Pointless Learning

Florence Clerc[1], Vincent Danos[2,4], Fredrik Dahlqvist[3], and Ilias Garnier[4(✉)]

[1] McGill University, Montreal, Canada
[2] ENS Paris/CNRS, Paris, France
[3] UCL, London, UK
[4] University of Edinburgh, Edinburgh, Scotland
ilias.gar@gmail.com

Abstract. Bayesian inversion is at the heart of probabilistic programming and more generally machine learning. Understanding inversion is made difficult by the pointful (kernel-centric) point of view usually taken in the literature. We develop a pointless (kernel-free) approach to inversion. While doing so, we revisit some foundational objects of probability theory, unravel their category-theoretical underpinnings and show how pointless Bayesian inversion sits naturally at the centre of this construction.

1 Introduction

The soaring success of Bayesian machine learning has yet to be matched with a proper foundational understanding of the techniques at play. These statistical models are fundamentally programs that manipulate probability distributions. Therefore, the semantics of programming languages can and should inform the semantics of machine learning. This point of view, upheld by the proponents of *probabilistic programming*, has given rise to a growing body of work on matters ranging from the computability of disintegrations [1] to operational and denotational semantics of probabilistic programming languages [12]. These past approaches have all relied on a pointful, *kernel-centric* view of the key operation in Bayesian learning, namely *Bayesian inversion*. In this paper, we show that a *pointless*, operator-based approach to Bayesian inversion is both more general, simpler and offers a more structured view of Bayesian machine learning.

Let us recall the underpinnings of Bayesian inversion in the finite case. Bayesian statistical inference is a method for updating subjective probabilities on an unknown random process as observations are collected. In a finite setting, this update mechanism is captured by Bayes' law:

$$P(d) \cdot P(h \mid d) = P(d \mid h) \cdot P(h) \tag{1}$$

F. Clerc—Research supported by NSERC, Canada.

F. Dahlqvist—Research partially supported by the ERC project *ProFoundNet - Probabilistic Foundations for Networks* (grant number 679127).

I. Garnier—Research supported by the ERC project *RULE* (grant number 320823).

J. Esparza and A.S. Murawski (Eds.): FOSSACS 2017, LNCS 10203, pp. 355–369, 2017.
DOI: 10.1007/978-3-662-54458-7_21

On the right-hand side, the *likelihood* $P(d \mid h)$ encodes a parameter-dependent probability over data, weighted by the *prior* $P(h)$ which corresponds to our current belief on which parameters best fit the law underlying the unknown random process. The left-hand side of Eq. 1 involves the *marginal likelihood* $P(d)$, which is the probability of observing the data d under the current subjective probability, and the *posterior* $P(h \mid d)$ which tells us how well the occurrence of d is explained by the parameter h. More operationally, the posterior tells us how we should revise our prior as a function of the observed data d. In a typical Bayesian setup, the prior and likelihood are given and the marginal likelihood can be computed from the two first ingredients. The only unknown is the *posterior* $P(h \mid d)$. Equation 1 allows one to compute the posterior from the two first ingredients–whenever $P(d) > 0$! This formulation emphasises the fundamental symmetry between likelihood and posterior, and hopefully makes clear why the process of computing the posterior is called *Bayesian inversion*. The key observation is that both the likelihood and posterior can be seen as matrices, and Eq. 1 encodes nothing more than a relation of adjunction between these matrices seen as (finite-dimensional) operators. This simple change of point of view, where one thinks no longer directly in terms of kernels (which transform probability measures forward), but in terms of their semantics as operators (which transform real-valued obervables backward) generalises well and gives us a much more comprehensive account of Bayesian learning as adjunction. If one thinks of observables as extended predicates, this change of point of view is nothing but a predicate transformer semantics of kernels: a well-established idea planted in the domain of probabilistic semantics by Kozen in the 80s [10]. The object of this paper is to develop in this setting a *pointless* approach to Bayesian inversion.

Our contributions are as follows. In Sect. 3, we recall how Bayesian inversion is formulated using the language of kernels, following the seminal work of [5] and our own preliminary elaboration of the ideas developed in the current paper [6]. The adequate setting is a category of *typed* kernels, i.e. measure-preserving kernels between probability spaces. We observe that Bayesian inversion fits somewhat awkwardly in this pointful setting. Drawing from domain-theoretic ideas [11], we develop in Sect. 4 a categorical theory of ordered Banach cones, including an adjunction theorem for L_p^+/L_q^+ cones taken from Reference [3]. In Sect. 5, we define a functorial operator interpretation of kernels in the category of Banach cones and prove that pointful Bayesian inversion corresponds through this functorial bridge to adjunction, expanding our recent result [6] to arbitrary L_p^+/L_q^+ cones. Unlike the pointful case, the pointless, adjunction-based approach works with arbitrary measurable spaces. Finally, in Sect. 6 we extract from the pointful and pointless approaches what we consider to be the essence of Bayesian inversion: a correspondence between *couplings* and linear operators. In this new light, adjunction (and therefore Bayesian inversion) is nothing more than a permutation of coordinates. We conclude with a sketch of some directions for future research where one could most profit of the superior agility and extension of the pointless approach.

Note that a long version of this article, containing all proofs, is available [4].

2 Preliminaries

We refer the reader to e.g. [2] for the concepts of measure theory and functional analysis used in this paper. For convenience, some basic definitions are recalled here.

A *measurable space* $(X, \varSigma)$ is given by a set X together with a σ-algebra of subsets of X denoted by $\varSigma$. Where unambiguous, we will omit the σ-algebra and denote a measurable space by its underlying set. We will also consider the measurable spaces generated from *Polish* (completely metrisable and separable) topological spaces, called *standard Borel spaces* [9]. A measurable function $f : (X, \varSigma) \to (Y, \varLambda)$ is a function $f : X \to Y$ such that for all $B \in \varLambda$, $f^{-1}(B) \in \varSigma$. The category of measurable spaces and measurable functions will be denoted by **Mes**. For B a measurable set, we denote by $\mathbb{1}_B$ the indicator function of that set. A *finite measure* μ over a measurable space $(X, \varSigma)$ is a σ-additive function $\mu : \varSigma \to [0, \infty)$ that verifies $\mu(X) < \infty$. Whenever $\mu(X) = 1$, μ is a *probability measure*. A pair (X, μ) with X a measurable space and μ a probability measure on X is called a *probability space*. A measurable set B will be qualified of *μ-null* if $\mu(B) = 0$.

The Giry endofunctor, denoted by $\mathsf{G} : \mathbf{Mes} \to \mathbf{Mes}$, maps each measurable space X to the space $\mathsf{G}(X)$ of probability measures over X. The measurable structure of $\mathsf{G}(X)$ is given by the initial σ-algebra for the family $\{ev_B\}_B$ of evaluation maps $ev_B(\mu) = \mu(B)$, where B ranges over measurable sets in X. The action of G on arrows is given by the pushforward (or image measure): for $f : X \to Y$ measurable, we have $\mathsf{G}(f) : \mathsf{G}(X) \to \mathsf{G}(Y)$ given by $\mathsf{G}(f)(\mu) = \mu \circ f^{-1}$. This functor admits the familiar monad structure (G, m, δ) where $m : \mathsf{G}^2 \Rightarrow \mathsf{G}$ and $\delta : Id \Rightarrow \mathsf{G}$ are natural transformations with components at X defined by $m_X(P)(B) = \int_{\mathsf{G}(X)} ev_B \, dP$ and $\delta_X(x)(B) = \delta_x(B)$. It is well-known that when restricted to standard Borel spaces, the Giry functor admits the same monad structure. See [7] for more details on this construction. The Kleisli category of the Giry monad, corresponding to Lawvere's category of probabilistic maps, will be denoted by $\mathcal{K}\ell$. The objects of $\mathcal{K}\ell$ correspond to those of **Mes** and arrows from X to Y correspond to so-called *kernels* $f : X \to \mathsf{G}(Y)$. Kleisli arrows will be denoted by $f : X \rightarrowtriangle Y$. For $f : X \rightarrowtriangle Y, g : Y \rightarrowtriangle Z$, the Kleisli composition is defined as usual by $g \circ' f = m_Z \circ \mathsf{G}(g) \circ f$. We distinguish deterministic Kleisli maps as those that can be factored as a measurable function followed by δ and denote these arrows $f : X \rightarrowtriangle_\delta Y$. We write 1 for the one element measurable space (which is the terminal object in **Mes**). Clearly the Homset $\mathcal{K}\ell(1, Y)$ is in bijection with the set of probabilities over Y. This justifies the following slight abuse of notation: if $\mu \in \mathsf{G}(X)$ is a probability and $f : X \rightarrowtriangle Y$ is a kernel, the pushforward of μ through f will be denoted $f \circ' \mu$. Observe that for $f : X \to Y$ an usual measurable map, $\mathsf{G}(f)(\mu) = (\delta_Y \circ f) \circ' \mu$, so the pushforward through a kernel extends the earlier definition.

Consider the full subcategory of $\mathcal{K}\ell$ restricted to finite spaces. In that setting, any kernel $f : X \rightarrowtriangle Y$ can be presented as a positive, real-valued matrix that we denote $T(f) = \{f(x)(y)\}_{x,y}$ with X rows, Y columns and where all rows sum to 1 (aka a stochastic matrix). Matrix multiplication corresponds to Kleisli composition: taking f, g as above, one has $T(g \circ' f) = T(f)T(g)$ (hence, this representation of kernels as matrices is *contravariant*). Such matrices act on vectors of dimension Y (*observables on* Y) and map them to observables on X: for $v \in \mathbb{R}^Y$, $T(f)v$ corresponds to the expectation of v according to f. This is the basis for the "operator interpretation" of kernels which we will extend to **Mes** below.

3 Bayesian Inversion in a Category of Typed Kernels

We introduce the category **Krn** of typed kernels and recall the statement of Bayesian inversion in this setting.

3.1 Definition of Krn

Our starting point is the under category $1 \downarrow \mathcal{K}\ell$, where 1 is the one-element measurable space. Objects of $1 \downarrow \mathcal{K}\ell$ are Kleisli arrows $\mu : 1 \multimap X$, i.e. probability spaces (X, μ) with $\mu \in \mathsf{G}(X)$; while typed kernels from (X, μ) to (Y, ν) are Kleisli arrows $f : X \multimap Y$ such that $f \circ' \mu = \nu$. We will call these arrows "kernels" for short. For a deterministic map $f_\delta : X \multimap_\delta Y$ (factoring as $f_\delta = \delta_Y \circ f$), the constraint boils down to $\nu = \mathsf{G}(f)(\mu)$. In other words, the subcategory of $1 \downarrow \mathcal{K}\ell$ consisting of deterministic maps is isomorphic to the usual category of probability spaces and measure-preserving maps. We define **Krn** to be the subcategory of $1 \downarrow \mathcal{K}\ell$ restricted to standard Borel spaces.

3.2 Bayesian Inversion in the Finite Subcategory of Krn

We translate the presentation of Bayesian inversion of Sect. 1 in the language of **Krn**. We are given finite spaces of data D and parameters H and it is assumed that there exists an unknown probability on D, called the "truth" and denoted τ in the following, that we wish to learn. The likelihood corresponds to a $\mathcal{K}\ell$ arrow $f : H \multimap D$, The prior is a probability $\mu \in \mathsf{G}(H)$ while the marginal likelihood $\nu \in \mathsf{G}(D)$ is obtained as $\nu = f \circ' \mu$. Thus the entire situation is captured by a **Krn** arrow $f : (H, \mu) \multimap (D, \nu)$. *If* our prior was perfect, we would have $\nu = \tau$ but of course (by assumption) this is not the case! The only access we have to the truth is through an infinite, independent family $\{d_n\}_{n \in \mathbb{N}}$ of random elements in D each distributed according to τ. The Bayesian update is the process of using this sequence of data (sometimes called *evidence*) to iteratively revise our prior. In this language, Bayes's law reads as follows:

$$\nu(d) \cdot f^\dagger(d)(h) = f(h)(d) \cdot \mu(h) \tag{2}$$

where $f^\dagger : (D, \nu) \multimap (H, \mu)$ denotes the sought posterior map, to be computed in function of μ and f. Observe that both the left and right hand side of Eq. 2 define the same *joint probability* $\gamma \in \mathsf{G}(H \times D)$ given by $\gamma(h, d) = f(h)(d) \cdot \mu(h) = \nu(d) \cdot f^\dagger(d)(h)$. Denoting π_H, π_D the left and right projections from $H \times D$, one easily verifies that $\mathsf{G}(\pi_H) = \mu$ and $\mathsf{G}(\pi_D) = \nu$. In other terms, γ is a *coupling* of μ and ν. We draw the attention of the reader to the following points.

- As hinted before, $f^\dagger(d)$ is uniquely defined only when $\nu(d) > 0$. Conversely, $f^\dagger$ does not depend on f on μ-null sets. These hurdles will be circumvented by considering equivalence classes of kernels up to null sets. This is the object of Sect. 3.3.
- Section 2 introduces a correspondence between (finite) kernels and Markov or stochastic matrices. This raises the following question: what is Bayesian inversion seen through that lens? The answer is *adjunction.* As we show in Sect. 5, this *pointless* point of view generalises to arbitrary measurable spaces and is better behaved than the pointful one.

We now proceed to the generalisation of this situation to the case of standard Borel spaces, i.e. to that of **Krn**.

3.3 Bayesian Inversion in Krn

Bayesian inversion in **Krn** relies crucially on the construction of an (almost sure) bijection between the **Krn** Homset $\mathbf{Krn}(X,\mu;Y,\nu)$ and the set of couplings $\mathbf{\Gamma}(X,\mu;Y,\nu)$ of μ and ν (to be defined next).

Couplings and kernels. To any pair of objects $(X,\mu)(Y,\nu)$, one can associate the space of *couplings* of μ and ν, i.e. the set of all probabilities $\gamma \in \mathsf{G}(X \times Y)$ such that $\mathsf{G}(\pi_X)(\gamma) = \mu$ and $\mathsf{G}(\pi_Y)(\gamma) = \nu$. We denote this set of couplings $\mathbf{\Gamma}(X,\mu;Y,\nu)$. It is a standard Borel space, as the set of couplings of two measures is a closed convex subset in $\mathsf{G}(X \times Y)$ for any choice of a Polish topology for X, Y. In order to construct a mapping from couplings to **Krn** arrows, we will need the *disintegration theorem*, which requires us to introduce some terminology. In the following, we denote $N(f,f') = \{x \mid f(x) \neq f'(x)\}$.

Lemma 1. *For all $f, f' : (X,\mu) \rightarrowtriangle (Y,\nu)$, $N(f,f')$ is measurable.*

Note that in more general measurable spaces, $N(f,f')$ is not necessarily a measurable set, as those spaces are not always countably generated. We can now introduce the disintegration theorem.

Theorem 1 (Disintegration ([8], Theorem 5.4)). *For all deterministic* **Krn** *arrow $f : (X,\mu) \rightarrowtriangle_\delta (Y,\nu)$, there exists $f^\dagger : (Y,\nu) \rightarrowtriangle (X,\mu)$ such that $f \circ' f^\dagger = id_{(Y,\nu)}$ and such that for all $h : (Y,\nu) \rightarrowtriangle (X,\mu)$ verifying $f \circ' h = id_{(Y,\nu)}$, $\nu(N(f^\dagger,h)) = 0$. In short, we say that $f^\dagger$ is the ν-almost surely unique kernel verifying $f \circ' f^\dagger = id_{(Y,\nu)}$.*

Disintegrations correspond to *regular conditional probabilities* (see e.g. [8]). The deterministic map $f : X \to Y$ along which the disintegration of μ is computed acts through its fibers as a parameterised family of subsets on each of which μ is conditioned, resulting in a measurable family of conditional probabilities parameterised by Y. Note that the characteristic property of disintegrations can be equivalently stated as the fact that $f^\dagger(y)$ is ν-almost surely supported by $f^{-1}(y)$.

Example 1. In the finite case, disintegration is simply the formula for conditional probabilities. Given X, Y finite and $f : (X, \mu) \rightarrow\!\!\triangleright_\delta (Y, \nu)$, for $y \in Y$ s.t. $\nu(y) = \mu(f^{-1}(y)) > 0$, it holds that $f^\dagger(y)(x) = \frac{\mu(x)}{\nu(y)}$. However, when $\nu(y) = 0$, the disintegration theorem does not constrain the value of $f^\dagger(y)$.

Disintegration establishes a bijective (up to null sets) correspondence between couplings and kernels. Let us make this formal.

Definition 1. *For fixed $(X, \mu)(Y, \nu)$, we define on $\mathbf{Krn}(X, \mu; Y, \nu)$ $\sim$ as the smallest equivalence relation such that $f \sim f'$ if $\mu(N(f, f')) = 0$. We denote $\mathbf{Krn}(X, \mu; Y, \nu)/\mu$ the set of $\sim$-equivalence classes of $\mathbf{Krn}(X, \mu; Y, \nu)$.*

Any **Krn** arrow $f : (X, \mu) \rightarrow\!\!\triangleright (Y, \nu)$ induces a measure on $X \times Y$, defined on measurable rectangles $B_X \times B_Y$ as:

$$I^{Y,\nu}_{X,\mu}(f)(B_X \times B_Y) = \int_{x \in B_X} f(x)(B_Y)\, d\mu. \tag{3}$$

Lemma 2. *$I^{Y,\nu}_{X,\mu}$ is a **Set** injection from $\mathbf{Krn}(X, \mu; Y, \nu)/\mu$ to $\mathbf{\Gamma}(X, \mu; Y, \nu)$.*

The second part of the bijection between couplings and quotiented **Krn** arrows relies crucially on disintegration.

Lemma 3. *There is a **Set** injection $D^{Y,\nu}_{X,\mu} : \mathbf{\Gamma}(X, \mu; Y, \nu) \to \mathbf{Krn}(X, \mu; Y, \nu)/\mu$. Moreover, $D^{Y,\nu}_{X,\mu}$ and $I^{Y,\nu}_{X,\mu}$ are inverse of one another.*

Bayesian inversion in **Krn**. Bayesian inversion corresponds to the composition of the bijections we just defined with the pushforward along the permutation map $\sigma : X \times Y \to Y \times X$.

Theorem 2 (Bayesian inversion). *Let $-^\dagger$ be defined as $f^\dagger = D^{X,\mu}_{Y,\mu} \circ \mathsf{G}(\sigma) \circ I^{Y,\nu}_{X,\mu}$. The map $-^\dagger : \mathbf{Krn}(X, \mu; Y, \nu)/\mu \to \mathbf{Krn}(Y, \nu; X, \mu)/\nu$ is a bijection.*

This section would be incomplete if we didn't address *learning* in its relation to Bayesian inversion. It is known that in good cases,[1] Bayesian inversion will make the sequence of marginal likelihoods converge to the truth in some appropriate topology. However, issues of convergence are not the subject of this paper and will not be discussed further.

3.4 Pointfulness Is Harmful

Let us take a critical look at the approach to Bayesian inversion developed so far. The fact that $-^\dagger$ is by construction $\sim$-invariant and yields $\sim$-equivalence classes of **Krn** arrows suggests that $-^\dagger$ would be better typed on a hypothetical quotient of **Krn** by $\sim$. This mismatch between the behaviour of $-^\dagger$ and its actual type already arises in the finite case where Bayes' rule yields kernels

[1] E.g. H, D finite and μ putting strictly positive measure on $f^{-1}(\tau)$.

only defined up to a null set (see discussion after Eq. 2), and is an inevitable consequence of the pointful point of view: kernels should respect the measures endogenous to their domain. Constructing the quotient of **Krn** w.r.t. $\sim$ would require proving that this equivalence relation is compatible with the composition of **Krn**. However, carrying out this approach successfully seems non-trivial:[2] our past attempts are riddled with obstructions stemming from accumulation of negligible sets–the very technical hurdles that make the theory of disintegration of measures so unintuitive in the first place, while moreover relying on standard Borel assumptions.

This improper typing obscures the categorical structure of Bayesian inversion. In the next sections, we leave the inhospitable world of kernels and relocate the theory of Bayesian inversion in a category of Banach cones and linear maps where these problems vanish, and the structure we seek for becomes manifest.

4 Banach Cones

Following [3,11], we introduce a category of Banach cones and ω-continuous linear maps, with the intent of interpreting Markov kernels as linear operators between well-chosen function spaces. In the subcategory corresponding to these function spaces, we develop a powerful adjunction theorem that will be used in Sect. 5 to implement pointless Bayesian inversion.

4.1 The Category Ban

A *Banach cone*, informally, corresponds to a normed convex cone of a Banach space which is ω-complete with respect to a particular order. Let us introduce these cones progressively.

Definition 2. *A normed, convex cone $(C, +, \cdot, 0, \|\cdot\|_C)$ of a normed vector space $(V, +, \cdot, 0, \|\cdot\|_V)$ is a subset $C \subseteq V$ that is closed under addition, convex combinations and multiplications by non-negative scalars, endowed with the restriction of the ambient norm, which must be monotone w.r.t. the partial order $u \leq_C v \Leftrightarrow \exists w \in C. u + w = v$.*

We require our Banach cones to be *ω-complete* with respect to this order, and to be subsets of Banach spaces.

Definition 3 (Banach cones). *A normed convex cone C is* ω-complete *if for all chain (i.e. $\leq_C$-increasing countable family) $\{u_n\}_{n\in\mathbb{N}}$ of bounded norm, the least upper bound $\bigvee_n u_n$ exists and $\|\bigvee_n u_n\|_C = \bigvee_n \|u_n\|_C$. A* Banach cone *is an ω-complete normed cone of a Banach space.*

Norm convergence and order convergence are related by the following result.

[2] Without additional assumptions the quotient is not compatible with precomposition, differently to what we mistakenly stated in ([6], Lemma 3).

Lemma 4 ([3], Lemma 2.12). *Let $\{u_n\}_{n\in\mathbb{N}}$ be a chain of bounded norm in a Banach cone. Then $\lim_{i\to\infty} \|\bigvee_n u_n - u_i\| = 0$.*

A prime example of Banach cones is given by the *positive cones* associated to classical L_p spaces of real-valued functions (see e.g. [2] for a definition of those spaces). In details: for (X, μ) a measure space with μ finite and $p \in [1, \infty]$, the set of elements $f \in L_p(X, \mu)$ which are non-negative μ-a.e. is closed under addition, multiplication by non-negative scalars and under linear combinations with non-negative coefficients. Equipped with the restriction of the norm of $L_p(X, \mu)$, this subset forms a normed convex cone that we denote $L_p^+(X, \mu)$. The partial order associated to these L_p^+ cones can be defined explicitly: for $f, g \in L_p^+(X, \mu)$, we write that $f \leq g$ if $f(x) \leq g(x)$ μ-a.e. One easily checks that this coincides with the definitional partial order.

Proposition 1 (ω-completeness of L_p^+ cones, [3]). *For all X measurable, $\mu \in \mathsf{G}(X)$ and $p \in [1, \infty]$, $L_p^+(X, \mu)$ is a Banach cone.*

This result is a direct consequence of the definition of suprema in $L_p^+(X, \mu)$. We are going to construct a category of all Banach cones and we thus have to specify what a morphism between such cones is. We consider only linear maps which are Scott-continuous, which in this case[3] boils down to commuting with supremas of increasing chains.

Definition 4. *Let C, C' be Banach cones and $A : C \to C'$ be a linear map. A is ω-continuous if for every chain $\{f_n\}_{n\in\mathbb{N}}$ such that $\bigvee_n f_n$ exists, $A(\bigvee_n f_n) = \bigvee_n A(f_n)$.*

The following example should help make ω-continuity less mysterious. Observe that for $Y = 1$ (the singleton set), all Banach cones $L_p^+(Y, \mu)$ (for μ nonzero, otherwise $L_p^+(Y, \mu) \cong \{0\}$) are isomorphic to $\mathbb{R}_{\geq 0}$ – therefore, $\mathbb{R}_{\geq 0}$ is a *bona fide* Banach cone.

Example 2. There exists a familiar linear map from $L_p^+(X, \mu)$ to $\mathbb{R}_{\geq 0}$, namely the Lebesgue integral $\int : L_p^+(X, \mu) \to \mathbb{R}_{\geq 0}$, taking $u \in L_p^+(X, \mu)$ to $\int_X u \, d\mu$. In this case, ω-continuity of the integral is simply the monotone convergence theorem!

Unless stated otherwise, all maps in the remainder of this section are ω-continuous. The property of ω-continuity is closed under composition and the identity function is trivially ω-continuous. This takes us to the following definition.

Definition 5 (Categories of Banach cones and of L_p^+ cones). *The category* **Ban** *has Banach cones as objects and ω-continuous linear maps as morphisms. We distinguish the full subcategory* **L** *having as objects all L_p^+-spaces (ranging over all $p \in [1, \infty]$). Further,* **L** *admits a family of full subcategories $\{\mathbf{Lp}\}_{\mathbf{p}\in[1,\infty]}$, each having as objects L_p^+ spaces (for fixed p).*

[3] These cones have the "countable sup property" [2]. Therefore, all directed sets admit a countable subset having the same least upper bound, and we can restrict our attention to chains.

Ban is itself a full subcategory of the category ω**CC** of ω-complete normed cones and ω-continuous maps, as defined in [3]. Let us denote by $\mathbf{Ban}(C, C')$ the set of ω-continuous linear maps from C to C'. Denoting $\|\cdot\|_C$ the norm of C, we recall that the operator norm of a linear map $A : C \to C'$ is given by $\|A\|_{op} = \inf\{K \geq 0 \mid \forall\, u \in C, \|Au\|_{C'} \leq K\|u\|_C\}$. A partial order on $\mathbf{Ban}(C, C')$ is given by $A \leq B$ *iff* for all $u \in C$, $A(u) \leq_{C'} B(u)$. Selinger proved in [11] that ω-continuous linear maps between ω-complete cones have automatically bounded norm (i.e. they are continuous in the usual sense), therefore we can and will abstain from asking continuity explicitly. The following result is a cone-theoretic counterpart to the well-known fact that the vector space of bounded linear operators between two Banach spaces forms a Banach space for the operator norm.

Proposition 2. *For all Banach cones C, C', the cone of ω-continuous linear maps $\mathbf{Ban}(C, C')$ is a Banach cone for the operator norm and the pointwise order.*

4.2 Duality in Banach Cones

We use a powerful Banach cone duality result initially proved in the supplementary material to [3]. We say that a pair (p, q) with $p, q \in [1, \infty]$ is *Hölder conjugate* if $\frac{1}{p} + \frac{1}{q} = 1$. For any Banach cone C, its dual C^* is by definition the Banach cone of ω-continuous linear functionals, i.e. the cone $C^* = \mathbf{Ban}(C, \mathbb{R}_{\geq 0})$. This operation defines a contravariant endofunctor $-^* : \mathbf{Ban} \to \mathbf{Ban}^{op}$ mapping each cone C to C^* and each map of cone $A : C \to C'$ to the map $A^* : C'^* \to C^*$ defined by $A^*(\varphi) = \varphi \circ A$, for $\varphi \in C'^*$. For Hölder conjugate (p, q), we have the following extension to the classical isomorphism of L_p^+ spaces.

Theorem 3 (L_p^+ cone duality [3]). *There is a Banach cone isomorphism $\varepsilon_p : L_p^+(X, \mu) \cong L_q^{+,*}(X, \mu)$.*

We won't reproduce the proof of this theorem here, which can be found in the supplementary material to [3]. Suffice it to say it is a Riesz duality type argument which relies entirely on the Radon-Nikodym theorem. Note that Theorem 3 implies in particular that $L_\infty^{+,*}(X, \mu) \cong L_1^+(X, \mu)$, which classically fails in the usual setting of L_p Banach spaces. It is instructive to study how ω-continuity wards off a classical counter-example to duality in the general Banach case.

Example 3 (Taken from [11]*).* Let μ be a probability measure on $\mathbb{N}$ with full support. We consider the cone $\ell_\infty^+ = L_\infty^+(\mathbb{N}, \mu)$ of bounded sequences of real numbers. Let $\mathcal{U}$ be a non-principal ultrafilter on $\mathbb{N}$ (i.e. an ultrafilter on the partial order of subsets of $\mathbb{N}$ without a least element). We define the function $\lim_{\mathcal{U}} : \ell_\infty^+ \to \mathbb{R}$ as $\lim_{\mathcal{U}}(\{x_n\}_{n\in\mathbb{N}}) = \sup\{y \mid \{n \mid x_n \geq y\} \in \mathcal{U}\}$. This function is linear and bounded. However, consider the chain $\{u^k \in \ell_\infty^+\}_{k\in\mathbb{N}}$ with $u_n^k = 1$ for all $n \leq k$ and $u_n^k = 0$ for all $n > k$. The supremum of this chain is the constant 1 sequence. On the other hand, we have $\lim_{\mathcal{U}}(u^k) = 0$ for all k, whereas $\lim_{\mathcal{U}}(\bigvee_k u^k) = 1$. Therefore, $\lim_{\mathcal{U}}(u^k)$ is not ω-continuous–i.e., $\lim_{\mathcal{U}} \notin \ell_\infty^{+,*}$.

It is useful to have a concrete representation of the isomorphism stated in Theorem 3. This theorem implies that for all $u \in L_p^+(X,\mu)$, there exists a unique ω-continuous linear functional $\varepsilon(u) \in L_q^{+,*}(X,\mu)$–which must therefore correspond to $\varepsilon(u)(v) = \int_X uv\, d\mu$. The pairing between L_p^+ and L_q^+ cones that we introduce below corresponds to the evaluation of such a functional against some argument.

Definition 6 (Pairing). *For Hölder conjugate (p,q), the* pairing *is the map $\langle \cdot,\cdot \rangle_X : L_p^+(X,\mu) \times L_q^+(X,\mu) \to \mathbb{R}_{\geq 0}$ defined by $\langle u, u' \rangle = \int uu'\, d\mu$.*

The pairing is bilinear, continuous and ω-continuous in each argument (consequences of the corresponding properties of the Lebesgue integral). We can now state the adjunction theorem.

4.3 Adjunctions Between Conjugate L_p^+ Cones

It is instructive to look at Theorem 3 under a slightly more general light. Observe that $L_p^+(X,\mu)$ is isomorphic to $\mathbf{Ban}(\mathbb{R}_{\geq 0}, L_p^+(X,\mu))$: indeed, any map A in this function space is entirely constrained by linearity by its value at 1. Therefore, Theorem 3 really states a Banach cone isomorphism between $\mathbf{Ban}(\mathbb{R}_{\geq 0}, L_p^+(X,\mu))$ and $\mathbf{Ban}(L_q^+(X,\mu), \mathbb{R}_{\geq 0})$. This isomorphism generalises to the case where $\mathbb{R}_{\geq 0}$ is replaced by an arbitrary conjugate pair of cones $L_p^+(Y,\mu), L_q^+(Y,\nu)$ (i.e. s.t. (p,q) are Hölder conjugate).

Theorem 4 (L_p^+/L_q^+ adjunction). *For (p,q) Hölder conjugate and for all $A : L_p^+(X,\mu) \to L_p^+(Y,\nu)$, $A^* : L_q^+(Y,\nu) \to L_q^+(X,\mu)$ is unique such that*

$$\forall\, u \in L_p^+(X,\mu), v \in L_q^+(Y,\nu),\ \langle v, A(u) \rangle_Y = \langle A^*(v), u \rangle_X. \tag{4}$$

The essence of the previous theorem is neatly captured as follows.

Corollary 1. *For all Hölder conjugate (p,q), the duality functor $-^* : \mathbf{Ban} \to \mathbf{Ban}^{op}$ restricts to an equivalence of categories $-^* : \mathbf{Lp} \to \mathbf{Lq}^{op}$.*

Figure 1 recapitulates the categories of Banach cones mentioned in this section along their relationships.

5 Pointless Bayesian Inversion

Krn arrows can be represented as linear maps between function spaces. This bridge allows one to manipulate Markov kernels both from the measure-theoretic side and from the functional-analytic side. Concretely, this linear interpretation of kernels is presented as a family of functors from **Krn** to **L**, the subcategory of **Ban** restricted to L_p^+ cones and ω-continuous linear maps. We show that pointful Bayesian inversion, whenever it is defined, coincides with adjunction.

$$\begin{tikzcd}
\mathbf{Lp} \arrow[dr, hook] \arrow[dd, bend left, "-^*"] & & \\
& \mathbf{L} \arrow[r, hook] & \mathbf{Ban} \\
\mathbf{Lq}^{op} \arrow[ur, hook] \arrow[uu, bend left, "(-^*)^{op}"] & &
\end{tikzcd}$$

Fig. 1. Categories of cones

$$\begin{tikzcd}
\mathbf{Krn}^{op} \arrow[r, "\mathsf{T}_p"] & \mathbf{AMK_p} \arrow[r] \arrow[d, bend left, "-^*"] & \mathbf{Lp} \arrow[d, bend left, "-^*"] \\
& \mathbf{MO_q}^{op} \arrow[r] \arrow[u, bend left, "(-^*)^{op}"] & \mathbf{Lq}^{op} \arrow[u, bend left]
\end{tikzcd}$$

Fig. 2. Kernels, AMKs and MOs

5.1 Representing Krn Arrows as AMKs

More precisely, kernels are associated to *abstract Markov kernels* (AMKs for short), which are a generalisation of stochastic matrices. Below, we denote by $\mathbb{1}_X$ the constant function equal to 1 on the space X. Since all measures we consider are finite, $\mathbb{1}_X \in L_p^+(X, \mu)$ for all $p \in [1, \infty]$.

Definition 7 (Abstract Markov kernels). *An* **Lp** *morphism* $A : L_p^+(Y, \nu) \to L_p^+(X, \mu)$ *is an* AMK *if* $A(\mathbb{1}_Y) = \mathbb{1}_X$ *and* $\|A\| = 1$*. Clearly, AMKs are closed under composition and the identity operator is trivially an AMK.* $\mathbf{AMK_p}$ *is the subcategory of* **Lp** *having the same objects and where morphisms are restricted to AMKs.*

Example 4. Let us look at the particular case where X and Y are finite discrete spaces and μ, ν finite measures with full support. Then $L_p^+(X, \mu) \cong \mathbb{R}^X$ and similarly for $L_p^+(Y, \nu)$. Therefore, A corresponds to an $Y \times X$ matrix. The constraint that $A(\mathbb{1}_Y) = \mathbb{1}_X$ amounts to asking that the rows of A sum to 1, i.e. that A is a stochastic matrix.

The adjoint of an AMK is in general *not* an AMK. In the finite case, this reflects the fact that the transpose of a stochastic matrix is not necessarily stochastic. Adjoints of AMKs are called *Markov operators* (MOs for short). Whereas AMKs pull back observables, an MO pushes densities forward. In the following, we make use of the fact that for $p \leq q$, any $u \in L_q^+(X, \mu)$ belongs to $L_p^+(X, \mu)$.

Definition 8 (Markov operators). *An arrow* $A : L_p^+(X, \mu) \to L_p^+(Y, \nu)$ *is an* MO *if for all* $u \in L_p^+(X, \mu)$, $\|A(u)\|_1 = \|u\|_1$ *and* $\|A\| = 1$*.* $\mathbf{MO_p}$ *is the subcategory of* **Lp** *having the same objects and where morphisms are restricted to MOs.*

Notice that we require an MO to be norm preserving for the L_1^+ norm. This is a mass preservation constraint in disguise. Adjunction maps AMKs to MOs and conversely.

Proposition 3. *The equivalence of categories* $-^* : \mathbf{Lp} \to \mathbf{Lq}^{op}$ *restricts to an equivalence of categories* $-^* : \mathbf{AMK_p} \to \mathbf{MO_q}^{op}$*.*

We now introduce a family of contravariant functors $\mathsf{T}_p : \mathbf{Krn}^{op} \to \mathbf{AMK_p}$. On objects, we set $\mathsf{T}_p(X,\mu) = L_p^+(X,\mu)$. For $f : (X,\mu) \multimap (Y,\nu)$ a **Krn** arrow, and for $v \in \mathsf{T}_p(Y,\nu) = L_p^+(Y,\nu)$, we define $\mathsf{T}_p(f)(v)(x) = \int_Y v \, df(x)$. The following theorem generalises the interpretation of kernels as stochastic matrices given in Sect. 2.

Theorem 5. T_p *is a functor from* $\mathbf{Krn}^{op}$ *to* $\mathbf{AMK_p}$*.*

The relationship between AMKs and MOs is summed up in Fig. 2. Notice that $\mathbf{AMK_p}$ and $\mathbf{MO_p}$ are subcategories of **Lp** which are not full.

5.2 Bayesian Inversion in Krn

Recall that Theorem 2 gives Bayesian inversion as a bijection

$$-^\dagger : \mathbf{Krn}(X,\mu;Y,\nu)/\mu \cong \mathbf{Krn}(Y,\nu;X,\mu)/\nu.$$

T_p is $\sim$-invariant, which allows us to apply it to $\sim$-equivalence classes of arrows.

Lemma 5. *Let* $f, f' : (X,\mu) \multimap (Y,\nu)$ *be such that* $f \sim f'$*. Then for all* $p \in [1,\infty]$, $\mathsf{T}_p(f) = \mathsf{T}_p(f')$*.*

Proof. Since $\mu(\{x \mid f(x) \neq f'(x)\}) = 0$, we have for all function $g : \mathsf{G}(Y) \to [0,\infty]$ that $\mu(\{x \mid g \circ f(x) \neq g \circ f'(x)\}) = 0$. Taking $g = ev_v(\lambda) = \int_Y v \, d\lambda$, the sought property follows. □

The following theorem states that pointful Bayesian inversion implements adjunction.

Theorem 6. *For all* **Krn** *arrow* $f : (X,\mu) \multimap (Y,\nu)$ *and all Hölder conjugate* (p,q), $\mathsf{T}_p(f^\dagger) = \mathsf{T}_q(f)^*$*.*

Proof. It is enough to prove that for all $u \in L_p^+(X,\mu), v \in L_q^+(Y,\nu)$, we have $\langle \mathsf{T}_p(f^\dagger)(u), v\rangle_Y = \langle u, \mathsf{T}_q(f)(v)\rangle_X$. We compute:

$$\begin{aligned}
\langle \mathsf{T}_p(f^\dagger)(u), v\rangle_Y &= \int_{y\in Y} v(y) \int_{x\in X} u(x)\, df^\dagger(y)\, d\nu \\
&= \int_{y\in Y} \int_{(x,-)\in X\times Y} u(x)v(y)\, d\pi_Y^\dagger(y)\, d\nu \quad (*) \\
&= \int_{(x,y)\in X\times Y} u(x)v(y)\, dI_{X,\mu}^{Y,\nu}(f) \\
&= \int_{x\in X} \int_{(-,y)\in X\times Y} u(x)v(y)\, d\pi_X^\dagger(x)\, d\mu \\
&= \int_{x\in X} u(x) \int_{y\in Y} v(y)\, df(x)\, d\mu \qquad (*) \\
&= \langle u, \mathsf{T}_q(f)(v)\rangle_X
\end{aligned}$$

This string of equations follows from the definition of $-^\dagger$ (Theorem 2). At the equations marked $(*)$ we used the characteristic property of disintegrations to move u (resp. v) in (resp. out of) the integral (see Theorem 1). □

This proves that Bayesian inversion is really just adjunction. However, performing Bayesian inversion in **Krn** relies on standard Borel assumptions, while adjunction does not! Most importantly, Bayesian inversion in $\omega\mathbf{CC}$ is better structured, as it corresponds to a categorical duality.

6 Pointless Bayesian Inversion Through Couplings

Under standard Borel assumptions, Bayesian models can be given equivalently either in terms of **Krn** arrows or more classically in terms of *joint probabilities* (i.e. couplings). The latter appears crucially in the definition of the pointful inverse, as demonstrated in Theorem 2. However, pointless Bayesian inversion seems *prima facie* to do away with these objects entirely. We conclude this work by shedding some light on the status of couplings w.r.t. pointless inversion: we show that the bijection $I_{X,\mu}^{Y,\nu} : \mathbf{Krn}(X,\mu;Y,\nu)/\mu \to \mathbf{\Gamma}(X,\mu;Y,\nu)$ defined in Sect. 3.3 generalises, for X, Y arbitrary measurable spaces, to a bijection from the Homset $\mathbf{MO}_\infty(L_\infty^+(X,\mu); L_\infty^+(Y,\nu))$ to the set of couplings $\mathbf{\Gamma}(X,\mu;Y,\nu)$. In this setting, we prove that Bayesian inversion amounts to permuting the coordinates of the coupling. Our first ingredient is a map from couplings to ω-continuous linear operators. The key observation is the following.

Proposition 4. *For all $p \in [1,\infty]$, any coupling $\gamma \in \mathbf{\Gamma}(X,\mu;Y,\nu)$ induces an ω-continuous linear operator $\mathsf{K}_p(\gamma) \in \mathbf{AMK_p}(X,\mu;Y,\nu)$ defined for $u \in L_p^+(X,\mu)$ and $v \in L_q^+(Y,\nu)$ (using $\varepsilon_p : L_p^+(Y,\nu) \cong L_q^{+,*}(Y,\nu)$ for (p,q) Hölder conjugate) as $\mathsf{K}_p(\gamma)(u)(v) = \int_{(x,y)\in X\times Y} u(x)v(y)\, d\gamma$.*

Dually to Proposition 4, any MO gives rise to a probability measure (but not necessarily a coupling!). For $A : L_p^+(X,\mu) \to L_p^+(Y,\nu)$ and $B_X \times B_Y$ a basic measurable rectangle in $X \times Y$, we define:

$$\mathsf{C}_p(A)(B_X \times B_Y) = \int_Y \mathbb{1}_{B_Y} A(\mathbb{1}_{B_X})\, d\nu. \tag{5}$$

Lemma 6. *For all MO $A : L_p^+(X,\mu) \to L_p^+(Y,\nu)$, $\mathsf{C}_p(A) \in \mathsf{G}(X \times Y)$.*

It is not obvious what a necessary and sufficient condition should be for $\mathsf{C}_p(A)$ to give rise to a coupling. However, we have the following reasonable sufficient condition.

Proposition 5. *For all MO $A : L_\infty^+(X,\mu) \to L_\infty^+(Y,\nu)$, $\mathsf{C}_\infty(A) \in \mathbf{\Gamma}(X,\mu;Y,\nu)$.*

C and K are the counterparts of respectively I and D in Sect. 3.3, with kernels replaced by respectively MOs and AMKs. However, no quotient is needed to obtain the following result, which states that pointless Bayesian inversion (i.e. adjunction) coincides in the world of couplings to the operation which permutes the coordinates (namely the isomorphism $\mathsf{G}(\sigma) : \mathsf{G}(X \times Y) \to \mathsf{G}(Y \times X)$).

Proposition 6. *For all MO $A : L_\infty^+(X,\mu) \to L_\infty^+(Y,\nu)$, $A^* = \mathsf{K}_1 \circ \mathsf{G}(\sigma) \circ \mathsf{C}_\infty(A)$.*

In order to close the circle, we prove that couplings are indeed in bijections with $\mathbf{MO}_\infty$ arrows (and hence, by duality, $\mathbf{AMK}_1$ arrows).

Theorem 7. *For all X, Y measurable and $\mu \in \mathsf{G}(X), \nu \in \mathsf{G}(Y)$, C_∞ defines a bijection $\mathbf{MO}_\infty(L_\infty^+(X,\mu); L_\infty^+(Y,\nu)) \cong \mathbf{\Gamma}(X,\mu;Y,\nu)$.*

The correspondence between adjunction and the permutation of coupling coordinates together with this last result show that couplings are really at the heart of the semantics of Bayesian inversion.

7 Conclusion

Pointless Bayesian inversion has several qualities which its pointful counterpart lacks: it does not rely on Polish assumptions on the underlying space, it is better typed (as it boils down to an equivalence of categories between abstract Markov kernels and Markov operators) and it admits a trivial and elegant computational interpretation in terms of couplings (as well as the structure of a self-duality on the category of couplings sketched above).

This pointless categorical approach to Bayesian inversion opens the way for exciting new research. First, one yearns to reinterpret previous constructions performed in a kernel-centric way in this new light, such as [12]. Also, the connection between our categories of operators and couplings hints at connections with the Kantorovich distance [13]. For instance, one could study issues of convergence of learning using the weak topology on the space of couplings, which suggests possibly fruitful connections with information geometry.

But chiefly, our more structured framework allows one to reason on the interactions between the approximation of Markov processes by averaging [3] and Bayesian inversion. For instance, we can now ask whether some properties of the Bayesian learning procedure are *profinite*, i.e. entirely characterised by considering the finite approximants (one thinks of issues of convergence of learning, for instance). More generally, we posit that pointless inversion is the right tool to investigate *approximate learning.*

References

1. Ackerman, N.L., Freer, C.E., Roy, D.M.: Noncomputable conditional distributions. In: Proceedings of the 26th Annual IEEE Symposium on Logic in Computer Science, LICS, Toronto, Ontario, Canada, pp. 107–116, 21–24 June 2011
2. Aliprantis, C., Border, K.: Infinite Dimensional Analysis, vol. 32006. Springer, Heidelberg (1999)
3. Chaput, P., Danos, V., Panangaden, P., Plotkin, G.: Approximating markov processes by averaging. J. ACM 61(1), 45 pages (2014)
4. Clerc, F., Dahlqvist, F., Danos, V., Garnier, I.: Pointless learning (long version) (2017)
5. Culbertson, J., Sturtz, K.: A categorical foundation for Bayesian probability. Appl. Categorical Struct. **22**(4), 647–662 (2012)

6. Dahlqvist, F., Danos, V., Garnier, I., Kammar, O.: Bayesian inversion by omega-complete cone duality (invited paper). In: Desharnais, J., Jagadeesan, R. (eds.) 27th International Conference on Concurrency Theory (CONCUR 2016), vol. 59 of Leibniz International Proceedings in Informatics (LIPIcs), pp. 1:1–1:15. Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik, Dagstuhl, Germany (2016)
7. Giry, M.: A categorical approach to probability theory. In: Banaschewski, B. (ed.) Categorical Aspects of Topology and Analysis. LNM, vol. 915, pp. 68–85. Springer, Heidelberg (1982). doi:10.1007/BFb0092872
8. Kallenberg, O.: Foundations of Modern Probability. Springer, New York (1997)
9. Kechris, A.S.: Classical Descriptive Set Theory. Graduate Text in Mathematics, vol. 156. Springer, New York (1995)
10. Kozen, D.: A probabilistic PDL. In: Proceedings of the Fifteenth Annual ACM Symposium on Theory of Computing, STOC 1983, pp. 291–297. ACM, New York (1983)
11. Selinger, P.: Towards a semantics for higher-order quantum computation. In: Proceedings of the 2nd International Workshop on Quantum Programming Languages, TUCS General Publication, vol. 33, pp. 127–143 (2004)
12. Staton, S., Yang, H., Heunen, C., Kammar, O., Wood, F.: Semantics for probabilistic programming: higher-order functions, continuous distributions, and soft constraints. CoRR, abs/1601.04943 (2016)
13. Villani, C.: Optimal Transport, Old and New. Grundlehren der mathematischen Wissenschaften. Springer, Heidelberg (2006)