PROBABILITY 2020: PART 1
PROBABILITY ON A FINITE SAMPLE SPACE

GIOVANNI PISTONE

Contents

1. Probability functions

1.1. **Convex sets.** Convex analysis is an important topic in applied probability. A standard reference is the monographs [1].

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 $x_0, \ldots, x_n \in V$ 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 $x_0, \ldots, x_n \in V$ the set of all $\lambda_0 x_0 + \cdots + \lambda_n x_n$ with $\lambda_0 + \cdots + \lambda_n = 1$ 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

Date: DRAFT March 31, 2020.

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).

1.2. **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 $e_j - e_m$, $j = 1, \dots, N-1$ 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 $\mathbb{R}^n \ni e_n \mapsto 0 \in \mathbb{R}^{n-1}$.

Example 1. 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/` .

Example 2. 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`.]

1.3. **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 $d(f \circ g)(x) = df(g(x)) \circ dg(x)$.

1.4. **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 $\lambda(\bar{\omega}) = 0$.

1.5. **Aside: convex functions.** If a convex set $A \in \mathbb{R}^m$ is open, then every straight line intersects A 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.*[1]

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 is not 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 convex set $\mathcal{O}$. 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

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

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 $\nabla\phi$ is monotone on an open convex 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 prove. observe that

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

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

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.[2]

Example 3 (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) $\mathbb{R}_+ \ni x \mapsto x\log x$.

1.6. **Inequalities for the expectation.** If $u_1, \ldots, u_n$ 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,

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

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

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

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

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*

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

that $1/a + 1/b = 1$*, it holds*

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

Example 4 (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

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

we obtain

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

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

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

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.[3]

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*

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

Minkovski inequality shows that $L(\Omega) \ni X \mapsto \mathbb{E}_p[|X|^a]^{1/a} = \|X\|_{p,a}$ *is a* norm *if* p *is strictly positive*. If p is not strictly positive, then it is a *semi-norm*.[4]

Example 5 (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 $b = a/(a - 1)$.

Example 6 (L^2-convergence and Weak LLN). Consider the Bernoulli n-scheme and define $S_n = X_1 + \cdots + X_n$. Then $\mathbb{E}_\theta[S_n/n] = \theta$ and $\mathbb{E}_\theta\left[\left(\frac{S_n}{n} - \theta\right)^2\right] = \frac{1}{n}\theta(1 - \theta) \to 0$ as $n \to \infty$.

Example 7 (Cramer inequality and Strong LLN). Let p be a probability function, X a real random variable, $c > 0$. For all $t > 0$,

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

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

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

[3] This proof is taken from [3, §3.2.16]

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

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.[5].

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

Every positive probability function is of the form $p(\omega) = \mathrm{e}^{V(\omega)}$. This simple remark is frequently used in many applications as it provides a way to avoid inequality contrains. We start with some examples.

Example 8 (The Bernoulli model as an exponential family). 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

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

For each ω the function $\theta \mapsto p(\omega;\theta)$ is called *likelihood* of ω.

The parameter $\theta \in]0,1[$ is the value of a probability or expected value, $\theta = P_p(X_j = 1) = \mathbb{E}_p[X_j]$. The new parameter $o = \theta/(1-\theta)$, $\theta = o/(1+o)$, represents the *odds*, and $o \in]0,+\infty[$. WE use the parameter log-*odds*, $\alpha = \log\left(\frac{\theta}{1-\theta}\right)$, $\theta = \mathrm{e}^\alpha/(1+\mathrm{e}^\alpha)$, $\alpha \in \mathbb{R}$ so that 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

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

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

$$\kappa'''(\alpha) = n\frac{\mathrm{e}^\alpha(1-\mathrm{e}^\alpha)}{(1+\mathrm{e}^\alpha)^3} = \theta(1-\theta)^2 - \theta^2(1-\theta) \ .$$

The log-*likelihood* at ω is

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

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

It is strictly concave with

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

in particular, $\ell'(\omega;0+) = T(\omega)$ and $\ell(+\infty) = -\infty$. The maximum obtains at $\hat{\theta}(\omega)$ such that $T(\omega) = \hat{\theta}(\omega)$ 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

$$\mathbb{P}_{p(\theta)}\left(\left|\hat{\theta}-\theta\right| \geqslant \epsilon\right) \leqslant \epsilon^{-2}\,\mathbb{E}_{p(\theta)}\left[\left(\hat{\theta}-\theta\right)^2\right] = \epsilon^{-2}\frac{\theta(1-\theta)}{n^2} \to 0 \quad \text{if} \quad n \to \infty.$$

The behaviour of the standardized error i.e., the CLT will be discussed later.

The exercise above provides the simplest example of classical Statistics and the simplest example of the expression of a parametrized probability function as an *exponential family*. The next exercise shows the use of weight functions.

Example 9. Consider the binomial probability function

$$p(k;\theta) = \binom{n}{k}\theta^k(1-\theta)^{m-k} \ , \quad k \in \{0,1,\ldots,n\} \ , \quad \theta \in]0,1[\ .$$

Here, the binomial factor does not depend on the parameter. It is convenient to consider the function $k \mapsto \binom{n}{k}$ as a *weight function* on the sample space and write the exponential expression as

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

Now, the interesting factor is the *density* with respect to the binomial weight, $\mathrm{e}^{\alpha K - \kappa(\alpha)}$, α being the log-odds.

A similar model has been considered in Statistical Physics a long time before its use in Statistics. In the next example, we use an *un-normalised probability function* e.i., a function $f\colon \Omega \to \mathbb{R}_+$. Given such a function, one computes the *normalizing constant* $Z = \sum_{\omega\in\Omega} f(\omega)$ and $p = f/Z$ is a probability function. The set of all un-normalized probability function is a *pointed cone* and the normalization is a mapping from the cone to the probability simplex.

Example 10 (Gibbs distribution). If Ω is a finite set of states of a physical system, and $\omega \mapsto U(\omega)$ is a non-negative function whose value is the energy of the state ω, the probability function

$$p(\omega;t) \propto f(\omega) = \exp\left(-\frac{U(\omega)}{t}\right) \ , \ t > 0 \ ,$$

provides a probability on the set of states Ω which is called *Gibbs distribution*. C.f. [4, §28]. The parameter t represents the absolute temperature.

The normalising constant is

$$Z(t) = \sum_{\omega\in\Omega}\exp\left(-\frac{U(\omega)}{t}\right) \ ,$$

so that the Gibbs probability function is

$$p(\omega;t)=\frac{\exp\left(-\frac{U(\omega)}{t}\right)}{\sum_{\omega\in\Omega}\exp\left(-\frac{U(\omega)}{t}\right)}=\exp\left(-\frac{U(\omega)}{t}-\log Z(t)\right)\ .$$

One has

$$\frac{d}{dt}\log Z(t)=\frac{\sum_{\omega\in\Omega}\exp\left(-\frac{U(\omega)}{t}\right)\frac{U(\omega)}{t^2}}{Z(t)}=\frac{1}{t^2}\sum_{\omega\in\Omega}U(\omega)p(\omega;t)=\frac{1}{t^2}\,\mathbb{E}_{p(t)}\left[U\right]$$

and

$$\frac{d}{dt}\log p(\omega;t)=\frac{1}{t^2}U(\omega)-\frac{1}{t^2}\,\mathbb{E}_{p(t)}\left[U\right]\ .$$

Other equations similar to those we have obtained for the Bernoulli distribution hold. This provides the basic formalism for this physical model. For example, the deviation of the energy from its mean value is

$$U-\mathbb{E}_{p(t)}\left[U\right]=t^2\frac{d}{dt}\log p(t)\ .$$

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 $\log p(\omega)=U(\omega)-\kappa(U)$, then U is identified up to a constant and, for any given U,

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

Example 11. 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

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

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

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

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

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

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

Such a way to express probability functions and the related formalism was initiated in Statistical Physics by J.W. Gibbs (1901).

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 base 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 constraint 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*

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

A similar, but more general and more useful computation, considers the case of an exponential density with respect to a weight function.

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., $\lim_{t\to\infty} \mathrm{e}^{tU-k(tU)}$.

Example 12 (Limit cases of the Gibbs distribution). If the energy in the Gibbs model is zero at some states Ω_0, it is possible to use the expression above to compute the limit of the probabilities as $t \to 0$ and $t \to \infty$.

3. Independence and conditioning

When the sample space has a factorial structure, $S = S_1 \times \cdots S_n$, we define the marginal projections $X_j \colon S \ni x = (x_1, \ldots, x_n) \mapsto x_j$. If $\gamma \in \Delta(S)$ we say that the probability function γ provides the joint distribution of the marginal projection that is,

$$x \mapsto \mathbb{P}_\gamma \left(X_1 = x_1, \ldots, X_n = x_n \right) = \gamma(x_1, \ldots, x_n) \ .$$

Given any $I \subset \{1, \ldots, n\}$, the I-marginal joint distribution is

$$x_I \mapsto \mathbb{P}_\gamma \left(X_I = x_I \right) = \sum_{x \colon X_I(x) = x_I} \gamma(x) = \gamma_I(x_I) \ .$$

In particular, the marginal distributions are the probability functions

$$x_j \mapsto \gamma_j(x_j) = \mathbb{P}_\gamma \left(X_j = x_j \right) = \sum_{x | X_j(x) = x_j} \gamma(x) \ .$$

In a slightly more general set-up, we have a sample space Ω, a probability function $p \in \Delta(\Omega)$, and n random variables $Y_j \colon \Omega \to S_j$. The image of p under $Y = (Y_1, \ldots, Y_n)$ is a probability function $p_Y = \lambda \in \Delta(S)$, $S = S_1 \times \cdots \times S_n$, and the above discussion applies.

For a generic real random variable $Z \colon \Omega$ we have $\mathbb{E}_p \left[Z \right] = \sum_{\omega \in \Omega} Z(\omega) p(\omega)$. The real random variables of the form $Z = \phi(Y)$ are said to be Y-measurable. In such a case, it is easy to verify the fundamental *change of variable* equation

$$\mathbb{E}_p \left[\phi(Y) \right] = \mathbb{E}_{p_Y} \left[\phi \right] \ .$$

3.1. **2 factors.** In case of two factors, let us write X, Y for the two marginal projection and μ, ν for the marginal probability function.

A *transition function* is a mapping

$$P \colon S = S_1 \times S_2 \ni (x, y) \mapsto P(y|x) \in [0, 1]$$

such that, for each fixed x, $y \mapsto P(y|x)$ is a probability function. If μ is a probability function on the first factor S_1, then $(x, y) \mapsto P(y|x)\mu(x)$ is a joint probability function. Conversely, given any joint probability function γ with margins μ and ν, there exists transition functions P and Q such that

$$\gamma(x, y) = P(y|x)\mu(x) = Q(x|y)\nu(y) \ .$$

Notice that $\mu(x) \neq 0$ implies $P(y|x) = \gamma(x, y)/\mu(x)$. Otherwise, $\mu(x) = 0$ implies $\gamma(x, y) = 0$ for each y and $P(.|x)$ is any probability function. Assume both the marginal probability functions μ and ν are both positive. Then

$$Q(x|y) = \frac{P(y|x)\mu(x)}{\nu(y)}$$

is called *Bayes formula*.

As for each x the function $y \mapsto P(y|x)$ is a probability function, we can compute the expectation of a real random variable $f \colon \S_2 \to \mathbb{R}$ with respect to the transition probability function as $\mathbb{E}_{P(\cdot|x)} \left[f \right] = \sum_{y \in \Omega_2} f(y) P(y|x)$.

More generally, if $f \colon S \to \mathbb{R}$, the mapping

$$S \ni (x, y) \mapsto \mathbb{E}_{P(\cdot|x)} \left[f(x, \cdot) \right]$$

is a real random variable on Ω which depends on x only and is called conditional expectation.

3.2. **General conditional expectation.** In the case of generic random variables, we have a sample space Ω with a probability function p and random variables $X\colon \Omega \to S_1$, $Y\colon \Omega \to S_2$. Let us consider the image p_{XY} and the representation

$$p_{XY}(x,y) = P(y|x)p_X(x) = Q(x|y)p_Y(y)\ .$$

p_{XY} is the joint probability function; p_X is the probability function of X and p_Y is the probability function of Y; P is the probability function of Y given X; Q is the probability function of X given Y.

If one of the conditional probability functions is constant say, $P(y|x) = P(y)$, then $\gamma(x,y) = p_X(x)p_Y(y)$ and we say that Y and Y are independent.

Let $Z = \phi(X,Y)$ be a real random variable in $L(X,Y)$. Then $\mathbb{E}_p[Z] = \mathbb{E}_{p_{XY}}[\phi]$ because of the change of variable formula. Consider the function

$$\hat{\phi}\colon x \mapsto \mathbb{E}_{P(\cdot|x)}[\phi(x,\cdot)] = \sum_{y\in S_2} \phi(x,y)P(y|x)\ .$$

and define

$$\mathrm{E}_p(\phi(X,Y)|X) = \hat{\phi}(X)$$

If X and Y are independent, then the conditional expectation is constant and equal to the expectation.

The conditional expectation computed above is characterized by the following two defining properties.

(1) $\mathrm{E}_p(Z|X)$ *is a function of X; and*
(2) $\mathbb{E}_p[ZY] = \mathbb{E}_p[\mathrm{E}_p(Z|X)\,Y]$ *for all Y which is a function of X .*

It follows that $\mathrm{E}_p(Z\phi(X)|X) = \phi(X)\,\mathrm{E}_p(Z|x)$.

Example 13. In the Bernoulli model with marginal projections X_j, $j = 1,\dots,n$, show that X_I is independent of X_J, $I \cap J = \varnothing$. Compute the marginal and joint distribution of $S_n = \sum_{j=1}^n X_j$ and $T_n = \inf\{k = 1,\dots,n \,|\, X_k = 1\}$. Compute all the relevant conditional quantities.

3.3. **Using matrices and tables.** A probability function γ on $S_1 \times S_2$ is commonly represented as a matrix $\Gamma = [\gamma(x,y)] \in \mathbb{R}^{S_1\times S_2}$. In this representation, the row vectors $\mu^t = \Gamma\mathbf{1}$ and $\nu = \mathbf{1}^t\Gamma$ represent the marginal probability functions as in

$$\Gamma\mathbf{1} = \begin{bmatrix} \gamma(1,1) & \gamma(1,2) & \gamma(1,3) \\ \gamma(2,1) & \gamma(2,2) & \gamma(2,3) \\ \gamma(3,1) & \gamma(3,2) & \gamma(3,3) \end{bmatrix} \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix} = \begin{bmatrix} \mu(1) \\ \mu(2) \\ \mu(3) \end{bmatrix}$$

Transition functions are commonly represented as matrices with elements $P(y|x)$, x being the row-index. When $S = S_1 = S_2$, the transition matrix is called Markov matrix i.e., a Markov matrix is a square matrix P with non-negative elements such that $P\mathbf{1} = \mathbf{1}$. Notice that this relation provides an eigen-value and an eigen-vector of P. If the first marginal probability function μ is represented as a row vector then the row vector $\nu = \mu P$ is the other marginal probability function of the joint distribution.

The probability matrix Γ, the marginal probability vectors μ and ν, and the two transition matrices P and Q are related as matrices by the equations

$$\mu = \mathbf{1}^t\Gamma^t, \quad \nu = \mathbf{1}^t\Gamma, \quad \Gamma = \mathrm{diag}(\mu)\,P = Q^t\,\mathrm{diag}(\nu)$$

for example,

$$\begin{bmatrix} \gamma(1,1) & \gamma(1,2) & \gamma(1,3) \\ \gamma(2,1) & \gamma(2,2) & \gamma(2,3) \\ \gamma(3,1) & \gamma(3,2) & \gamma(3,3) \end{bmatrix} = \begin{bmatrix} \mu(1) & 0 & 0 \\ 0 & \mu(2) & 0 \\ 0 & 0 & \mu(3) \end{bmatrix} \begin{bmatrix} P(1|1) & P(2|1) & P(3|1) \\ P(1|2) & P(2|2) & P(3|2) \\ P(1|3) & P(2|3) & P(3|3) \end{bmatrix}$$

Example 14 (Reversibility). A probability function γ on $S^2 = S \times S$ is reversible (or, symmetric) if $\gamma(x,y) = \gamma(y,x)$ that is, if $\Gamma = \Gamma^t$. When γ is reversible, then the two margins are equal, $\mu = \mathbf{1}^t\Gamma^t = \mathbf{1}^t\Gamma = \nu$. Moreover, $\operatorname{diag}\mu P = \operatorname{diag}\mu Q$, so that $P = Q$ when $\mu > 0$. If $N = \#S$, then a generic joint probability function belongs to the (N^2-1)-simplex $\Delta(S^2)$. The set of reversible probability function is the subset defined by the $\binom{N}{2} = N(N-1)/2$ symmetry constraints then has $(N^2-1)-N(N-1)/2 = N(N+1)/2-1$ degrees of freedom. It is a bounded polyhedron hence a polytope. It is interesting to find its $\binom{N}{2} = N(N-1)/2$ vertexes. Map all off-diagonal positions (x,y) into the set $\{x,y\}$. Consider $\alpha(c)$, $c \in \binom{S}{2}$, such that $\alpha(c) \geqslant 0$, $\sum_c \alpha(c) \leqslant 1$. This provides a parametrization of the off-diagonal elements of of Γ. Then split the remaining mass on the diagonal.

Example 15 (Earth-mover problem). A transition function P from S_1 to S_2 could be seen as a rule to move a fraction $P(y|x) = P_{x,y}$ of the mass at $x \in S_1$ into the position $y \in S_2$. In this way, a total mass μ on S_1 is moved in a mass $\nu = \mu P$ on S_2. There is a joint distribution $\gamma(x,y) = P(y|x)\mu(x)$ giving the mass from x that is moved to y. The earth-mover has given initial μ and final ν and looks for a feasible transport plan P. The set of feasible transport plans is convex and is better represented as a convex subset of $\Delta(S_1 \times S_2)$ via the joint distribution. Assume the transport from x to y has a cost $c(x,y)$ for the earth-mover. The total cost of the transport plan is $\sum_{x,y} c(x,y)P(y|x)\pi(x)$. The constrained optimization problem has an elementary solution in some cases e.g., $S_1 = S_2 = S$, $\#S = 3$, and the cost is a distance.

3.4. n **factors and Markov chains.** The case were there is a finite number of random variables does not present any special new feature.

Consider a finite set S and the sample space $\Omega = S^{n+1}$, $\Omega \ni \omega = (x_0, x_1, \dots, x_n)$, with marginal projections $X_j(\omega) = x_j$. The sequence $I = (0, \dots, n)$ is thought of as a sequence of times and the sequence $\omega = (x_0, x_1, \dots, x_n)$ is a trajectory of something evolving in S.

Given a probability function $\pi_0 \colon S$ and a transition function P on S, we define the joint probability function

$$\gamma(x_1, \dots, x_n) = \left(\prod_{j=1}^{n} P(x_j|x_{j-1}) \right) \pi_0(x_0) = \pi_0(x_0) \prod_{j=1}^{n} P_{x_{j-1},x_j} \ .$$

Notice the probabilistic notation (middle side) and the matrix notation (right hand side).

The structure $(\Omega = S^{n+1}, \gamma, (X_j)_{j=0}^n)$ is a (canonical) *Markov chain with initial distribution* π_0 *and stationary transition probability* P. This is a constructive definition. An equivalent (non-canonical) definition shows the intrinsic property of such a structure:

$(\Omega, \mathbb{P}, (X_j)_{j=0}^n))$ *is a Markov chain with initial probability* π_0 *and stationary transition probability* P *if and only if*

(1) $X_0 \sim \pi_0$.
(2) *For each* $k = 1, \dots, n$ *it holds*

$$\mathbb{P}\left(X_k = x_k | X_{k-1} = x_{k-1}, \dots X_0 = x_0\right) = P(x_k|x_{k-1}) \ .$$

The notation $X_0 \sim \pi_0$ means that the image of the base probability with X_0 has probability function π_0.The proof of equivalence is a simple algebraic check.

The distribution of each X_k is, using the matrix notation for transitions,

$$\pi_k(x_k) = \mathbb{P}_\gamma\left(X_k = x_k\right) = \sum_{x_0,\ldots,x_{k-1}} \sum_{x_{k+1},\ldots,x_n} \pi_0(x_0) \prod_{j=1}^{n} P(x_{j-1}, x_j) =$$

$$\sum_{x_0,\ldots,x_{k-1}} \pi_0(x_0) \prod_{j=1}^{k} P(x_{j-1}, x_j) = \sum_{x_0} \pi_0(x_0) \sum_{x_1,\ldots,x_{k-1}} \prod_{j=1}^{k} P(x_{j-1}, x_j) =$$

$$\pi_0 P^k(x_k)$$

An important special case appears when the initial probability function is *invariant*, $\sum_x P(y|x)\pi_0(x) = \pi_0$, or $\pi_0 = \pi_0 P$. Note that π_0 is a left eigen-value of P and $\pi_k = \pi_0$.

Example 16 (Binary Markov chain). Let

$$P = \begin{bmatrix} 1-\alpha & \alpha \\ \beta & 1-\beta \end{bmatrix} \quad \alpha, \beta \in [0, 1]$$

be the generic Markov matrix on the two elements set $S = \{0, 1\}$. An invariant probability function is $\pi = \begin{bmatrix} p & 1-p \end{bmatrix}$ such that

$$\begin{cases} p = p(1-\alpha) + (1-p)\beta \\ 1-p = p\alpha + (1-p)(1-\beta) \end{cases} .$$

The two equations are dependent because the rank of $P - I$ is 1. It follows

$$p(\alpha + \beta) = \beta \quad \text{and} \quad (1-p)(\alpha+\beta) = \alpha \ .$$

If $\alpha+\beta = 0$ i.e., $P = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$, then all probability functions are invariant. In the following, we assume $\alpha + \beta > 0$. In such a case, the invariant probability function is

$$\pi = \begin{bmatrix} \frac{\beta}{\alpha+\beta} & \frac{\alpha}{\alpha+\beta} \end{bmatrix} .$$

For example, the invariant probability of both $P = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}$ and $P = \begin{bmatrix} 1/2 & 1/2 \\ 1/2 & 1/2 \end{bmatrix}$ is $\pi = \begin{bmatrix} \frac{1}{2} & \frac{1}{2} \end{bmatrix}$. Note that the first example produces a "deterministic" process while the second produces and "independent" process.

The characteristic equation of P is

$$\det\left(P - \lambda I\right) = \det \begin{bmatrix} (1-\alpha) - \lambda & \alpha \\ \beta & (1-\beta) - \lambda \end{bmatrix} = \lambda^2 - (2 - \alpha - \beta)\lambda + (1 - \alpha - \beta) = 0 \ .$$

One solution is $\lambda_1 = 1$, while the other is $\lambda_2 = 1 - \alpha - \beta$. The first eigen-vector is a vector

$$\boldsymbol{u}_1 = \begin{bmatrix} u_{11} \\ u_{21} \end{bmatrix} \quad \text{such that} \quad \begin{bmatrix} -\alpha & \alpha \\ -\beta & \beta \end{bmatrix} \begin{bmatrix} u_{11} \\ u_{21} \end{bmatrix} = 0$$

e.g., $\boldsymbol{u}_1 = \begin{bmatrix} 1 & 1 \end{bmatrix}^*$. The second eigen-vector is a vector

$$\boldsymbol{u}_2 = \begin{bmatrix} u_{12} \\ u_{22} \end{bmatrix} \quad \text{such that} \quad \begin{bmatrix} \beta & \alpha \\ \beta & \alpha \end{bmatrix} \begin{bmatrix} u_{12} \\ u_{22} \end{bmatrix} = 0$$

e.g., $\boldsymbol{u}_2 = \begin{bmatrix} -\alpha & \beta \end{bmatrix}^*$. It follows that

$$P = U \begin{bmatrix} 1 & 0 \\ 0 & 1 - \alpha - \beta \end{bmatrix} U^{-1} \quad \text{with} \quad U = \begin{bmatrix} \boldsymbol{u}_1 & \boldsymbol{u}_2 \end{bmatrix} = \begin{bmatrix} 1 & -\alpha \\ 1 & \beta \end{bmatrix}$$

because $\det U = \alpha + \beta > 0$. It follows that

$$P^n = U \begin{bmatrix} 1 & 0 \\ 0 & (1-\alpha-\beta)^n \end{bmatrix} U^{-1} \ .$$

Assume $\alpha + \beta \neq 2$, so that $-1 < 1 - \alpha - \beta < 1$. We have

$$\lim_{n\to\infty} P^n =$$

$$U \lim_{n\to\infty} \begin{bmatrix} 1 & 0 \\ 0 & (1-\alpha-\beta)^n \end{bmatrix} U^{-1} = \frac{1}{\alpha+\beta} \begin{bmatrix} 1 & -\alpha \\ 1 & \beta \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} \beta & \alpha \\ -1 & 1 \end{bmatrix} =$$

$$\begin{bmatrix} \frac{\beta}{\alpha+\beta} & \frac{\alpha}{\alpha+\beta} \\ \frac{\beta}{\alpha+\beta} & \frac{\alpha}{\alpha+\beta} \end{bmatrix} = \begin{bmatrix} \pi \\ \pi \end{bmatrix} \ .$$

In conclusion: if $\alpha + \beta = 0$ all probability functions are invariant and there is no convergence to the invariant probability; If $\alpha + \beta > 0$ there is a unique invariant probability. If moreover $\alpha + \beta \neq 2$ there is convergence. Another interesting case happens when $\alpha = 0$ while $\beta > 0$.

3.5. **Aside: graphs.** A *relation* on a set V is a subset $\mathcal{E}$ of $V \times V$ and its indicator function presented as a matrix is its *adjacency matrix* E. The couple $(V, \mathcal{E})$ is called a *graph* with vertexes V and edges $\mathcal{E}$. A graph is *undirected* if the relation is symmetric. A *path* from x to y of length n is a sequence $x = x_0, \ldots, x_n = y$ such that $(x_{i-1}, x_i) \in \mathcal{E}$, $i = 1, \ldots, n$. A path is a cycle if $x = y$. A graph is *connected* if for all $x, y \in V$ there is a path from x to y. In a connected graph, the minimum length of a path connecting two vertexes v and w is a distance. A *tree* is a connected graph without cycles. A *rooted tree* is a tree with a distinguished vertex v_0. In a rooted tree $(V, \mathcal{E}, v_0)$ all vertexes are classified according their distance from the root. A *rooted directed tree* is a rooted tree where the relation is restricted in such a way that the distance from the root increases in the direction of the edges. This directed relation allows to define the *childs* and the *parent* of a vertex. The root has no parent. The *leaves* have no childs. A *situation* is a non-leaf vertex. *Outcome* is another name for leaf.

In a rooted directed tree each vertex of depth k can be coded as $v_0 v_1 \cdots v_k$ where v_j range in a set of codes for the k-layer. Such a tree is associated to a space of events as follows. Let Ω be the set of all leaves. Then each situation is naturally associated with a set of leaves. Notice that any vertex of the tree can be chosen as a root. Different root correspond to different event trees which correspond to different "causal explanations" of the leaves.

Example 17 (Probability tree). If interested, have a look to [2]. Each directed edge (v, w) of a rooted tree can be decorated with a probability $p(w|v) > 0$ in such a way that the sum of probabilities on each floret is 1, $\sum_{w \in C(v)} p(w|v) = 1$. For each leaf ω there is a unique path from the root and the product of all decorations provides a probability function on the set of leaves. The decoration are conditional probabilities for the corresponding descendant leaves. It is possible to consider dynamic models where individuals move down the tree at different speeds and, when do move, choose a child according to the assigned probabilities. All the construction is clarified on simple examples.

Example 18 (Conditional independence). The properties

$$\mathbb{P}(A|B \cap C) = \mathbb{P}(A|B) \ ,$$
$$\mathbb{P}(C|B \cap A) = \mathbb{P}(C|B) \ ,$$
$$\mathbb{P}(A \cap C|B) = \mathbb{P}(A|B)\,\mathbb{P}(C|B) \ ,$$

are equivalent. The property in the first two equations is called *sufficiency* and the property in the last equation is called *conditional independence*.

Example 19. The Markov property is symmetric in the direction of time. If $X_0, \dots, X_n$ is a MC, then the time-reversed process $Y_h = X_{n-h}$ is a Markov process with transitions

$$\mathbb{P}\left(Y_{h+1} = x | Y_h = y\right) = \mathbb{P}\left(X_{n-h-1} = x | X_{n-h} = y\right) = \frac{\mathbb{P}\left(X_{n-h-1} = x, X_{n-h} = y\right)}{\mathbb{P}\left(X_{n-h} = y\right)} = \frac{\mathbb{P}\left(X_{n-h} = y | X_{n-h-1} = x\right)\mathbb{P}\left(X_{n-h-1} = x\right)}{\mathbb{P}\left(X_{n-h} = y\right)} = \frac{P_{x,y}\pi_{n-h-i}(x)}{\pi_{n-h}(y)} \ .$$

If moreover the MC is stationary that is $\pi_t = \pi$, then the time-reversed process is a MC with the same invariant distribution and transitions

$$Q_{y,x} = \frac{\pi(x)P_{x,y}}{\pi(y)} \ .$$

Equivalently, we can say that the 2-dimensional distribution are given by

$$\mathbb{P}\left(X_s = x, X_{s+1} = y\right) = \pi(x)P_{x,y} = \pi(y)Q_{y,x} \ .$$

A stationary Markov chain is *reversible* if $Q_{y,x} = P_{x,y}$. Equivalently, if π is a probability function such that

$$\pi(x)P_{x,y} = \pi(y)P_{y,x} \ ,$$

we sum the previous relation over x to get

$$\sum_{x\in S} \pi(x)P_{x,y} = \pi(y)\sum_{x\in S} P_{y,x} = \pi(y) \ ,$$

so that π is indeed an invariant probability and the MC constructed from π and P is reversible.

Given a Markov matrix P, if there exists a positive function $\kappa\colon S$ such that $\kappa(x)P_{x,y} = \kappa(y)P_{y,x}$ then we can normalize κ. In such a case we have an immediate way to compute the invariant probability.

Example 20. Let $G = (S, \mathcal{E})$ be a graph. For each vertex $x \in S$ the *degree* of x, $\deg x$, is the number of edges from x. Let E be the *adjacency matrix* of G. The degree as a row vector is $E\mathbf{1}$. Define the Markov matrix

$$P = \operatorname{diag}\left(E\mathbf{1}\right)^{-1} E \ .$$

i.e., the transitions

$$P_{x,y} = \begin{cases} \frac{1}{\deg x} & \text{if } y \text{ is connected with } x, \\ 0 & \text{if } y \text{ is not connected with } x. \end{cases}$$

Observe that x is connected to y if, and only if, y is connected to x, hence

$$(\deg x)P_{x,y} = (x \to y) = (y \to x) = (\deg y)P_{y,x} \ .$$

It follows that the invariant probability is

$$\pi(x) = \frac{\deg x}{\sum_{y\in S} \deg y} \ .$$

and the MC is reversible.

Example 21 (Hastings-Metropolis). Consider the following problem: Given a Markov matrix Q on a finite S and a probability function π on S, define the matrix

$$P_{x,y} = \begin{cases} Q_{x,y}\alpha(x,y) & \text{if } x \neq y \\ Q_{x,x} + \sum_{z \neq x} Q_{x,z}(1 - \alpha(x,z)) & \text{if } x = y \end{cases} ,$$

where $0 \leqslant \alpha(x,y) \leqslant 1$ Notice that $P_{x,y} \geqslant 0$ and

$$\sum_{y \in S} P_{x,y} = \sum_{y \neq x} Q_{x,y}\alpha(x,y) + Q_{x,x} + \sum_{z \neq x} Q_{x,z}(1 - \alpha(x,z)) = Q_{x,x} + \sum_{z \neq x} Q_{x,z} = 1 .$$

The Markov matrix P is reversible with invariant probability π if

$$\pi(x)Q_{x,y}\alpha(x,y) = \pi(y)Q_{y,x}\alpha(y,x) , \quad x \neq y .$$

One possible choice is

$$\alpha(x,y) = 1 \wedge \frac{\pi(y)Q_{y,x}}{\pi(x)Q_{x,y}} .$$

References

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

[2] Rodrigo A. Collazo, Christiane Görgen, and Jim Q. Smith, *Chain event graph*, Computer Science and Data Analysis Series, CRC Press, 2018.

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

[4] Lev D. Landau and Eugenij M. Lifshits, *Course of theoretical physics. statistical physics.*, 3rd ed., vol. V, Butterworth-Heinemann, 1980.

[5] Paul Malliavin, *Integration and probability*, Graduate Texts in Mathematics, vol. 157, Springer-Verlag, 1995, With the collaboration of Héléne Airault, Leslie Kay and Gérard Letac, Edited and translated from the French by Kay, With a foreword by Mark Pinsky. MR MR1335234 (97f:28001a)

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

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

Collegio Carlo Alberto Room 203a
Email address: giovanni.pistone@carloalberto.org
URL: https://www.giannidiorestino.it/