PROBABILITY 2019: PART 1
PROBABILITY ON A FINITE SAMPLE SPACE

GIOVANNI PISTONE

Contents

Traditionally, a first course in probability presents a well established set of topics in probability on finite sets and probability on real numbers. Popular examples are [5, Ch. 1–3] and [3, Ch 3]. The present course takes on the same set of topics but the mathematical treatment is more sophisticated than usual. This is required in a number of contemporary applications.

1. Elementary probability from an advanced viewpoint

1.1. **Probability.** Given a finite set Ω, a *probability function* is a mapping $p\colon \Omega \to \mathbb{R}_+$ such that $\sum_{\omega\in\Omega} p(\omega) = 1$. In such a set-up, a *random variable* is a generic function $f\colon \Omega \to S$, where S is any set.[1] An *event* is any subset $A \subset \Omega$. The set of all events $\mathcal{A}$ is a *field* i.e., contains $\varnothing$ and Ω, and is closed under complement, union, intersection. A *proposition* is a mapping $a\colon \Omega \to \{\text{FALSE}, \text{TRUE}\}$. The set of all propositions and the field of events are in 1-to-1 correspondence $\omega \in A \leftrightarrow a(\omega) = \text{TRUE}$ and set operation translate into logical operations, A^c is $\neg a$, $A \cup B$ is $a \vee b$, $A \cap B$ is $a \wedge b$. Another equivalent presentation of events uses *indicator functions* $A \leftrightarrow \mathbf{1}_A$. In this case $\mathbf{1}_{A\cap B} = \mathbf{1}_A \mathbf{1}_B$ and

Date: DRAFT February 16, 2019.

[1] A more sophisticated set-up i.e., measure theory, is needed when the sample space is not finite

$\mathbf{1}_{A\cup B} = \mathbf{1}_A + \mathbf{1}_B - \mathbf{1}_A\mathbf{1}_B$. The last presentation is frequently easier to andle when algebraic computations are required.

The probability function p induces a *probability measure* $P\colon \mathcal{A} \to [0,1]$ by setting $P(A) = \sum_{\omega\in A} p(\omega)$. The probability measure P has the following properties:

(1) $P(\Omega) = 1$;
(2) $P(A \cup B) = P(A) + P(B) - P(A \cup B)$.

Conversely, any function on $P\colon \mathcal{A}$ satisfying the conditions above comes from the probability function $p(\omega) = P(\{\omega\})$.

Here are some notable cases.

- The constant probability function $f(\omega) = 1/\#\Omega$ induces the *uniform probability measure*.
- The probability function such that $p(\omega) = 1$ if $\omega = a$ induces the Dirac probability measure δ_a such that $\delta_a(A) = 1$ if and only if $\omega \in A$.
- Given probability functions p and q and a number $\lambda \in [0,1]$, then $(1-\lambda)p + \lambda q$ is a probability function, a *mixture of p and q*.

Exercise 1 (Finite state spaces). Consider an example of finite set with an interesting structure and compute the uniform probability function. [Permutations, partitions, subsets with a given number of points, graphs, trees, tables with given margins, ...]

1.2. **Random variables.** Consider a random variable $f\colon \Omega \to S$, where S is any set.[2] A *σ-field* $\mathcal{S}$ on S is a family of subsets of S which is a field, and moreover, is closed under numerable intersections and numerable unions.

Given a probability P on Ω the equation

$$Q(B) = f_{\#}P(B) = P(f^{-1}(B)) , \quad B \in \mathcal{S} ,$$

defines a *probability measure* on the *measurable space* $(S, \mathcal{S})$, that is a mapping $Q\colon \mathcal{S} \to [0,1]$ such that

(1) $Q(S) = 1$;
(2) $Q(\cup B_\alpha) = \sum_\alpha Q(B_\alpha)$ for each numerable family of disjoint sets $\{Q_\alpha\}$.

$f_{\#}P$ is the *image of P under f* or the *distribution of f under P*. Another notation is P_f. If S is actually finite and $\mathcal{S}$ is the set of parts of S, then the image probability function is defined by $p_f(s) = \sum_{\omega | f(\omega)=s} p(\omega)$.

1.3. **Real random variables.** Consider now the case of a random variable $u\colon \Omega \to \mathbb{R}$. The set of real random variables is denoted by $L(\Omega)$. It is a real vector space, an algebra with unity, and it is closed under $\wedge = \max$ and $\vee = \min$. It is also denoted by $\mathbb{R}^\Omega$. If $\Omega = \{1, \dots, N\}$, then $L(\Omega)$ can be identified with the set of all column vectors in $\mathbb{R}^{N\times 1}$.

The *expected value* of $u \in L(\Omega)$ under the probability function p is $\mathbb{E}_p[f] = \sum_\omega u(\omega)p(\omega)$. Notice that $u = \mathbf{1}_A$ gives the special case $\mathbb{E}_p[\mathbf{1}_A] = P_p(A)$, where P_p is the probability measure induced by p. If both p and u are represented as vectors $\boldsymbol{p}, \boldsymbol{u} \in \mathbb{R}^{N\times 1}$, then $\mathbb{E}_p[u] = \boldsymbol{p}^t\boldsymbol{u}$.

Here are some properties of the expectation operator.

(1) $\mathbb{E}_p\colon L(\Omega) \to \mathbb{R}$ *is a linear operator*: $\mathbb{E}_p[\alpha u + \beta v] = \alpha\,\mathbb{E}_p[u] + \beta\,\mathbb{E}_p[v]$.
(2) $\mathbb{E}_p$ *is positive and faithful*: If $u \geqslant 0$, then $\mathbb{E}_p[u] \geqslant 0$. In the case $\mathbb{E}_p[u] = 0$, then $u = 0$ on $\mathrm{Supp}\,(p) = \{\omega \in \Omega | p(\omega) > 0\}$.
(3) *Change of variable*: Let $f\colon \Omega \to S$ with S finite. Then $\mathbb{E}_p[u \circ f] = \mathbb{E}_{p_f}[u]$.

[2] A fully finite treatment is not possible because we want real valuer random variables

Exercise 2. Show that for all non-negative random variable u, all $c \geqslant 0$, and all $a \geqslant 1$, one has $P(u \geqslant c) \leqslant \frac{1}{c^a}\mathbb{E}(u^a)$.[Use indicator functions and positivity]

Exercise 3. Let $\Omega = \{0,1\}^n$, $\omega = x_1 \dots x_n$. Define $f(\omega) = \sum_{j=1}^n x_j$. For each $\theta \in [0,1]$ show that $p(\omega) = \theta^{f(\omega)}(1-\theta)^{n-f(\omega)}$ is a probability function on Ω. Compute p_f, $m = \mathbb{E}_p[f]$, $\mathbb{E}_p[(f-m)^2]$.

1.4. **Simulation.** Let be given an infinite sequence $\omega_1, \omega_2, \dots$ and assume that the limits $p(\omega) = \lim_{n\to\infty} \frac{1}{n}\sum_{k=1}^n (\omega_k = \omega)$ exist for all $\omega \in \omega$. Then $p(\omega) \geqslant 0$ and $\sum_{\omega\in\Omega} p(\omega) = 1$. We say that the probability function p is the distribution of the sequence. Conversely, the sequence is a *simulation* of p. *Every probability function admits a simulation.* In fact, for $\Omega = \{1, \dots, N\}$, consider a partition of the of $]0,1[$ into intervals $I_1, \dots, I_n$ such that the length of each I_k is p_k. There exists a sequence $x_1, x_2, \dots$ of real numbers in $]0,1[$ such that sequence defined by $\omega_n = k$ if $x_n \in I_k$ is a simulation of p.[3]

A simulation allows to compute expected values as $\mathbb{E}_p[u] = \lim_{n\to\infty} \frac{1}{n}\sum_{k=1}^n u(\omega_k)$. In particular this applies to probability measures and we can talk about the simulation of a probability measures. *If $(\omega_n)_n$ is a simulation of P, then $(f(\omega_n))_n$ is a simulation of $f_\# P$.* The full structure of probability on finite sample space could be presented as a topic in simulation.

1.5. **Aside: convex sets.** [4]

A subset H of a vector space V is an *affine space* if $\{x - y | x, y \in H\}$ is a sub-vector space of V which is called the vector subspace parallel to H. The dimension of the affine space H is the dimension its parallel vector subspace. Given $x_0, \dots, x_n \in V$ the set of all vectors of the form $x_0 + \sum_{j=1}^n \lambda_j x_j$, $\lambda_j \in \mathbb{R}$, is the affine space generated by the given vectors. An affine space of dimension $n-1$ in $\mathbb{R}^n$ is an *hyper-plane*,

A subset C of the vector space V is *convex* if for all $x, y \in C$ all of the segment $(1-\lambda)x + \lambda y$, $\lambda \in [0,1]$ belongs to C. The intersection of two convex sets is convex. Given $x_0, \dots, x_n \in V$ the set of all $\lambda_0 x_0 + \dots + \lambda_n x_n$ with $\lambda_0 + \dots + \lambda_n = 1$ is the convex set generated by the given vectors. Such a set is called a *polytope* (or convex polytope). Notice that $\sum_{j=0}^n \lambda_j x_j = (1 - \sum_{j=1}^n \lambda_j)x_0 + \sum_{j=1}^n \lambda_j x_j = x_0 + \sum_{j=1}^n \lambda_j (x_j - x_0)$ that is, the polytope is a part of the affine space generated. A notable example of convex set is the *half-space* of $v \in V$ such that $\langle c, v\rangle \leqslant b$ with $c \in V$ and $b \in \mathbb{R}$. A finite intersection of half-spaces is a convex set called a *polyhedron*. A bounded polyhedron is a polytope.

The vectors $x_0, \dots, x_m$ are *affinely independent* if the vectors $x_1 - x_0, \dots, x_m - x_0$ are linearly independent. They form a vector basis of the sub-space parallel to the generated polytope which in this case is called a *simplex*. Two simplexes of the same dimension can be mapped one onto the other by an affine transformation that map their respective generators (the vertexes).

1.6. **Affine geometry of the probability simplex.** Let λ be a probability function on Ω. As $\lambda \in \mathbb{R}^\Omega$, we can write $\lambda = \sum_{x\in\Omega} \lambda(x)\delta_x$, so that the set $\Delta(\Omega)$ is the convex set generated by the probability functions associated to the Dirac probability measures. Let us code Ω as $\{1, \dots, N\}$ and write $\lambda = \sum_{j=1}^n \lambda_j e_j$. The vectors $e_j - e_m$, $j = 1, \dots, N-1$

[3]The proof of the existence of such a sequence is not given here. It follows for example from the Strong Law of Large Numbers, to be discussed later. Other proofs are bases on arguments from Ergodic Theory or Fourier Analysis. For example, irrational rotations of the torus in $\mathbb{C}$ have such a property. This example originally suggested to von Neumann the algorithms for the computer generation of random numbers. For a simple presentation of this topic, check S.H. Ross *Simulation* on Google Books.

[4]Convex analysis is an important topic in applied probability. Standard references are the monographs [4, 1]

are linearly independent so that $\Delta(\Omega)$ is a special simplex which is called the *probability simplex*. The parallel vector space is the vector space of the vectors of the form $\sum_{j=1}^n \alpha_j(e_j - e_1)$ that is of the form $\sum_{j=1}^n \alpha_j e_j$ with $\sum_{j=1}^n \alpha_j = 0$. These are the vectors which are orthogonal to the constant vectors.

The set of probability functions with support $\Omega_1 \subset \Omega$ form a simplex of dimension $\#\Omega_1 - 1$. If $\#\Omega_1 = n - 1$ this sub-simplex is a *face* of $\Delta(\Omega)$.

There is another simplex that represents the probability simplex $\Delta(\Omega)$ namely, the *solid probability simplex*. In fact, we can represent a probability function by its $n - 1$ values $\lambda_j, \dots, \lambda_{n-1}$ which form a vector in $\mathbb{R}^{n-1}$ satisfying the conditions $\lambda_j \geqslant 0$ and $\sum_{j=1}^{n-1} \lambda_j \leqslant 1$. The vectors $e_1, \dots, e_{n-1}, 0 \in \mathbb{R}^{n-1}$ are affinely independent and generate a simplex of dimension $n - 1$ as $\sum_{J=1}^{n-1} \lambda_j e_j + \lambda_n 0$. The mapping between the two representations is given by $\mathbb{R}^n \ni e_j \mapsto e_j \in \mathbb{R}^{n-1}$ for $j = 1, \dots, n-1$ and $\mathbb{R}^n \ni e_n \mapsto 0 \in \mathbb{R}^{n-1}$.

Exercise 4. Study the probability simplex $\Delta(\{1, 2, 3\})$. In particular, construct the solid simplex and show it is a polyhedron. Consider the representation as an equilateral triangle. [Check for example `http://henr.in/crumbs/simplex/` .

Exercise 5. Study the probability simplex on $\Omega = \{0, 1\}^2$. It is a simplex of dimension 3 and it is interesting to consider its graphical representations. [Check the Wikipedia entry `https://en.wikipedia.org/wiki/Simplex`.]

1.7. **Aside: differentials.** Let $f \colon \mathcal{O} \to \mathbb{R}^n$, where $\mathcal{O}$ is an open sub-set of $\mathbb{R}^m$. The function is differentiable at $\bar{x} \in \mathcal{O}$ if there exists a linear mapping $df(\bar{x}) \in L(\mathbb{R}^m, \mathbb{R}^n)$ such that

$$f(\bar{x} + h) - f(\bar{x}) - df(\bar{x})[h] = \mathrm{o}(h) \ .$$

The matrix representing the linear operator $df(\bar{x})$ is called the Jacobian matrix of f, $Jf(\bar{x})$, whose elements are the partial derivatives

$$Jf(\bar{x}) = \left[\tfrac{\partial}{\partial x_j} f_i(x_1, \dots, x_n)\right]_{i=1,\dots,n;j=1,\dots m}$$

The derivative of the composite function $f \circ g$ at x is $df \circ g(x) = df(g(x)) \circ dg(x)$.

1.8. **Differentiability on the probability simplex.** Let $I \ni \theta \mapsto \lambda(\theta)$ be a curve in the probability simplex which is differentiable in $\mathbb{R}^\Omega$. The derivative

$$\lambda'(\theta) = \lim_{h \to 0} h^{-1}(\lambda(\theta + h) - \lambda(\theta)$$

belongs to the subspace parallel to the simplex. If $\lambda(\bar{\omega}; \bar{\theta}) = 0$, then the real differentiable function $\theta \mapsto \lambda(\bar{\omega}, \theta)$ has a minimum at $\theta = \bar{\theta}$, so that $\lambda'(\bar{\omega}, \bar{\theta}) = 0$ and $\lambda'(\bar{\theta})$ belong to the space parallel to the face of the simplex characterised by $\lambda(\bar{\omega}) = 0$.

Exercise 6. The function $H(\lambda) = -\sum_\omega \lambda(\omega) \log \lambda(\omega)$ is defined on the convex set $\Delta^\circ(\Omega)$ of strictly positive probability functions. As the function $x \mapsto \phi(x) = -x \log x$ is concave, so that

$$\frac{1}{\#\Omega} H(\lambda) = \frac{1}{\#\Omega} \sum_{\omega \in \Omega} \phi(\lambda(\omega)) \leqslant \phi\left(\frac{1}{\#\Omega} \sum_{\omega \in \Omega}\right) = \phi\left(\frac{1}{\#\Omega}\right)$$

and the uniform probability function is a maximum of the entropy. Let us show that this maximum is unique. Assume there is a $\bar{\lambda}$ which is a maximum for the entropy and let $\theta \mapsto \lambda(\theta)$ be a differentiable curve in $\Delta^0(\Omega)$ such that $\lambda(0) = \bar{\lambda}$. Let us compute the derivative

$$\frac{d}{d\theta} H(\lambda(\theta))\bigg|_{\lambda=0} = -\sum_{\omega \in \Omega} (\log \lambda(\omega; \theta) + 1) \lambda'(\omega; \theta)\bigg|_{\theta=0} = -\sum_{\omega \in \Omega} (\log \bar{\lambda}(\omega) + 1) \lambda'(\omega; 0) = 0 \ .$$

As $\bar{\lambda}$ is in the $\Delta^0(\Omega)$, for each v in the space parallel to the simplex we can consider the curve $\theta \mapsto \bar{\lambda} + \theta v$ whose derivative atv $\theta = 0$ is v. It follows that for each v we have

$$\sum_{\omega\in\Omega} (\log \bar{\lambda}(\omega) + 1)v(\omega) = 0$$

hence, $\log \bar{\lambda}$ is constant that is, $\bar{\lambda}$ is constant $\bar{\lambda}(\omega) = 1/\#\Omega$.

1.9. **Aside: convex functions.** If a convex set $A \in \mathbb{R}^m$ is open, then every straight line intersects C in an open interval or an empty interval. For example, the subset of the solid probability simplex consisting of strictly positive probability functions is an open convex set. The closure $\overline{A}$ of an open convex set A is a convex set. The difference $\overline{A}\backslash A$ is the boundary of the convex set. Let x be a point of the boundary. A unit vector u applied at x enters A if there is a $y \in A$ such that $u = (y - x)/\|y - x\|$. The set of all entering vectors cannot contain two antipodal elements so that there is a unit vector w such that $\langle w, u\rangle < 0$ for all entering unit vector. This argument leads to the proof of the following Isolation Theorem: *Let A be an open convex set in $\mathbb{R}^m$ and let x be in the border of A. There exists a unit vector w such that $\langle w, y - x\rangle < 0$ for all $y \in A$ that is, the half-space contains the convex set.*[5]

A function ϕ defined on $\mathbb{R}^n$ with values in $\overline{\mathbb{R}} = \mathbb{R} \cup \{+\infty\}$ is convex if the *epigraph* $\operatorname{epi}(\phi) = \{(x, t)|x \in \operatorname{dom}(\phi), t \in \mathbb{R}, \phi(x) \leqslant t\}$ is a convex subset of $\mathbb{R}^{n+1}$. We define $\operatorname{dom}(\phi)$ to be the set where ϕ takes finite values. *If ϕ is convex, then $\operatorname{dom}(\phi)$ is a convex subset of $\mathbb{R}^n$.* If $x_1, x_2 \in \operatorname{dom}(\phi)$, then there exist $(x_1, t_1), (x_2, t_2) \in \operatorname{epi}(\phi)$ and for all $\lambda \in [0, 1]$ it holds $((1 - \lambda)x_1 + \lambda x_2, (1 - \lambda)t_1 + \lambda t_2) \in \operatorname{epi}(\phi)$. In particular, $\phi((1-\lambda)x_1+\lambda x_2) < +\infty$. *If ϕ is convex, then $(1-\lambda)\phi(x_1)+\lambda\phi(x_2) \leqslant \phi((1-\lambda)x_1+\lambda x_2)$ for all $x_1, x_2 \in \mathbb{R}^n$ and $\lambda \in [0, 1]$.* If any of x_1, x_2 in nor in $\operatorname{dom}(\phi)$ the inequality is trivially satisfied. Otherwise, it is the same computation as above. Conversely, if $\phi\colon \operatorname{dom}(\phi) \to \mathbb{R}$ and $(1 - \lambda)\phi(x_1) + \lambda\phi(x_2) \leqslant \phi((1 - \lambda)x_1 + \lambda x_2)$ for all $x_1, x_2 \in \operatorname{dom}(\phi)$ and $\lambda \in [0, 1]$, then the function extended with value $+\infty$ outside the domain is convex.

Let ϕ be convex, and define the strict epigraph be open convex set

$$\{(x, t)|x \in \operatorname{dom}(\phi), t \in \mathbb{R}, \phi(x) < t\} \ .$$

Assume that at a point $(x, \phi(x))$ the entering unit vectors are not all horizontal. Then the Isolation Theorem implies that there exist at least a *supporting hyper-plane.* In such a case, ϕ on all such points ϕ *is the point-wise maximum of the supporting affine functions.* In the differentiable case, the tangent plane is the unique supporting hyperplane. If $\phi \in C^2(\mathcal{O})$ then the Hessian matrix is non-negative definite.

Let ϕ be convex and let ϕ be differentiable on an open $\mathcal{O}$. Then $\nabla\phi\colon \mathcal{O} \to \mathbb{R}^n$ is monotone *i.e., $\langle \nabla\phi(x) - \nabla\phi(y), x - y\rangle \geqslant 0$ for $x, y \in \mathcal{O}$.* We can re-write the basic inequality as

$$\lambda^{-1}(\phi(x + \lambda(y - x)) - \phi(x)) \leqslant \phi(y) - \phi(x) \ .$$

If $\lambda \to 0$.

$$\langle \nabla\phi(x), y - x\rangle \leqslant \phi(y) - \phi(x) \ .$$

By adding the inequality with x and y exchanged we obtain the monotonicity.

Conversely, *if ϕ is differentiable and monotone on an open set $\mathcal{O}$, then ϕ is convex on $\mathcal{O}$.* Write $z = (1 - \lambda)x + \lambda y$ and assume $0 < \lambda < 1$ because otherwise there is nothing to

[5] See a full proof in [1, p 45-46].

prove. observe that

$$\phi(z)-\phi(x)=\int_0^1\langle\nabla\phi(x+t(z-x)),z-x\rangle\ dt=$$
$$\int_0^1\langle\nabla\phi(x+t(z-x))-\nabla\phi(z),z-x\rangle\ dt+\langle\nabla\phi(z),z-x\rangle\leqslant$$
$$\langle\nabla\phi(z),z-x\rangle=\lambda\langle\nabla\phi(z),y-x\rangle\ .$$

In fact, $z-x$ and $(x+t(x-z))-z$ are proportional with factor $-(1-t)\leqslant 0$. In a similar way,

$$\phi(y)-\phi(z)=\int_0^1\langle\nabla\phi(z+t(y-z)),y-z\rangle\ dt=$$
$$\int_0^1\langle\nabla\phi(z+t(y-z))-\nabla\phi(z),y-z\rangle\ dt+\langle\nabla\phi(z),y-z\rangle\geqslant$$
$$\langle\nabla\phi(z),y-z\rangle=(1-\lambda)\langle\nabla\phi(z),y-x\rangle\ ,$$

as $y-z$ and $(z+t(y-z))-z$ are proportional with a factor $t\geqslant 0$. We rearrange the two inequalities as

$$\phi((1-\lambda)x+\lambda y)\leqslant\phi(x)+\lambda\langle\nabla\phi(z),y-x\rangle$$
$$\phi((1-\lambda)x+\lambda y)\leqslant\phi(y)+(1-\lambda)\langle\nabla\phi(z),y-x\rangle$$

and take the convex combination to conclude the proof.[6]

Exercise 7 (Examples of convex functions). Show that the following functions are convex and compute the gradient mapping if it exists.

(1) $\mathbb{R}^n\ni x\mapsto\sum_{j=1}^n|x_j|^a=\|x\|_a^a$, $a\geqslant 1$.
(2) $\mathbb{R}^n\ni x\mapsto\exp(\langle a,x\rangle)$, $a\in\mathbb{R}^n$.
(3) $\mathbb{R}^n\ni x\mapsto-\log(\langle a,x\rangle)$, $a\in\mathbb{R}^n$.
(4) $\mathbb{R}_+\ni x\mapsto x\log x$.

1.10. **Inegalities for the expectation.** If $u_1,\dots,u_n$ are real random variables, and u denotes the corresponding random variable with values in $\mathbb{R}^n$, then for each probability function p the vector $\mathbb{E}_p[u]=\sum_{\omega\in\Omega}p(\omega)u(\omega)$ is well defined. The operator $\mathbb{E}_p$ is linear and affine, namely for vector random variables u, v, reals α, β, and constant b, it holds

$$\mathbb{E}_p[\alpha u+\beta v+c]=\alpha\,\mathbb{E}_p[u]+\beta\,\mathbb{E}_p[v]+b\ .$$

The basic convexity inequality is Jensen Inequality. *Let p be a probability function and let u be a vector random variable. If ϕ is a convex function on C and C contains the image of u, then* $\mathbb{E}_p[\phi\circ u]\leqslant\phi(\mathbb{E}_p[u])$. Here are two proofs, both interesting. First, observe that the convexity inequality can be easily generalised to any number of terms,

$$\phi\left(\sum_{j=1}^n\lambda_jx_i\right)\leqslant\sum_{j=1}^n\phi(x_j)\ ,\quad\lambda_j\geqslant 0\ ,\sum_{j=1}^n\lambda_j=1\ ,$$

which is exactly the Jensen inequality written differently. Proof by recurrence. Second, let $x\mapsto a^tx+b$ be an affine function which is bounded by ϕ. Then $a^t\,\mathbb{E}_p[u]+b=\mathbb{E}_p[a^tu+b]\leqslant\mathbb{E}_p[\phi\circ u]$. Now take the supporting affine function at $\mathbb{E}_p[u]$ that is, choose a and b such that $a^t\,\mathbb{E}_p[u]+b=\phi(\mathbb{E}_p[u])$.

[6]This proof is taken from [4, p. 26]

The most common example of application is with $\phi(x) = \sum_{j=1}^m |x_j|^a = \|x\|_a^a$, $a \geqslant 1$. It follows that

$$\mathbb{E}_p\left[\sum_{j=1}^n |u_j|^a\right] \geqslant \sum_{j=1}^n \mathbb{E}_p\left[|u_j|\right]^a \ .$$

Another inequality of interest is the Hölder Inequality: *For all probability function p, all couple of random variables X and Y, and all couple of positive numbers a and b such that $1/a + 1/b = 1$, it holds*

$$\mathbb{E}_p\left[XY\right] \leqslant \mathbb{E}_p\left[|X|^a\right]^{1/a} \mathbb{E}_p\left[|Y|^b\right]^{1/b}$$

Exercise 8 (A proof of the Hölder inequality). Here is a proof involving computations of independent interest. From the convexity of $x \mapsto \mathrm{e}^x$, that is

$$\left(\mathrm{e}^u\right)^{1/a}\left(\mathrm{e}^v\right)^{1/b} = \mathrm{e}^{\frac{1}{a}u+\frac{1}{b}v} \leqslant \frac{1}{a}\mathrm{e}^u + \frac{1}{b}\mathrm{e}^v \ ,$$

we obtain

$$\mathbb{E}_p\left[\left(\mathrm{e}^u\right)^{1/a}\left(\mathrm{e}^v\right)^{1/b}\right] \leqslant \frac{1}{a}\mathbb{E}_p\left[\mathrm{e}^u\right] + \frac{1}{b}\mathbb{E}_p\left[\mathrm{e}^v\right] \ .$$

Let U, V be strictly positive random variables and define u and v by $\mathrm{e}^u = U^a/\mathbb{E}_p\left[U^a\right]$ and $\mathrm{e}^v = V^b/\mathbb{E}_p\left[V^b\right]$, respectively. Notice that now $\mathbb{E}_p\left[\mathrm{e}^u\right] = \mathbb{E}_p\left[\mathrm{e}^v\right] = 1$. The inequality above becomes

$$\mathbb{E}_p\left[\left(U^a/\mathbb{E}_p\left[U^a\right]\right)^{1/a}\left(V^b/\mathbb{E}_p\left[V^b\right]\right)^{1/b}\right] \leqslant \frac{1}{a}\mathbb{E}_p\left[\mathrm{e}^u\right] + \frac{1}{b}\mathbb{E}_p\left[\mathrm{e}^v\right] = 1 \ .$$

A little algebra produces the Hólder inequality for strictly positive random variable. Now consider $U+\epsilon$, $V+\epsilon$ and the limit $\epsilon \to 0$ to prove the inequality for non-negative random variables. Finally, take $U = |X|$ and $V = |Y|$ and observe that $XY \leqslant |X|\,|Y|$ to conclude the proof.[7]

Another classical inequality is the Minkovski Inequality: *For all probability function p, all couple of random variables X and Y, and $a \geqslant 1$, it holds*

$$\mathbb{E}_p\left[|X+Y|^a\right] \leqslant \mathbb{E}_p\left[|X|^a\right] + \mathbb{E}_p\left[|Y|^a\right] \ .$$

Minkovski inequality shows that $L(\Omega) \ni X \mapsto \mathbb{E}_p\left[|X|^a\right]^{1/a} = \|X\|_{p,a}$ *is a* norm *if p is strictly positive*. If p is not strictly positive, then it is a *semi-norm*.[8]

Exercise 9 (Proof of Minkovski inequality). The case $a = 1$ has an immediate proof. If $a > 1$ use $(X+Y)^a = X(X+Y)^{a-1} + Y(X+Y)^{a-1}$ and Hölder inequality. Notice that $1/a + 1/b = 1$ if and only if $b = a/(a-1)$.

Exercise 10 (L^2-convergence and Weak LLN). Consider the Bernoulli n-scheme and define $S_n = X_1 + \cdots + X_n$. Then $\mathbb{E}_\theta\left[S_n/n\right] = \theta$ and $\mathbb{E}_\theta\left[\left(\frac{S_n}{n} - \theta\right)^2\right] = \frac{1}{n}\theta(1-\theta) \to 0$ as $n \to \infty$.

Exercise 11 (Cramer inequality and Strong LLN). Let p be a probability function, X a real random variable, $c > 0$. For all $t > 0$,

$$P_p(X \geqslant c) = P_p(tX \geqslant ct)P_p(\mathrm{e}^{tX} \geqslant \mathrm{e}^{ct}) \leqslant \frac{1}{\mathrm{e}^{ct}}\mathbb{E}_p\left[\mathrm{e}^{tX}\right] = \exp\left(-\left(ct - \log\mathbb{E}_p\left[\mathrm{e}^{tX}\right]\right)\right) \ .$$

[7] This proof is taken from [2, §3.2.16]

[8] It is interesting to compare this statement with the corresponding statement as seen in Measure Theory

The function $\kappa\colon t \mapsto \log \mathbb{E}_p\left[\mathrm{e}^{tX}\right]$ is convex with

$$\kappa'(t) = \frac{\mathbb{E}_p\left[X\mathrm{e}^{tX}\right]}{\mathbb{E}_p\left[\mathrm{e}^{tX}\right]}$$

and

$$\kappa''(t) = \frac{\mathbb{E}_p\left[X^2\mathrm{e}^{tX}\right]\mathbb{E}_p\left[\mathrm{e}^{tX}\right] - \mathbb{E}_p\left[X\mathrm{e}^{tX}\right]^2}{\mathbb{E}_p\left[\mathrm{e}^{tX}\right]^2} = \frac{\mathbb{E}_p\left[(X - \mathbb{E}_p\left[X\right])^2\mathrm{e}^{tX}\right]}{\mathbb{E}_p\left[\mathrm{e}^{tX}\right]} > 0 \ .$$

To get the optimal inequality we look for

$$\sup_{t\geqslant 0} ct - \kappa(t) = \sup_{t\in\mathbb{R}} ct - \kappa(t) \ .$$

If $\hat{t}$ is the solution of $c\hat{t} = \kappa'(\hat{t})$, then

$$P_p(X \geqslant c) \leqslant \mathrm{e}^{-(c\hat{t} - \log \mathbb{E}_p\left[\mathrm{e}^{\hat{t}X}\right])}$$

In particular, if X is binomial, then

$$\mathbb{E}_p\left[\mathrm{e}^{tX}\right] = \sum_{k=0}^{n} \mathrm{e}^{tk}\binom{n}{k}\theta^k(1-\theta)^{n-k} = (\theta\mathrm{e}^t + (1-\theta))^n$$

so that

$$\kappa(t) = n\log\left(\theta\mathrm{e}^t + (1-\theta)\right) \ , \quad \kappa'(t) = n\frac{\theta\mathrm{e}^t}{\theta\mathrm{e}^t + (1-\theta)} \ .$$

The optimum value for the inequality is explicitly computable.[9].

2. Exponential expression of the open simplex $\Delta^\circ(\Omega)$

Exercise 12. The Bernoulli model

$$p(\omega;\theta) = \theta^{T(\omega)}(1-\theta)^{n-T(\omega)}$$

with $\theta \in]0,1[$, $(X_1(\omega),\dots,X_n(\omega)) = \omega \in \Omega = \{0,1\}^n$, $X_j(\omega) = x_j$, $T(\omega) = \sum_{j=1}^n X_j(\omega)$, can be written as

$$p(\omega;\theta) = \exp\left(\log\left(\frac{\theta}{1-\theta}\right)T(\omega) + n\log(1-\theta)\right) \qquad \theta \in]0,1[\ .$$

If $\alpha = \log\left(\frac{\theta}{1-\theta}\right)$, we can write the Bernoulli model in the form

$$p(\omega;\alpha) = \exp\left(\alpha T(\omega) - \kappa(\alpha)\right) \ , \quad \kappa(\alpha) = n\log\left(1+\mathrm{e}^\alpha\right) \ .$$

The function κ is strictly convex with

$$\kappa'(\alpha) = \frac{\mathrm{e}^\alpha}{1+\mathrm{e}^\alpha} = n\theta = \mathbb{E}_{p(\theta)}\left[T\right] \ ;$$
$$\kappa''(\alpha) = \frac{\mathrm{e}^\alpha}{(1+\mathrm{e}^\alpha)^2} = n\theta(1-\theta) = \mathbb{E}_{p(\theta)}\left[(T-n\theta))^2\right] \ ;$$
$$\kappa'''(\alpha) = \frac{\mathrm{e}^\alpha(1-\mathrm{e}^\alpha)}{(1+\mathrm{e}^\alpha)^3} = \theta(1-\theta)^2 - \theta^2(1-\theta) \ .$$

The log-likelihood

$$\alpha \mapsto \log p(\omega;\alpha) = \alpha T(\omega) - \kappa(\alpha)$$

[9]The relation with the Strong LLN appears when evaluating the dependence on n. To be discussed later

is strictly concave and the maximum obtains if

$$\frac{d}{d\alpha}\log p(\omega;\alpha) = T(\omega) - \kappa'(\alpha) = T(\omega) - n\theta = 0$$

that is, $\hat{\theta}(\omega) = T(\omega)/n$. The random variable $\hat{\theta}$ is the *maximum likelihood estimator* of the parameter θ. This estimator is *unbiased* because $\mathbb{E}_{p(\theta)}\left[\hat{\theta}\right] = \theta$ and it is *weakly consistent* because $\mathbb{E}_{p(\omega)}\left[\left(\hat{\theta}-\theta\right)^2\right] = \theta(1-\theta)/n \to 0$ if $n \to \infty$. This exercise provides the simplest example of classical Statistics and the simplest example of the exponential expression of a probability function.

2.1. **Positive probability functions.** In general, if the probability function $p : \Omega$ is positive, it is always possible to write it as $p(u) = \exp\left(U(\omega) - \kappa(U)\right)$, where U is a random variable and $\psi(U)$ is constant depending on U. In fact, if $\log p(\omega) = U(\omega) - \kappa(U)$, so then U is identified up to a constant and, for any given U,

$$1 = \sum_{\omega\in\Omega} p(\omega) = \mathrm{e}^{-\kappa(U)} \sum_{\omega\in\Omega} \mathrm{e}^{U(\omega)} \quad \text{so that,} \quad \kappa(U) = \log\left(\sum_{\omega\in\Omega} \mathrm{e}^{U(\omega)}\right) \ .$$

Exercise 13. Condider the binomial probability function $p(k) = \binom{n}{k}\theta^k(1-\theta)^{n-k}$, $k = 0,\dots,n$, $\theta \in]0,1[$, we can write

$$p(k) = \binom{n}{k}(1-\theta)^n\left(\frac{\theta}{1-\theta}\right)^k = \exp\left(\log\binom{n}{k} + k\log\left(\frac{\theta}{1-\theta}\right) + n\log(1-\theta)\right)$$

that is, $U(k) = \log\binom{n}{k} + k\log\left(\frac{\theta}{1-\theta}\right)$ and $\kappa(U) = -n\log(1-\theta)$.

It is convenient to change the parameter: if $\alpha = \log\left(\frac{\theta}{1-\theta}\right)$, then $\theta = \frac{\mathrm{e}^\alpha}{1+\mathrm{e}^\alpha}$ and

$$p(k) = \exp\left(\alpha k + \log\binom{n}{k} - n\log(1+\mathrm{e}^\alpha)\right) \ .$$

In fact, $U(k) = \alpha k + \log\binom{n}{k}$, hence

$$\kappa(U) = \log\left(\sum_{k=0}^{n} \mathrm{e}^{U(k)}\right) = \log\left(\sum_{k=0}^{n}\binom{n}{k}\alpha^k\right) = n\log(1+\mathrm{e}^\alpha) \ .$$

Such a way to express probability functions and the related formalism was initiated in Statistical Physics by J.W. Gibbs. Because of that, the function U is sometimes called a *potential*. In Physics the quantity $\sum_{\omega\in\Omega}\mathrm{e}^{U(\omega)}$ is called *partition function*.

The mapping $U \mapsto p = \mathrm{e}^{U-\kappa(U)}$ cannot be injective because the vector space of random variables has dimension $\#\Omega$ while the convex set of probability functions has dimension $\#\Omega - 1$. Precisely,

$$\mathrm{e}^{U(\omega)-\kappa(U)} = \mathrm{e}^{V(\omega)-\kappa(V)} \quad \Rightarrow U(\omega) - V(\omega) = \kappa(V) - \kappa(U) \ .$$

The function e^U is a generic positive function and the set of positive functions is a cone. A basis of this cone is the open probability simplex and the normalization is a projection onto this basis.

There are many ways to add a one-dimensional constrain to obtain a 1-to-1 function.

2.2. **Potentials in the space parallel to the simplex.** *For each positive probability function there is a unique potential U such that $\sum_{\omega\in\Omega} U(\omega) = 0$.* Assume $p = \exp(U - \kappa(U))$ with $\sum_{\omega\in\Omega} U(\omega) = 0$. Then

$$\sum_{\omega\in\Omega} \log p(\omega) = \sum_{\omega\in\Omega} U(\omega) - \kappa(\omega) = N\kappa(U) \ , \quad N = \#\Omega \ ,$$

that is, $\kappa(U) = \frac{1}{N}\sum_{\omega\in\Omega} \log p(\omega)$. Conversely, given any positive probability function, we can define $U = \log p - \frac{1}{N}\sum_{\omega\in\Omega} \log p(\omega)$ so that, $\sum_\omega U(\omega) = 0$. Moreover,

$$p = \exp(\log p) = \exp\left(U + \frac{1}{N}\sum_{\omega\in\Omega} \log p(\omega)\right) = \exp(U - \kappa(U))$$

with $k(U) = -\frac{1}{N}\sum_{\omega\in\Omega} \log p(\omega)$. In conclusion: *Let B_0 denote the vector space parallel to the simplex. The mapping $B_0 \ni U \mapsto \mathrm{e}^{U-k(U)}$ with $\kappa(U) = -\frac{1}{N}\sum_{\omega\in\Omega} \mathrm{e}^{U(\omega)}$ is 1-to-1. The inverse of this mapping is*

$$\Delta^\circ(\omega) \ni p \mapsto \log p - \frac{1}{N}\log p(\omega) \ .$$

2.3. **Potential centered at the probability function.** Consider now the mapping

$$\Delta^\circ(\Omega) \ni p \mapsto V = \log p - \mathbb{E}_p[\log p] \ .$$

Notice that

$$-\mathbb{E}_p[\log p] = -\sum_{\omega\in\Omega} p(\omega)\log p(\omega) = H(p)$$

is the entropy of p, that is in this case $p = \mathrm{e}^{V+H(p)}$.

2.4. **Non-negative potential.** Consider the set $\mathcal{U}$ of all non-negative real functions $U : \Omega$ such that $\min U = 0$. Notice the peculiar shape of such a sub-set on $\mathbb{R}^\Omega$: it is a pointed non-convex cone that is, if $U \in \mathcal{U}$ then $\rho U \in \mathcal{Y}$ for all $\rho \geqslant 0$ and moreover it is contained in the half-space associate to B_0.

The expression is unique, because of all the U's such that $p = \mathrm{e}^{U-\kappa(U)}$ only one belongs to $\mathcal{U}$.

In this expression, if $\Omega_0 = \{\omega \in \Omega | U(\omega) = 0\}$ and $\Omega_+ = \{\omega \in \Omega | U(\omega) > 0\}$, then

$$\kappa(U) = \log \sum_{\omega\in\Omega} \mathrm{e}^{U(\omega)} = \log\left(\#\Omega_0 + \sum_{\omega\in\Omega_+} \mathrm{e}^{U(\omega)}\right)$$

and

$$p(\omega) = \mathrm{e}^{U(\omega)-\kappa(U)} = \begin{cases} \dfrac{1}{\#\Omega_0 + \sum_{\omega\in\Omega_+} \mathrm{e}^{U(\omega)}} & \text{if } \omega \in \Omega_0, \\ \dfrac{\mathrm{e}^{U(\omega)}}{\#\Omega_0 + \sum_{\omega\in\Omega_+} \mathrm{e}^{U(\omega)}} & \text{if } \omega \in \Omega_+, \end{cases}$$

The previous expression allows to compute limit cases e.g., $\lim_{t\to\infty} \mathrm{e}^{tU-k(tU)}$.

Next,

3. Independence and conditioning

4. The infinite Bernoulli scheme

References

[1] Alexander Barvinok, *A course in convexity*, Graduate Studies in Mathematics, vol. 54, American Mathematical Society, Providence, RI, 2002.

[2] Didier Dacunha-Castelle and Marie Duflo, *Probabilités et statistiques. tome 1: Problèmes à temps fixe*, Collection Mathématiques Appliqées pour la Matrise, Masson, 1982.

[3] Ian J. Goodfellow, Yoshua Bengio, and Aaron Courville, *Deep learning*, MIT Press, Cambridge, MA, USA, 2016, `http://www.deeplearningbook.org`.

[4] R. Tyrrell Rockafellar, *Convex analysis*, Princeton Mathematical Series, No. 28, Princeton University Press, 1970. MR MR0274683 (43 #445)

[5] Sheldon M. Ross, *Introduction to Probability Models*, 10th ed., Academic Press, 2010.

Collegio Carlo Alberto Room 203a

E-mail address: `giovanni.pistone@carloalberto.org`

URL: `https://www.giannidiorestino.it/`