# Introduction to Bayesian Methods - 3

*Edoardo Milotti*

Università di Trieste and INFN-Sezione di Trieste

# 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 (MaxEnt and Jeffreys')
- here we discuss

1. The Maximum Entropy Method
2. Jeffreys' method
3. Reference priors

# 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. This means that choosing a uniform distribution as the "least informative" distribution is not enough, *unless we specify which variate should be uniformly distributed*.

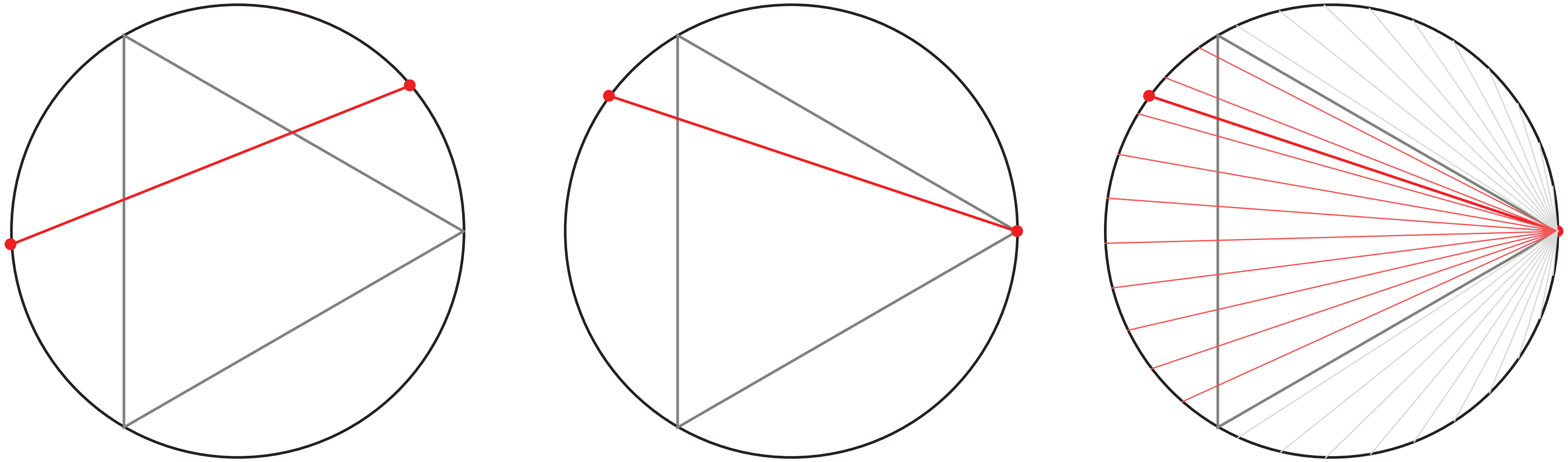
- How can we "objectively" choose a prior distribution???

# Bertrand’s paradox and the ambiguities of probability models (and how physics can point to a way out)

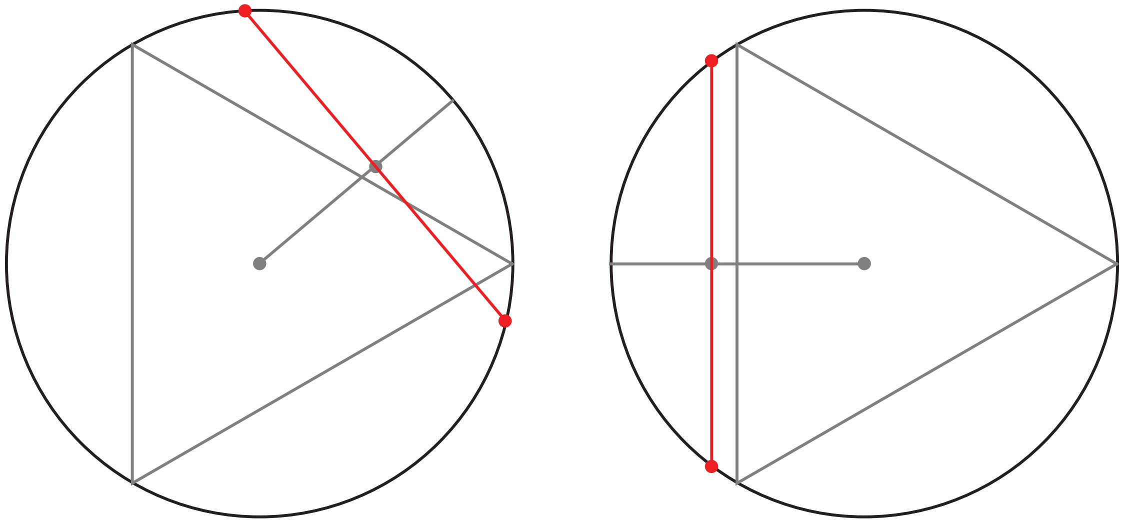
Bertrand’s paradox goes as follows:

“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?”

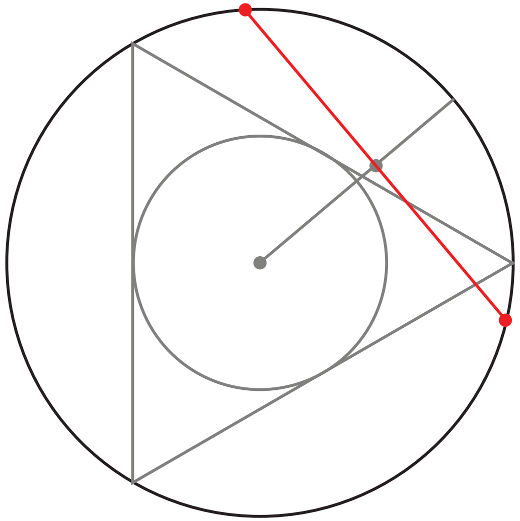
(Bertrand, 1889)



**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**.



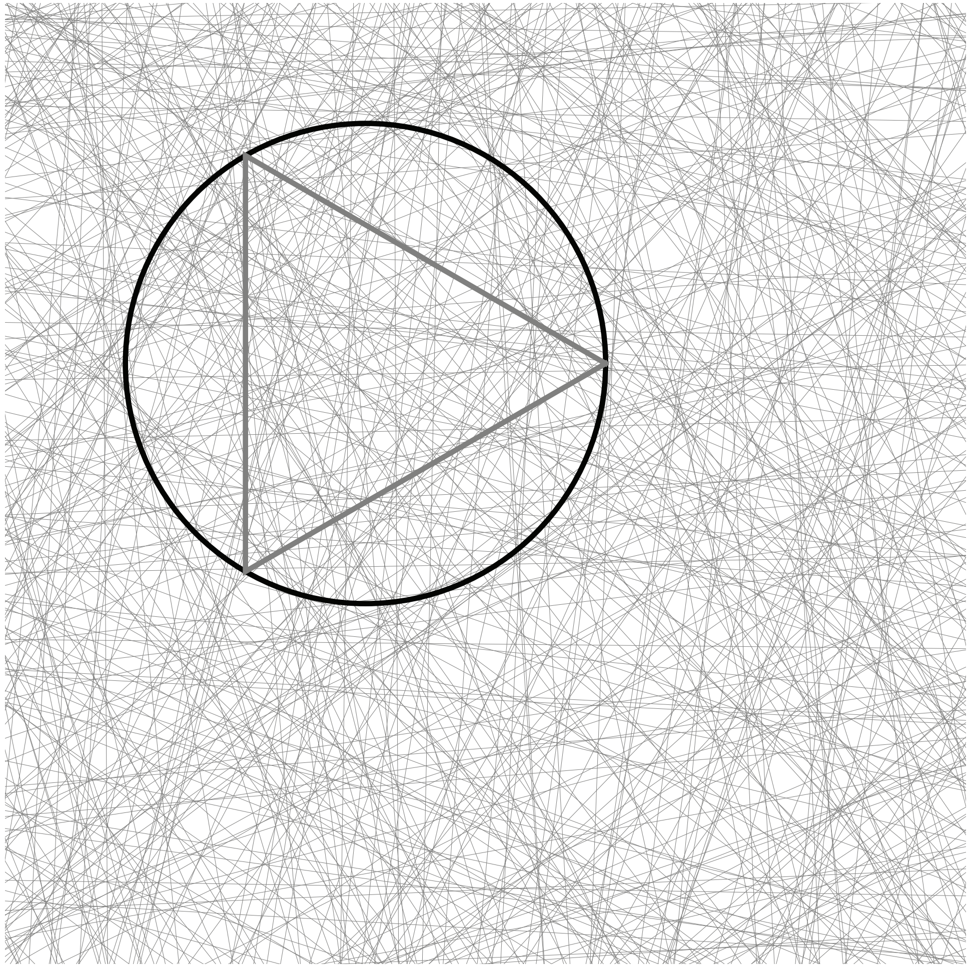
**Solution 2**: we take first a random radius, and next we choose a random point on this random radius. Then, we take the chord through this point and perpendicular to the radius. When we rotate the triangle so that the radius is perpendicular to one of the sides, we see that half of the points give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

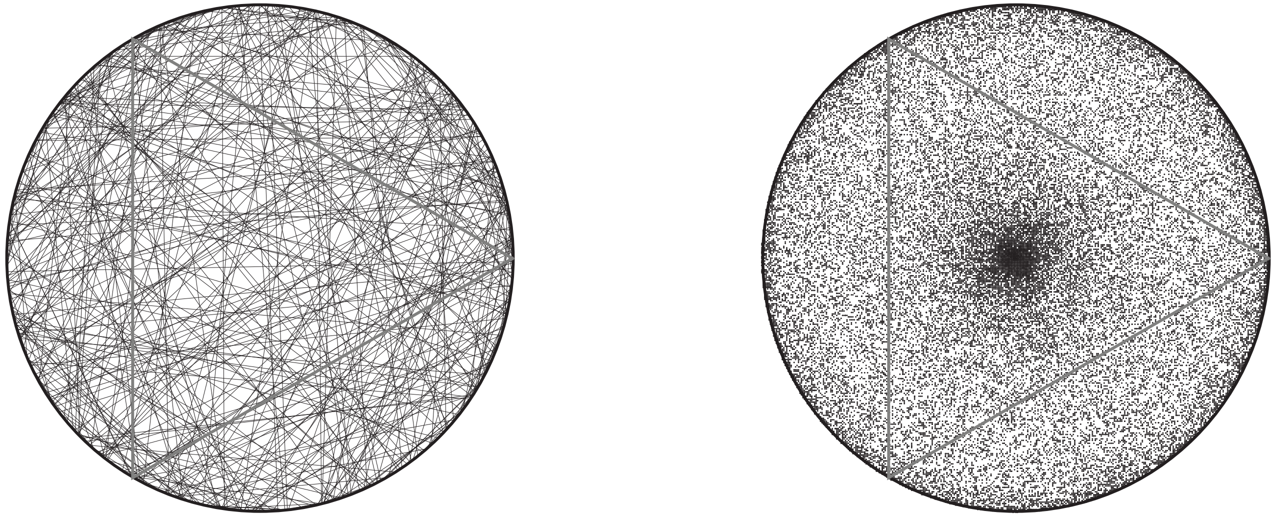


**Solution 3**: we take the chord midpoints located inside the circle inscribed in the triangle, and we obtain chords that are longer than one side of the triangle. Since the ratio of the areas of the two circles is 1/4, we find that now the probability of drawing a long chord is just 1/4.

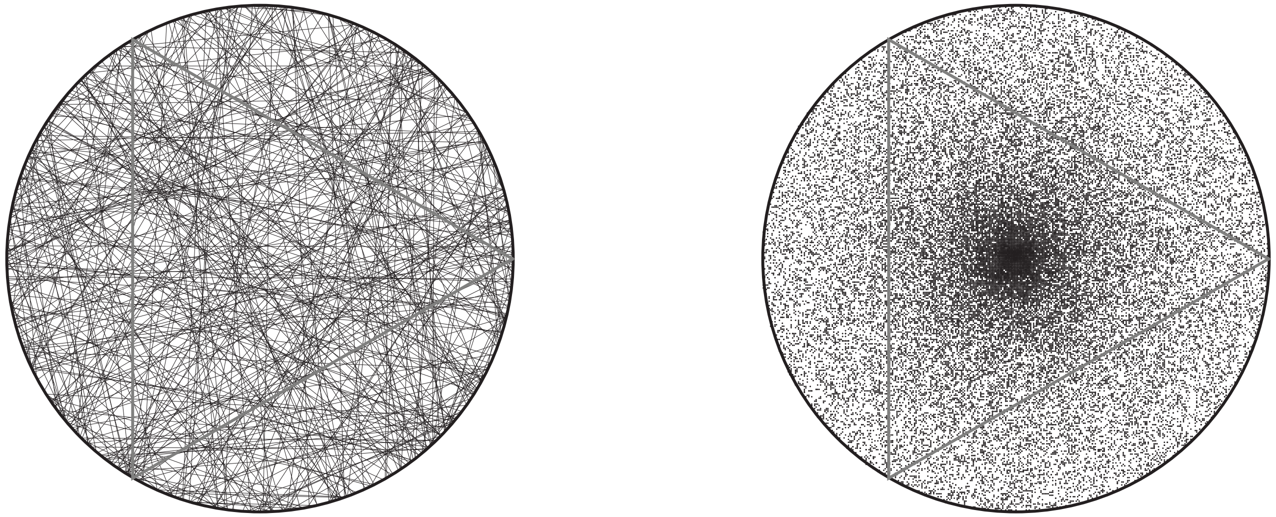
*At least 3 different “solutions”: which one is correct, and why?*

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



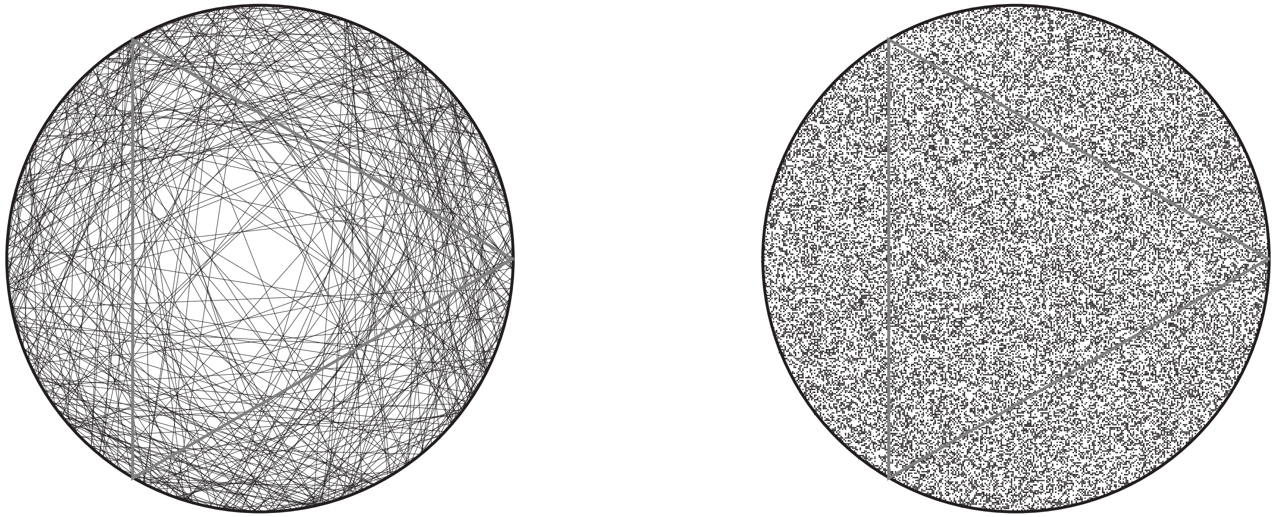


**Distribution 1**: distribution of chords (left panel) and of midpoints (right panel) in the first solution of Bertrand’s paradox (the left panel shows 400 chords, the right panel shows 100000 midpoints).

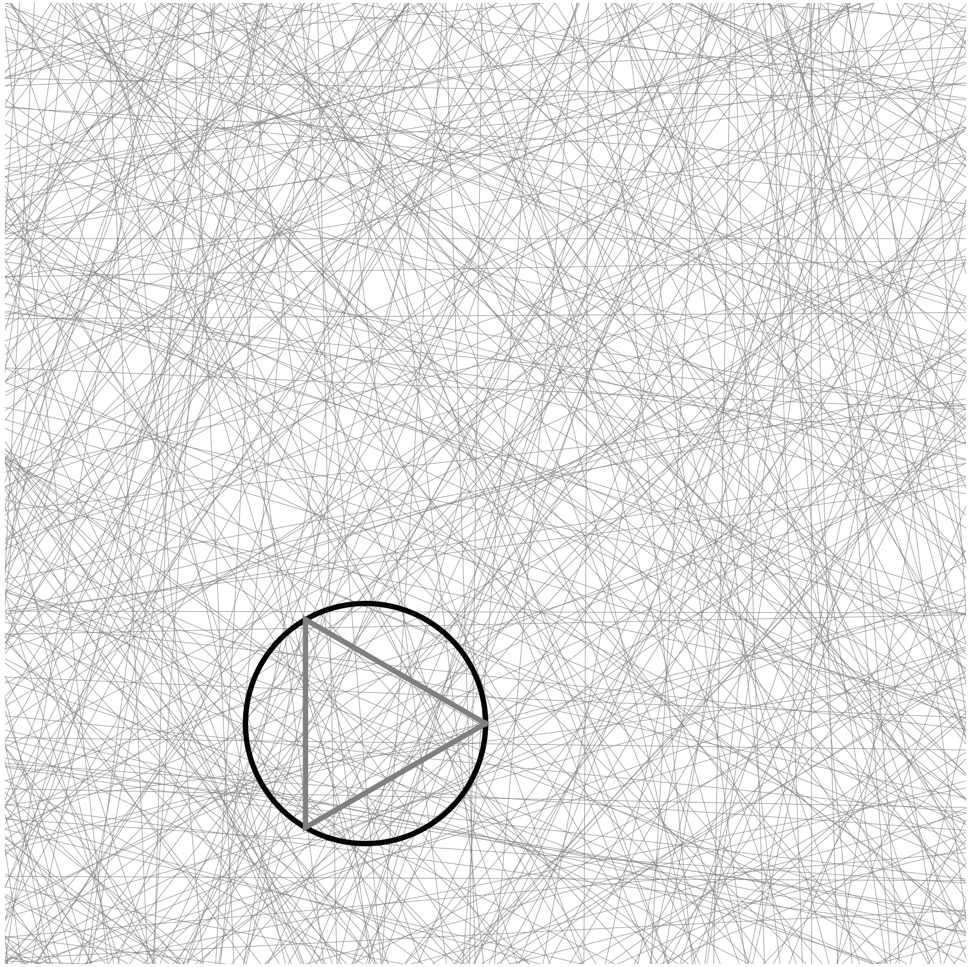


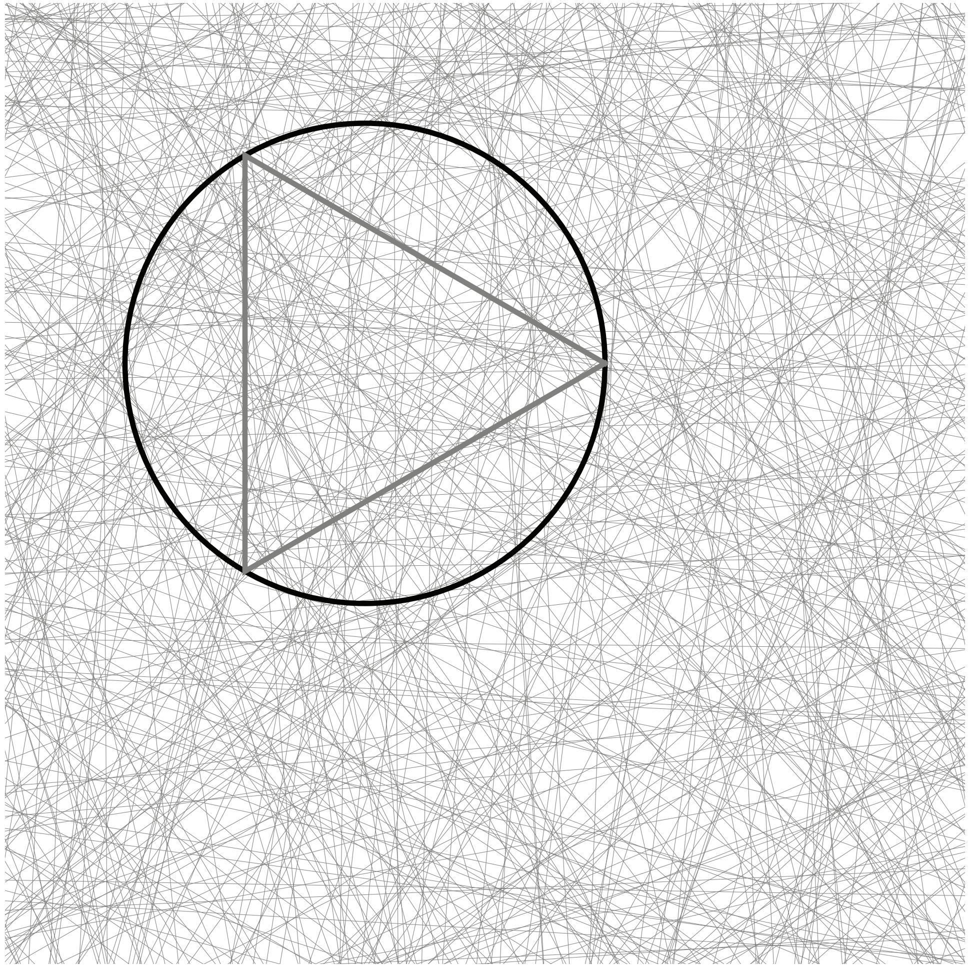
**Distribution 2**: Distribution of chords (left panel) and of midpoints (right panel) in the second solution of Bertrand’s paradox (the left panel shows 400 chords, the right panel shows 100000 midpoints).

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):

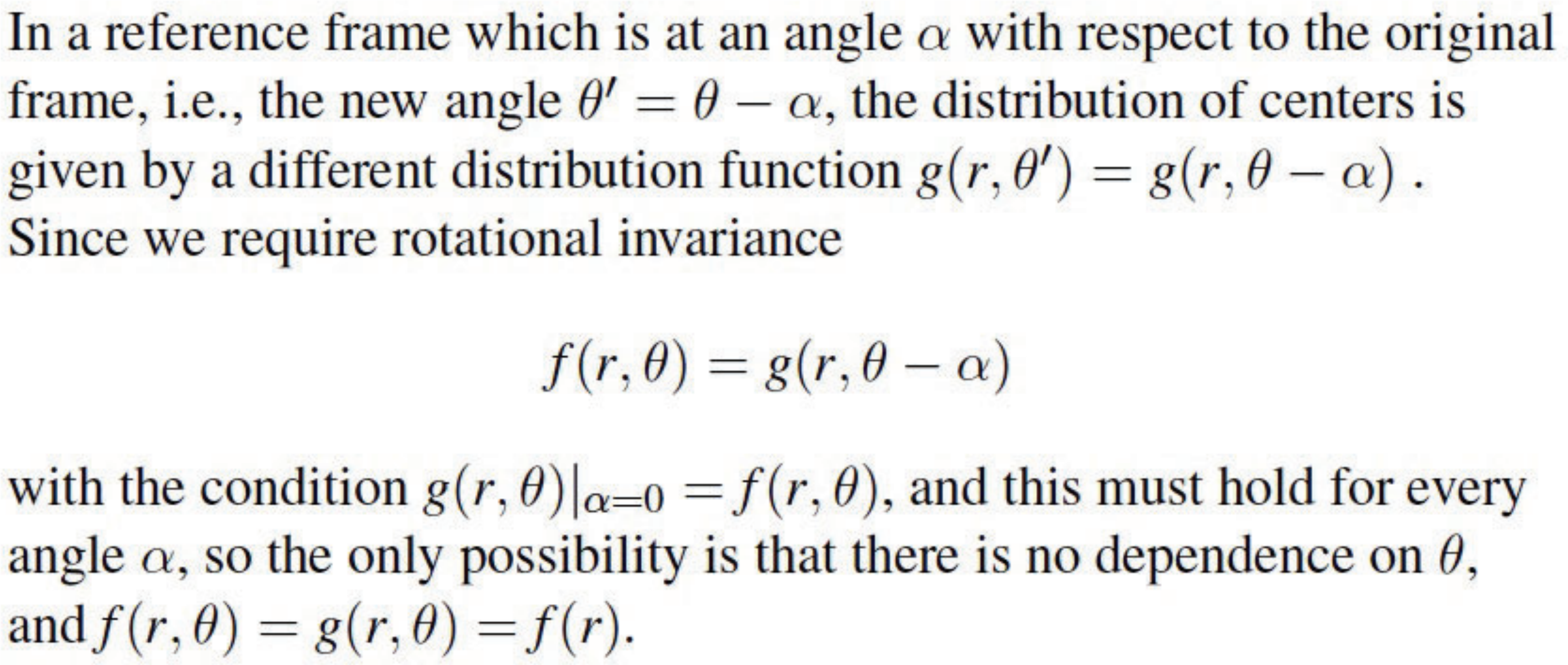
- rotational invariance
- scale invariance
- translational invariance

Now let

be the probability density of chord centers

$$f(r, \theta)$$

# Rotational invariance

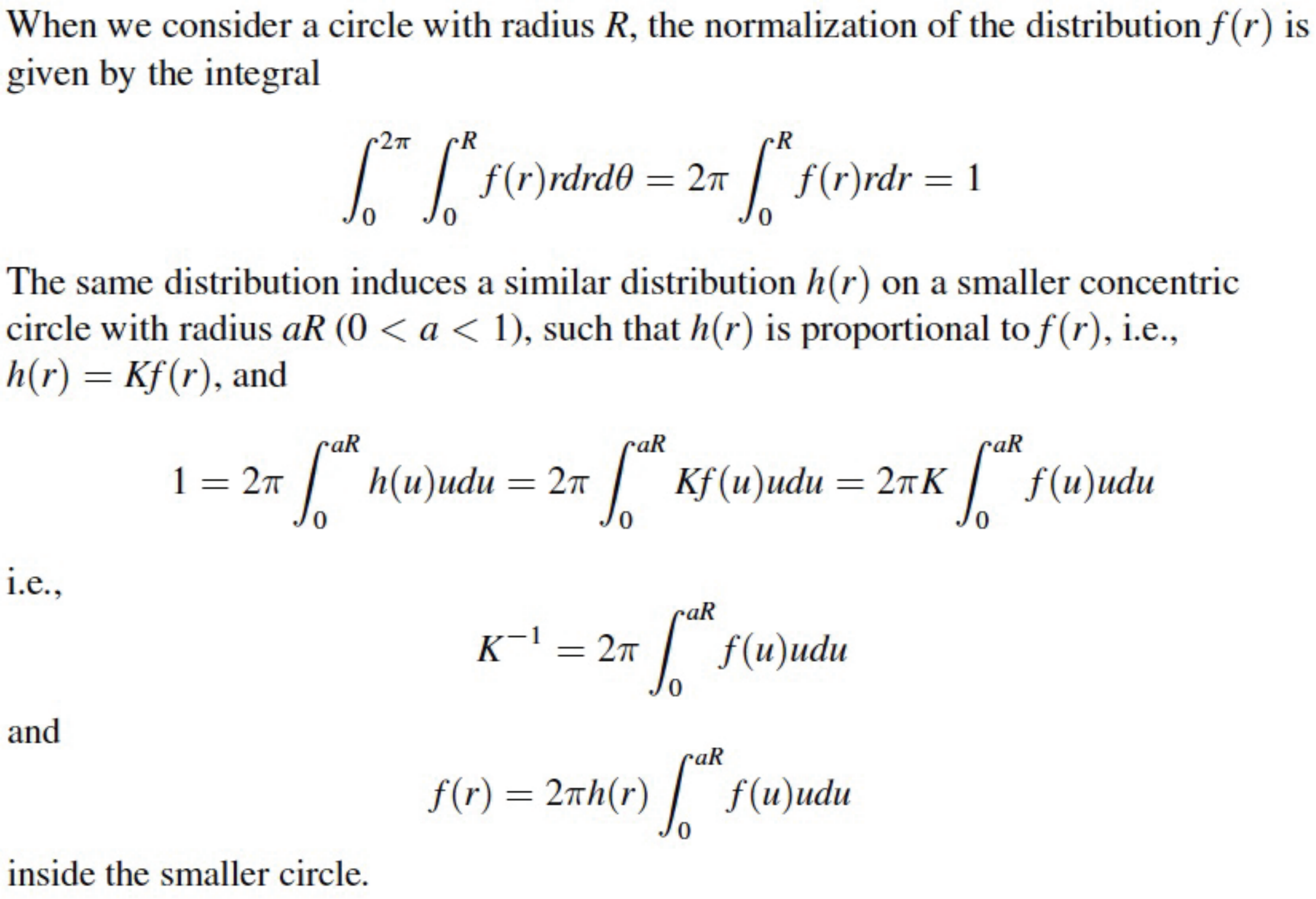


In a reference frame which is at an angle $\alpha$ with respect to the original frame, i.e., the new angle $\theta' = \theta - \alpha$, the distribution of centers is given by a different distribution function $g(r,\theta') = g(r,\theta-\alpha)$ . Since we require rotational invariance

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

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

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

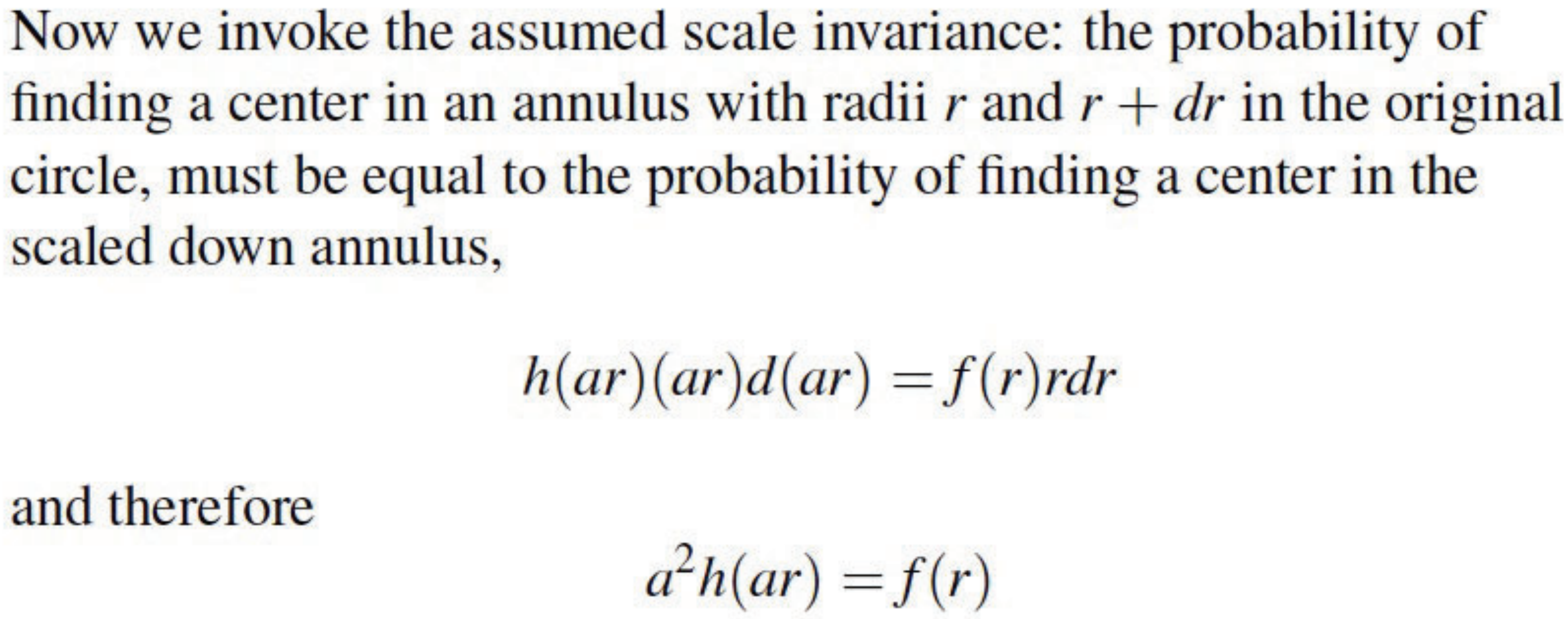
i.e.,

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

and

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

inside the smaller circle.

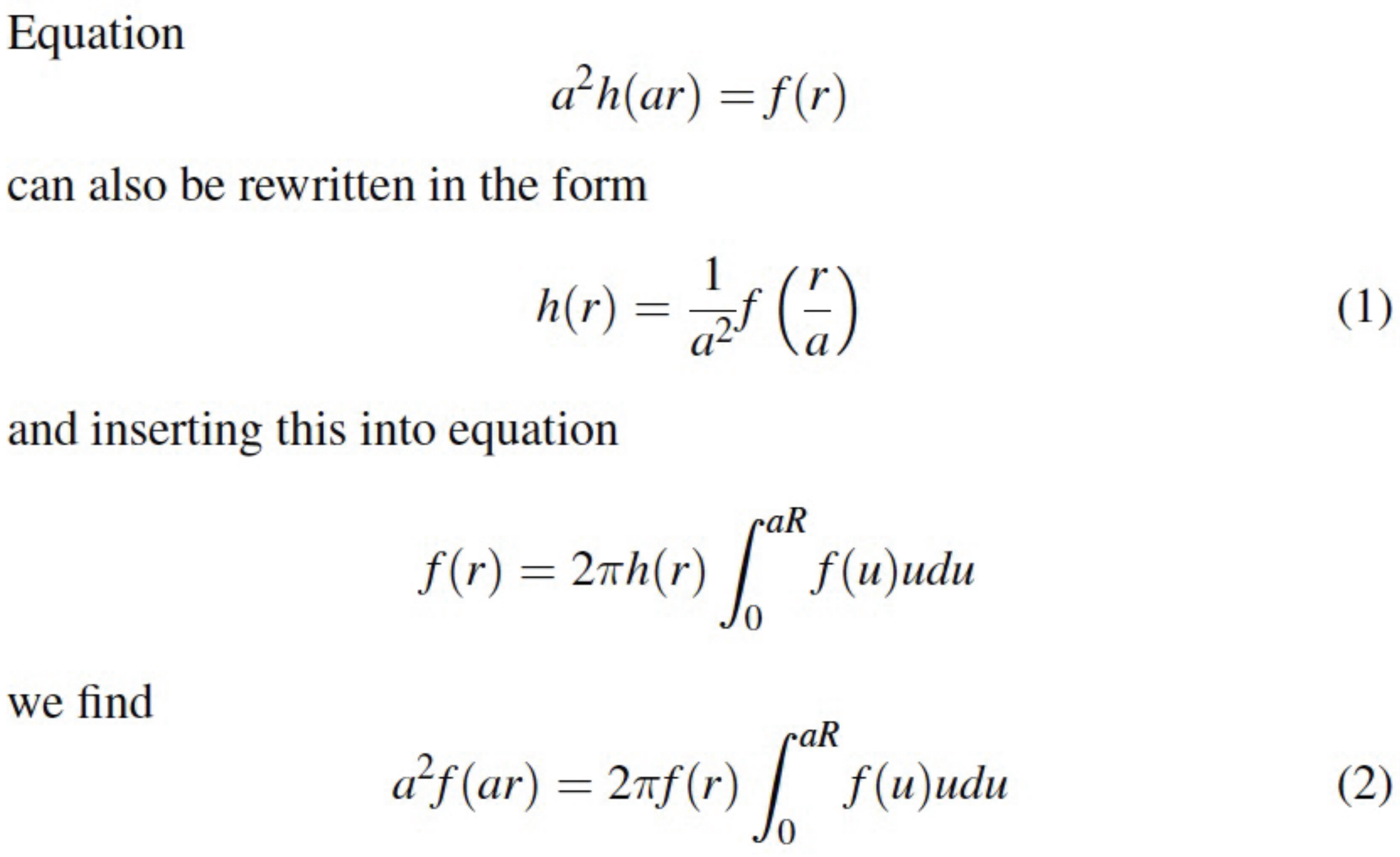


Now we invoke the assumed scale invariance: the probability of finding a center in an annulus with radii $r$ and $r + dr$ in the original circle, must be equal to the probability of finding a center in the scaled down annulus,

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

and therefore

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



Equation

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

can also be rewritten in the form

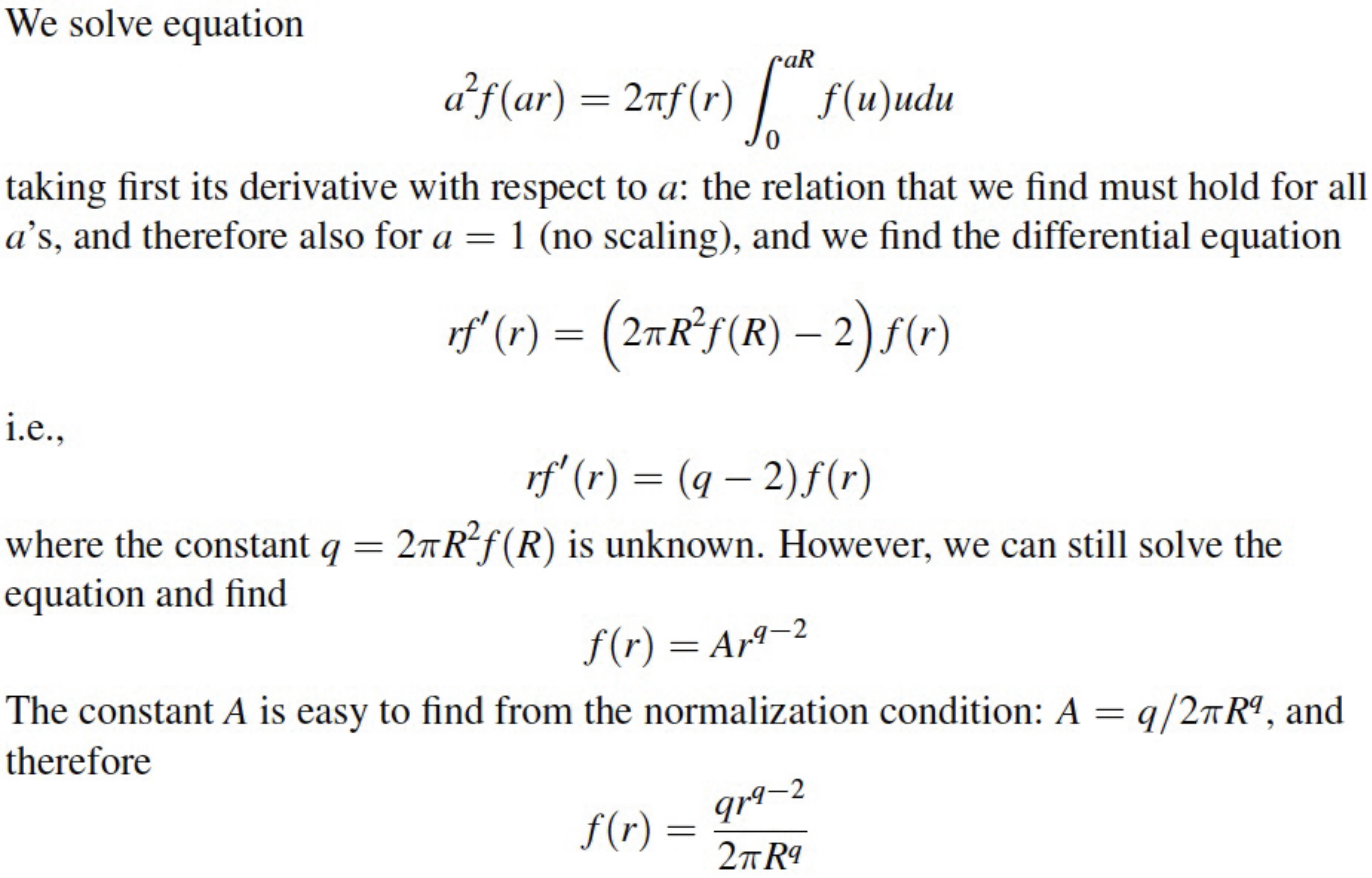
$$h(r) = \frac{1}{a^2} f\left(\frac{r}{a}\right) \tag{1}$$

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.,

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

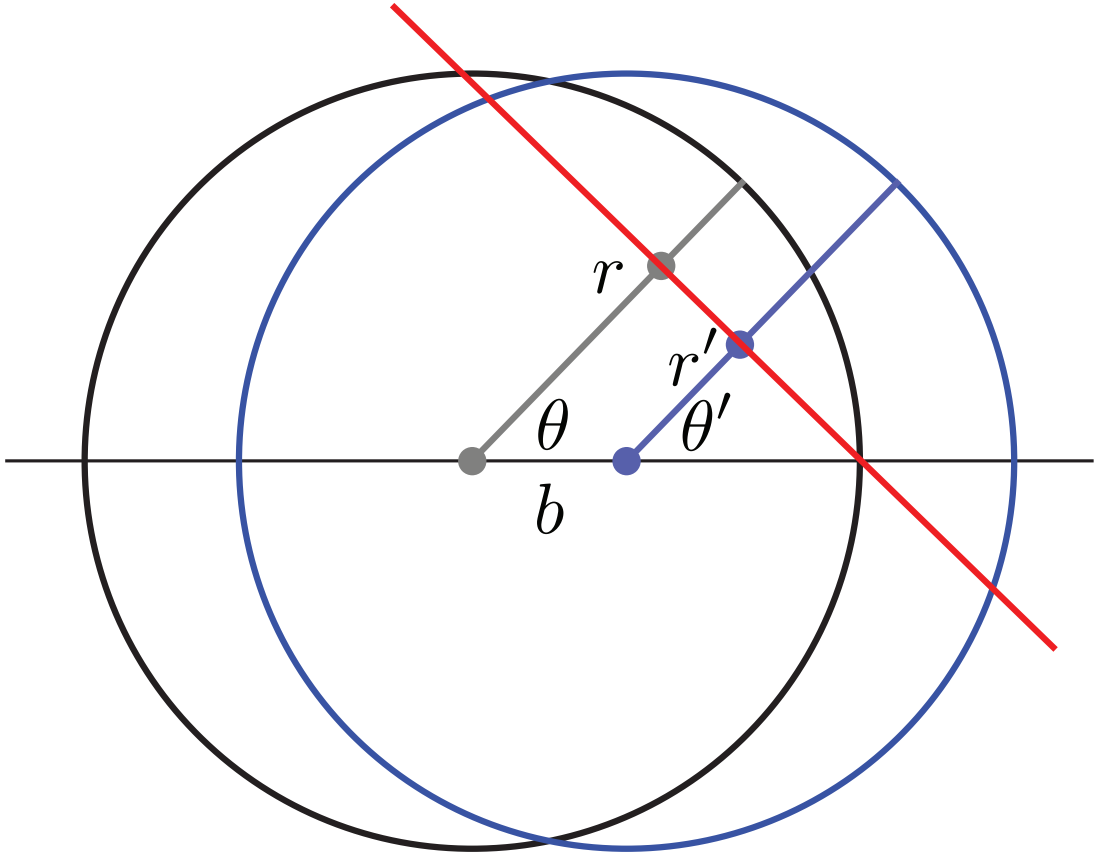
where the constant $q = 2\pi R^2 f(R)$ is unknown. However, we can still solve the equation and find

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

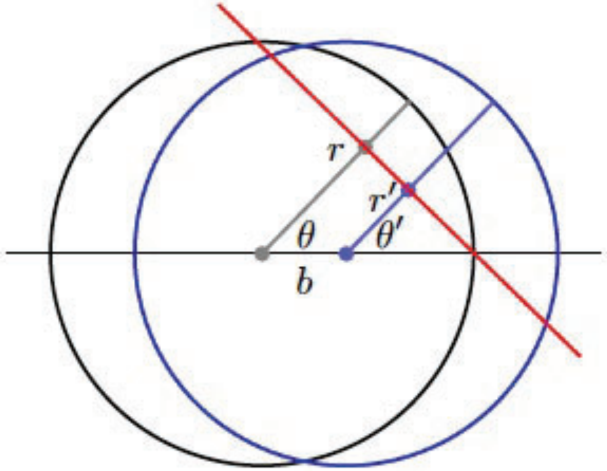
The constant $A$ is easy to find from the normalization condition: $A = q/2\pi R^q$, and therefore

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

# Translational invariance



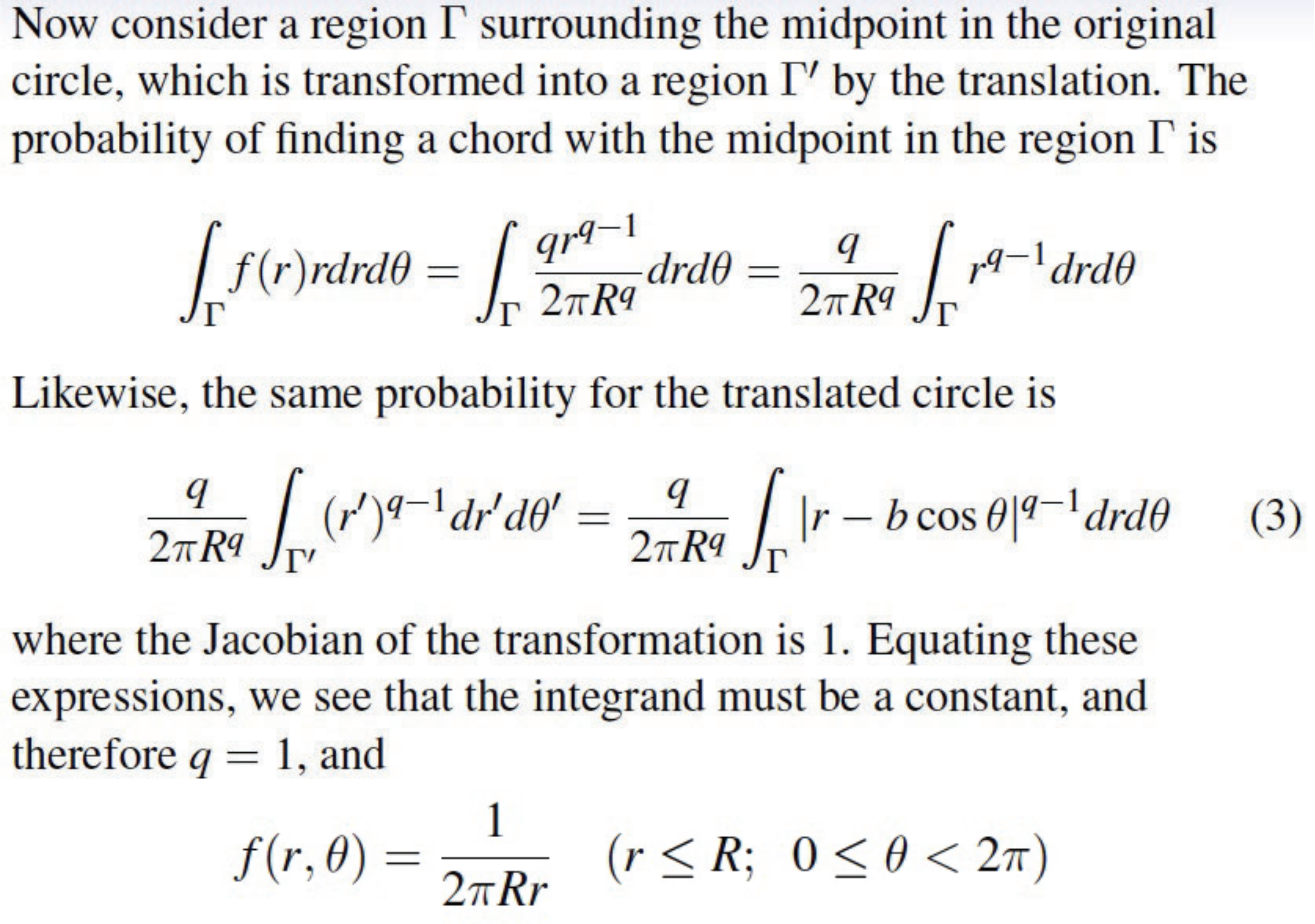
Geometrical construction for the discussion of translational invariance. The original circle (black) is crossed by a straight line (red) which defines the chord. The translated circle is shown in blue.



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

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

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



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

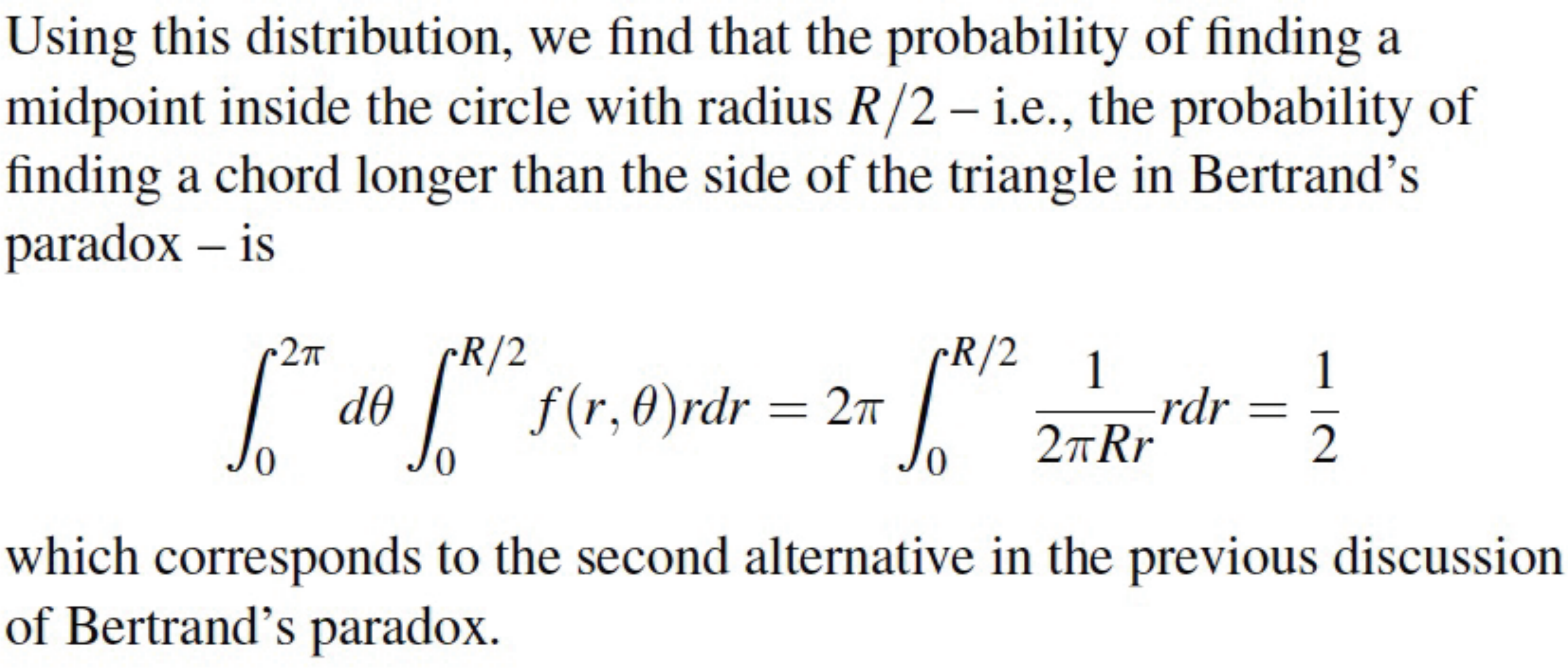
$$f(r,\theta) = \frac{1}{2\pi R r} \quad (r \le R; \;\; 0 \le \theta < 2\pi)$$

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 based on real-world constraints, derived from our own experience.

# Proof of the Bartlett identities for a parametric family of pdf's

- pdf normalization

$$\int_{\{x\}} p(x,\theta) dx = 1$$

- derivation of normalization formula

$$\frac{\partial}{\partial\theta} \int_{\{x\}} p(x,\theta) dx = \int_{\{x\}} \frac{\partial p(x,\theta)}{\partial\theta} dx = 0$$

- further manipulation of the previous result

$$\begin{aligned} 0 &= \int_{\{x\}} \frac{\partial p(x,\theta)}{\partial\theta} dx = \int_{\{x\}} \frac{1}{p(x,\theta)} \frac{\partial p(x,\theta)}{\partial\theta} p(x,\theta) dx \\ &= \int_{\{x\}} \frac{\partial \ln p(x,\theta)}{\partial\theta} p(x,\theta) dx \\ &= \mathbf{E}\left[\frac{\partial \ln p(x,\theta)}{\partial\theta}\right] \end{aligned}$$

**First Bartlett identity**

# Proof of the Bartlett identities for a parametric family of pdf's (ctd.)

- derivation of the first identity $$\int_{\{x\}} \frac{\partial \ln p(x,\theta)}{\partial \theta} p(x,\theta) dx = 0$$

- further manipulation of the previous result

$$
\begin{aligned}
0 &= \frac{\partial}{\partial \theta} \int_{\{x\}} \frac{\partial \ln p(x,\theta)}{\partial \theta} p(x,\theta) dx = \int_{\{x\}} \left[ \frac{\partial^2 \ln p(x,\theta)}{\partial \theta^2} p(x,\theta) + \frac{\partial \ln p(x,\theta)}{\partial \theta} \frac{\partial p(x,\theta)}{\partial \theta} \right] dx \\
&= \int_{\{x\}} \left[ \frac{\partial^2 \ln p(x,\theta)}{\partial \theta^2} + \frac{1}{p(x,\theta)} \frac{\partial \ln p(x,\theta)}{\partial \theta} \frac{\partial p(x,\theta)}{\partial \theta} \right] p(x,\theta) dx \\
&= \int_{\{x\}} \left\{ \frac{\partial^2 \ln p(x,\theta)}{\partial \theta^2} + \left[ \frac{\partial \ln p(x,\theta)}{\partial \theta} \right]^2 \right\} p(x,\theta) dx \\
&= \mathbf{E} \left[ \frac{\partial^2 \ln p(x,\theta)}{\partial \theta^2} \right] + \mathbf{E} \left[ \left( \frac{\partial \ln p(x,\theta)}{\partial \theta} \right)^2 \right]
\end{aligned}
$$

**Second Bartlett identity**

# The Cramér-Rao-Fisher bound and the Fisher information

- using both identities $$\mathrm{var}\left[\frac{\partial \ln L(x,\theta)}{\partial \theta}\right] = \mathbf{E}\left[\left(\frac{\partial \ln L(x,\theta)}{\partial \theta}\right)^2\right] = -\mathbf{E}\frac{\partial^2 \ln L(x,\theta)}{\partial \theta^2}$$

- expectation value of ML estimator $$\mathbf{E}[\hat{\theta}(D)] = \int_{\{x\}} \hat{\theta}(x) L(x,\theta) dx = \theta + b_n$$

- derivative of the previous expression

$$\begin{aligned}\frac{\partial}{\partial \theta}\mathbf{E}[\hat{\theta}(D)] &= 1 + \frac{\partial b_n}{\partial \theta} = \int_{\{x\}} \hat{\theta}(x) \frac{\partial L(x,\theta)}{\partial \theta} dx = \int_{\{x\}} \hat{\theta}(x) \frac{\partial \ln L(x,\theta)}{\partial \theta} L(x,\theta) dx \\ &= \mathbf{E}\left[\hat{\theta}(D)\frac{\partial \ln L(D,\theta)}{\partial \theta}\right] = \mathrm{cov}\left[\hat{\theta}(D), \frac{\partial \ln L(D,\theta)}{\partial \theta}\right] + \mathbf{E}\left[\hat{\theta}(D)\right]\mathbf{E}\left[\frac{\partial \ln L(D,\theta)}{\partial \theta}\right] \\ &= \mathrm{cov}\left[\hat{\theta}(D), \frac{\partial \ln L(D,\theta)}{\partial \theta}\right]\end{aligned}$$

# The Cramér-Rao-Fisher bound and the Fisher information (ctd.)

- we use Schwartz's inequality for covariance $[\mathrm{cov}(x,y)]^2 \leq \sigma_x^2 \sigma_y^2$

- we apply the inequality to the previous result $1 + \frac{\partial b_n}{\partial \theta} = \mathrm{cov}\left[\hat{\theta}(D), \frac{\partial \ln L(D,\theta)}{\partial \theta}\right]$ and find

$$\left(1 + \frac{\partial b_n}{\partial \theta}\right)^2 = \left\{\mathrm{cov}\left[\hat{\theta}(D), \frac{\partial \ln L(D,\theta)}{\partial \theta}\right]\right\}^2 \leq \mathrm{var}[\hat{\theta}(D)]\mathrm{var}\left[\frac{\partial \ln L(D,\theta)}{\partial \theta}\right]$$

- rearranging terms, we obtain the **Cramér-Rao-Fisher bound**

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

Definition of Fisher Information. A very concentrated pdf is very informative. Therefore, the smaller the variance, the greater the "information".

Thus, from the (unbiased, consistent) Cramér-Rao-Fisher bound

one is led to the Fisher Information

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

# 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

$$S_B = k_B \ln \Omega = k_B \ln \frac{N!}{N_1!N_2!\ldots N_M!} \approx \left[ (N \ln N - N) - \sum_i (N_i \ln N_i - N_i) \right]$$

$$= k_B \left[ N \ln N - \sum_i N p_i (\ln p_i + \ln N) \right] = k_B N \sum_i p_i \ln \frac{1}{p_i}$$

# From Boltzmann's entropy to Shannon's entropy

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

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

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

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

$$\Rightarrow \quad \boxed{p_j = \exp(\lambda \ln 2 - 1) = 1/N}$$

all probabilities have the same value

# The Kullback-Leibler divergence

The obvious extension of the Shannon entropy to continuous distributions

$$S = \int_{-\infty}^{+\infty} p(x)dx \log_2 \frac{1}{p(x)dx}$$

does not work, because it diverges.

A solution is suggested again by statistical mechanics ...

# Boltzmann entropy with degeneracy number attached to each level

$$\Omega = \frac{N!}{N_1!N_2!\dots N_M!} g_1^{N_1} g_2^{N_2} \dots g_M^{N_M}$$

$$\begin{aligned}\ln\Omega &= \ln N! - \sum_{k=1}^{M} \ln N_k! + \sum_{k=1}^{M} N_k \ln g_k \\ &= -N \sum_{k=1}^{M} (N_k/N) \ln \frac{(N_k/N)}{g_k} \\ &= -N \sum_{k=1}^{M} p_k \ln \frac{p_k}{g_k}\end{aligned}$$

Kullback-Leibler divergence

$$I_{KL} = \sum_{k=1}^{M} p_k \ln \frac{p_k}{g_k}$$

# Properties of the Kullback-Leibler divergence

- extreme value when $p_k = g_k$.
  Indeed, using again a Lagrange multiplier we must consider the auxiliary function

$$I_{KL} + \lambda \sum_k p_k$$

and we find the extreme value at

$$p_k = g_k e^{\lambda - 1} \underset{\substack{\uparrow \\ \text{normalization}}}{=} g_k$$

**(homework!)**

The KL divergence is a measure of the number of excess bits that we must use when we take a distribution of symbols which is different from the reference distribution

$$\begin{aligned} I_{KL} &= \sum_{k=1}^{M} p_k \ln \frac{p_k}{g_k} \\ &= \sum_{k=1}^{M} p_k \ln \frac{1}{g_k} - \sum_{k=1}^{M} p_k \ln \frac{1}{p_k} \end{aligned}$$

# The KL divergence for continuous distributions does not diverge

$$
\begin{aligned}
I_{KL} &= \sum_k p_k \ln \frac{p_k}{g_k} \\
&\to \int_{-\infty}^{+\infty} p(x)dx \ln \frac{p(x)dx}{g(x)dx} \\
&= \int_{-\infty}^{+\infty} p(x) \ln \frac{p(x)}{g(x)} \, dx
\end{aligned}
$$

# The KL divergence is non-negative

Notice first that when we define $\phi(t) = t\ln t$ we find

$$\phi(t) = \phi(1) + \phi'(1)(t-1) + \frac{1}{2}\phi''(h)(t-1)^2 = (t-1) + \frac{1}{2h}(t-1)^2$$

where $t < h < 1$ and therefore

$$\begin{aligned} I_{KL} &= \int_{-\infty}^{+\infty} p(x)\ln\frac{p(x)}{g(x)}dx = -\int_{-\infty}^{+\infty} \frac{p(x)}{g(x)}\ln\frac{p(x)}{g(x)}g(x)dx = \int_{-\infty}^{+\infty}\phi\left(\frac{p(x)}{g(x)}\right)g(x)dx \\ &= \int_{-\infty}^{+\infty}\left[\left(\frac{p(x)}{g(x)}-1\right) + \frac{1}{2h}\left(\frac{p(x)}{g(x)}-1\right)^2\right]g(x)dx = \int_{-\infty}^{+\infty}\frac{1}{2h}\left(\frac{p(x)}{g(x)}-1\right)^2 g(x)dx \\ &= \int_{-\infty}^{+\infty}\frac{1}{2h}\frac{(p(x)-g(x))^2}{g(x)}dx \geq 0 \end{aligned}$$

# The KL divergence is a quasi-metric (however a local version of the KL divergence is the Fisher information, which is a true metric)

The KL divergence can be used to measure the “distance” between two distributions.

**Example**: the KL divergence

$$I_{KL}(p,q) = \int_{-\infty}^{+\infty} p(x) \ln \frac{p(x)}{q(x)} dx$$

for the distributions

$$\begin{aligned} p(x) &= \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{x^2}{2\sigma^2}\right) \\ q(x) &= \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) \end{aligned} \quad \Rightarrow \quad I_{KL}(p,q) = \frac{\mu^2}{2\sigma^2}$$

Now consider a family of parametric distributions and evaluate the KL divergence between two close elements of the family

$$I_{KL}\left(p(x,\theta),p(x,\theta+\epsilon)\right) = \int_{-\infty}^{+\infty} p(x,\theta) \ln \frac{p(x,\theta)}{p(x,\theta+\epsilon)} dx$$
$$= \mathbf{E}\left(\ln p(x,\theta) - \ln p(x,\theta+\epsilon)\right)$$

Since

$$\ln p(x,\theta+\epsilon) \approx \ln p(x,\theta) + \frac{\partial \ln p(x,\theta)}{\partial\theta}\epsilon + \frac{1}{2}\frac{\partial^2 \ln p(x,\theta)}{\partial\theta^2}\epsilon^2$$

we find, using the first Bartlett identity,

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

i.e., locally the KL divergence is just the Fisher information

# The KL divergence can be transformed into a true distance between pdf's

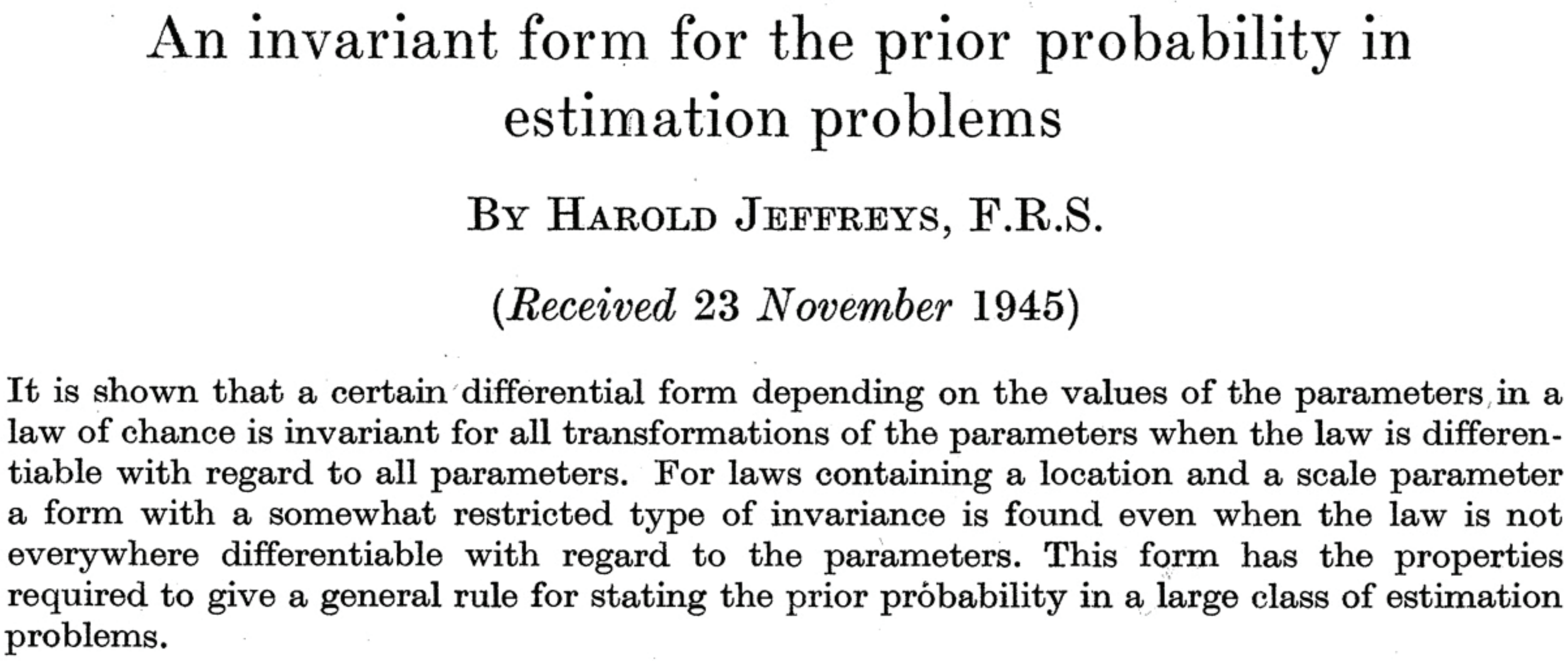
- Jeffreys' distance

$$I_J(p,q) = \frac{1}{2} I_{KL}(p,q) + \frac{1}{2} I_{KL}(q,p)$$

- Jensen-Shannon distance

$$I_{\mathrm{JS}}(p,q) = \frac{1}{2} I_{KL}\left(p, \frac{p+q}{2}\right) + \frac{1}{2} I_{KL}\left(q, \frac{p+q}{2}\right)$$

# 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.

The KL divergence is invariant with respect to generic random variable transformations.

From the definition of KL divergence, and from the transformation formula for pdf's we find

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

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

In this case, our random variables are the parameter estimates, therefore the KL divergence is invariant with respect to parameter (random variable transformations), therefore the associated Fisher Information from the local expansion of the KL divergence is also invariant with respect to parameter transformations.

From the equation that relates KL divergence and Fisher Information, we find a corresponding pdf as follows. Equation

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

means that the KL divergence depends quadratically on small changes of the expansion parameter and that the KL divergence remains constant if the term on the r.h.s. remains constant.

Dimensionally, the Fisher information is quadratic with respect to a pdf, therefore, we have to take the square root of that constant to define a pdf, i.e.,

This must be normalized to obtain a pdf that is invariant with respect to parameter transformations.

**Example**: a simple Gaussian Likelihood for *n* datapoints, with known variance

➡

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

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, with known mean

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

This power-law pdf is another improper prior.

**Example**: Poisson distribution

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

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

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

This power-law pdf is yet another improper prior.

## **Example**: binomial distribution

$$L(D|\theta) = \binom{N}{n} \theta^n (1-\theta)^{N-n}$$

➡

$$\ln L(D|\theta) \sim n \ln \theta + (N-n) \ln(1-\theta)$$

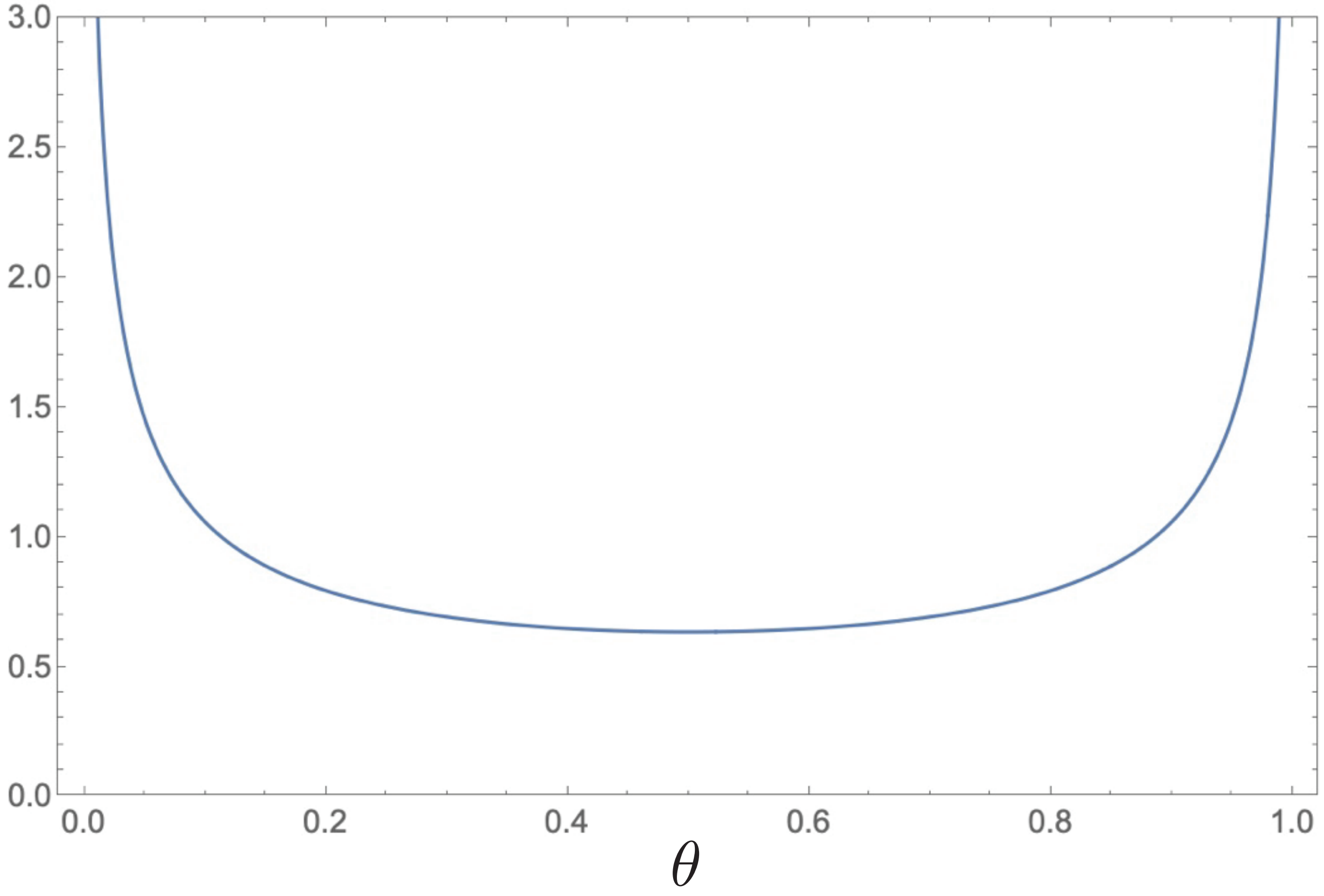
➡

$$\begin{aligned} \mathbf{E}\left[-\frac{\partial^2 \ln L(D|\theta)}{\partial \theta^2}\right] &\sim \frac{N\theta}{\theta^2} + \frac{N - N\theta}{(1-\theta)^2} \\ &= \frac{N}{\theta} + \frac{N}{1-\theta} \\ &= \frac{N}{\theta(1-\theta)} \end{aligned}$$

➡

$$I(\theta) \sim \frac{1}{\theta(1-\theta)}$$

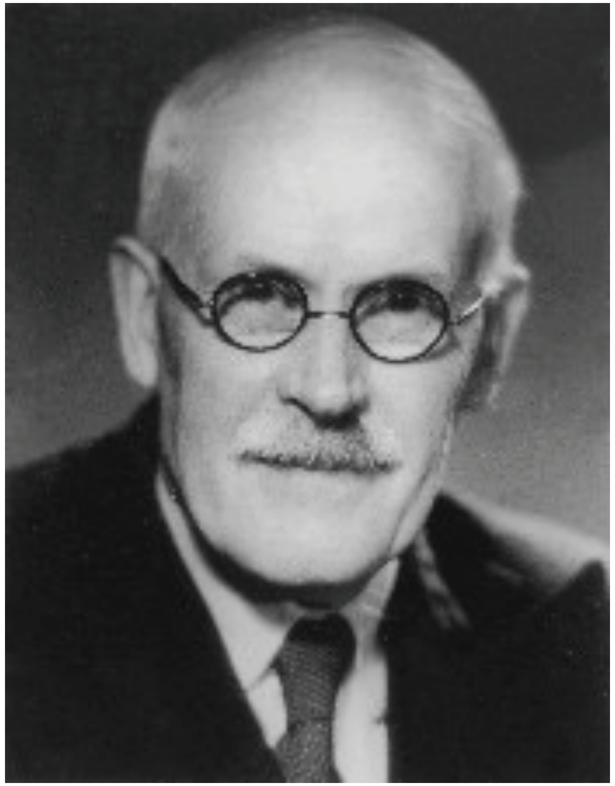
⇒ $$\sqrt{I(\theta)} \sim \frac{\theta^{1/2}(1-\theta)^{1/2}}{B(1/2,1/2)}$$



# A lesson learned from Jeffreys priors

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

Well ... no, the Likelihood is also constructed using prior information (obviously!). So, in this approach the Likelihood and the priors are both determined using the available prior information.



Harold Jeffreys
(1891-1989)

# Reference priors

In this case we need the definition of "sufficient statistic"

*A statistic t is sufficient with respect to a statistical model and its associated unknown parameter if "no other statistic that can be calculated from the same sample provides any additional information as to the value of the parameter" (Fisher, 1920)*

Given the data **D**, a statistic $t$ = T(**D**) is sufficient with respect to the parameter if it contains all the information needed to estimate the parameter.

Examples:

- the sample mean is sufficient for the mean of a normal distribution with known variance. Once the sample mean is known, no further information about the mean can be obtained from the sample itself.
- for an arbitrary distribution the median is not sufficient for the mean: even if the median of the sample is known, knowing the sample itself would provide further information about the population mean.

The idea behind a reference prior is that it must be such that data affect our posterior distribution the most. We can formalize this by means of the KL divergence by requiring that the KL divergence between prior and posterior be maximal.

To proceed, we utilize a posterior that depends on a sufficient statistic instead of the original data

$$I_{KL} = \int_\Theta p(\theta|t) \ln \frac{p(\theta|t)}{p(\theta)} d\theta$$

then, its expection value over the statistic is

$$\begin{aligned}\mathbf{E}\left[I_{KL}\right]_t &= \int_T p(t) \int_\Theta p(\theta|t) \ln \frac{p(\theta|t)}{p(\theta)} d\theta \; dt \\ &= \int_T \int_\Theta p(\theta|t)p(t) \ln \frac{p(\theta|t)p(t)}{p(\theta)p(t)} d\theta \; dt \\ &= \int_T \int_\Theta p(\theta, t) \ln \frac{p(\theta, t)}{p(\theta)p(t)} d\theta \; dt\end{aligned}$$

Mutual information between the two distributions

A reference prior is a pdf that maximizes the mutual information

$$\int_T \int_\Theta p(\theta, t) \ln \frac{p(\theta, t)}{p(\theta)p(t)} d\theta \; dt$$

This is a complex variational problem, and we shall not pursue it further in this course.