{A}|K) \in \mathbf{S}_K \equiv \left\{ \mathsf{S}(\mathsf{A}|K) : \int_{\mathbf{A}} \mathsf{g}(\mathsf{A}) \mathsf{S}(\mathsf{A}|K)\, \mathrm{d}\mathsf{A} \leq 0 \right\} \neq \emptyset, \tag{33}$$

where $\mathsf{g}(\cdot)$ specifies a finite-dimensional real image of $\mathbf{A}$.

(ii) $\mathsf{A}(x|K) \in \mathbf{A} \neq \emptyset$ *is unconstrained.*

(iii)

$$\mathsf{M}_{\mathsf{I}}(x, \mathsf{A}|\mathsf{S}, K) \equiv \mathsf{A}(x|K)\mathsf{S}_{\mathsf{I}}(\mathsf{A}|K).$$

Theorem 3 (FPD for S under Functional Constraints). *$\mathcal{I}$'s hierarchical probability model, M, is defined in (1) and Definition 2. The FPD principle for optimal design of $\mathsf{S}(\mathsf{A}|K)$ is*

$$\mathsf{S}^o(\mathsf{A}|K) = \arg\min_{\mathsf{S} \in \mathbf{S}_K} \mathcal{D}(\mathsf{S}||\mathsf{S}_{\mathsf{I}}). \tag{34}$$

Its solution, when processing knowledge, K, defined in Remark 8, is

$$\mathsf{S}^o(\mathsf{A}|K) \propto \mathsf{S}_{\mathsf{I}}(\mathsf{A}|K) \exp[-\mu^T \mathsf{g}(\mathsf{A})], \tag{35}$$

where μ is the unique vector of Kuhn-Tucker (non-negative) multipliers, of compatible dimension with g, chosen so that $\mathsf{S}^o(\mathsf{A}|K) \in \mathbf{S}_{\mathbf{K}}$ (33). The FPD-optimal hierarchical model, via Definition 2, and using (35), is therefore

$$\mathsf{M}^o(x, \mathsf{A}|\mathsf{S}^o, K) = \mathsf{A}(x|K)\mathsf{S}^o(\mathsf{A}|K). \tag{36}$$

Proof: $\mathcal{D}(\mathsf{S}||\mathsf{S_I})$ (34) is a convex functional of S in the non-empty convex set $\mathbf{S}_K$ (33). Therefore, existence of the minimizer is guaranteed, and is found as a minimizer of the corresponding Kuhn-Tucker functional [27]:

$$\mathsf{L}(\mathsf{S}, \mathsf{S_I}, \mu) \equiv \mathcal{D}(\mathsf{S}||\mathsf{S_I}) + \mu^T \int_{\mathbf{A}} \mathsf{g}(\mathsf{A})\mathsf{S}(\mathsf{A}|K)\ \mathrm{d}\mathsf{A}. \tag{37}$$

The multiplier, μ—with dimension chosen such that $\mu^T\mathsf{g}(\mathsf{A})$ is a valid scalar product—has entries which are either zero (in the case where the corresponding functional constraint is inactive) or strictly positive and unique (where the corresponding functional constraint is tightly satisfied by the optimum). From the definition of the KLD,

$$\begin{aligned}\mathsf{L}(\mathsf{S}, \mathsf{S_I}, \mu) &= \int_{\mathbf{A}} \ln\left(\frac{\mathsf{S}(\mathsf{A}|K)}{\mathsf{S_I}(\mathsf{A}|K)\exp(-\mu^T\mathsf{g}(\mathsf{A}))}\right)\mathsf{S}(\mathsf{A}|K)\ \mathrm{d}\mathsf{A} \\ &= \mathcal{D}(\mathsf{S}||\mathsf{S}^o) - \ln c_{\mathsf{S}^o},\end{aligned}$$

where S^o is given by (35). By the properties of the KLD, S^o is therefore the minimizer of (37), and, therefore, it minimizes the KLD under the constraints (33), satisfying the FPD principle. M^o (36) follows immediately from (3), by letting $\mathsf{S} = \mathsf{S}^o$ (35). □

This setting of the FPD principle, for the purposes of optimal design of the Bayesian hierarchical model in Definition 2, is important for the following reasons:

- There is freedom to process knowledge of A expressed via constraints on nonlinear functionals, $\mathsf{g}(\mathsf{A})$, of A (33), rather than via the conventional linear functional constraints on A (28).
- Once again, a 'randomized' inference of unknown $\mathsf{A}(x|K)$ is provided. It relaxes the point estimate, $\mathsf{A}^o(x|K)$ (32), arising from deterministic design of A under FPD, and its analogous design under the MXE principle (30).
- The FPD-optimal design in (35) is consistent with Bayesian conditioning in the case where the ideal is replaced by a prior choice, $\mathsf{S_P}$ (see Remark 7). This consistency follows from the fact that the processed knowledge , K, in this case (Remark 8) conforms with the axioms of the MXE principle, whose (usual) consistency was demonstrated in [4].

Importantly, the applicability of FPD to the processing of more generally specified knowledge structures, K, has been shown in [23], and its axiomatic consistency was proved in [25], as already noted in Section 2. This allows us to explore other consistent relaxations of the deterministic FPD design (32) in the hierarchical context of Definition 2 (as in the Žirafa theorem, which follows). For this purpose, we formulate functional constraints that are consistent with $\mathcal{I}$'s hierarchical model, and that satisfy the following desiderata:

Remark 9 (K in the Žirafa theorem).

(i) constraints are imposed on $\mathcal{I}$'s knowledge of x, as in (28), and not on their knowledge of A; constraining beliefs about x rather than beliefs about A is a natural knowledge structure for $\mathcal{I}$ to adopt, since x is the "unknown quantity of interest" (Section 2) for which the epistemic hierarchy is elicited;

(ii) $\mathcal{I}$'s uncertainty about A is explicitly modelled via $\mathsf{S}(\mathsf{A}|K)$, thereby relaxing the constraint—imposed by conventional FPD (32) and MXE (30) designs—that A be deterministic;

(iii) the constraints mentioned in (i) should induce a non-empty convex optimization domain, $\mathbf{S}_K$, in (4), and guarantee the existence and uniqueness of the solution of this optimization problem.

This knowledge specification implies the following constrained set for S:

$$\mathsf{S} \in \mathbf{S}_K \equiv \{\mathsf{S}(\mathsf{A}|K) : \mathsf{E}_\mathsf{M}[\mathsf{g}] \leq 0\} \neq \emptyset. \tag{38}$$

Here, M is the hierarchical probability model (3), $\mathsf{g}(x)$ is a known, finite-dimensional, real mapping from $\boldsymbol{x}$, and

$$\begin{aligned} \mathsf{E}_\mathsf{M}[\mathsf{g}] &\equiv \int_{\mathbf{A}_K} \int_{\boldsymbol{x}} \mathsf{g}(x)\mathsf{A}(x|K)\mathsf{S}(\mathsf{A}|K)\, \mathrm{d}x\, \mathrm{d}\mathsf{A} \\ &= \int_{\boldsymbol{x}} \mathsf{g}(x) \int_{\mathbf{A}_K} \mathsf{A}(x|K)\mathsf{S}(\mathsf{A}|K)\, \mathrm{d}A\, \mathrm{d}x = \int_{\boldsymbol{x}} \mathsf{g}(x)\hat{\mathsf{A}}(x|K)\, \mathrm{d}x, \end{aligned} \tag{39}$$

and where

$$\hat{\mathsf{A}}(x|K) \equiv \mathsf{E}_\mathsf{S}[\mathsf{A}(x|K)] \equiv \int_{\mathbf{A}_K} \mathsf{A}(x|K)\mathsf{S}(\mathsf{A}|K)\, \mathrm{d}\mathsf{A} = \mathsf{M}(x|\mathsf{S}, K), \tag{40}$$

where the latter identity follows by marginalizing the hierarchical (joint) model, $\mathsf{M}(x, \mathsf{A}|\mathsf{S}, K)$ (3) over A. (40) demonstrates that the marginal distribution of x—in the hierarchical context—equals the expected distribution, $\hat{\mathsf{A}}$, of x under $\mathsf{S}(\mathsf{A}|K)$, as also noted in (13). Hence, (38) imposes functional constraints on *marginal* $\mathsf{M}(x|K)$ (3), or, equivalently, on the expected distribution, $\hat{\mathsf{A}}(x|K) \in \mathbf{A}_K$ (28). The set $\mathbf{S}_K$ (38) requires this. We will strengthen this requirement and insist that supports of all $\mathsf{S} \in \mathbf{S}_K$ are themselves in $\mathbf{A}_K$. Note, therefore, that A is integrated over $\mathbf{A}_K$ in (39) and (40). The FPD-optimal hierarchy implied by the knowledge structure above is revealed by the following theorem.

Theorem 4 (Žirafa). *$\mathcal{I}$'s hierarchical probability model, M, is given by (3), in the extended measurable space, $(\boldsymbol{x} \times \mathbf{A}_K,\ \sigma(\boldsymbol{x} \times \mathbf{A}_K))$, where $\mathbf{A}_K$ is given by (28). Consider fully probabilistic design (4) in the case where it processes the knowledge, K, specified in Remark 9; i.e. $\mathsf{S} \in \mathbf{S}_K$, as defined in (38). $\mathcal{I}$ expresses preferences about* both *factors of the hierarchy, and these are quantified by the hierarchical ideal distribution in (6). Then, the FPD minimizer, defined by (4) is*

$$\mathsf{S}^o(\mathsf{A}|K) \propto \mathsf{S}_{\mathsf{I}}(\mathsf{A}|K) \exp\left[-\mathcal{D}(\mathsf{A}||\mathsf{A}^o)\right], \tag{41}$$

where $\mathsf{A}^o(x|K)$ is given by (32).

Proof: The objective functional (4) is convex in S, and must be optimized on the non-empty convex support $\mathbf{S}_K$ (38). This guarantees that a minimizer exists. This is found as the minimizer of the associated Kuhn-Tucker functional constructed from (5) and (38) [27]:

$$\mathsf{L}(\mathsf{M}, \mathsf{M}_{\mathsf{I}}, \mu) \equiv \mathcal{D}(\mathsf{M}||\mathsf{M}_{\mathsf{I}}) + \mu^T \int_{\mathbf{A}_K} \int_{\boldsymbol{x}} \mathsf{g}(x) \mathsf{A}(x|K) \mathsf{S}(\mathsf{A}|K)\ \mathrm{d}x\ \mathrm{d}\mathsf{A}. \tag{42}$$

The multipliers, μ, are zero when corresponding to inactive constraints in (38), and unique and positive [29] when an optimum satisfies the correspond-

ing equality in (38). Therefore, expanding $\mathcal{D}(\mathsf{M}||\mathsf{M}_\mathsf{I})$ as in (10):

$$\mathsf{L}(\mathsf{M},\mathsf{M}_\mathsf{I},\mu) = \int_{\mathbf{A}_K} \tag{43}$$

$$\left[\int_{\boldsymbol{x}} \left(\ln\left(\frac{\mathsf{A}(x|K)}{\mathsf{A}_\mathsf{I}(x|K)}\right) + \mu^T \mathsf{g}(x)\right) \mathsf{A}(x)\ \mathrm{d}x + \ln\left(\frac{\mathsf{S}(\mathsf{A}|K)}{\mathsf{S}_\mathsf{I}(\mathsf{A}|K)}\right)\right] \mathsf{S}(\mathsf{A}|K) d\mathsf{A} =$$

$$\int_{\mathbf{A}_K} \left[\int_{\boldsymbol{x}} \ln\left(\frac{\mathsf{A}(x|K)}{\mathsf{A}_\mathsf{I}(x|K)\exp(-\mu^T\mathsf{g}(x))}\right) \mathsf{A}(x)\ \mathrm{d}x + \ln\left(\frac{\mathsf{S}(\mathsf{A}|K)}{\mathsf{S}_\mathsf{I}(\mathsf{A}|K)}\right)\right] \mathsf{S}(\mathsf{A}|K)\ \mathrm{d}\mathsf{A}$$

$$\underbrace{=}_{(32)} \int_{\mathbf{A}_K} \left[\mathcal{D}\left[\mathsf{A}(x|K)||\mathsf{A}^o(x|K)\right] + \ln\left(\frac{\mathsf{S}(\mathsf{A}|K)}{\mathsf{S}_\mathsf{I}(\mathsf{A}|K)}\right)\right] \mathsf{S}(\mathsf{A}|K)\ \mathrm{d}\mathsf{A} - \ln c_{\mathsf{A}^o}$$

$$= \int_{\mathbf{A}_K} \ln\left(\frac{\mathsf{S}(\mathsf{A}|K)}{\mathsf{S}_\mathsf{I}(\mathsf{A}|K)\exp\left(-\mathcal{D}\left[\mathsf{A}(x|K)||\mathsf{A}^o(x|K)\right]\right)}\right) \mathsf{S}(\mathsf{A}|K)\ \mathrm{d}\mathsf{A} - \ln c_{\mathsf{A}^o}$$

$$\underbrace{=}_{(41)} \mathcal{D}\left[\mathsf{S}(\mathsf{A}|K)||\mathsf{S}^o(\mathsf{A}|K)\right] - \ln c_{\mathsf{A}^o} c_{\mathsf{S}^o},$$

where c_{A^o} and c_{S^o} are the normalizing constants (see Remark 4) of the probability distributions, (32) and (41), respectively. Hence, $\mathsf{S}^o(\mathsf{A}|K) \in \mathbf{S}_K$ is a minimizer of (42) and, therefore, the FPD minimizer, defined in (4), under the marginal functional constraints (38). □

The FPD-optimal hierarchical model follows from (3):

$$\mathsf{M}^o(x,\mathsf{A}|\mathsf{S}^o,K) = \mathsf{A}(x|K)\mathsf{S}^o(\mathsf{A}|K),$$

with $\mathsf{S}^o(\mathsf{A}|K)$ given by (41). The optimal model, $\mathsf{A}^o(x|K)$, is again equal to the expected distribution under S^o, via (13).

We note the following:

- (41) confers a fully Bayesian relaxation of the deterministic estimate of $\mathsf{A}(x|K)$ resulting from the conventional application of the FPD principle (32) and the analogous MXE design in the case of an assumed prior. It replaces the strategy of optimizing A (as A^o) with a fully Bayesian strategy for optimal design of the *distribution* of A.

- This Bayesian relaxation casts A^o in the role of a point estimate, and allows uncertainty measures to be associated with, and computed for, this estimate.

- Consider the case where $\mathcal{I}$ adopts the deterministic specialization of the hierarchy in Definition 2, such that $\mathsf{S}(\mathsf{A}|K) \equiv \delta(\mathsf{A} - \hat{\mathsf{A}})$; i.e. $\mathcal{I}$'s

hierarchical uncertainty model is singular at some unknown choice, $\hat{\mathsf{A}} \in \mathbf{A}_K$. Then, the Žirafa theorem implies that $\hat{\mathsf{A}} = \mathsf{A}^o$ as given in (32).

- Within the general setting of the Žirafa theorem, $\mathsf{A}^o(x|K)$ acts as a centering distribution for $\mathsf{S}^o(\mathsf{A}|K)$. Indeed, if $\mathcal{I}$ chooses S_I with mode at A^o, then, from (41), this is sufficient for the FPD-optimal modal distribution of x to be A^o. (41) provides the optimal compromise between A^o of the deterministic setting, and the specification of an ideal, S_I, for A, in the Bayesian hierarchical setting of Definition 2.

- If all the functional constraints (38) are inactive, then $\mu = 0$ in (42). Inserting $\mu = 0$ into (41) yields (8) as the unconstrained FPD-optimum, establishing Theorem 1 as a corollary of the Žirafa theorem iff $\mathsf{A} \in \mathbf{A}_K$ (28).

For completeness, we now specialize the Žirafa theorem to a number of cases of interest.

Corollary 2 (of the Žirafa theorem). *The FPD problem specified in the Žirafa theorem is solved under each of following incomplete specifications of the ideal:*
(a) *If*

$$\mathsf{M}_\mathsf{I}(x, \mathsf{A}|\mathsf{S}, K) \equiv \mathsf{A}_\mathsf{I}(x|\mathsf{S}, K)\mathsf{S}(\mathsf{A}|K),$$

then the FPD minimizer, defined in (4), is

$$\mathsf{S}^o(A|K) = \delta(\mathsf{A} - \mathsf{A}^o), \tag{44}$$

where A^o is given by (32). The FPD-optimal hierarchical model is then

$$\mathsf{M}^o(x, \mathsf{A}|\mathsf{S}^o, K) = \mathsf{A}^o(x|K)\delta(\mathsf{A} - \mathsf{A}^o).$$

(b) *If*

$$\mathsf{M}_\mathsf{I}(x, \mathsf{A}|K) \equiv \mathsf{A}(x|K)\mathsf{S}_\mathsf{I}(\mathsf{A}|K),$$

then the FPD minimizer, defined in (4), is

$$\mathsf{S}^o(\mathsf{A}|K) \propto \mathsf{S}_\mathsf{I}(\mathsf{A}|K) \exp\left[-\mu^T \mathsf{E}_\mathsf{A}[\mathsf{g}]\right], \tag{45}$$

where $\mathsf{E}_\mathsf{A}[\mathsf{g}] \equiv \int_{\boldsymbol{x}} \mathsf{g}(x)\mathsf{A}(x|K)\, \mathrm{d}x$.

Proof:
(a) Taking $\mathsf{S_I} = \mathsf{S}$, then the second term in (43) is zero, and so

$$\mathsf{L}(\mathsf{M}, \mathsf{M_I}, \mu) = \int_{\mathbf{A}_K} \mathcal{D}(\mathsf{A}||\mathsf{A}^o)\mathsf{S}(\mathsf{A}|K)\ \mathrm{d}\mathsf{A} - \ln c_{\mathsf{A}^o},$$

using (32). The first term is zero for the choice in (44), confirming that it is a minimizer of L.
(b) Taking $\mathsf{A_I} = \mathsf{A}$, then, from (43):

$$\mathsf{L}(\mathsf{M}, \mathsf{M_I}, \mu) = \int_{\mathbf{A}_K} \left[\mathsf{E}_\mathsf{A}[\mathsf{g}] + \ln\left(\frac{\mathsf{S}(\mathsf{A}|K)}{\mathsf{S_I}(\mathsf{A}|K)}\right)\right] \mathsf{S}(\mathsf{A}|K)\ \mathrm{d}\mathsf{A}.$$

Manipulations of the kind in the proof of the Žirafa theorem, above, then lead to the result. □

In case **(a)**, the FPD-optimal hierarchy is degenerate at the deterministic design, A^o (32), of unknown A. In case **(b)**, the nonlinear function, $\mathsf{g}(x)$ (28), is replaced by its expectation under unknown A.

3.2. Functionally Constrained Designs with Uniform Ideals

In common with Section 2.1, we briefly consider the functionally constrained designs of the hierarchy for the case of uniform ideals. Considering, firstly, the fully specified ideals of the Žirafa theorem, then, adopting $\mathsf{A_I}(x|K) \equiv \mathsf{U}(x)$ in (32),

$$\mathsf{A}^o(x|K) \propto \exp\left(-\mu^T \mathsf{g}(x)\right). \tag{46}$$

This is the maximum entropy design of deterministic A under functional constraints (28) [13], a widely adopted specialization of the minimum cross-entropy principle to the case of a uniform prior (understood here as $\mathcal{I}$'s ideal distribution). Inserting this into (41), and adopting $\mathsf{S_I}(\mathsf{A}|K) \equiv \mathsf{U}(\mathsf{A})$,

$$\mathsf{S}^o(\mathsf{A}|K) \propto \exp\left(-\mathcal{D}\left[\mathsf{A}||\left(c_{\mathsf{A}^o}^{-1}\exp(-\mu^T\mathsf{g})\right)\right]\right),$$

which again provides the fully probabilistic relaxation of the maximum entropy solution (46) for the case of uncertain and hierarchically modelled A.

In case **(a)** of the Žirafa corollary above, with $\mathsf{A_I}(x|K) \equiv \mathsf{U}(x)$ and $\mathsf{S_I}$ unspecified, then an FPD-optimal design of S is given by (44), i.e. it is

degenerate at the maximum entropy design, A^o (46), of A. Finally, in case **(b)**, with $\mathsf{S}_\mathsf{I}(\mathsf{A}|K) \equiv \mathsf{U}(\mathsf{A})$ and A_I unspecified, then, from (45),

$$\mathsf{S}^o(\mathsf{A}|K) \propto \exp\left[-\mu^T \mathsf{E}_\mathsf{A}[\mathsf{g}]\right]. \tag{47}$$

Comparing (47) with (46), the elegance of the relaxation of the classical maximum entropy principle [13] to the fully probabilistic setting is revealed: $\mathsf{E}_\mathsf{A}[\mathsf{g}]$ now enters the FPD-optimal stochastic model for A, in place of $\mathsf{g}(x)$ which shaped the deterministic maximum entropy design of A (46).

3.3. Hierarchical Functional Constraints

The Žirafa theorem processes marginal functional constraints relating to the hierarchy in Definition 2, as already noted in the comments following (40), and this establishes a special rôle for $\hat{\mathsf{A}}(x|K)$ (40), the expected distribution of x under $\mathsf{S}(\mathsf{A}|K)$. This is consistent with knowledge specification (i) in Remark 9, which asserts that $\mathcal{I}$'s "natural knowledge structure" should constrain their beliefs about x and not A. The key insight of the Žirafa theorem is that—in the fully Bayesian context of Definition 2—this must be processed as a constraint on S (38), and yields a unique FPD-optimum (41).

While this knowledge structure is, indeed, natural in many contexts, it is interesting to relax the marginal functional constraints (38), and enquire into the FPD-optimal design of S under *joint* functional constraints on the hierarchy, M:

$$\mathsf{S} \in \mathbf{S}_K \equiv \{\mathsf{S}(\mathsf{A}|K) : \mathsf{E}_\mathsf{M}[\mathsf{g}_{\boldsymbol{x}\mathbf{A}}] \leq 0\} \neq \emptyset. \tag{48}$$

Here, $\mathsf{g}_{\boldsymbol{x}\mathbf{A}}(x, \mathsf{A})$ is a known, finite-dimensional, real mapping from the product set $\boldsymbol{x} \times \mathbf{A}$ of the joint model (3). Also,

$$\begin{aligned} \mathsf{E}_\mathsf{M}[\mathsf{g}_{\boldsymbol{x}\mathbf{A}}] &\equiv \int_\mathbf{A} \int_x \mathsf{g}_{\boldsymbol{x}\mathbf{A}}(x, \mathsf{A}) \mathsf{A}(x|K) \mathsf{S}(\mathsf{A}|K) \, \mathrm{d}x \, \mathrm{d}\mathsf{A} \\ &= \int_\mathbf{A} \mathsf{g}_\mathsf{A}(\mathsf{A}) \mathsf{S}(\mathsf{A}|K) \, \mathrm{d}\mathsf{A} \equiv \mathsf{E}_\mathsf{S}[\mathsf{g}_\mathbf{A}], \end{aligned} \tag{49}$$

where

$$\mathsf{g}_\mathsf{A}(\mathsf{A}) \equiv \mathsf{E}_\mathsf{A}[\mathsf{g}_{\boldsymbol{x}\mathbf{A}}].$$

The ideal specification (6) can also be relaxed, to allow fully hierarchical dependence:

$$\mathsf{M}_\mathsf{I}(x, \mathsf{A}|K) \equiv \mathsf{A}_\mathsf{I}(x|\mathsf{A}, K) \mathsf{S}_\mathsf{I}(\mathsf{A}|K). \tag{50}$$

This differs from (6) in allowing $\mathsf{A_I}$ to be chosen dependently on candidate values of the unknown distribution, A (recall the note in Section 2 on the freedom between preferences and beliefs). The processing of this knowledge structure, K, to yield an FPD-optimal design of the hierarchy, is provided by the next theorem.

Theorem 5 (Hierarchical Functional Constraints). *Consider the FPD problem specified in Theorem 4, conditioned on the fully hierarchical knowledge structure, K, defined in (48), with ideals as specified in (50). Then, the FPD minimizer, defined in (4), is*

$$\mathsf{S}^o(\mathsf{A}|K) \propto \mathsf{S_I}(\mathsf{A}|K) \exp\left[-\mathcal{D}(\mathsf{A}||\mathsf{A}^o)\right], \tag{51}$$

where

$$\mathsf{A}^o(x|\mathsf{A},K) \propto \mathsf{A_I}(x|\mathsf{A},K) \exp[-\mu^T \mathsf{g}_{\boldsymbol{x}\mathbf{A}}(x,\mathsf{A})], \tag{52}$$

with normalizing constant,

$$c_{\mathsf{A}^o} \equiv \int_{\boldsymbol{x}} \mathsf{A_I}(x|\mathsf{A},K) \exp[-\mu^T \mathsf{g}_{\boldsymbol{x}\mathbf{A}}(x,\mathsf{A})]\ \mathrm{d}x.$$

Proof: Similar to the proof of the Žirafa theorem, replacing (39) with (49). □

We note the following:

- The FPD design (51) differs from that in (41) in that the centering distribution (52) is now uncertain, being a functional of A. It replaces the deterministic centering distribution, $\mathsf{A}^o(x|K)$ (32), in the Žirafa theorem.
- This generalization (51) may again be useful in applications that require the processing of knowledge (functional constraints) relating to nonlinear transformations of the unknown distribution, A, of x, via (48).
- Letting $\mathsf{g}_{\boldsymbol{x}\mathbf{A}}(x,\mathsf{A}) \equiv \mathsf{g}(x)$ in (48), and $\mathsf{A_I}(x|\mathsf{A},K) \equiv \mathsf{A_I}(x|K)$ in (50), then the FPD design in (52) becomes $\mathsf{A}^o(x|\mathsf{A},K) = \mathsf{A}^o(x|K)$ (32), and so the FPD-optimal design, S^o (51), is the same as the one in (41). This confirms that the Žirafa theorem is a corollary of Theorem 5.

- The specialization of Theorem 5 to the case which processes *marginal* knowledge and preferences of $\mathsf{A} \sim \mathsf{S}$ can be found by letting $\mathsf{g}_{\boldsymbol{x}\mathbf{A}}(x, \mathsf{A}) = \mathsf{g}(\mathsf{A})$ and $\mathsf{A}_{\mathsf{I}}(x|\mathsf{A}, K) = \mathsf{A}(x|K)$, as in (18), leaving the design of A unconstrained. Then, the FPD-optimal design, S^o (51), is given by (35), and so Theorem 3 is also a special case of Theorem 5.

Notwithstanding these last two points, (41) provides the direct Bayesian relaxation of the long-established and widely adopted deterministic MXE-type design of A (32) [38]. For this reason, the Žirafa theorem is the primary technical development of the paper.

4. Relationship to Bayesian Nonparametric Design

Throughout this paper, we have assumed that the unknown quantity, x, has a finite-state set, $\boldsymbol{x}$ (case (b) of Section 2), and so we have been able to adopt standard probability calculus (3) with respect to probability models on spaces measurable with respect to Lebesgue or counting measure (Theorem 1). In particular, the KLD—which must be optimized within the FPD strategy (4)—can be defined in the usual way (5), given this assumption. The domain of S (and therefore of S^o (4)) is $\mathbf{A} \equiv \boldsymbol{\Delta}$ in this case, as noted in (b) of Section 2. This restriction means that further hyper-modelling of S may be required in particular applications in order to achieve robustness to choice of S. As already noted in Section 2, the relaxation to nonparametric A *can* be truncated at the level of the nonparametric process model, $\mathcal{S}$, as in case (c) of Section 2 [1]. Also as noted in Section 2, finite-state x can be viewed as the image of continuous y under a specified irreversible finite projection (quantization),

$$\mathsf{Q}_{\mathcal{P}(\boldsymbol{y})} : \boldsymbol{y} \to \boldsymbol{x}, \tag{53}$$

with $|\boldsymbol{y}| \not< \infty$ and $|\boldsymbol{x}| < \infty$. Here, the quantization schedule, $\mathsf{Q}_{\mathcal{P}(\boldsymbol{y})}$, is defined in terms of a specified finite, measurable partition, $\mathcal{P}(\boldsymbol{y})$, of $\boldsymbol{y}$. $\mathcal{I}$ models y via infinite-dimensional, unknown (i.e. nonparametric) distribution, $\mathsf{A}(y)$. Then, $\mathsf{A}_{\mathcal{P}(\boldsymbol{y})} \in \boldsymbol{\Delta}$ is the probability mass function induced on this *specific* $\mathcal{P}(\boldsymbol{y})$-indexed image, $\boldsymbol{x}$, of $\boldsymbol{y}$, by A [34]. FPD-optimal designs, $\mathsf{S}^o_{\mathcal{P}(\boldsymbol{y})}(\mathsf{A}_{\mathcal{P}(\boldsymbol{y})}|K)$ ((8), (41), *etc*) are also indexed by $\mathcal{P}(\boldsymbol{y})$, and require, as appropriate, specification of the ideal distribution, $\mathsf{A}_{\mathsf{I},\mathcal{P}(\boldsymbol{y})}$, on the image, x, and specification of the associated ideal, $\mathsf{S}_{\mathsf{I},\mathcal{P}(\boldsymbol{y})}$, on $\boldsymbol{\Delta}$.

Importantly, the various FPD-optimal designs of the hierarchy (3) for x (the image of y under quantization schedule, $\mathsf{Q}_{\mathcal{P}(\boldsymbol{y})}$ (53)), place no special restriction on the schedule, and are therefore valid for all such schedules. Hence, following [8], this leads to the following definition of the nonparametric generalization of these FPD-optimal designs.

Definition 3 (FPD-optimal Bayesian nonparametric processes). *Consider the Bayesian nonparametric hierarchy,*

$$\begin{array}{rcll} y|\mathsf{A}, K & \sim & \mathsf{A}(y|K), & y \in \boldsymbol{y}, \quad \mathsf{A} \in \mathbf{A}, \\ \mathsf{A}|\mathcal{S}, K & \sim & \mathcal{S}(\mathsf{A}|K), & \mathcal{S} \in \boldsymbol{\mathcal{S}}, \end{array} \tag{54}$$

and let $\mathsf{Q}_{\mathcal{P}(\boldsymbol{y})}$ (53) be any finite, measurable partition of $\boldsymbol{y}$.
(a) *Let $\mathcal{I}$ specify an ideal distribution, $\mathcal{S}_{\mathsf{I}}$, for A (54), and an ideal distribution A_{I}, for $y \in \boldsymbol{y}$. Then, the FPD-optimal Bayesian nonparametric process, conditioned on this K, is*

$$\mathsf{A}|K \sim \mathcal{S}^o(\mathcal{S}_{\mathsf{I}}, \mathsf{A}_{\mathsf{I}}),$$

defined such that the distribution of the measure induced by A under any *$\mathsf{Q}_{\mathcal{P}(\boldsymbol{y})}$ is given by the FPD-optimal design in (8). In the latter, S_{I} is the measure induced by $\mathcal{S}_{\mathsf{I}}$ on $\boldsymbol{\Delta}$, under $\mathsf{Q}_{\mathcal{P}(\boldsymbol{y})}$. Also in (8), the ideal A is the measure induced by A_{I} under $\mathsf{Q}_{\mathcal{P}(\boldsymbol{y})}$.*
(b) *Let K be further enriched, to include the constraint, $\mathsf{A} \in \mathbf{A}_K \subset \mathbf{A}$, where*

$$\mathbf{A}_K \equiv \left\{ \mathsf{A}(y|K) : \int_{\boldsymbol{y}} \mathsf{g}(y) \mathsf{A}(y|K) \, \mathrm{d}y \leq 0 \right\} \neq \emptyset,$$

and where $\mathsf{g}(y)$ is a specified finite-dimensional real function of $y \in \boldsymbol{y}$. Then, the FPD-optimal Bayesian nonparametric process, conditioned on this K, is $\mathsf{A}|K \sim \mathcal{S}^o(\mathcal{S}_{\mathsf{I}}, \mathsf{A}_{\mathsf{I}}, \mathsf{g})$, defined such that the distribution of the measure induced by A under any $\mathsf{Q}_{\mathcal{P}(\boldsymbol{y})}$ is given by the FPD-optimal design (41), where $\mathsf{A}^o(x|K)$ is the measure induced under the projection $\mathsf{Q}_{\mathcal{P}(\boldsymbol{y})}$ by the (deterministic) FPD design on $\boldsymbol{y}$ (32):

$$\mathsf{A}^o(y|K) \propto \mathsf{A}_{\mathsf{I}}(y|K) \exp[-\mu^T \mathsf{g}(y)].$$

Nonparametric relaxations of the other FPD designs in this paper may also be defined in a similar way. We omit a formal demonstration that the $\mathcal{S}^o(\cdot)$ defined respectively in **(a)** and **(b)** above are, indeed, FPD-optimal designs of the hierarchy (54) under the stated knowledge structures and ideals.

5. Discussion and Conclusion

The FPD-optimal design, $\mathsf{S}^o(\mathsf{A}|K)$, in (8) is a Gibbs distribution [10], assuming that S_I is, itself, chosen as Gibbs. Indeed, (8) may be re-written as

$$\mathsf{S}^o(\mathsf{A}|K) \propto \mathsf{S}_\mathsf{I}(\mathsf{A}) \exp\left(-\frac{1}{\tau}\mathcal{D}(\mathsf{A}||\mathsf{A}_\mathsf{I})\right), \tag{55}$$

being the FPD-optimal distribution when $\mathcal{I}$'s ideal model in (6) is adapted as follows:

$$\mathsf{M}_\mathsf{I}(x, \mathsf{A}|\mathsf{S}, \tau, K) \propto [\mathsf{A}_\mathsf{I}(x|\mathsf{S}_\mathsf{I}, K)]^{\frac{1}{\tau}}\, \mathsf{S}_\mathsf{I}(\mathsf{A}|K). \tag{56}$$

$\tau > 0$ acts as the 'temperature' parameter in the Gibbs distribution (55). A similar adaptation of the FPD-optimal nonparametric measure, $\mathsf{A}|K \sim \mathcal{S}^o(\mathcal{S}_\mathsf{I}, \mathsf{A}_\mathsf{I})$, in Definition 3**(a)** above, is also possible. The role of τ is evident from (56): it explores a class of ideal distributions for x, ranging from uniform on $\boldsymbol{x}$ ($\tau \to \infty$) to ones that are tightly distributed around $\arg\max_{\boldsymbol{x}} \mathsf{A}_\mathsf{I}(x|\mathsf{S}_\mathsf{I}, K)$ ($\tau \to 0$). In applications involving sequential processing of knowledge, K, a schedule for varying τ can be adopted to control the influence of A_I, providing a fully Bayesian counterpart to simulated annealing in classical estimation [40].

In the FPD-optimal Gibbs formulation (55), the KLD acts as the free energy of A relative to A_I. In the specialization (25), it is the differential entropy of A, i.e. $\ln|\boldsymbol{x}| - \mathcal{H}_\mathsf{A}$ (23), that acts as the free energy. These FPD optima lead to interesting interpretations of the information-theoretic quantities, $\mathcal{D}$ and $\mathcal{H}$, and they also reveal the conditions under which their use in various applications is justified. For instance, (25) has been widely adopted as a smoothness prior for regularizing the solution of inverse problems involving measures. In particular, it has been successfully adopted in maximum entropy image restoration [39]. Section 2.1 confirms that this prior is optimal, in the FPD sense, under hierarchical modelling of unknown A, when the ideal hierarchy is specified as uniform on $\boldsymbol{x} \times \mathbf{A}$.

Fully probabilistic design of hierarchical Bayesian models is a powerful and flexible principle for optimal model design, and can provide solutions for a wide range of problems in inference and decision-making. These include the following:

Merging of external data knowledge The observer, $\mathcal{I}$, models data d, by a parametric model $\mathsf{F}_\theta \equiv \mathsf{F}(d|\theta)$, $\theta \in \boldsymbol{\theta}$ and may also have access to external data-based knowledge, expressed via the distribution, $\mathsf{F} \equiv$

$\mathsf{F}(d)$. [26] proposed the following way to modify $\mathcal{I}$'s prior distribution, $\mathsf{F}_{\mathsf{P}}(\theta)$, in order to merge this external data-based knowledge, K:

$$\mathsf{F}^{o}(\theta|K) \propto \mathsf{F}_{\mathsf{P}}(\theta) \exp[-\mathcal{D}(\mathsf{F}||\mathsf{F}_{\theta})].$$

This formula was heuristically designed in [26], but can be shown to be a consequence of Corollary 2**(a)**. In [20], the practical usefulness of this handling of probabilistic data-based knowledge, F, was demonstrated.

Approximate recursive learning and stabilised forgetting Recursive learning modifies an already computed distribution,

$$\mathsf{F}_{t-1}(\theta) \equiv \mathsf{F}(\theta|d_{t-1}, \ldots, d_1, K),$$

of an unknown parameter, θ, with a parametric model, $\mathsf{F}(d|\theta)$, via Bayes' rule. Typically, the unreduced $\mathsf{F}_t(\theta)$ is to be approximated by a tractable distribution, $\mathsf{G}_t(\theta)$, on $\boldsymbol{\theta}$. The sequential approximation process causes the exact specification of F_t to be replaced by knowledge that it lies inside a ball $\mathcal{D}(\mathsf{F}_t||\mathsf{G}_t) \leq \kappa_t < \infty$. Application of FPD with this *non-linear constraint* on the true distribution, F_t, achieves the form of stabilized forgetting and counteracts the accumulation of approximation errors that otherwise occurs [17]. The consistency and optimality of this procedure are verified by the findings in this paper. A recent application to the recursive estimation of high-order Markov chains is provided in [18].

Cooperation of interacting agents Complex inference problems involve distributed processing of knowledge and distributed decision-making. Sharing of knowledge and searching for a compromise between differing objectives can be handled consistently within FPD by processing (merging) knowledge and objectives (preferences) described by probability distributions, as in this paper. To achieve good merging requires comparison of the KLD from the processed distributions to a proposed compromise; i.e. again to use FPD with nonlinear constraints [37]. An application in local adaptive control design for multi-agent cooperative systems is presented in [24].

As noted, key applications such as these require knowledge constraints to be formulated via nonlinear functionals of the unknown distribution, A, of x, as in (33). This paper has shown that such knowledge constraints arise

in, and are optimally processed by, fully probabilistic design of the hierarchical model. By demonstrating consistency with the axioms of the minimum cross-entropy principle in the case where ideals replace priors, *all* the FPD designs in this paper—including those that process nonlinear functionals of the unknown distribution—are therefore verified as being consistent also with Bayesian conditioning [4].

In adopting FPD for optimal design of $\mathsf{S}(\mathsf{A}|K)$, we have avoided point estimation of A and any requirement for a prior choice or estimate of A, as necessitated by the classical maximum entropy and minimum cross-entropy principles in their usual settings (see Section 3). In this way, we have avoided the limitations of such point estimation, providing, for instance, measures of uncertainty around these classical distributional estimates. Importantly, the computation of FPD optima, S^o, such as (8) involves random draws from $\mathsf{A} \in \mathbf{A}$. This can be computationally far less onerous than the task of optimizing $\mathsf{A} \in \mathbf{A}$, and it allows $\mathcal{I}$'s intrinsic uncertainty about A to be represented in their inference of x, as fomalized in (13).

Finally, the hierarchical setting of FPD has allowed us to relate the designs, S^o, to important classes of distribution, including Gibbs, and related entropic distributions, as explained earlier in this section. In particular, it has inspired an FPD-optimal design of a distribution for nonparametric processes, $\mathsf{A}|K \sim \mathcal{S}^o$, in which knowledge and preferences about x and its unknown distribution—being nonparametric A—are optimally processed.

Acknowledgement

This research has been supported by SFI grant 10/RFP/MTH2877 and by GAČR grant 13-13502S.

References

[1] C.E. Antoniak. Mixtures of Dirichlet processes with applications to Bayesian nonparametric problems. *The Annals of Statistics*, 2(6) (1974) 1152–1174.

[2] J.O. Berger. *Statistical Decision Theory and Bayesian Analysis*. Springer, New York, (1985).

[3] J.M. Bernardo. Expected information as expected utility. *The Annals of Statistics*, 7(3) (1979) 686–690.

[4] J.M. Van Campenhout and T.M. Cover. Maximum entropy and conditional probability. *IEEE Tran. on Inf. Theory*, 27(4) (1981) 483–489.

[5] L. Chen and P. Pu. Survey of preference elicitation methods. Technical Report IC/2004/67, Human Computer Interaction Group Ecole Politechnique Federale de Lausanne (EPFL), CH-1015 Lausanne, Switzerland, 2004.

[6] T. Cover and J. Thomas. *Elements of Information Theory*. Wiley, (1991). 2nd edition.

[7] R.T. Cox. Probability, frequency and reasonable expectation. *American Journal of Physics*, 14(1) (1946) 1–13.

[8] T.S. Ferguson. A Bayesian analysis of some nonparametric problems. *The Annals of Statistics*, 1 (1973) 209–230.

[9] P.H. Garthwaite, J.B. Kadane, and A. O´Hagan. Statistical methods for eliciting probability distributions. *J. of the American Statistical Association*, 100(470) (2005) 680–700.

[10] S. Geman and D. Geman. Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, PAMI-6(6) (1984) 721–741.

[11] L. Groarke. *An Aristotelian Account of Induction: Creating Something From Nothing*. McGill-Queen's University Press, (2009).

[12] E.T. Jaynes. Information theory and statistical mechanics. *Physical Review Series II*, 106(4) (1957) 620–630.

[13] E.T. Jaynes. *Probability Theory: The Logic of Science*. Cambridge University Press, (2003).

[14] H Jeffreys. *Theory of Probability*. Oxford University Press, 3 edition, (1961).

[15] H. Jeffreys. *Theory of Probability*. Oxford University Press, (1998).

[16] M. Kárný. On approximate fully probabilistic design of decision making strategies. In T.V. Guy and M. Kárný, editors, *Proceedings of the 3rd International Workshop on Scalable Decision Making, ECML/PKDD 2013*. UTIA AV ČR, Prague, 2013. ISBN 978-80-903834-8-7.

[17] M. Kárný. Approximate Bayesian recursive estimation. *Information Sciences*, 289 (2014) 100–111. DOI 10.1016/j.ins.2014.01.048.

[18] M. Kárný. Recursive estimation of high-order Markov chains: approximation by finite mixtures. *Information Sciences*, 326 (2016) 188–201.

[19] M. Kárný, J. Andrýsek, A. Bodini, T. V. Guy, J. Kracík, P. Nedoma, and F. Ruggeri. Fully probabilistic knowledge expression and incorporation. Technical Report 8-10MI, Istituto di Matematica Applicata e Tecnologie Informatiche, 2008.

[20] M. Kárný, A. Bodini, T.V. Guy, J. Kracík, P. Nedoma, and F. Ruggeri. Fully probabilistic knowledge expression and incorporation. *Statistics and Its Interface*, 7(4) (2014) 503–515.

[21] M. Kárný, J. Böhm, T. V. Guy, L. Jirsa, I. Nagy, P. Nedoma, and L. Tesař. *Optimized Bayesian Dynamic Advising: Theory and Algorithms*. Springer, (2006).

[22] M. Kárný and T.V. Guy. Preference elicitation in fully probabilistic design of decision strategies. In *Proc. of the 49th IEEE Conference on Decision and Control*, 2010.

[23] M. Kárný and T.V. Guy. On support of imperfect Bayesian participants. In T.V. Guy, M. Kárný, and D.H. Wolpert, editors, *Decision Making with Imperfect Decision Makers*, volume 28. Springer, Berlin, 2012. Intelligent Systems Reference Library.

[24] M. Kárný and R. Herzallah. Scalable harmonization of complex networks with local adaptive controllers. *IEEE Trans. on Systems, Man and Cybernetics: Systems*. DOI 10.1109/TSMC.2015.2502427 (11 pages).

[25] M. Kárný and T. Kroupa. Axiomatisation of fully probabilistic design. *Information Sciences*, 186(1) (2012) 105–113.

[26] J. Kracík and M. Kárný. Merging of data knowledge in Bayesian estimation. In J. Filipe, J. A. Cetto, and J. L. Ferrier, editors, *Proc. of the Second Int. Conference on Informatics in Control, Automation and Robotics*, pages 229–232, Barcelona, 2005. INSTICC.

[27] H.W. Kuhn and A.W. Tucker. Nonlinear programming. In *Proc. of 2nd Berkeley Symposium*, pages 481–492. Univ. of California Press, 1951.

[28] S. Kullback and R. Leibler. On information and sufficiency. *Annals of Mathematical Statistics*, 22 (1951) 79–87.

[29] J. Kyparisis. On uniqueness of Kuhn-Tucker multipliers in nonlinear programming. *Mathematical Programming*, 32 (1985) 242–246.

[30] P. Laplace. *Theorie Analytique des Probabilités*. Courcier, (1812).

[31] D.J.C. MacKay. *Information Theory, Inference, and Learning Algorithms*. Cambridge University Press, (2003).

[32] P. Müller and F.A. Quintana. Nonparametric Bayesian data analysis. *Statist. Sci.*, 19(1) (2004) 95–110.

[33] V. Peterka. Bayesian system identification. In P. Eykhoff, editor, *Trends and Progress in System Identification*, pages 239–304. Pergamon Press, Oxford, 1981.

[34] A. Quinn and M. Kárný. Learning for nonstationary Dirichlet processes. *Int. J. of Adaptive Control and Signal Processing*, 21(10) (2007) 827–855.

[35] M.M. Rao. *Measure Theory and Integration*. John Wiley, NY, (1987).

[36] L.J. Savage. *Foundations of Statistics*. Wiley, NY, (1954).

[37] V. Sečkárová. On supra-Bayesian weighted combination of available data determined by Kerridge inaccuracy and entropy. *Pliska Stud. Math. Bulgar.*, 22 (2013) 159–168.

[38] J. Shore and R. Johnson. Axiomatic derivation of the principle of maximum entropy & the principle of minimum cross-entropy. *IEEE Trans. on Inf. Th.*, 26(1) (1980) 26–37.

[39] J. Skilling and S.F. Gull. Maximum entropy method in image processing. *Proceedings of IEE*, 131 (1984) 646.

[40] B. Suman and P. Kumar. A survey of simulated annealing as a tool for single and multiobjective optimization. *J. of the Operational Research Society*, 57 (2005) 1143–1160.

[41] A. Zellner. Models, prior information, and Bayesian analysis. *Journal of Econometrics*, 75(1) (1996) 51–68.