Introduction to Bayesian Statistics - 3

Edoardo Milotti

Università di Trieste and INFN-Sezione di Trieste

At the end of the last lesson we found the following expression for the "Bayesian line fit"

$$p\left(a,b \mid \mathbf{y},\mathbf{x},\sigma\right)=\frac{\exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{N}\left(y_i-ax_i-b\right)^2\right]}{\left\{\int_{-\infty}^{+\infty}da\int_{-B/\cos\varphi}^{B/\cos\varphi}db\,\exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{N}\left(y_i-ax_i-b\right)^2\right]\cdot\frac{1}{\left(1+a^2\right)^{3/2}}\right\}}\cdot\frac{1}{\left(1+a^2\right)^{3/2}}$$

$$\approx\frac{\frac{1}{\left(1+a^2\right)^{3/2}}\exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{N}\left(y_i-ax_i-b\right)^2\right]}{\left\{\int_{-\infty}^{+\infty}\frac{da}{\left(1+a^2\right)^{3/2}}\int_{-\infty}^{+\infty}db\,\exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{N}\left(y_i-ax_i-b\right)^2\right]\right\}}$$

This expression has a partly Gaussian structure, and now we rearrange the quadratic expression in the exponential

$$
\begin{aligned}
\sum_{i=1}^{N}\left(y_i-ax_i-b\right)^2 &= \sum_{i=1}^{N}\left[\left(y_i-ax_i\right)^2-2b\left(y_i-ax_i\right)+b^2\right] \\
&= \sum_{i=1}^{N}\left(y_i-ax_i\right)^2-2b\sum_{i=1}^{N}\left(y_i-ax_i\right)+Nb^2 \\
&= N\left\{\left[b^2-2b\frac{1}{N}\sum_{i=1}^{N}\left(y_i-ax_i\right)+\left(\frac{1}{N}\sum_{i=1}^{N}\left(y_i-ax_i\right)\right)^2\right]+\frac{1}{N}\sum_{i=1}^{N}\left(y_i-ax_i\right)^2-\left(\frac{1}{N}\sum_{i=1}^{N}\left(y_i-ax_i\right)\right)^2\right\} \\
&= N\left\{\left(b-\frac{1}{N}\sum_{i=1}^{N}\left(y_i-ax_i\right)\right)^2+\frac{1}{N}\sum_{i=1}^{N}\left(y_i-ax_i\right)^2-\left(\frac{1}{N}\sum_{i=1}^{N}\left(y_i-ax_i\right)\right)^2\right\} \\
&= N\left(b-\frac{1}{N}\sum_{i=1}^{N}\left(y_i-ax_i\right)\right)^2+N\left(\frac{1}{N}\sum_{i=1}^{N}y_i^2-2a\frac{1}{N}\sum_{i=1}^{N}x_iy_i+a^2\frac{1}{N}\sum_{i=1}^{N}x_i^2\right)-N\left(\frac{1}{N}\sum_{i=1}^{N}y_i-a\frac{1}{N}\sum_{i=1}^{N}x_i\right)^2 \\
&= N\left(b-\frac{1}{N}\sum_{i=1}^{N}\left(y_i-ax_i\right)\right)^2+N\left(\operatorname{var}y-2a\operatorname{cov}(x,y)+a^2\operatorname{var}x\right)
\end{aligned}
$$

therefore the normalization integral becomes

$$
\begin{aligned}
&\int_{-\infty}^{+\infty}\frac{da}{\left(1+a^2\right)^{3/2}}\exp\left[-\frac{N}{2\sigma^2}\left(\operatorname{var}y-2a\operatorname{cov}(x,y)+a^2\operatorname{var}x\right)\right]\int_{-\infty}^{+\infty}db\,\exp\left[-\frac{N}{2\sigma^2}\left(b-\frac{1}{N}\sum_{i=1}^{N}\left(y_i-ax_i\right)\right)^2\right] \\
&= \sqrt{\frac{2\pi\sigma^2}{N}}\int_{-\infty}^{+\infty}\frac{da}{\left(1+a^2\right)^{3/2}}\exp\left[-\frac{N}{2\sigma^2}\left(\operatorname{var}y-2a\operatorname{cov}(x,y)+a^2\operatorname{var}x\right)\right]
\end{aligned}
$$

For the next step we use *Laplace's method* (this is the *saddle-point method* – also called the *method of steepest descent* – in the real domain) for the evaluation of the integral of a unimodal function

$$Z = \int_{-\infty}^{+\infty} p(x)dx = \int_{-\infty}^{+\infty} e^{\Phi(x)}dx$$

where

$$\Phi(x) = \ln p(x) \approx \ln p(x_0) - \frac{1}{2s}(x - x_0)^2$$

where x_0 is the modal value and

$$\frac{1}{s} = -\frac{\partial^2 \ln p(x)}{\partial x^2}$$

therefore

$$\boxed{Z \approx \int_{-\infty}^{+\infty} p(x_0)e^{-\frac{(x-x_0)^2}{2s}}dx = p(x_0)\sqrt{2\pi s}}$$

Approximate integration of the remaining integral

$$\int_{-\infty}^{+\infty} \frac{da}{\left(1+a^2\right)^{3/2}} \exp\left[-\frac{N}{2\sigma^2}\left(\operatorname{var} y - 2a\operatorname{cov}(x,y) + a^2 \operatorname{var} x\right)\right]$$

We evaluate this integral using Laplace's method.

As usual in this method, we start with the logarithm of the integrand, we find its maximum and we Taylor expand about the maximum

$$\Phi(a) = -\frac{3}{2}\ln\left(1+a^2\right) - \frac{N}{2\sigma^2}\left(\operatorname{var} y - 2a\operatorname{cov}(x,y) + a^2 \operatorname{var} x\right)$$

$$\Phi(a) = -\frac{3}{2}\ln\left(1+a^2\right) - \frac{N}{2\sigma^2}\left(\operatorname{var} y - 2a\operatorname{cov}(x,y) + a^2\operatorname{var} x\right)$$

$$\frac{d\Phi}{da} = -\frac{3a}{1+a^2} + \frac{N}{\sigma^2}\left(\operatorname{cov}(x,y) - a\operatorname{var} x\right) = 0$$

we find *a* from this cubic equation

note that when N>>1 the peak is at position $a_0 \approx \frac{\operatorname{cov}(x,y)}{\operatorname{var} x}$

We use the Newton-Raphson method for the solution of the cubic equation:

$$f(a_0) = -\frac{3a_0}{1+a_0^2}$$

$$f'(a_0) = -3\frac{1-a_0^2}{(1+a_0^2)^2} - \frac{N}{\sigma^2}\operatorname{var} x \approx -\frac{N}{\sigma^2}\operatorname{var} x$$

then

$$\delta a_1 = -\frac{3a_0}{1+a_0^2}\frac{\sigma^2}{N\,\mathrm{var}\,x} \qquad\qquad a_1 = a_0 - \frac{3a_0}{1+a_0^2}\frac{\sigma^2}{N\,\mathrm{var}\,x} \tag{1}$$

Now, to complete the expansion, we must evaluate the second derivative at a_1:

$$\frac{d^2\Phi}{da^2} = -3\frac{1-a_1^2}{(1+a_1^2)^2} - \frac{N}{\sigma^2}\mathrm{var}\ x = -\frac{1}{\sigma_1^2} \tag{2}$$

$$\Phi(a) \approx \Phi(a_1) + \frac{1}{2}\left.\frac{d^2\Phi}{da^2}\right|_{a_1}(a-a_1)^2 = \Phi(a_1) - \frac{(a-a_1)^2}{2\sigma_1^2}$$

we find this by using equations (1) and (2)

Now we complete the evaluation of the integral

$$\int_{-\infty}^{+\infty} \frac{da}{\left(1+a^2\right)^{3/2}} \exp\left[-\frac{N}{2\sigma^2}\left(\operatorname{var} y - 2a\operatorname{cov}(x,y) + a^2 \operatorname{var} x\right)\right]$$

$$= \int_{-\infty}^{+\infty} \exp\left[\Phi(a)\right] da$$

$$\approx \int_{-\infty}^{+\infty} \exp\left[\Phi(a_1) - \frac{\left(a-a_1\right)^2}{2\sigma_1^2}\right] da = \sqrt{2\pi\sigma_1^2}\exp\left[\Phi(a_1)\right]$$

and finally we find the posterior distribution.

Moreover

$$p(a,b\,|\,\mathbf{y},\mathbf{x},\sigma) \propto \frac{1}{\left(1+a^2\right)^{3/2}} \exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{N}\left(y_i - a x_i - b\right)^2\right]$$

$$\approx \exp\left[-\Phi(a_1) - \frac{\left(a-a_1\right)^2}{2\sigma_1^2}\right] \exp\left[-\frac{N}{2\sigma^2}\left(b - \frac{1}{N}\sum_{i=1}^{N}\left(y_i - a_1 x_i\right)\right)^2\right]$$

and thus we see that:

$$\langle a\rangle = a_1; \quad \operatorname{var} a = \sigma_1^2;$$

$$\langle b\rangle = \frac{1}{N}\sum_{i=1}^{N}\left(y_i - a_1 x_i\right); \quad \operatorname{var} b = \frac{\sigma^2}{N}$$

Prior distributions

The choice of prior distribution is an important aspect of Bayesian inference

- prior distributions are one of the main targets of frequentists: how much do posteriors differ when we choose different priors?
- there are two main “objective” methods for the choice of priors

1. Jeffreys' method
2. The Maximum Entropy Method

Random variable transformations and prior distributions

$$p_x(x)dx = p_x\left(x(y)\right)\left|\frac{dx}{dy}\right| dy = p_y(y)dy$$

$$\Rightarrow \quad p_y(y) = p_x\left(x(y)\right)\left|\frac{dx}{dy}\right|$$

In general, if the first pdf is uniform, the other one is not.

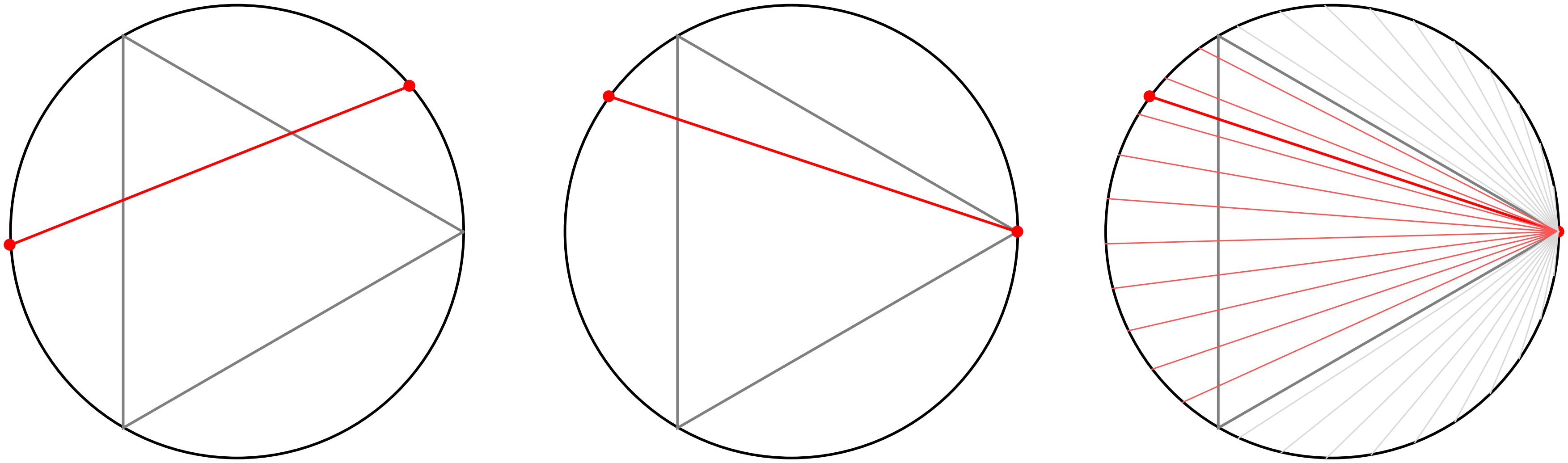
How can we "objectively" choose a prior distribution???

Bertrand’s paradox and the ambiguities of probability models

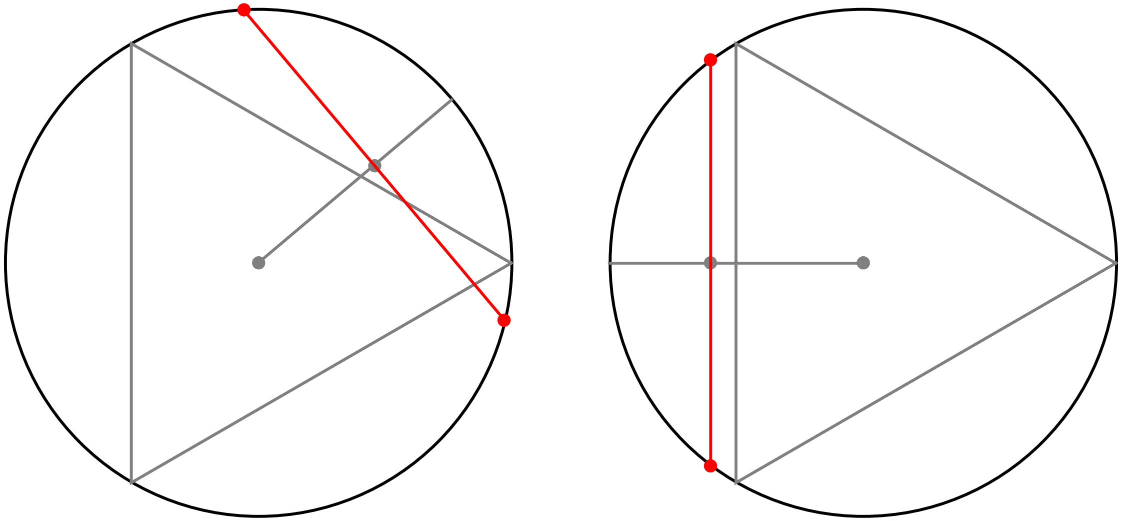
Bertrand’s paradox goes as follows:

“consider an equilateral triangle inscribed inside a circle, and suppose that a chord is chosen at random. What is the probability that the chord is longer than a side of the triangle?”

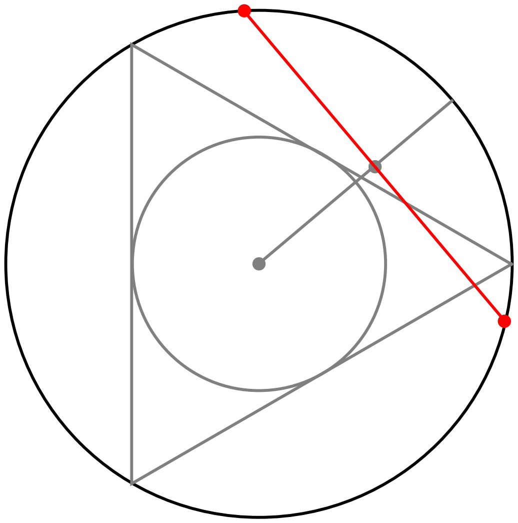
(Bertrand, 1889)



Solution: we take two random points on the circle (radius *R)*, then we rotate the circle so that one of the two points coincides with one of the vertices of the inscribed triangle. Thus a random chord is equivalent to taking the first point that defines the chord as one vertex of the triangle while the other is taken “at random” on the circle. Here “at random” means that it is uniformly distributed on the circumference. Then only those chords that cross the opposite side of the triangle are actually longer than each side. Since the subtended arc is 1/3 of the circumference, the probability of drawing a random chord that is longer than one side of the triangle is 1/3.



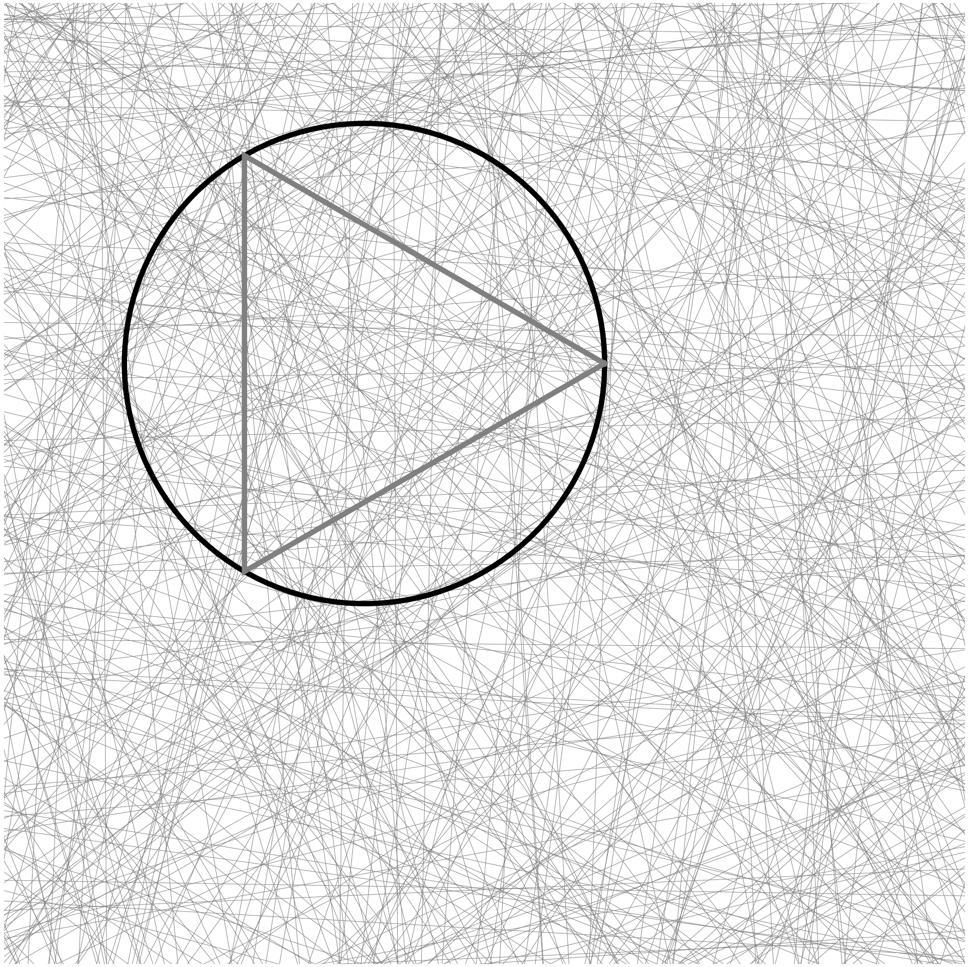
Solution 2: we take first a random radius, and next we choose a random point on this random radius. Then, we take the chord through this point and perpendicular to the radius. When we rotate the triangle so that the radius is perpendicular to one of the sides, we see that half of the points give chords longer than one side of the triangle, therefore the probability is 1/2.

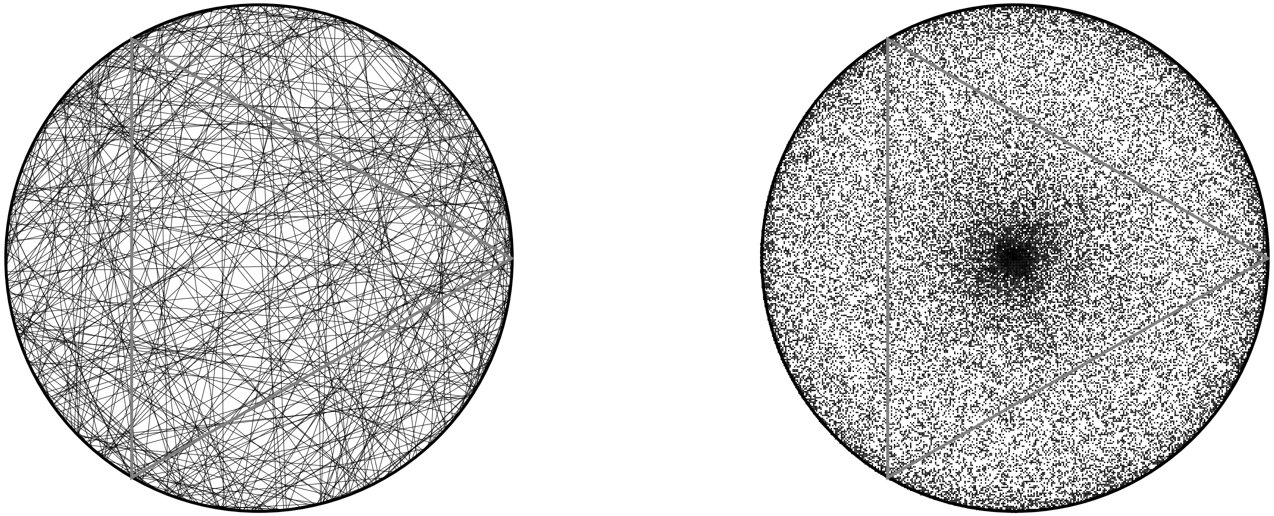


Solution 3: we take the chord midpoints located inside the circle inscribed in the triangle, and we obtain chords that are longer than one side of the triangle. Since the ratio of the areas of the two circles is 1/4, we find that now the probability of drawing a long chord is just 1/4.

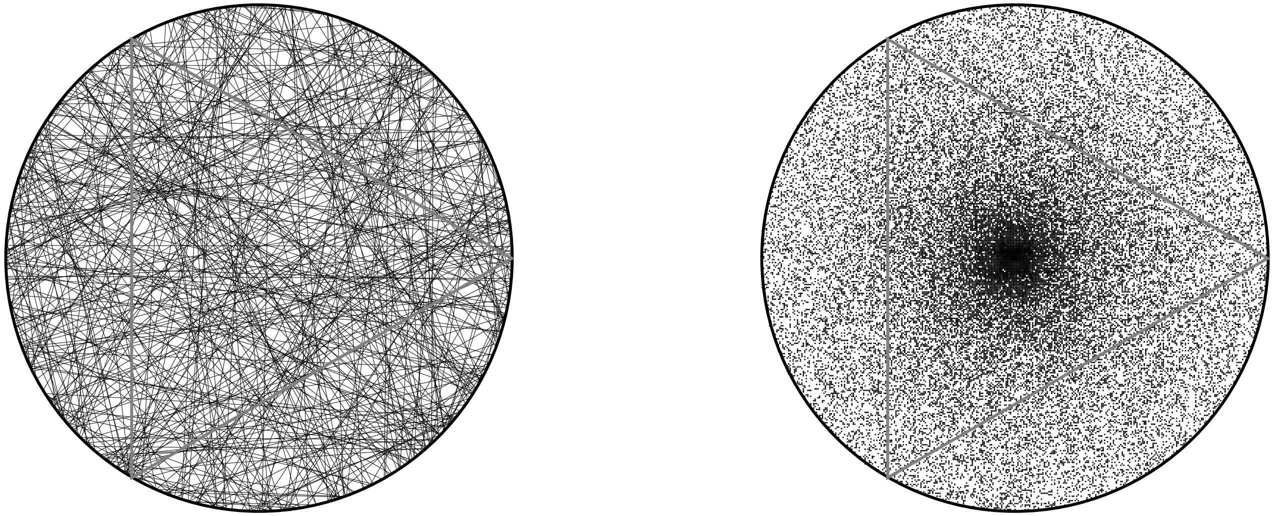
At least 3 different "solutions": which one is correct, and why?

Now we widen the scope of the problem and we consider the distribution of chords in the plane



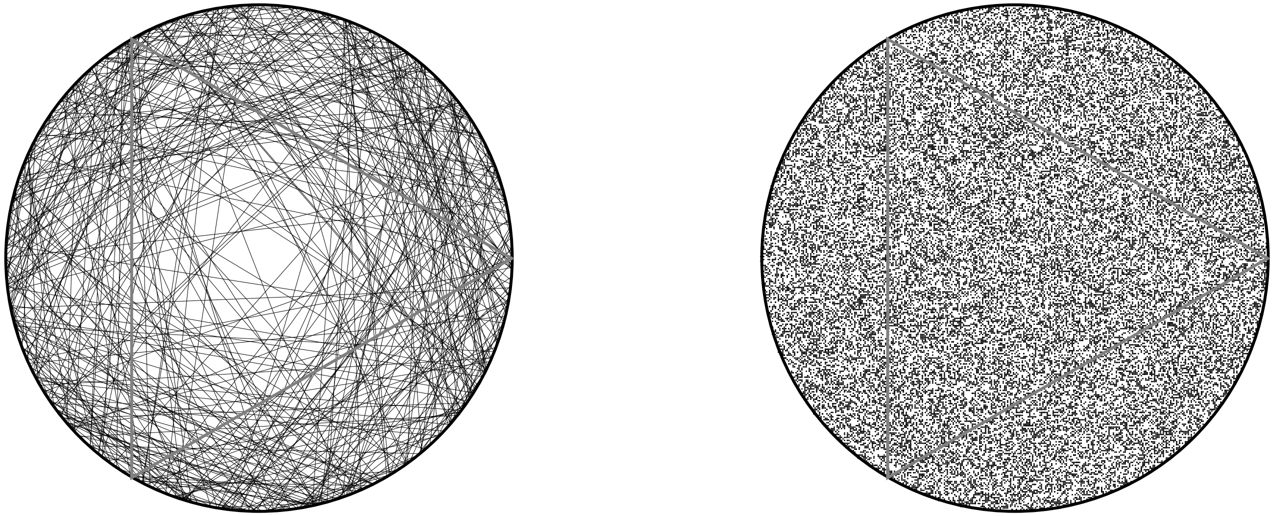


Distribution 1: distribution of chords (left panel) and of midpoints (right panel) in the first solution of Bertrand’s paradox (the left panel shows 400 chords, the right panel shows 100000 midpoints).

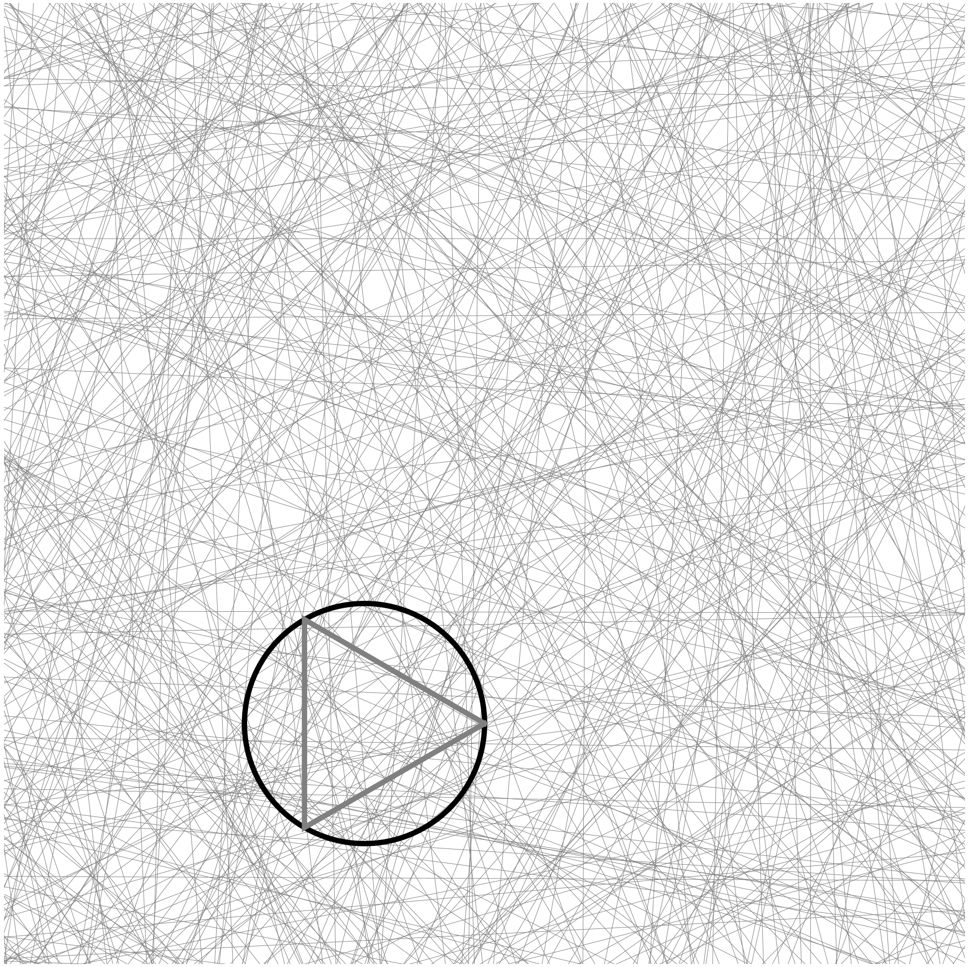


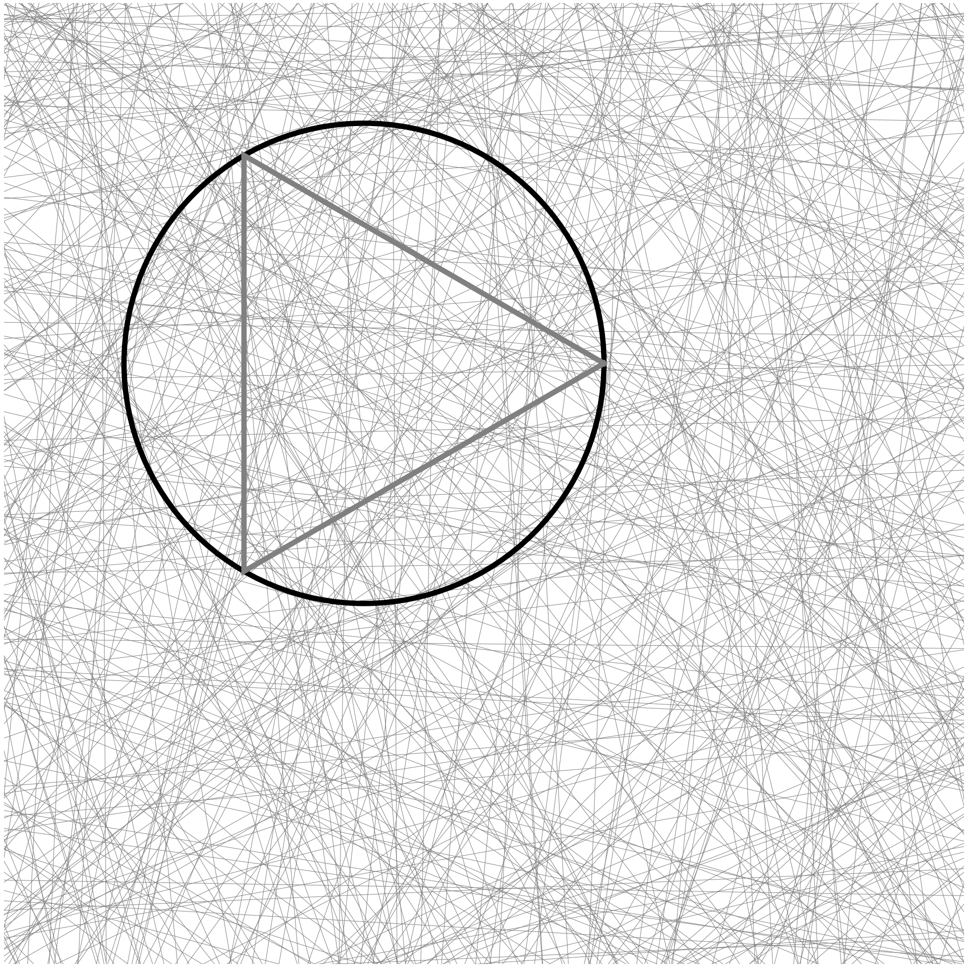
Distribution 2: Distribution of chords (left panel) and of midpoints (right panel) in the second solution of Bertrand’s paradox (the left panel shows 400 chords, the right panel shows 100000 midpoints).

In this case it is very easy to find the radial density function of chord centers, since here we take first a random radius, and next we choose a random point (the center) on this random radius.



Distribution 3: Distribution of chords (left panel) and of midpoints (right panel) in the third solution of Bertrand’s paradox (the left panel shows 400 chords, the right panel shows 100000 midpoints). Notice that while the distribution of midpoints is uniform, the distribution of the resulting chords is distinctly non-uniform.





Hidden assumptions (Jaynes):

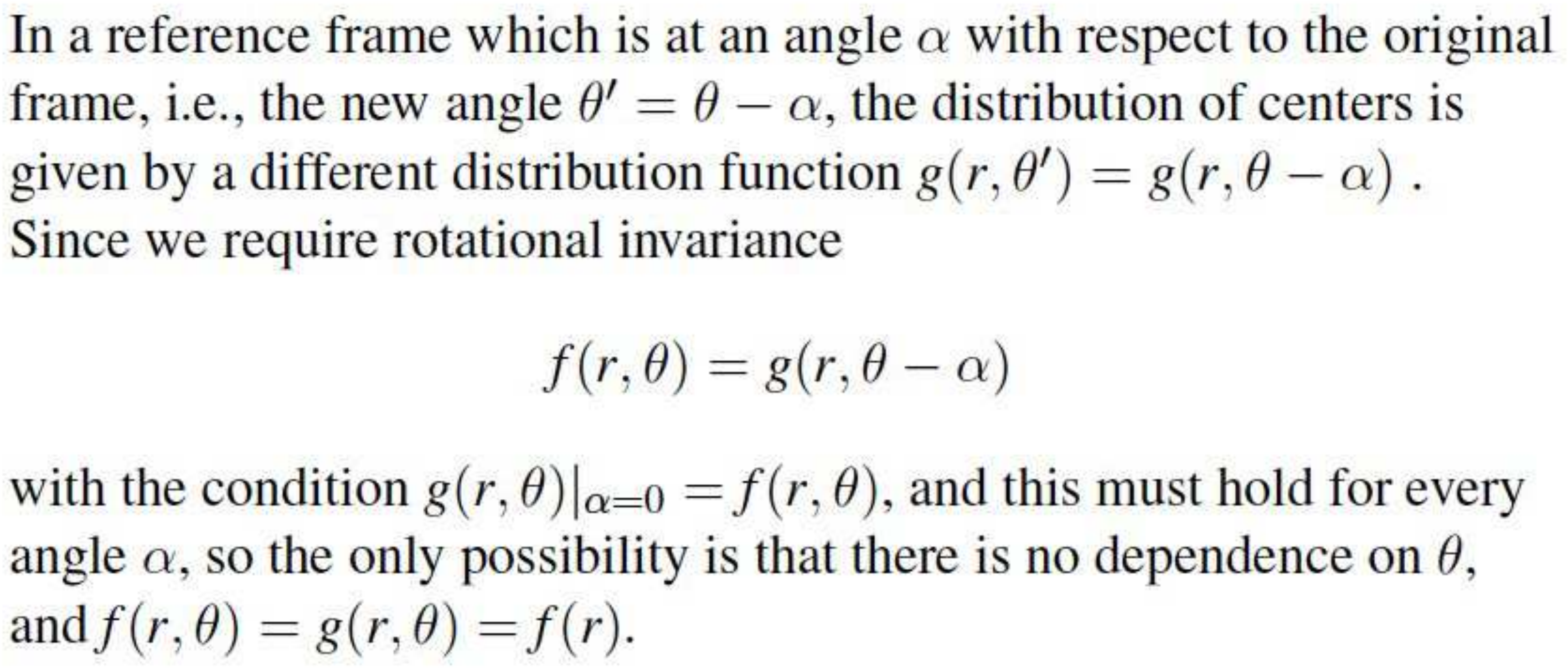
- rotational invariance
- scale invariance
- translational invariance

Now let

$$f(r,\theta)$$

be the probability density of chord centers

Rotational invariance

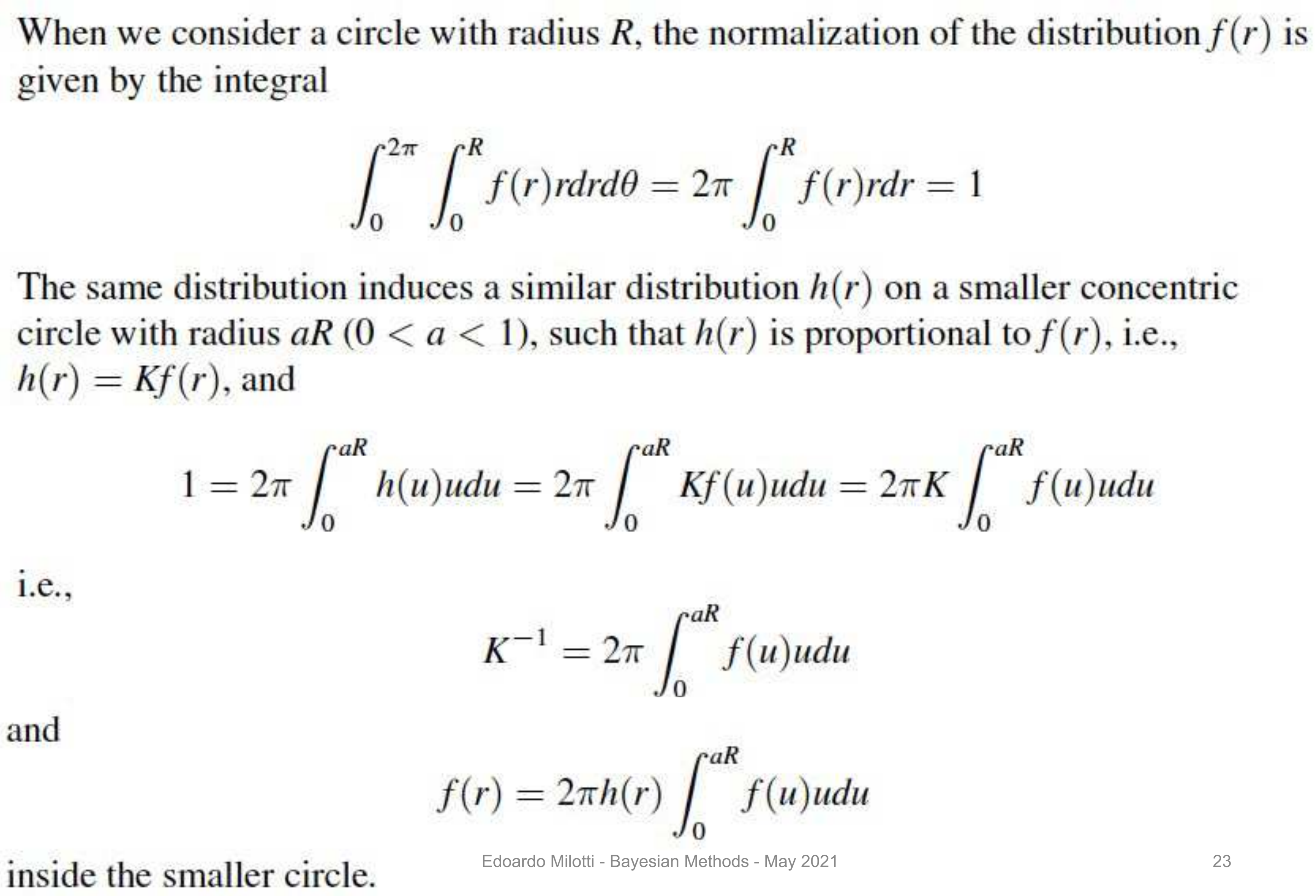


In a reference frame which is at an angle α with respect to the original frame, i.e., the new angle $\theta' = \theta - \alpha$, the distribution of centers is given by a different distribution function $g(r, \theta') = g(r, \theta - \alpha)$. Since we require rotational invariance

$$f(r, \theta) = g(r, \theta - \alpha)$$

with the condition $g(r, \theta)|_{\alpha=0} = f(r, \theta)$, and this must hold for every angle α, so the only possibility is that there is no dependence on θ, and $f(r, \theta) = g(r, \theta) = f(r)$.

Scale invariance



When we consider a circle with radius R, the normalization of the distribution $f(r)$ is given by the integral

$$\int_0^{2\pi}\int_0^R f(r) r dr d\theta = 2\pi \int_0^R f(r) r dr = 1$$

The same distribution induces a similar distribution $h(r)$ on a smaller concentric circle with radius aR ($0 < a < 1$), such that $h(r)$ is proportional to $f(r)$, i.e., $h(r) = Kf(r)$, and

$$1 = 2\pi \int_0^{aR} h(u) u du = 2\pi \int_0^{aR} K f(u) u du = 2\pi K \int_0^{aR} f(u) u du$$

i.e.,

$$K^{-1} = 2\pi \int_0^{aR} f(u) u du$$

and

$$f(r) = 2\pi h(r) \int_0^{aR} f(u) u du$$

inside the smaller circle.

Now we invoke the assumed scale invariance: the probability of finding a center in an annulus with radii r and $r+dr$ in the original circle, must be equal to the probability of finding a center in the scaled down annulus,

$$h(ar)(ar)d(ar) = f(r)rdr$$

and therefore

$$a^2 h(ar) = f(r)$$

Equation

$$a^2 h(ar) = f(r)$$

can also be rewritten in the form

$$h(r) = \frac{1}{a^2} f\left(\frac{r}{a}\right) \tag{1}$$

and inserting this into equation

$$f(r) = 2\pi h(r) \int_0^{aR} f(u) u du$$

we find

$$a^2 f(ar) = 2\pi f(r) \int_0^{aR} f(u) u du \tag{2}$$

We solve equation

$$a^2 f(ar) = 2\pi f(r) \int_0^{aR} f(u) u du$$

taking first its derivative with respect to a: the relation that we find must hold for all a's, and therefore also for $a = 1$ (no scaling), and we find the differential equation

$$rf'(r) = \left(2\pi R^2 f(R) - 2\right) f(r)$$

i.e.,

$$rf'(r) = (q - 2) f(r)$$

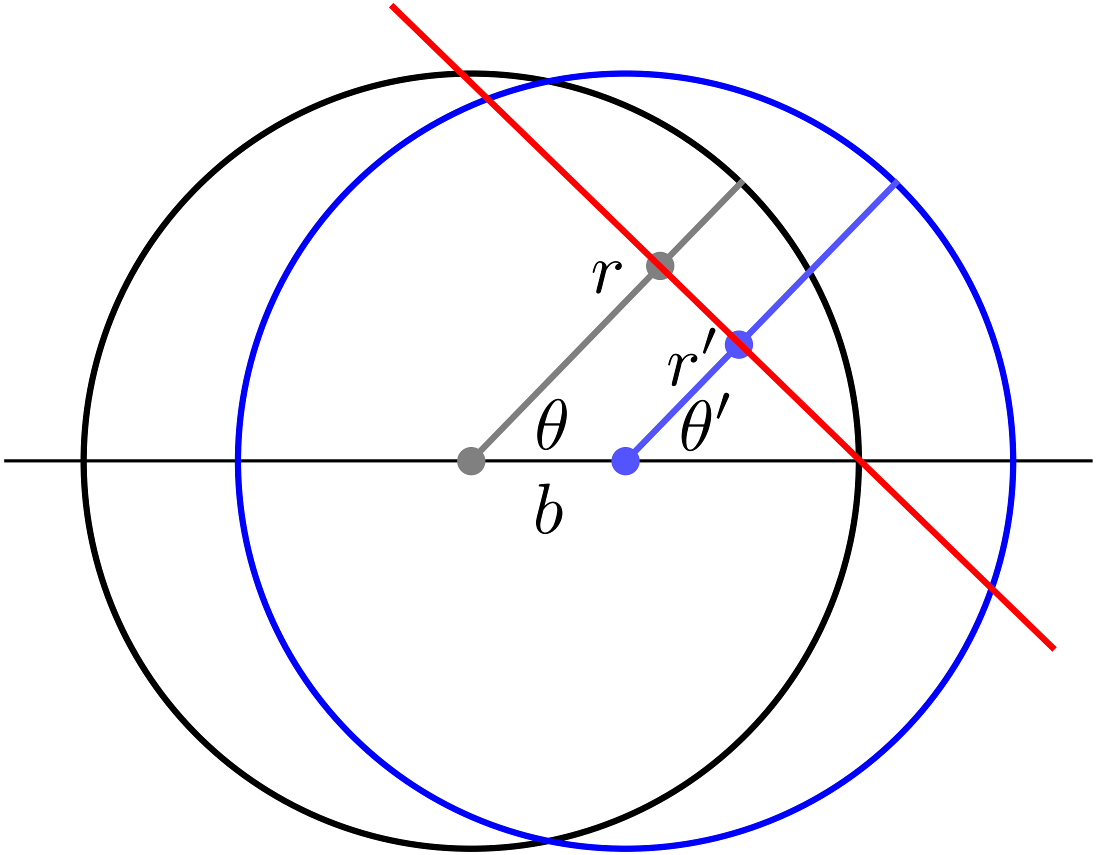
where the constant $q = 2\pi R^2 f(R)$ is unknown. However, we can still solve the equation and find

$$f(r) = Ar^{q-2}$$

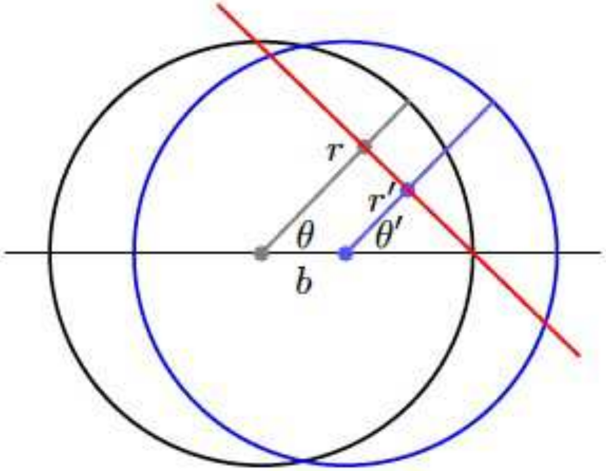
The constant A is easy to find from the normalization condition: $A = q/2\pi R^q$, and therefore

$$f(r) = \frac{qr^{q-2}}{2\pi R^q}$$

Translational invariance



Geometrical construction for the discussion of translational invariance. The original circle (black) is crossed by a straight line (red) which defines the chord. The translated circle is shown in blue.



This circle is displaced by the amount b, and the new radius and angle that define the midpoint of the chord are

$$r' = |r - b\cos\theta|$$

$$\theta' = \theta \ \ (\text{if } r \geq b\cos\theta) \quad \text{or} \quad \theta' = \theta + \pi \ \ (\text{if } r < b\cos\theta)$$

Now consider a region Γ surrounding the midpoint in the original circle, which is transformed into a region Γ' by the translation. The probability of finding a chord with the midpoint in the region Γ is

$$\int_\Gamma f(r) r dr d\theta = \int_\Gamma \frac{q r^{q-1}}{2\pi R^q} dr d\theta = \frac{q}{2\pi R^q} \int_\Gamma r^{q-1} dr d\theta$$

Likewise, the same probability for the translated circle is

$$\frac{q}{2\pi R^q} \int_{\Gamma'} (r')^{q-1} dr' d\theta' = \frac{q}{2\pi R^q} \int_\Gamma |r - b\cos\theta|^{q-1} dr d\theta \qquad (3)$$

where the Jacobian of the transformation is 1. Equating these expressions, we see that the integrand must be a constant, and therefore $q = 1$, and

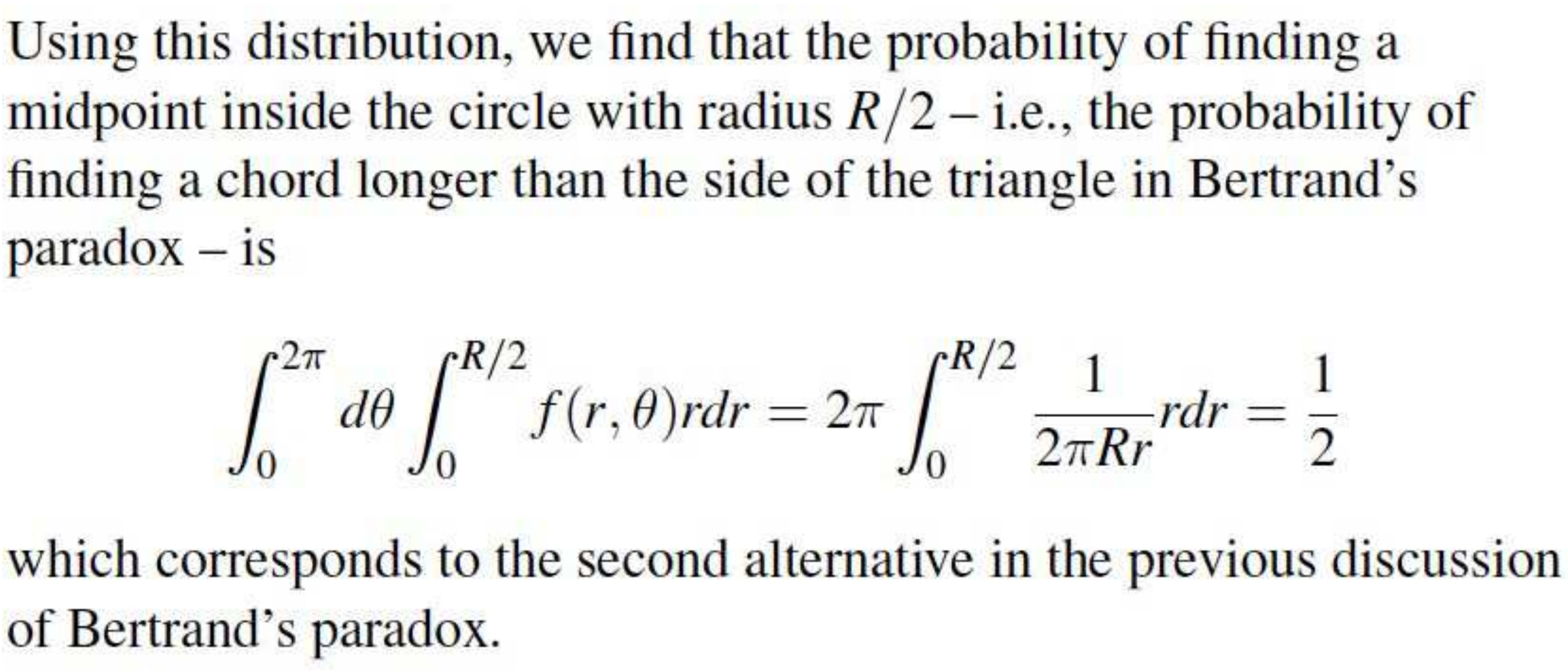
$$f(r, \theta) = \frac{1}{2\pi R r} \quad (r \leq R; \;\; 0 \leq \theta < 2\pi)$$

Therefore

$$f(r,\theta) = f(r) = C/r$$

$$\Rightarrow \quad \text{(normalization)} \quad 1 = \int_C f(r) 2\pi r dr = 2\pi C R$$

$$\Rightarrow \quad f(r) = \frac{1}{2\pi r R}$$



Using this distribution, we find that the probability of finding a midpoint inside the circle with radius $R/2$ – i.e., the probability of finding a chord longer than the side of the triangle in Bertrand's paradox – is

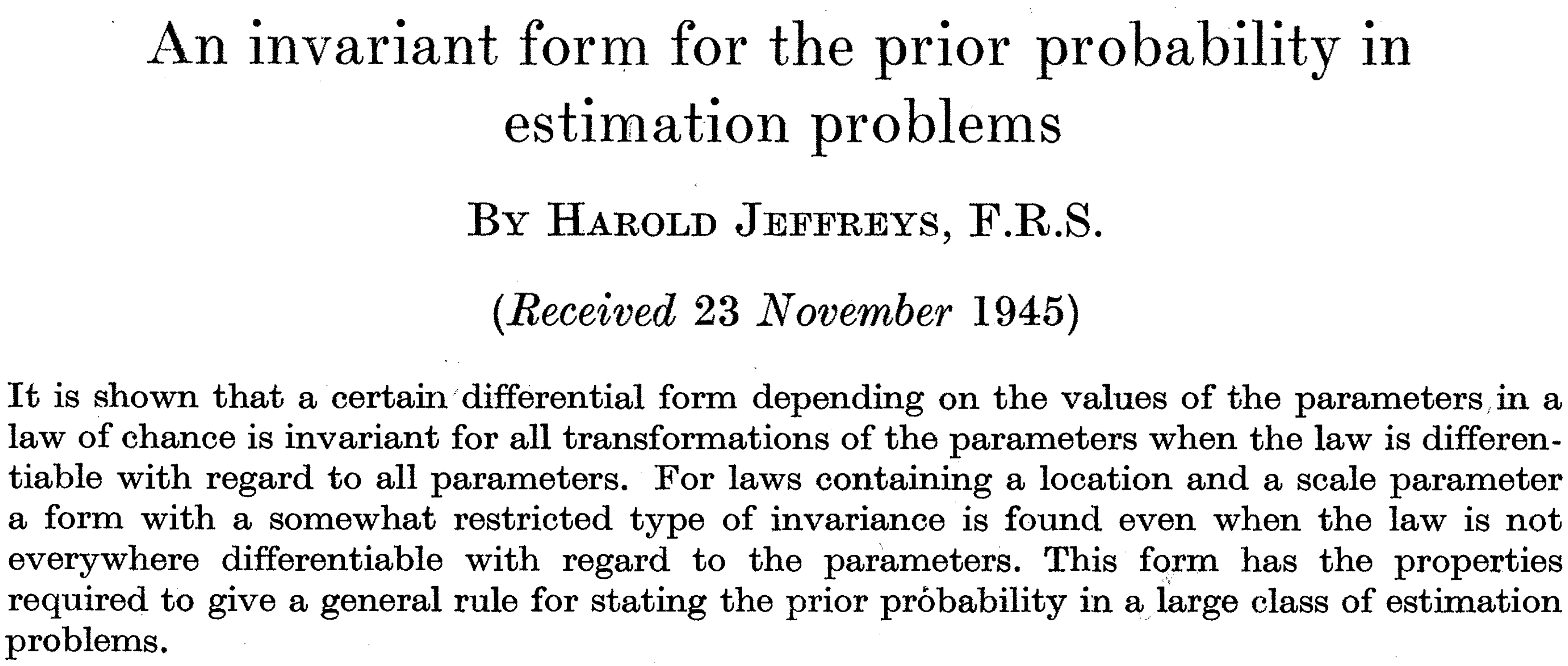
$$\int_0^{2\pi} d\theta \int_0^{R/2} f(r,\theta) r dr = 2\pi \int_0^{R/2} \frac{1}{2\pi R r} r dr = \frac{1}{2}$$

which corresponds to the second alternative in the previous discussion of Bertrand's paradox.

Lesson drawn from Bertrand's paradox:

probability models depend on physical assumptions, they are not God-given. We define the elementary events on the basis of real-world constraints, derived from our own experience.

A way forward to "objective" priors: Jeffreys' priors



An invariant form for the prior probability in estimation problems

By Harold Jeffreys, F.R.S.

(*Received* 23 *November* 1945)

It is shown that a certain differential form depending on the values of the parameters in a law of chance is invariant for all transformations of the parameters when the law is differentiable with regard to all parameters. For laws containing a location and a scale parameter a form with a somewhat restricted type of invariance is found even when the law is not everywhere differentiable with regard to the parameters. This form has the properties required to give a general rule for stating the prior probability in a large class of estimation problems.

Starting remark: here we concentrate on a problem of parametric statistics.

We seek the parameter that best adapts our theory to data. This means that we search among incompatible hypotheses that are defined by different parameter values.

The different hypotheses (and therefore, the different parameters) correspond to different pdf's

$$p(x|\theta)$$

Step 1: Bartlett identities for a parametric pdf family

$$\mathbf{E}\left[\frac{\partial \ln p(x|\theta)}{\partial\theta}\right] = 0$$

$$\mathbf{E}\left[\frac{\partial^2 \ln p(x|\theta)}{\partial\theta^2}\right] = -\mathbf{E}\left[\left(\frac{\partial \ln p(x|\theta)}{\partial\theta}\right)^2\right]$$

Step 2: a parameter-dependent Likelihood is a family of pdf's that represent the distribution of the data, given the value of the parameter(s).

It can be shown that the following inequality holds

$$\mathrm{var}[\hat{\theta}(D)] \geq= \frac{1}{\mathbf{E}\left[\left(\frac{\partial \ln L(D,\theta_0)}{\partial \theta_0}\right)^2\right]} = \frac{1}{-\mathbf{E}\frac{\partial^2 \ln L(D,\theta_0)}{\partial \theta_0^2}}$$

where θ_0 is the "true" value of the parameter, and $\hat{\theta}(D)$ is the ML estimator (Cramer-Rao-Fisher bound).

Step 3: definition of Fisher Information. A very concentrated pdf is very informative. Therefore, the smaller the variance, the greater the "information".

Thus, from the Cramer-Rao-Fisher bound

$$\mathrm{var}[\hat{\theta}(D)] \geq= \frac{1}{\mathbf{E}\left[\left(\frac{\partial \ln L(D,\theta_0)}{\partial \theta_0}\right)^2\right]} = \frac{1}{-\mathbf{E}\frac{\partial^2 \ln L(D,\theta_0)}{\partial \theta_0^2}}$$

one is led to the Fisher Information

$$I(\theta) = \mathbf{E}\left[\left(\frac{\partial \ln p(x,\theta)}{\partial \theta}\right)^2\right] = -\mathbf{E}\frac{\partial^2 \ln p(x,\theta)}{\partial \theta^2}$$

Step 4: it can be shown that the Fisher Information is a local (and symmetrical) form of the Kullback-Leibler divergence (see below)

$$I_{KL}\left(p(x|\theta), p(x|\theta+\epsilon)\right) = -\frac{1}{2}\mathbf{E}\left[\frac{\partial^2 \ln p(x|\theta)}{\partial\theta^2}\right]\epsilon^2 = \frac{1}{2}I(\theta)\epsilon^2$$

From this, and from the properties of the KL divergence, we see that the Fisher Information behaves like a (squared) distance between distributions.

Step 5: the KL divergence is invariant with respect to random variable transformations, and therefore also to parameter transformations. From the definition of KL divergence, and from the transformation formula for pdf's we find

$$\int_{-\infty}^{+\infty} p_y(y) \ln\left(\frac{p_y(y)}{q_y(y)}\right) dy = \int_{-\infty}^{+\infty} p_x(x) \ln\left(\frac{p_x(x)\left|\frac{dx}{dy}\right|}{q_x(x)\left|\frac{dx}{dy}\right|}\right) dx$$

$$= \int_{-\infty}^{+\infty} p_x(x) \ln\left(\frac{p_x(x)}{q_x(x)}\right) dx$$

Therefore, the Fisher Information is also invariant with respect to parameter transformations.

Step 6: from the equation that relates KL divergence and Fisher Information, we find a corresponding pdf:

$$I_{KL}\left(p(x|\theta), p(x|\theta+\epsilon)\right) = -\frac{1}{2}\mathbf{E}\left[\frac{\partial^2 \ln p(x|\theta)}{\partial\theta^2}\right]\epsilon^2 = \frac{1}{2}I(\theta)\epsilon^2$$

this means that for small fluctuations of the parameter, Fisher's information changes quadratically. Then, we recover linear changes when we take the square root.

Finally, by defining the pdf

$$f(\theta) \sim \sqrt{I(\theta)}$$

we obtain a pdf that is invariant with respect to parameter transformations. We apply this to likelihoods, that define parametric pdf families

Example: a simple Gaussian Likelihood for *n* datapoints

$$L(D|\mu) = \prod_n \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(- \frac{(x_n - \mu)^2}{2\sigma^2} \right)$$

⇨ $$\ln L(D|\mu) \sim \sum_n \left(- \ln \sigma - \frac{(x_n - \mu)^2}{2\sigma^2} \right)$$ fixed sigma

⇨ $$I(\mu) = \mathbf{E}\left[- \frac{\partial^2 \ln L(D|\mu)}{\partial \mu^2} \right] \sim \text{constant}$$

This points to a uniform prior for *μ*. In general, this uniform prior is an improper prior.

Example: a simple Gaussian Likelihood for *n* datapoints (ctd.)

$$L(D|\mu) = \prod_n \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x_n-\mu)^2}{2\sigma^2}\right)$$

$$I(\sigma) = \mathbf{E}\left[-\frac{\partial^2 \ln L(D|\sigma)}{\partial\sigma^2}\right] \sim \frac{1}{\sigma^2}$$

fixed mu

$$\sqrt{I(\sigma)} \sim \frac{1}{\sigma}$$

This power-law prior is another improper prior.

Example: Poisson distribution

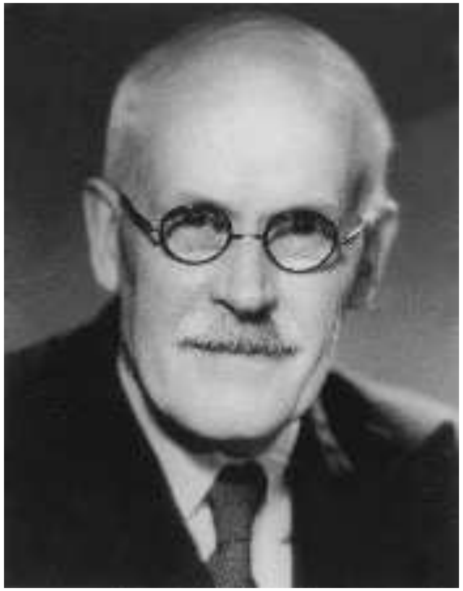
$$L(D|\mu) = \prod_n \frac{a^{k_n}}{k_n!} e^{-a}$$

➡ $$I(a) = \mathbf{E}\left[-\frac{\partial^2 \ln L(D|a)}{\partial a^2} \right] \sim \frac{1}{a}$$

➡ $$\sqrt{I(a)} \sim \frac{1}{\sqrt{a}}$$

This power-law prior is yet another improper prior.

A lesson learned from Jeffreys priors



Harold Jeffreys
(1891-1989)

Jeffreys priors are tuned to the Likelihood, but doesn't this sound strange? Shouldn't the prior information be tied to the prior distribution alone?

NO, the Likelihood is also constructed using prior information (obviously!). So, in a sense, Likelihood and the selected priors are related.

A short refresher on (Boltzmann's) entropy in statistical mechanics

- consider a system where states *n* are occupied by N_n identical particles (*n, n=1, ... , M*).

- the number of ways to fill these states is given by

$$\Omega = \frac{N!}{N_1!N_2!\ldots N_M!}$$

- then Boltzmann's entropy is

$$S_B = k_B \ln\Omega = k_B \ln\frac{N!}{N_1!N_2!\ldots N_M!} \approx k_B\left(\left(N\ln N - N\right) - \sum_n \left(N_n \ln N_n - N_n\right)\right)$$

$$= k_B\left(N\ln N - \sum_n N p_n\left(\ln p_n + \ln N\right)\right) = k_B \sum_n p_n \ln\frac{1}{p_n}$$

$$S_B = k_B \sum_i p_i \ln \frac{1}{p_i}$$

probability of physical states

Boltzmann’s entropy is just like Shannon’s entropy

this logarithmic function is the information carried by the *i*-th symbol

$$S_I = \sum_i p_i \log_2 \frac{1}{p_i}$$

probability of source symbols

Shannon’s entropy is the average information output by a source of symbols

Examples:

- just two symbols, 0 and 1, same source probability

$$S_I = -2\left(\frac{1}{2}\log_2\frac{1}{2}\right) = 1\ \mathrm{bit}$$

there are 2 equal terms

average information conveyed by each symbol

the result is given in pseudounit “bits” (for natural logarithms this is “nats”)

- just two symbols, 0 and 1, probabilities ¼ and ¾ , respectively

$$S_I = -\frac{1}{4}\log_2\frac{1}{4} - \frac{3}{4}\log_2\frac{3}{4} \approx 0.81 \text{ bit}$$

- 8 symbols, equal probabilities

$$S_I = -\sum_1^8 \frac{1}{8}\log_2\frac{1}{8} = \log_2 8 = 3 \text{ bit}$$

The Shannon entropy is additive for independent sources.

If symbols are emitted simultaneously and independently by two sources, the joint probability distribution is

$$p(j,k) = p_1(j)p_2(k)$$

and therefore the joint entropy is

$$\begin{aligned} S &= -\sum_{j,k} p(j,k)\log_2 p(j,k) = -\sum_{j,k} p_1(j)p_2(k)\log_2[p_1(j)p_2(k)] \\ &= -\sum_j p_1(j)\log_2 p_1(j) - \sum_k p_2(k)\log_2 p_2(k) \\ &= S_1 + S_2 \end{aligned}$$

The Shannon entropy is at a maximum for the uniform distribution.

This is an easy result that follows using one Lagrange multiplier to keep probability normalization into account

$$
\begin{aligned}
S + \lambda \sum_{k=1}^{N} p_k &= -\sum_{k=1}^{N} p_k \log_2 p_k + \lambda \sum_{k=1}^{N} p_k \\
&= -\frac{1}{\ln 2} \sum_{k=1}^{N} p_k \ln p_k + \lambda \sum_{k=1}^{N} p_k
\end{aligned}
$$

$$\frac{\partial}{\partial p_j}\left(S + \lambda \sum_{k=1}^{N} p_k\right) = -\frac{1}{\ln 2}(\ln p_j + 1) + \lambda = 0$$

$$p_j = \exp(\lambda \ln 2 - 1) = 1/N$$

all probabilities have the same value