10

Bayesian Nonparametric Learning: Expressive Priors for Intelligent Systems

Michael I. Jordan

1 Introduction

One of the milestones in the development of artificial intelligence (AI) is the embrace of uncertainty and inductive reasoning as primary concerns of the field. This embrace has been a surprisingly slow process, perhaps because the naive interpretation of "uncertain" seems to convey an image that is the opposite of "intelligent." That the field has matured beyond this naive opposition is one of the singular achievements of Judea Pearl. While the pre-Pearl AI researcher tended to focus on mimicking the deductive capabilities of human intelligence, a post-Pearl researcher has been sensitized to the inevitable uncertainty that intelligent systems face in any realistic environment, and the need to explicitly represent that uncertainty so as to be able to mitigate its effects. Not only does this embrace of uncertainty accord more fully with the human condition, but it also recognizes that the first artificially intelligent systems—necessarily limited in their cognitive capabilities—will be if anything *more* uncertain regarding their environments than us humans. It is only by embracing uncertainty that a bridge can be built from systems of limited intelligence to those having robust human-level intelligence.

A computational perspective on uncertainty has two aspects: the explicit representation of uncertainty and the algorithmic manipulation of this representation so as to transform and (often) to reduce uncertainty. In his seminal 1988 book, *Probabilistic Reasoning in Intelligent Systems*, Pearl showed that these aspects are intimately related. In particular, obtaining a compact representation of uncertainty has important computational consequences, leading to efficient algorithms for marginalization and conditioning. Moreover, marginalization and conditioning are the core inductive operations that tend to reduce uncertainty. Thus, by developing an effective theory of the representation of uncertainty, Pearl was able to also develop an effective computational approach to probabilistic reasoning.

Uncertainty about an environment can also be reduced by simply observing that environment; i.e., by learning from data. Indeed, another response to the early focus on deduction in AI has been to emphasize learning as a pathway to the development of intelligent systems. In the 1980's, concurrently with Pearl's work on probabilistic expert systems, this perspective was taken up in earnest, building on an earlier tradition in pattern recognition (which itself built on even earlier traditions in statistics). The underlying inductive principle was essentially the law of large numbers, a principle of probability theory which states that the statistical aggregation of independent, identically distributed samples yields a decrease of uncertainty that goes (roughly speaking) at a rate inversely proportional to the square root of the number of samples. The question has been how to perform this "aggregation," and the learning field has been avidly empirical, exploring a variety of computational architectures, including extremely simple representations (e.g., nearest neighbor), ideas borrowed from deductive traditions (e.g., decision trees), ideas closely related to classical statistical models (e.g., boosting and the support vector machine), and architectures motivated at least in part by complex biological and physical systems (e.g., neural networks). Several of these architectures have factorized or graphical representations, and numerous connections to graphical models have been made.

A narrow reader of Pearl's book might wish to argue that learning is not distinct from the perspective on reasoning presented in that book; in particular, observing the environment is simply a form of conditioning. This perspective on learning is indeed reasonable if we assume that a learner maintains an explicit probabilistic model of the environment; in that case, making an observation merely involves instantiating some variable in the model. However, many learning researchers do not wish to make the assumption that the learner maintains an explicit probabilistic model of the environment, and many algorithms developed in the learning field involve some sort of algorithmic procedure that is not necessarily interpretable as computing a conditional probability. These procedures are instead justified in terms of their unconditional performance when used again and again on various data sets.

Here we are of course touching on the distinction between the Bayesian and the frequentist approaches to statistical inference. While this is not the place to develop that distinction in detail, it is worth noting that statistics—the field concerned with the theory and practice of inference—involves the interplay of the conditional (Bayesian) and the unconditional (frequentist) perspectives and this interplay also underlies many developments in AI research. Indeed, the trend since Pearl's work in the 1980's has been to blend reasoning and learning: put simply, one does not need to learn (from data) what one can infer (from the current model). Moreover, one does not need to infer what one can learn (intractable inferential procedures can be circumvented by collecting data). Thus learning (whether conditional or not) and reasoning interact. The most difficult problems in AI are currently being approached with methods that blend reasoning with learning. While the extremes of classical expert systems and classical tabula rasa learning are still present and still have their value in specialized situations, they are not the centerpieces of the field. Moreover, the caricatures of probabilistic reasoning and statistical inference that fed earlier ill-informed debates in AI have largely vanished. For this we owe much to Judea Pearl.

There remain, however, a number of limitations—both perceived and real—of probabilistic and statistical approaches to AI. In this essay, I wish to focus on some of these limitations and provide some suggestions as to the way forward.

It is both a perception and reality that to use probabilistic methods in AI one is generally forced to write down long lists of assumptions. This is often a helpful exercise, in that it focuses a designer to bring hidden assumptions to the foreground. Moreover, these assumptions are often qualitative in nature, with the quantitative details coming from elicitation methods (i.e., from domain experts) and learning methods. Nonetheless, the assumptions are not always well motivated. In particular, independence assumptions are often imposed for reasons of computational convenience, not because they are viewed as being true, and the effect on inference is not necessarily clear. More subtly, and thus of particular concern, is the fact that the tail behavior of probability distributions is often not easy to obtain (from elicitation or from data), and choices of convenience are often made.

A related issue is that probabilistic methods are often not viewed as sufficiently expressive. One common response to this issue has involved trying to bring ideas from first-order logic to bear on probabilistic modeling. This line of work has, however, mainly involved using logical representations as a high-level interface for model specification and then compiling these representations down to flat probabilistic representations for inference. It is not yet clear how to bring together the powerful inferential methods of logic and probability into an effective computational architecture.

In the current paper, we will pursue a different approach to expressive probabilistic representation and to a less assumption-laden approach to inference. The idea is to move beyond the simple fixed-dimensional random variables that have been generally used in graphical models (multinomials, Gaussians and other exponential family distributions) and to consider a wider range of probabilistic representations. We are motivated by the ubiquity of flexible data structures in computer science—the field is based heavily on objects such as trees, lists and collections of sets that are able to expand and contract as needed. Moreover, these data structures are often associated with combinatorial and algebraic identities that lead to efficient algorithms. We would like to mimic this flexibility within the world of probabilistic representations.

In fact, the existing field of *stochastic processes* provides essentially this kind of flexibility. Recall that a stochastic process is an indexed collection of random variables, where the index set can be infinite (countably infinite or uncountably infinite) [Karlin and Taylor 1975]. Within the general theory of stochastic processes it is quite natural to define probability distributions on objects such trees, lists and collections of sets. It is also possible to define probability distributions on spaces of probability distributions, yielding an appealing recursivity. Moreover, many stochastic processes have interesting ties to combinatorics (and to other areas of mathematics concerned with compact structure, such as algebra). Probability theorists have spent many decades developing these ties and a rich literature on "combinatorial stochastic processes" has emerged [Pitman 2002]. It is natural to take this literature as a point of departure for the development of expressive data structures for computationally efficient reasoning and learning.

One general way to use stochastic processes in inference is to take a Bayesian perspective and replace the parametric distributions used as priors in classical Bayesian analysis with stochastic processes. Thus, for example, we could consider a model in which the prior distribution is a stochastic process that ranges over trees of arbitrary depth and branching factor. Combining this prior with a likelihood, we obtain a posterior distribution that is also a stochastic process that ranges over trees of arbitrary depth and branching factor. Bayesian learning amounts to updating one flexible representation (the *prior stochastic process*) into another flexible representation (the *posterior stochastic process*).

This idea is not new, indeed it is the core idea in an area of research known as *Bayesian nonparametrics*, and there is a small but growing community of researchers who work in the area. The word "nonparametrics" needs a bit of explanation. The word does not mean "no parameters"; indeed, many stochastic processes can be usefully viewed in terms of parameters (often, infinite collections of parameters). Rather, it means "not parametric," in the sense that Bayesian nonparametric inference is not restricted to objects whose dimensionality stays fixed as more data is observed. The spirit of Bayesian nonparametrics is that of flexible data structures—representations can grow as needed. Moreover, stochastic processes yield a much broader class of probability distributions than the class of exponential family distributions that is the focus of the graphical model literature. In this sense, Bayesian nonparametric learning is less assumption-laden than classical Bayesian parametric learning.

In this paper we offer an invitation to Bayesian nonparametrics. Our presentation is meant to evoke Pearl's presentation of Bayesian networks in that our focus is on foundational representational issues. As in the case of graphical models, if the representational issues are handled well, then there are favorable algorithmic consequences. Indeed, the parallel is quite strong—in the case of graphical models, these algorithmic consequences are combinatorial in nature (they involve the combinatorics of sums and products), and in the case of Bayesian nonparametrics favorable algorithmic consequences also arise from the combinatorial properties of certain stochastic process priors.

2 De Finetti's theorem and the foundations of Bayesian inference

A natural point of departure for our discussion is a classical theorem due to Bruno De Finetti that is one of the pillars of Bayesian inference. This core result not only suggests the need for prior distributions in statistical models but it also leads directly to the consideration of stochastic processes as Bayesian priors.

Consider an infinite sequence of random variables, $(X_1, X_2, \ldots)$. To simplify our discussion somewhat, let us assume that these random variables are discrete. We say that such a sequence is *infinitely exchangeable* if the joint probability distribution of any finite subset of those random variables is invariant to permutation. That is, for any N, we have $p(x_1, x_2, \ldots, x_N) = p(x_{\pi(1)}, x_{\pi(2)}, \ldots, x_{\pi(N)})$, where π is a permutation and p is a probability mass function. De Finetti's theorem states that $(X_1, X_2, \ldots)$ are infinitely exchangeable if and only the joint probability distribution of any finite subset can be written as a marginal probability in the following way:

$$p(x_1, x_2, \ldots, x_N) = \int \prod_{i=1}^{N} p(x_i \,|\, G) P(dG). \tag{1}$$

In one direction this theorem is straightforward: If the joint distribution can be written as an integral in this way, then we clearly have invariance to permutation (because the product is invariant to permutation). It is the other direction that is non-trivial. It states that for exchangeable random variables, there necessarily exists an underlying random element G, and a probability distribution P, such that the random variables X_i are conditionally independent given G, and such that their joint distribution is obtained by integrating over the distribution P. If we view G as a "parameter," then this theorem can be interpreted as stating that exchangeability implies the existence of an underlying parameter and a prior distribution on that parameter. As such, De Finetti's theorem is often viewed as providing foundational support for the Bayesian paradigm.

We placed "parameter" in quotes in the preceding paragraph because there is no restriction that G should be a finite-dimensional object. Indeed, the full import of De Finetti's theorem is clear when we realize that in many instances G is in fact an infinite-dimensional object, and P defines a stochastic process.

Let us give a simple example. The *Pólya urn model* is a simple probability model for sequentially labeling the balls in an urn. Consider an empty urn and a countably infinite collection of colors. Pick a color at random according to some fixed distribution G_0 and place a ball having that color in the urn. For all subsequent balls, either choose a ball from the urn (uniformly at random) and return that ball to the urn with another ball of the same color, or choose a new color from G_0 and place a ball of that color in the urn. Mathematically, we have:

$$p(X_i = k \,|\, x_1, \ldots x_{i-1}) \propto \begin{cases} n_k & \text{if } x_j = k \text{ for some } j \in \{1, \ldots, i-1\} \\ \alpha_0 & \text{otherwise,} \end{cases} \tag{2}$$

where $\alpha_0 > 0$ is a parameter of the process.

It turns out that the Pólya urn model is exchangeable. That is, even though we defined the model by picking a particular ordering of the balls, the resulting distribution is independent of the order. This is proved by writing the joint distribution $p(x_1, x_2, \ldots, x_N)$ as a product of conditionals of the form in Eq. (2) and noting (after some manipulation) that the resulting expression is independent of order.

While the Pólya urn model defines a distribution on labels, it can also be used to induce a distribution on partitions. This is achieved by simply partitioning the balls into groups that have the same color. This distribution on partitions is known as the *Chinese restaurant process* [Aldous 1985]. As we discuss in more detail in Section 4, the Chinese restaurant process and the Pólya urn model can be used as the basis of a Bayesian nonparametric model of clustering where the random partition provides a prior on clusterings and the color associated with a given cell can be viewed as a parameter vector for a distribution associated with a given cluster.

The exchangeability of the Pólya urn model implies—by De Finetti's theorem—the existence of an underlying random element G that renders the ball colors conditionally independent. This random element is not a classical fixed-dimension random variable; rather, it is a stochastic process known as the *Dirichlet process.* In the following section we provide a brief introduction to the Dirichlet process.

3 The Dirichlet process

In thinking about how to place random distributions on infinite objects, it is natural to begin with the special case of the positive integers. A distribution $\pi = (\pi_1, \pi_2, \ldots)$ on the integers can be viewed as a sequence of nonnegative numbers that sum to one. How can we obtain *random* sequences that sum to one?

One solution to this problem is provided by a procedure known as "stick-breaking." Define an infinite sequence of independent random variables as follows:

$$\beta_k \sim \text{Beta}(1, \alpha_0) \qquad\qquad k = 1, 2, \ldots, \tag{3}$$

where $\alpha_0 > 0$ is a parameter. Now define an infinite random sequence as follows:

$$\pi_1 = \beta_1, \qquad\qquad \pi_k = \beta_k \prod_{l=1}^{k-1} (1 - \beta_l) \qquad\qquad k = 2, 3, \ldots. \tag{4}$$

It is not difficult to show that $\sum_{k=1}^{\infty} \pi_k = 1$ (with probability one).

We can exploit this construction to generate a large class of random distributions on sets other than the integers. Consider an arbitrary measurable space Ω and let G_0 be a probability distribution on Ω. Draw an infinite sequence of points $\{\phi_k\}$ independently from G_0. Now define:

$$G = \sum_{k=1}^{\infty} \pi_k \delta_{\phi_k}, \tag{5}$$

where δ_{ϕ_k} is a unit mass at the point ϕ_k. Clearly G is a measure. Indeed, for any measurable subset B of Ω, $G(A)$ just adds up the values π_k for those k such that $\phi_k \in B$, and this process satisfies the countable additivity needed in the definition of a measure. Moreover, G is a probability measure, because $G(\Omega) = 1$.

Note that G is random in two ways—the weights π_k are obtained by a random process, and the locations ϕ_k are also obtained by a random process. While it seems clear that such an object is not a classical finite-dimensional random variable, in what sense is G is a stochastic process; i.e., an indexed collection of random variables? The answer is that G is a stochastic process where the indexing variables are the measurable subsets of Ω. Indeed, for any fixed $A \subseteq \Omega$, $G(A)$ is a random variable. Moreover (and this is not an obvious fact), ranging over sets of subsets, $\{A_1, A_2, \ldots, A_K\}$, the joint distributions on the collections of random variables $\{G(A_i)\}$ are consistent with each other. This shows, via an argument in the spirit of the Kolmogorov theorem, that G is a stochastic process. A more concrete understanding of this fact can be obtained by specializing to sets $\{A_1, A_2, \ldots, A_K\}$ that form a partition of Ω. In this case, the random vector $(G(A_1), G(A_2), \ldots, G(A_K))$ can be shown to have a classical finite-dimensional Dirichlet distribution:

$$(G(A_1), \ldots, G(A_K)) \sim \mathrm{Dir}(\alpha_0 G_0(A_1), \ldots, \alpha_0 G_0(A_K)), \tag{6}$$

from which the needed consistency properties follow immediately from classical properties of the Dirichlet distribution. For this reason, the stochastic process defined by Eq. (5) is known as a *Dirichlet process*. Eq. (6) can be summarized as saying that a Dirichlet process has Dirichlet marginals.

Having defined a stochastic process G, we can now turn De Finetti's theorem around and ask what distribution is induced on a sequence $(X_1, X_2, \ldots, X_N)$ if we draw these variables independently from G and then integrate out G. The answer: the Pólya urn. We say that the Dirichlet process is the De Finetti mixing distribution underlying the Pólya urn.

In the remainder of this chapter, we denote the stochastic process defined by Eq. (5) as follows:

$$G \sim \mathcal{DP}(\alpha_0, G_0). \tag{7}$$

The Dirichlet process has two parameters, a *concentration parameter* α_0, which is proportional to the probability of obtaining a new color in the Pólya urn, and the *base measure* G_0, which is the source of the "atoms" ϕ_k.

The set of ideas introduced in this section emerged slowly over several decades. The basic definition of the Dirichlet process as a stochastic process is due to Ferguson [1973], based on earlier work by Freedman [1963]. The fact that the Dirichlet process is the De Finetti mixing distribution underlying the Pólya urn model is due to Blackwell and MacQueen [1973]. The stick-breaking construction of the Dirichlet process was presented by Sethuraman [1994]. The application of these ideas to Bayesian modeling and inference required some additional work as described in the following section.

The Dirichlet process and the stick-breaking process are essential tools in Bayesian nonparametrics. It is as important for a Bayesian nonparametrician to master them as it is for a graphical modeler to master Pearl's book. See Hjort et al. [2010] for a book-length treatment of the Dirichlet process and related ideas.

4 Dirichlet process mixtures

With an interesting class of stochastic process priors in hand, let us now describe an application of these priors to a Bayesian nonparametric modeling problem. In particular, as alluded to in the previous section, the Dirichlet process defines a prior on partitions of objects, and this prior can be used to develop a Bayesian nonparametric approach to clustering. A notable aspect of this approach is that one does not have to fix the number of clusters a priori.

Let $(X_1, X_2, \ldots, X_N)$ be a sequence of random vectors, whose realizations we want to model in terms of an underlying set of clusters. We treat these variables as exchangeable (i.e., as embedded in an infinitely-exchangeable sequence) and, as suggested by De Finetti's theorem, treat these variables as conditionally independent given an underlying random element G. In particular, letting G be a draw from a Dirichlet process, we define a *Dirichlet process mixture model* (DP-MM) [Antoniak 1974; Lo 1984] as follows:

$$\begin{aligned} G &\sim \mathcal{DP}(\alpha_0, G_0) \\ \theta_i \,|\, G &\sim G, \qquad i = 1, \ldots, N \\ x_i \,|\, \theta_i &\sim p(x_i \,|\, \theta_i), \qquad i = 1, \ldots, N, \end{aligned}$$

where $p(x_i \,|\, \theta_i)$ is a cluster-specific distribution (e.g., a Gaussian distribution, where θ_i is a mean vector and covariance matrix). This probabilistic specification is indeed directly related to De Finetti's theorem—the use of the intermediate variable θ_i is simply an expanded way to write the factor $p(x_i \,|\, G)$ in Eq. (1). In particular, G is a sum across atoms, and thus θ_i is simply one of the atoms in G, chosen with probability equal to the weight associated with that atom.

We provide a graphical model representation of the DP-MM in Figure 1. As this figure suggests, it is entirely possible to use the graphical model formalism to display Bayesian nonparametric models. Nodes in such a graph are associated with general random elements, and the distributions on these random elements can be general stochastic processes. By going to stochastic process priors we have not strayed beyond probability theory, and all of the conditional independence semantics of graphical models continue to apply.

5 Inference for Dirichlet process mixtures

Inference with stochastic processes is an entire topic of its own, and we limit ourselves here to a brief description of one particular Markov chain Monte Carlo (MCMC) inference procedure for the DP-MM. This particular procedure is due to Escobar [1994], and its virtue is simplicity of exposition, but it should not be viewed as the state of the art. See Neal [2000] for a discussion of a variety of other MCMC inference procedures for DP-MMs.

We begin by noting that the specification in Eq. (8) induces a Pólya urn marginal distribution on $\theta = (\theta_1, \theta_2, \ldots, \theta_N)$. The joint distribution of θ and $X = (X_1, X_2, \ldots, X_N)$

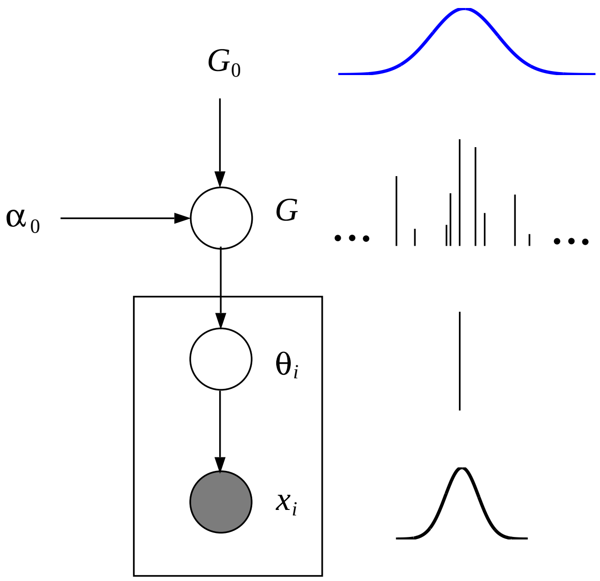

Figure 1. A graphical model representation of the Dirichlet process mixture model. Recall that the plate representation means that the parameters θ_i are drawn independently conditional on G. On the right side of the figure we have depicted specific instantiations of the random elements G and θ_i and the distribution of the observation x_i.

is thus the following product:

$$p(\theta, x) = p(\theta_1, \theta_2, \ldots, \theta_N) \prod_{i=1}^{N} p(x_i \,|\, \theta_i), \tag{8}$$

where the first factor is the Pólya urn model. This can be viewed as a product of a prior (the first factor) and a likelihood (the remaining factors).

The variable x is held fixed in inference (it is the observed data) and the goal is to sample θ. We develop a Gibbs sampler for this purpose. The main problem is to sample a particular component θ_i while holding all of the other components fixed. It is here that the property of exchangeability is essential. Because the joint probability of $(\theta_1, \ldots, \theta_N)$ is invariant to permutation, we can permute the vector to move θ_i to the end of the list. But the prior probability of the last component given all of the preceding variables is given by the urn model specification in Eq. (2). We multiply each of the distributions in this expression by the likelihood $p(x_i \,|\, \theta)$ and integrate with respect to θ. (We are assuming that G_0 and the likelihood are conjugate that this integral can be done in closed form.) The result is the conditional distribution of θ_i given the other components and given x_i. This conditional is sampled to yield the updated value of θ_i. This is done for all of the indices $i \in \{1, \ldots, N\}$ and the process iterates.

This link between exchangeability and an efficient inference algorithm is an important one. In other more complex Bayesian nonparametric models, while we may no longer assume exchangeability, we generally aim to maintain some weaker notion (e.g., partial exchangeability) so as to have some hope of tractable inference.

6 Hierarchical Dirichlet processes

The spirit of the graphical model formalism—in particular the Bayesian network formalism based on directed graphs—is that of hierarchical Bayesian modeling. In a hierarchical Bayesian model, the joint distribution of all of the variables in the model is obtained as a product over conditional distributions, where each conditional may depend on other variables in the model. While the graphical model literature has focused almost exclusively on parametric hierarchies—where each of the conditionals is a finite-dimensional distribution—it is also possible to build hierarchies in which the components are stochastic processes. In this section we consider how to do this for the Dirichlet process.

One of the simplest and most useful ways in which hierarchies arise in Bayesian models is in the form of a conditional independence motif in which a set of variables, $(\theta_1, \theta_2, \ldots, \theta_m)$, are coupled via an underlying variable θ_0. For example, θ_i might be a Gaussian variable whose mean is equal to θ_0, which is also Gaussian; moreover, the θ_i are conditionally independent given θ_0. The inferential effect of this construction is to "shrink" the posterior distributions of θ_i towards each other. This is often a desirable effect, particularly when m is large relative to the number of observed data points.

The same tying of distributions can be done with Dirichlet processes. Recall that a Dirichlet process, $G_i \sim \mathcal{DP}(\alpha_0, G_0)$, is a random measure G_i that has a "parameter" G_0 that is itself a measure. If we treat G_0 as itself a draw from a Dirichlet process, and let the measures $\{G_1, G_2, \ldots, G_m\}$ be conditionally independent given G_0, we obtain the following hierarchy:

$$\begin{aligned} G_0 \,|\, \gamma, H &\sim \mathcal{DP}(\gamma, H) \\ G_i \,|\, \alpha, G_0 &\sim \mathcal{DP}(\alpha_0, G_0) \qquad i = 1, \ldots, m, \end{aligned}$$

where γ and H are concentration and base measure parameters at the top of the hierarchy. This construction—which is known as a *hierarchical Dirichlet process* (HDP)—yields an interesting kind of "shrinkage." Recall that G_0 is a discrete random measure, with its support on a countably infinite set of atoms. Drawing $G_i \sim \mathcal{DP}(\alpha_0, G_0)$ means that G_i will also have its support on the same set of atoms, and this will be true for each of $\{G_1, G_2, \ldots, G_m\}$. Thus these measures will share atoms. They will differ in the weights assigned to these atoms. The weights are obtained via conditionally independent stick-breaking processes.

One application of this sharing of atoms is to share mixture components across multiple clustering problems. Consider in particular a problem in which we have

m groups of data, $\{(x_{11}, x_{12}, \ldots, x_{1N_1}), \ldots, (x_{m1}, x_{m2}, \ldots x_{mN_m})\}$, where we wish to cluster the points $\{x_{ij}\}$ in the ith group. Suppose, moreover, that we view the groups as related, and we think that clusters discovered in one group might also be useful in other groups. To achieve this, we define the following *hierarchical Dirichlet process mixture model* (HDP-MM):

$$\begin{aligned}
G_0 \,|\, \gamma, H &\sim \mathcal{DP}(\gamma, H) \\
G_i \,|\, \alpha, G_0 &\sim \mathcal{DP}(\alpha_0, G_0) \qquad i = 1, \ldots, m, \\
\theta_{ij} \,|\, G_i &\sim G_i \qquad j = 1, \ldots, N_i, \\
x_{ij} \,|\, \theta_{ij} &\sim F(x_{ij}, \theta_{ij}) \qquad j = 1, \ldots, N_i.
\end{aligned}$$

This model is shown in graphical form in Figure 2. To see how the model achieves our goal of sharing clusters across groups, recall that the Dirichlet process clusters points within a single group by assigning the same parameter vector to those points. That is, if $\theta_{ij} = \theta_{ij'}$, the points x_{ij} and $x_{ij'}$ are viewed as belonging to the same cluster. This equality of parameter vectors is possible because both θ_{ij} and $\theta_{ij'}$ are drawn from G_i, and G_i is a discrete measure. Now if G_i and $G_{i'}$ share atoms, as they do in the HDP-MM, then points in different groups can be assigned to the same cluster. Thus we can share clusters across groups.

The HDP was introduced by Teh, Jordan, Beal and Blei [2006] and it has since appeared as a building block in a variety of applications. One application is to the class of models known as *grade of membership models* [Erosheva 2003], an instance of which is the *latent Dirichlet allocation* (LDA) model [Blei, Ng, and Jordan 2003]. In these models, each entity is associated not with a single cluster but with a set of clusters (in LDA terminology, each "document" is associated with a set of "topics"). To obtain a Bayesian nonparametric version of these models, the DP does not suffice; rather, the HDP is required. In particular, the topics for the ith document are drawn from a random measure G_i, and the random measures G_i are drawn from a DP with a random base measure G_0; this allows the same topics to appear in multiple documents.

Another application is to the hidden Markov model (HMM) where the number of states is unknown a priori. At the core of the HMM is the transition matrix, each row of which contains the conditional probabilities of transitioning to the "next state" given the "current state." Viewing states as clusters, we obtain a set of clustering problems, one for each row of the transition matrix. Using a DP for each row, we obtain a model in which the number of next states is open-ended. Using an HDP to couple these DPs, the same pool of next states is available from each of the current states. The resulting model is known as the *HDP-HMM* [Teh, Jordan, Beal, and Blei 2006]. Marginalizing out the HDP component of this model yields an urn model that is known as the *infinite HMM* [Beal, Ghahramani, and Rasmussen 2002].

Similarly, it is also possible to use the HDP to define an architecture known as

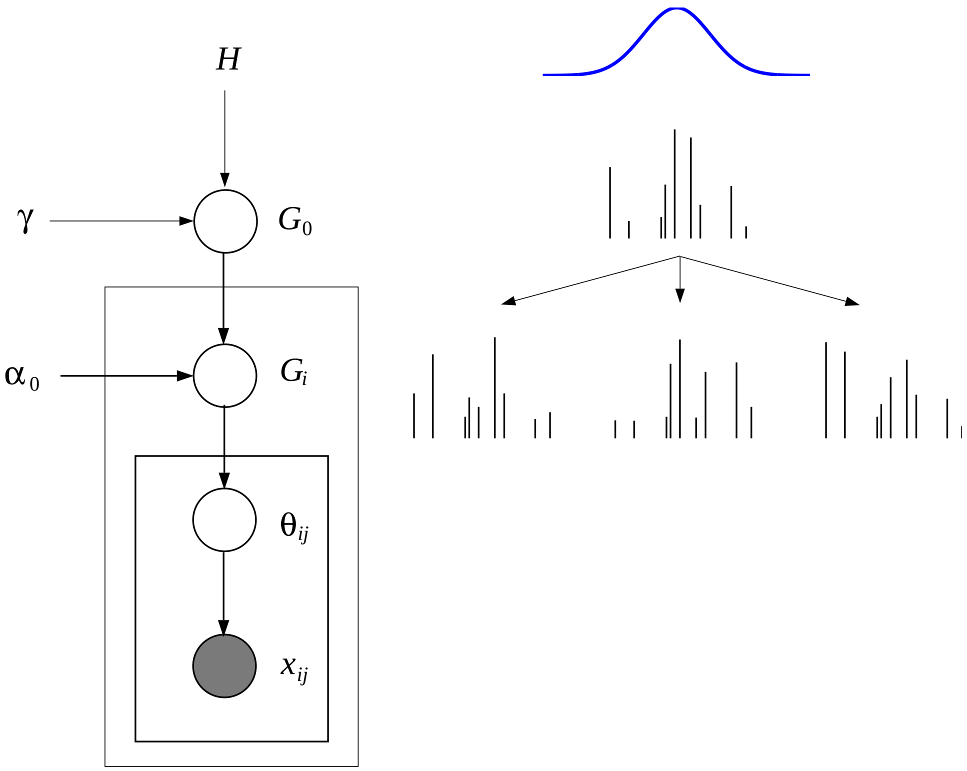

Figure 2. A graphical model representation of the hierarchical Dirichlet process mixture model. The nested plate representation means that G_0 is first drawn and held fixed, then the random measures $\{G_i\}$ are drawn independently (conditional on G_0), and finally the parameters $\{\theta_{ij}\}$ are drawn independently (conditional on G_i). On the right side of the figure we have depicted draws from G_0 and the $\{G_i\}$. Note that the atoms in these measures are at the same locations; only the weights associated with the atoms differ.

the *HDP hidden Markov tree* (HDP-HMT), a Markovian tree in which the number of states at each node in the tree is unknown a priori and the state space is shared across the nodes. The HDP-HMT has been shown to be useful in image denoising and scene recognition problems [Kivinen, Sudderth, and Jordan 2007].

Let us also mention that the HDP can be also used to develop a Bayesian nonparametric approach to probabilistic context free grammars. In particular, the HDP-PCFG of Liang, Jordan and Klein [2010] involves an HDP-based lexicalized grammar in which the number of nonterminal symbols is open-ended and inferred from data (see also Finkel, Grenager and Manning [2007] and Johnson, Griffiths and Goldwater [2007]). When a new nonterminal symbol is created at some location in a parse tree, the tying achieved by the HDP makes this symbol available at other locations in the parse tree.

There are other ways to connect multiple Dirichlet processes. One broadly useful idea is to use a Dirichlet process to define a distribution on Dirichlet processes. In particular, let $\{G_1^*, G_2^*, \ldots\}$ be independent draws from a Dirichlet process, $\mathcal{DP}(\gamma, H)$, and then let G be equal to G_k^* with probability π_k, where the weights $\{\pi_k\}$ are drawn from the stick-breaking process in Eq. (4). This construction (which can be extended to multiple levels) is known as a *nested Dirichlet process* [Rodríguez, Dunson, and Gelfand 2008]. Marginalizing over the Dirichlet processes the resulting urn model is known as the *nested Chinese restaurant process* [Blei, Griffiths, and Jordan 2010], which is a model that can be viewed as a tree of Chinese restaurants. A customer enters the tree at a root Chinese restaurant and sits at a table. This points to another Chinese restaurant, where the customer goes to dine on the following evening. The construction then recurses. Thus a given customer follows a path through the tree of restaurants, and successive customers tend to follow the same paths, eventually branching off.

These nested constructions differ from the HDP in that they do not share atoms among the multiple instances of lower-level DPs. That is, the draws $\{G_1^*, G_2^*, \ldots\}$ involve disjoint sets of atoms. The higher-level DP involves a choice among these disjoint sets.

A general discussion of some of these constructions involving multiple DPs and their relationships to directed graphical model representations can be found in Welling, Porteous and Bart [2008]. Finally, let us mention the work of MacEachern [1999], whose *dependent Dirichlet processes* provide a general formalism for expressing probabilistic dependencies among both the stick-breaking weights and the atom locations in the stick-breaking representation of the Dirichlet process.

7 Completely random measures

The Dirichlet process is not the only tool in the Bayesian nonparametric toolbox. In this section we briefly consider another class of stochastic processes that significantly expands the range of models that can be considered.

From the graphical model literature we learn that probabilistic independence of random variables has desirable representational and computational consequences. In the Bayesian nonparametric setting, random variables arise by evaluating a random measure G on subsets of a measurable space Ω; in particular, for fixed subsets A_1 and A_2, $G(A_1)$ and $G(A_2)$ are random variables. If A_1 and A_2 are disjoint it seems reasonable to ask that $G(A_1)$ and $G(A_2)$ be independent. Such an independence relation would suggest a divide-and-conquer approach to inference.

The class of stochastic processes known as *completely random measures* are characterized by this kind of independence—for a completely random measure the random masses assigned to disjoint subsets of the sample space Ω are independent [Kingman 1967]. Note that the Dirichlet process is *not* a completely random measure—the fact that the total mass is one couples the random variables $\{G(A_i)\}$.

The Dirichlet process provides a latent representation for a clustering problem, where each entity is assigned to one and only cluster. This couples the cluster assignments and suggests (correctly) that the underlying stochastic process is not completely random. If, on the other hand, we consider a latent trait model—one in which entities are described via a set of non-mutually-exclusive binary traits—it is natural to consider completely random processes as latent representations. In particular, the *beta process* is a completely random measure in which a draw consists of a countably infinite collection of atoms, each associated with a probability, where these probabilities are independent [Hjort 1990; Thibaux and Jordan 2007]. In effect, a draw from a beta process yields an infinite collection of independent coins. Tossing these coins once yields a binary featural representation for a single entity. Tossing the coins multiple times yields an exchangeable featural representation for a set of entities.

The beta process arises via the following general construction. Consider the product space $\Omega \otimes (0,1)$. Place a product measure on this space, where the measure associated with Ω is the *base measure* B_0, and the measure associated with $(0,1)$ is obtained from the improper beta density, $cp^{-1}(1-p)^{c-1}$, where $c > 0$ is a parameter. Treating this product measure as a rate measure for a nonhomogeneous Poisson process, draw a set of points $\{(\omega_i, p_i)\}$ in the product space $\Omega \otimes (0,1)$. From these points, form a random measure on Ω as follows:

$$B = \sum_{i=1}^{\infty} p_i \delta_{\omega_i}. \tag{9}$$

The fact that we obtain an infinite collection of atoms is due to the fact that we have used a beta density that integrates to infinity. This construction is depicted graphically in Figure 3.

If we replace the beta density in this construction with other densities (generally defined on the positive real line rather than the unit interval (0,1)), we obtain other completely random measures. In particular, we obtain the *gamma process* by using an improper gamma density in place of the beta density. The gamma process provides a natural latent representation for models in which entities are

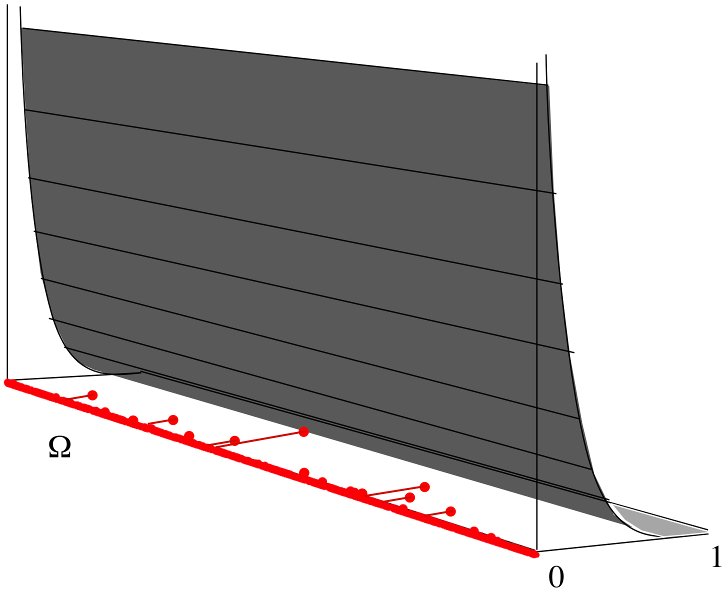

Figure 3. The construction of the beta process from a Poisson process. In this example, Ω is a bounded interval. The rate measure for the Poisson process is the shaded surface—it is the product of a uniform distribution on Ω with an improper beta distribution on $(0, 1)$. Sampling the Poisson process yields the red points in the plane, and these points are connected by line segments to the Ω-axis interval to form the random measure $B = \sum_{i=1}^{\infty} p_i \delta_{\omega_i}$.

represented by a countably infinite set of counts or rates. It is also worth noting that the Dirichlet process can be obtained by normalizing the gamma process.

Recall from our discussion in Section 2 that the Chinese restaurant process can be obtained by integrating out the Dirichlet process in a conditional independence hierarchy. In the other direction, the Dirichlet process is the random measure that is guaranteed (by exchangeability and De Finetti's theorem) to underlie the Chinese restaurant process. Given the importance of the latter model in Bayesian nonparametric modeling and computation, it is of interest to ask if there is a corresponding probability law on binary matrices obtained by integrating out the beta process. As shown by Thibaux and Jordan [2007], the answer is yes, where the probability law is the *Indian buffet process* (IBP) of Griffiths and Ghahramani [2006].

To describe the IBP, consider an Indian buffet with a countably infinite number of dishes. Let N customers arrive in sequence in the buffet line. Let Z denote a binary-valued matrix in which the rows are customers and the columns are the dishes, and where $Z_{n,k} = 1$ if customer n samples dish k. Customer n samples dish k with probability m_k/n, where m_k is the number of customers who have previously sampled dish k; that is, $Z_{n,k} \sim \text{Ber}(m_k/n)$. (Note that this rule can be interpreted in terms of classical Bayesian analysis as sampling the predictive distribution obtained from a sequence of Bernoulli draws based on an improper beta prior.) Having sampled from the dishes previously sampled by other customers, customer n then goes on to sample an additional number of new dishes determined by a draw from a $\text{Poiss}(\alpha/n)$ distribution.

The connection to the beta process delineated by Thibaux and Jordan [2007] is as follows (see Teh and Jordan [2010] for an expanded discussion). Dishes in the IBP correspond to atoms in the beta process, and the independent beta/Bernoulli updating of the dish probabilities in the IBP reflects the independent nature of the atoms in the beta process. Moreover, the fact that a Poisson distribution is adopted for the number of dishes in the IBP reflects the fact that the beta process is defined in terms of an underlying Poisson process. The exchangeability of the IBP (which requires considering equivalence classes of matrices if argued directly on the IBP representation) follows immediately from the beta process construction (by the conditional independence of the rows of Z given the underlying draw from the beta process).

It is also possible to define *hierarchical beta processes* for models involving multiple beta processes that are tied in some manner [Thibaux and Jordan 2007]. This is done by simply letting the base measure for the beta process itself be drawn from the beta process:

$$\begin{aligned} B_0 &\sim \text{BP}(c_0, B_{00}) \\ B &\sim \text{BP}(c, B_0), \end{aligned}$$

where $\text{BP}(c, B_0)$ denotes the beta process with concentration parameter c and base measure B_0. This construction can be used in a manner akin to the hierarchical Dirichlet process; for example, we can use it to model groups of entities that are described by sparse binary vectors, where we wish to share the sparsity pattern among groups.

8 Conclusions

Judea Pearl's work on probabilistic graphical models yielded a formalism that was significantly more expressive than existing probabilistic representations in AI, but yet retained enough mathematical structure that it was possible to design efficient computational procedures for a wide class of useful models. In this short article, we have argued that Bayesian nonparametrics provides a framework in which this agenda can be taken further. By replacing the traditional parametric prior distributions of Bayesian analysis with stochastic processes, we obtain a rich vocabulary, encompassing probability distributions on objects such as trees of infinite depth, partitions, subsets of features, measures and functions. We also obtain natural notions of recursion. In addition to this structural expressiveness, the Bayesian nonparametric framework also permits a wide range of distributional shapes. Finally, although we have devoted little attention to computation in this article, the stochastic processes that have been used in Bayesian nonparametrics have properties (e.g., exchangeability, independence of measure on disjoint sets) that permit the design of efficient inference algorithms. Certainly the framework is rich enough to design some intractable models, but the same holds true for graphical models. The point is that the Bayesian nonparametric framework opens the door to a richer class of useful models for AI. The growing list of successful applications of Bayesian nonparametrics testifies to the practical value of the framework [Hjort, Holmes, Mueller, and Walker 2010].

A skeptical reader might question the value of Bayesian nonparametric modeling given that for any given finite data set the posterior distribution of a Bayesian nonparametric model will concentrate on a finite set of degrees of freedom, and it would be possible in principle to build a parametric model that mimics the nonparametric model on those degrees of freedom. While this skepticism should not be dismissed out of hand—and we certainly do not wish to suggest that parametric modeling should be abandoned—this skeptical argument has something of the flavor of a computer scientist arguing that data structures such as linked lists and heaps are not needed because they can always be mimicked by fixed-dimension arrays. The nonparametric approach can lead to conceptual insights that are only available at the level of an underlying stochastic process. Moreover, by embedding a model for a fixed number of data points in a sequence of models for a growing number of data points, one can often learn something about the statistical properties of the model—this is the spirit of nonparametric statistics in general. Finally, infinite limits often lead to simpler mathematical objects.

In short, we view Bayesian nonparametrics as providing an expressive, useful language for probabilistic modeling, one which follows on directly from the tradition of graphical models. We hope and expect to see Bayesian nonparametrics have as broad of an effect on AI as that of graphical models.

References

Aldous, D. (1985). Exchangeability and related topics. In *Ecole d'Eté de Probabilités de Saint-Flour XIII–1983*, pp. 1–198. Springer, Berlin.

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

Beal, M. J., Z. Ghahramani, and C. E. Rasmussen (2002). The infinite hidden Markov model. In *Advances in Neural Information Processing Systems*, Volume 14, Cambridge, MA. MIT Press.

Blackwell, D. and J. B. MacQueen (1973). Ferguson distributions via Pólya urn schemes. *Annals of Statistics 1*, 353–355.

Blei, D. M., T. L. Griffiths, and M. I. Jordan (2010). The nested Chinese restaurant process and Bayesian inference of topic hierarchies. *Journal of the ACM 57*.

Blei, D. M., A. Y. Ng, and M. I. Jordan (2003). Latent Dirichlet allocation. *Journal of Machine Learning Research 3*, 993–1022.

Erosheva, E. A. (2003). Bayesian estimation of the grade of membership model. In *Bayesian Statistics*, Volume 7, Oxford, UK, pp. 501–510. Oxford University Press.

Escobar, M. D. (1994). Estimating normal means with a Dirichlet process prior. *Journal of the American Statistical Association 89*, 268–277.

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

Finkel, J. R., T. Grenager, and C. D. Manning (2007). The infinite tree. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, Prague, Czech Republic.

Freedman, D. (1963). On the asymptotic behavior of Bayes estimates in the discrete case. *Annals of Mathematical Statistics 34*, 1386–1403.

Griffiths, T. L. and Z. Ghahramani (2006). Infinite latent feature models and the Indian buffet process. In *Advances in Neural Information Processing Systems*, Volume 18, Cambridge, MA. MIT Press.

Hjort, N., C. Holmes, P. Mueller, and S. Walker (2010). *Bayesian Nonparametrics: Principles and Practice*. Cambridge, UK: Cambridge University Press.

Hjort, N. L. (1990). Nonparametric Bayes estimators based on beta processes in models for life history data. *Annals of Statistics 18*, 1259–1294.

Johnson, M., T. L. Griffiths, and S. Goldwater (2007). Adaptor grammars: A framework for specifying compositional nonparametric Bayesian models. In *Advances in Neural Information Processing Systems*, Volume 19, Cambridge, MA. MIT Press.

Karlin, S. and H. M. Taylor (1975). *A First Course in Stochastic Processes*. New York, NY: Springer.

Kingman, J. F. C. (1967). Completely random measures. *Pacific Journal of Mathematics 21*, 59–78.

Kivinen, J., E. Sudderth, and M. I. Jordan (2007). Learning multiscale representations of natural scenes using Dirichlet processes. In *IEEE International Conference on Computer Vision (ICCV)*, Rio de Janeiro, Brazil.

Liang, P., M. I. Jordan, and D. Klein (2010). Probabilistic grammars and hierarchical Dirichlet processes. In *The Handbook of Applied Bayesian Analysis*, Oxford, UK. Oxford University Press.

Lo, A. (1984). On a class of Bayesian nonparametric estimates: I. Density estimates. *Annals of Statistics 12*, 351–357.

MacEachern, S. (1999). Dependent nonparametric processes. In *Proceedings of the Section on Bayesian Statistical Science*. American Statistical Association.

Neal, R. M. (2000). Markov chain sampling methods for Dirichlet process mixture models. *Journal of Computational and Graphical Statistics 9*, 249–265.

Pitman, J. (2002). Combinatorial stochastic processes. Technical Report 621, Department of Statistics, University of California at Berkeley.

Rodríguez, A., D. B. Dunson, and A. E. Gelfand (2008). The nested Dirichlet process. *Journal of the American Statistical Association 103*, 1131–1154.

Sethuraman, J. (1994). A constructive definition of Dirichlet priors. *Statistica Sinica 4*, 639–650.

Teh, Y. W. and M. I. Jordan (2010). Hierarchical Bayesian nonparametric models with applications. In *Bayesian Nonparametrics: Principles and Practice*. Cambridge, UK: Cambridge University Press.

Teh, Y. W., M. I. Jordan, M. J. Beal, and D. M. Blei (2006). Hierarchical Dirichlet processes. *Journal of the American Statistical Association 101*, 1566–1581.

Thibaux, R. and M. I. Jordan (2007). Hierarchical beta processes and the Indian buffet process. In *Proceedings of the International Workshop on Artificial Intelligence and Statistics*, Volume 11, San Juan, Puerto Rico.

Welling, M., I. Porteous, and E. Bart (2008). Infinite state Bayesian networks for structured domains. In *Advances in Neural Information Processing Systems*, Volume 20, Cambridge, MA. MIT Press.