STAT C206B: Probability and Convexity

Alex Fu

Spring 2024

This set of course notes is based on Professor Steve Evans' lectures for STAT C206B: Probability and Convexity at UC Berkeley in the spring semester of 2024. All mistakes are my own.

Contents

Chapter 1

Dirichlet Processes

1.1 Introduction

We start with the notion of a convex set in a linear space. Recall that a **convex combination** (in a real vector space) is a linear combination of vectors whose coefficients are nonnegative and sum to 1. We say that a set is **convex** if it is closed under taking convex combinations. A standard example of a convex set is the set of all convex combinations of a given set of vectors S, which is called the **convex hull** of S and denoted by $\operatorname{hull}(S)$.

Let K be a compact convex subset of $\mathbb{R}^d$.

Definition 1.1.1. An **extreme point** of K is a point in K that cannot be written as a nontrivial convex combination of other points. $\lrcorner$

Fact 1.1.2. Any point in K can be represented by a convex combination of the extreme points in K.

This representation is not necessarily unique. However, it is unique whenever K is a **simplex**. (For example, take K to be an interval in dimension $d = 1$, a triangle in dimension 2, a tetrahedron in dimension 3, and so on.)

Question 1.1.3. How does this extend to infinite dimensions? (And what does this have to do with probability?)

One straightforward observation is that convex combinations are sums whose weights form discrete probability distributions, but we will want something a bit more exciting than that. The idea of **convex representations** shows up a lot in probability; we will look at many examples of this phenomenon.

1.2 Dirichlet distributions

The first example that we will consider is the *Dirichlet distribution* on the convex set of all probability distributions on a finite set.

Notation 1.2.1. Let Δ^n denote the n-dimensional *unit simplex*, $\{\vec{v} \in \mathbb{R}^n \mid v_i \geq 0 \text{ for all } i = 1, \ldots, n, \text{ and } \sum_{i=1}^n v_i \leq 1\}$. Let $\mathrm{A}^n \subset \Delta^n$ denote the *probability simplex* or *standard simplex* in $\mathbb{R}^n$, namely

$$\mathrm{A}^n := \operatorname{hull}(\{\vec{e}_1, \ldots, \vec{e}_n\}) = \left\{ \vec{v} \in \mathbb{R}^n \;\middle|\; v_i \geq 0 \text{ for all } i = 1, \ldots, n, \text{ and } \sum_{i=1}^n v_i = 1 \right\},$$

where $\vec{e}_i = \mathbf{e}_i$ denotes the ith coordinate vector in $\mathbb{R}^n$ for each $i = 1, \ldots, n$. Let $[n]$ denote the set $\{1, \ldots, n\}$.

With this notation, we can equivalently consider the set of all probability distributions on $[n]$ to be A^n. It follows that a distribution on A^n determines a *random probability measure* on $[n]$.

Definition 1.2.2. Given $\alpha > 0$ and $\beta > 0$, we say that a $\mathbb{R}^{\geq 0}$-valued random variable X has the **gamma distribution** with *shape parameter* α and *scale parameter* β if it has probability density

$$\frac{\beta^\alpha x^{\alpha-1}}{\Gamma(\alpha)} e^{-\beta x}, \quad \forall x \geq 0,$$

where Γ is the usual gamma function defined by $\Gamma(\alpha) = \int_0^\infty x^{\alpha-1} e^{-x}\, \mathrm{d}x$. ⌟

Lemma 1.2.3. *Let $X_1, \dots, X_n$ be independent random variables, where each X_i has distribution $\Gamma(\alpha_i, \beta)$. Then $X_1 + \cdots + X_n$ has distribution $\Gamma(\alpha_1 + \cdots + \alpha_n, \beta)$.*

For the sake of simplicity, we will be leaving many proofs to the appendix, including routine checks, solutions to exercises, and proofs that we do not deem relevant or necessary for the reader to know.

Definition 1.2.4. Given $\alpha_1, \dots, \alpha_{n+1} > 0$, we say that the random vector $(V_1, \dots, V_{n+1})$ has the **Dirichlet distribution** with parameters $(\alpha_1, \dots, \alpha_{n+1})$ if $V_1 + \cdots + V_{n+1} = 1$ and the random vector $(V_1, \dots, V_n)$ has probability density

$$\frac{\Gamma(\alpha_1 + \cdots + \alpha_{n+1})}{\Gamma(\alpha_1) \cdots \Gamma(\alpha_n)} v_1^{\alpha_1 - 1} \cdots v_n^{\alpha_n - 1} \left(1 - \sum_{i=1}^n v_i\right)^{\alpha_{n+1}-1} \cdot \mathbb{1}\{(v_1, \dots, v_n) \in \Delta^n\}.$$

This definition is motivated by the following equivalent but more constructive definition: ⌟

Definition 1.2.5. Let $X_1, \dots, X_{n+1}$ be independent random variables such that $X_i \sim \Gamma(\alpha_i, \beta)$ for each $i = 1, \dots, n+1$. Let $S_i = X_1 + \cdots + X_i$ and $V_i = X_i / S_{n+1}$ for each $1 \leq i \leq n+1$. Then $(V_1, \dots, V_{n+1})$ and S_{n+1} are independent, and we say that $(V_1, \dots, V_{n+1})$ has the **Dirichlet distribution** with parameters $(\alpha_1, \dots, \alpha_{n+1})$. In this case, we write

$$(V_1, \dots, V_{n+1}) \sim \mathrm{Dir}_{(\alpha_1, \dots, \alpha_{n+1})}.$$

Note that this definition does not depend on the choice of scale parameter β. For convenience, we will often take $\beta = 1$. Also, we may allow some but not all of the α_i to be 0, where we adopt the convention that a random variable with distribution $\Gamma(0, \beta)$ is identically 0. ⌟

Remark 1.2.6. It is somewhat of a miracle that the normalized vector $(V_1, \dots, V_{n+1})$ is independent of the normalizer S_{n+1}. This property in fact characterizes the gamma distribution; see Lukacs' proportion-sum independence theorem [1, 2]. (Also see [3] for a related characterization of the Dirichlet distribution.)

A key property of the Dirichlet distribution is the following:

Lemma 1.2.7 (Aggregation)**.** *Suppose that V has the Dirichlet distribution with parameters $(\alpha_1, \dots, \alpha_{n+1})$. For some $1 \leq r \leq n$, let $W_i = V_i$ for every $1 \leq i \leq r$, and let $W_{r+1} = V_{r+1} + \cdots + V_{n+1}$. Then W has the Dirichlet distribution with parameters $(\alpha_1, \dots, \alpha_r, \beta_{r+1})$, where $\beta_{r+1} = \alpha_{r+1} + \cdots + \alpha_{n+1}$.*

Iteratively applying this result, we see that "clumping together" entries in a Dirichlet random vector gives another Dirichlet random vector, whose parameters are given by "clumping together" the original parameters in the same manner. That is, if $\phi \colon [n+1] \to [m+1]$ is a surjective function and we put $U_j = \sum_{i:\phi(i)=j} V_i$ for every $1 \leq j \leq m+1$, then U has the Dirichlet distribution with parameters $(\gamma_1, \dots, \gamma_{m+1})$ where $\gamma_j = \sum_{i:\phi(i)=j} \alpha_i$.

It is also worth noting that the Dirichlet distribution (with $n+1$ parameters) is the multivariate generalization of the beta distribution:

Definition 1.2.8. Given $\alpha, \beta > 0$, we say that a random variable X has the **beta distribution** with parameters α and β if it takes values in $[0,1]$ and has probability density function

$$\frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1} (1-x)^{\beta-1}.$$

In this case, we write $X \sim \mathrm{B}(\alpha, \beta)$. ⌟

Note that the first component of a random vector with distribution $\mathrm{Dir}_{(\alpha,\beta)}$ has distribution $\mathrm{B}(\alpha,\beta)$. More generally, the marginal distributions of a Dirichlet random vector are beta distributions:

Lemma 1.2.9. *If* $(V_1,\dots,V_{n+1}) \sim \mathrm{Dir}_{(\alpha_1,\dots,\alpha_{n+1})}$, *then* $V_i \sim \mathrm{B}(\alpha_i, (\sum_{j=1}^{n+1}\alpha_j) - \alpha_i)$ *for each* $i = 1,\dots,n+1$.

Informally, we remark that the aggregation property and the marginal distributions of the Dirichlet distribution make it in some sense "self-similar", like the multivariate Gaussian distribution.

In the next section, we will generalize Dirichlet distributions to the infinite-dimensional case. Such *Dirichlet processes* or **Dirichlet measures** form a class of distributions of random probability measures on a general measurable space $(\mathscr{X},\Sigma)$. These distributions have applications in statistics, for example, in Bayesian nonparametrics.

1.3 Construction of Dirichlet processes

Definition 1.3.1 (Bayesian nonparametrics)**.** Suppose X is a random variable, representing "data" taking values in a measurable space $(\mathscr{X},\Sigma)$. Let the unknown distribution of X be P. Then P is the *parameter* in the nonparametric problem, and it takes values in $\mathscr{P}$, the collection of all probability measures on $(\mathscr{X},\Sigma)$. Now let $\mathfrak{C}$ be the σ-algebra on $\mathscr{P}$ that is generated by sets of the form

$$\{P \in \mathscr{P} \mid P(A) < r\}, \quad \forall A \in \Sigma, \quad \forall r \in [0,1].$$

Then $(\mathscr{P},\mathfrak{C})$ is a measurable space. A probability measure ν on $(\mathscr{P},\mathfrak{C})$ can be used as a *prior distribution* for P. The Bayesian solution is to compute the *posterior distribution* ν^X of P given X, and use it for decision making. ⌟

We also define a *measurable partition* of $\mathscr{X}$ to be a partition of $\mathscr{X}$ into measurable subsets.

Definition 1.3.2. Let $\mathscr{M}$ be the class of nonzero finite measures on $(\mathscr{X},\Sigma)$, and let $\alpha \in \mathscr{M}$. We say that a probability distribution ν on $(\mathscr{P},\mathfrak{C})$ is a **Dirichlet measure** with parameter α if for every measurable partition $\{B_1,\dots,B_k\}$ of $\mathscr{X}$ into finitely many subsets, we have

$$(P(B_1),\dots,P(B_k)) \sim \mathrm{Dir}_{(\alpha(B_1),\dots,\alpha(B_k))} \tag{1.3.3}$$

under ν (i.e., if $P \sim \nu$). In this case, we denote ν by Dir_α. ⌟

Observe that the finite-dimensional Dirichlet distribution is a special case of the Dirichlet measure by the aggregation property. That is, if α is a nonzero finite measure on $[k]$, then the Dirichlet measure Dir_α coincides with the k-dimensional Dirichlet distribution $\mathrm{Dir}_{(\alpha(\{1\}),\dots,\alpha(\{k\}))}$. However, we have yet to show that Dirichlet measures exist in the general case.

Remark 1.3.4. Dirichlet processes were first formally introduced by Ferguson [4], who gave a direct proof of their existence via the Kolmogorov Consistency Theorem. We will be following a later proof by Sethuraman [5], a more constructive approach in terms of the so-called "stick-breaking" process.

To motivate the constructive definition, let us state three main properties of Dirichlet measures that make them useful in Bayesian nonparametrics:

Proposition 1.3.5 (Properties of the Dirichlet measure)**.**

P1. Dir_α *is a probability measure on* $(\mathscr{P},\mathfrak{C})$.

P2. Dir_α *assigns probability* 1 *to the subset of all discrete probability measures on* $(\mathscr{X},\Sigma)$.

P3. *The posterior distribution* Dir_α^X *is the Dirichlet measure* $\mathrm{Dir}_{\alpha+\delta_X}$, *where* δ_X *denotes the degenerate probability distribution on* $(\mathscr{X},\Sigma)$ *localized at* X.

The posterior distribution Dir_α^X may seem slightly convoluted, so let us elaborate on its definition. We start with $P \sim \mathrm{Dir}_\alpha$, a random element of $(\mathscr{P}, \mathfrak{C})$. This gives rise to $X \sim P$, a random element of $(\mathscr{X}, \Sigma)$. After observing X, the posterior Dir_α^X is the conditional distribution of P given X. This posterior exists because P and X are defined on a common underlying probability space $(\Omega, \mathscr{F}, \mathbb{Q})$, as we will see shortly.

Notation 1.3.6. Given a finite measure α on $(\mathscr{X}, \Sigma)$, we denote its total variation $\alpha(\mathscr{X})$ by $\|\alpha\|$.

Definition 1.3.7 (Constructive definition of the Dirichlet measure [5])**.** Let α be a nonzero finite measure on $(\mathscr{X}, \Sigma)$, and let $\beta = \alpha / \|\alpha\|$ be the normalized probability distribution arising from α. Let $\theta_1, \theta_2, \dots$ be i.i.d. random variables with common distribution $\mathrm{B}(1, \|\alpha\|)$. Define a probability distribution on $\mathbb{Z}^+$ by

$$\begin{aligned} p_1 &= \theta_1, \\ p_n &= \theta_n \prod_{m=1}^{n-1} (1 - \theta_n) \quad \forall n \geq 2. \end{aligned}$$

We say that the sequence $(p_1, p_2, \dots)$ is constructed via *stick-breaking* from the proportions or weights $(\theta_1, \theta_2, \dots)$. Now let $Y_1, Y_2, \dots$ be i.i.d. random variables with common distribution β, independent of $(\theta_1, \theta_2, \dots)$. Then define a random probability measure P on $(\mathscr{X}, \Sigma)$ by

$$P(B) = P(\boldsymbol{\theta}, \mathbf{Y}; B) := \sum_{n=1}^{\infty} p_n \delta_{Y_n}(B). \tag{1.3.8}$$

The distribution of P is the **Dirichlet measure** with parameter α.

Observe that properties **P1** and **P2** are satisfied by design. In order to give a proof of property **P3**, we introduce an additional random variable I as follows. Let $(\Omega, \mathscr{F}, \mathbb{Q})$ be a probability space supporting a collection of random variables $(\boldsymbol{\theta}, \mathbf{Y}, I) = ((\theta_n, Y_n), 1 \leq n \leq I)$. Define θ_n, p_n, Y_n, $\forall n \geq 1$ as before. We define I by

$$\mathbb{Q}(I = n \mid (\boldsymbol{\theta}, \mathbf{Y})) = p_n \quad \forall n \geq 1.$$

This gives a valid probability distribution on $\mathbb{Z}^+$ since $\sum_{m=1}^n p_m = 1 - \prod_{m=1}^n (1 - \theta_m) \to 1$ holds with $\mathbb{Q}$-probability 1. We will be using I in the next section. ⌟

Now, the first order of business is to prove the following:

Proposition 1.3.9. *The distribution of P is* Dir_α.

TODO: To be continued …

Bibliography

[1] Eugene Lukacs. A characterization of the gamma distribution. *The Annals of Mathematical Statistics*, 26(2):319–324, 1955.

[2] James E. Mosimann. On the compound multinomial distribution, the multivariate β-distribution, and correlations among proportions. *Biometrika*, 49(1–2):65–82, 1962.

[3] Ian R. James and James E. Mosimann. A new characterization of the Dirichlet distribution through neutrality. *The Annals of Statistics*, 8(1):183–189, 1980.

[4] Thomas S. Ferguson. A Bayesian analysis of some nonparametric problems. *The Annals of Statistics*, 1:209–230, 1973.

[5] Jayaram Sethuraman. A constructive definition of Dirichlet priors. *Statistica Sinica*, 4(2):639–650, 1994.

[6] David Blackwell and James B. MacQueen. Ferguson distributions via Pólya urn schemes. *The Annals of Statistics*, 1(2):353–355, 1973.

Appendix A

Proofs

A.1 Chapter 1: Dirichlet Processes

Lemma 1.2.3. *Let $X_1, \dots, X_n$ be independent random variables, where each X_i has distribution $\Gamma(\alpha_i, \beta)$. Then $X_1 + \cdots + X_n$ has distribution $\Gamma(\alpha_1 + \cdots + \alpha_n, \beta)$.*

Proof. It suffices to prove the case of $n = 2$, after which the general result follows by induction. So, let $X_1 \sim \Gamma(\alpha_1, \beta)$ and $X_2 \sim \Gamma(\alpha_2, \beta)$ be independent random variables. The probability density function of their sum is

$$\begin{aligned} f(x) &= \frac{\beta^{\alpha_1}\beta^{\alpha_2}}{\Gamma(\alpha_1)\Gamma(\alpha_2)} \int_0^x t^{\alpha_1 - 1}(x-t)^{\alpha_2 - 1} e^{-\beta t} e^{-\beta(x-t)} \, \mathrm{d}t \\ &= \frac{\beta^{\alpha_1+\alpha_2}}{\Gamma(\alpha_1)\Gamma(\alpha_2)} e^{-\beta x} \int_0^1 (ux)^{\alpha_1-1}((1-u)x)^{\alpha_2-1} x \, \mathrm{d}u \\ &= \frac{\beta^{\alpha_1+\alpha_2}}{\Gamma(\alpha_1)\Gamma(\alpha_2)} e^{-\beta x} x^{\alpha_1+\alpha_2-1} \int_0^1 u^{\alpha_1-1}(1-u)^{\alpha_2-1} \, \mathrm{d}u. \end{aligned}$$

From the definition of the beta distribution, we find that the last integral evaluates to $\Gamma(\alpha_1)\Gamma(\alpha_2)/\Gamma(\alpha_1+\alpha_2)$. Thus, the probability density function of $X_1 + X_2$ simplifies to

$$f(x) = \frac{\beta^{\alpha_1+\alpha_2} x^{\alpha_1+\alpha_2-1}}{\Gamma(\alpha_1+\alpha_2)} e^{-\beta x},$$

which is precisely the probability density function of a random variable with distribution $\Gamma(\alpha_1 + \alpha_2, \beta)$. □

Lemma 1.2.7 (Aggregation)**.** *Suppose that V has the Dirichlet distribution with parameters $(\alpha_1, \dots, \alpha_{n+1})$. For some $1 \le r \le n$, let $W_i = V_i$ for every $1 \le i \le r$, and let $W_{r+1} = V_{r+1} + \cdots + V_{n+1}$. Then W has the Dirichlet distribution with parameters $(\alpha_1, \dots, \alpha_r, \beta_{r+1})$, where $\beta_{r+1} = \alpha_{r+1} + \cdots + \alpha_{n+1}$.*

Proof. This follows directly from Lemma 1.2.3 and Definition 1.2.5. □

Lemma 1.2.9. *If $(V_1, \dots, V_{n+1}) \sim \mathrm{Dir}_{(\alpha_1, \dots, \alpha_{n+1})}$, then $V_i \sim \mathrm{B}(\alpha_i, (\sum_{j=1}^{n+1} \alpha_j) - \alpha_i)$ for each $i = 1, \dots, n+1$.*

Proof. Assume $i = 1$ without loss of generality. We write α for $\alpha_1 + \cdots + \alpha_{n+1}$. By the aggregation property, $(V_1, V_2 + \cdots + V_{n+1})$ has distribution $\mathrm{Dir}_{(\alpha_1, \alpha - \alpha_1)}$. From this, we see that the marginal distribution of V_1 is $\mathrm{B}(\alpha_1, \alpha - \alpha_1)$. □

TODO: To be continued …