# MultiDimensional Distributions

## Concepts, Bayes Theorem, and Direct Sampling

Sachin Shanbhag

Department of Scientific Computing
Florida State University,
Tallahassee, FL 32306.

# Orientation

For the most part, we will focus on 2D distributions, the simplest multidimensional distribution.

This will allow us to introduce some terminology, including:

- Joint PDF
- Marginal PDF
- Conditional PDF

These concepts are connected in one of the most famous applications of MCMC: the Bayes Theorem.

Bayes Theorem usually leads to a PDF, which for non-trivial cases, is explored using MC.

The accept-reject direct Monte Carlo method is perfectly sufficient for simple PDFs, and completely inadequate for complicated PDFs.

# Example: 2D Normal Distribution

General 2D normal or Gaussian distribution:

$$
\begin{aligned}
f(x,y) &= A\exp\left(-BC\right) \\
A &= \frac{1}{2\pi\sigma_x\sigma_y\sqrt{1-\rho^2}} \\
B &= \frac{1}{2(1-\rho^2)} \\
C &= \frac{(x-\mu_x)^2}{\sigma_x^2} + \frac{(y-\mu_y)^2}{\sigma_y^2} - \frac{2\rho(x-\mu_x)(y-\mu_y)}{\sigma_x\sigma_y}
\end{aligned}
$$

It describes how $x$ and $y$ covary.

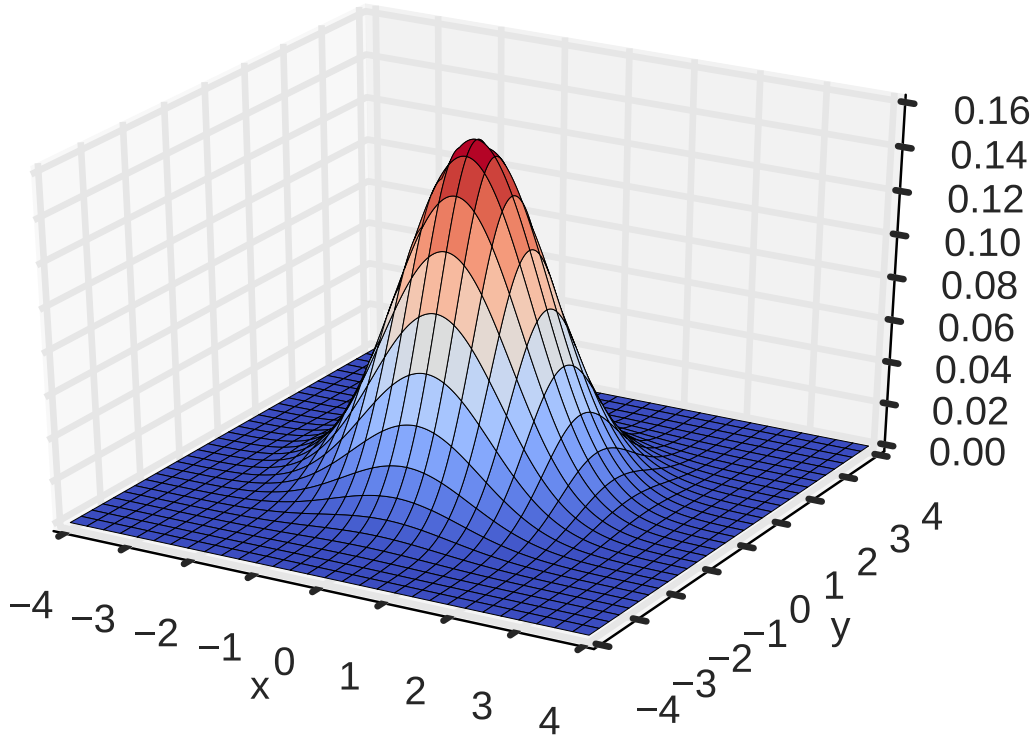
A PDF which describes the variation of multiple random variables together is called a joint probability distribution function.

# Example: 2D Normal Distribution

Lets visualize $f(x, y)$ for different values of $\sigma_x$, $\sigma_y$, and $\rho$;

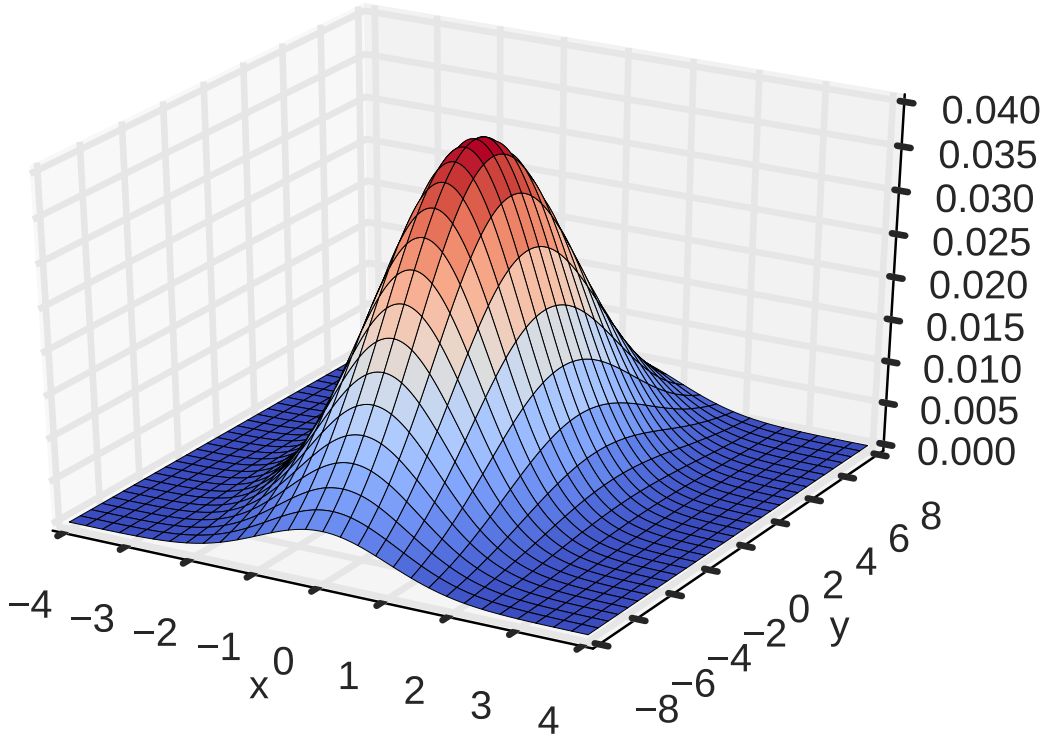
$-1 < \rho < 1$ is called the correlation coefficient

$$\sigma_x = \sigma_y = 1.0; \quad \rho = \mu_x = \mu_y = 0$$



# Example

$$\sigma_x = 1.0; \sigma_y = 4.0; \quad \rho = 0$$

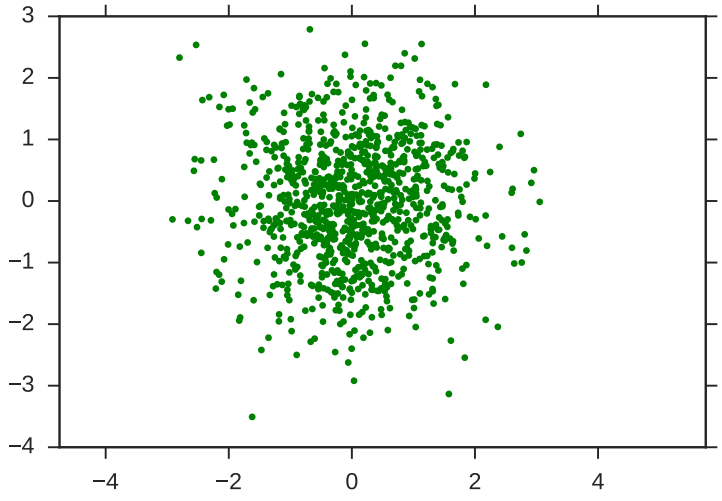



Note the increased spread in the $y$-direction.

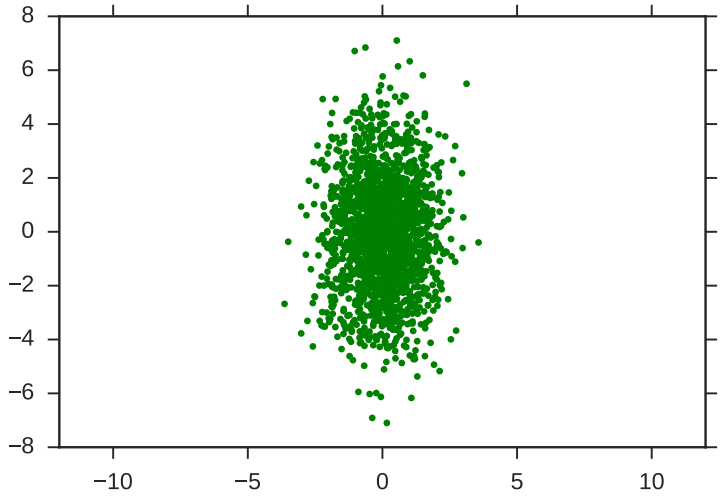
Let us draw samples from such "uncorrelated" distributions.

# Samples

$\sigma_x = 1.0; \sigma_y = 1.0; \rho = 0$




$\sigma_x = 1.0; \sigma_y = 2.0; \rho = 0$



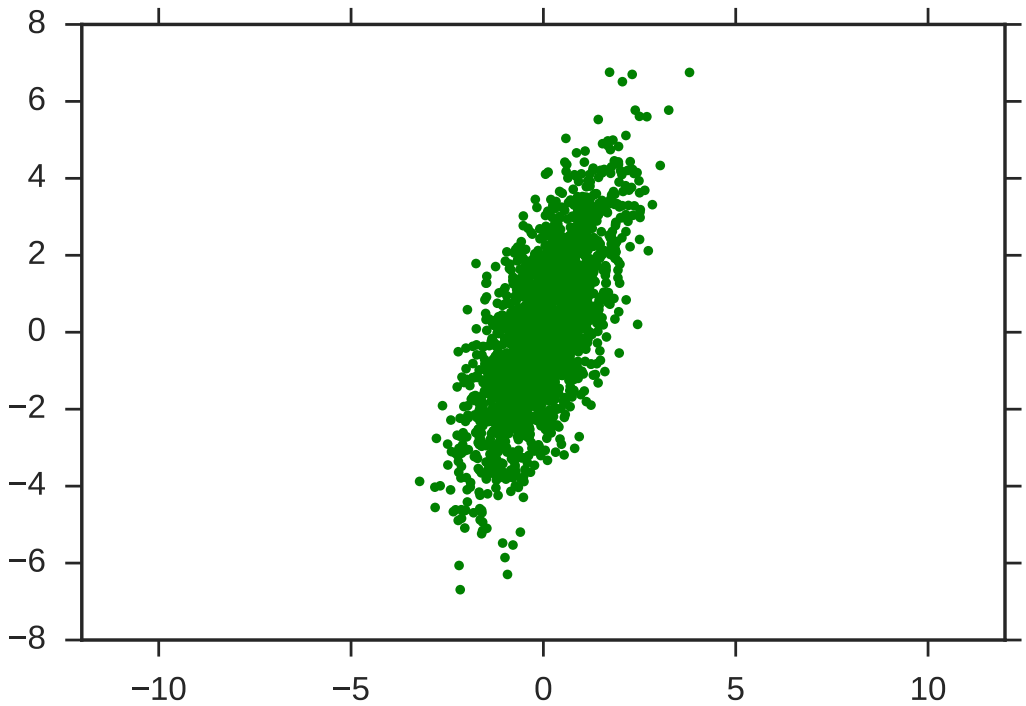

Note that the scale for the two plots is different; $\sigma_x$ and $\sigma_y$ control the scale or dispersion.

We have set $\mu_x = \mu_y = 0$, in these visualizations; changing them simply changes the location of the "center" or the "peak".

Finally, let us change $\rho$ to something non-zero.

# Sample: Nonzero correlation

$$\sigma_x = 1.0; \sigma_y = 2.0; \rho = 0.7$$




$\rho$ controls "tilt".

Practically, such a PDF may correspond to something like the joint PDF of height $(x)$ and weight $(y)$.

# Correlated Random Variables

Generally tall people are heavier than short people, and there is a positive correlation between height and weight ($\rho = 0.7$).

Exercise: Give other examples from everyday life of quantities that are positively and negatively correlated.

i.e., $x$ and $y$ are correlated. Knowing the value of one *tells you something* about the other.[1]

Two random variables $x$ and $y$ are uncorrelated if,

$$E(xy) = E(x)E(y),$$

and, thus their covariance

$$\mathrm{Cov}(x, y) = E(xy) - E(x)E(y) = 0$$

[1] The relationship between independent and uncorrelated variables is discussed in this handout from CMU Advanced Data Analysis

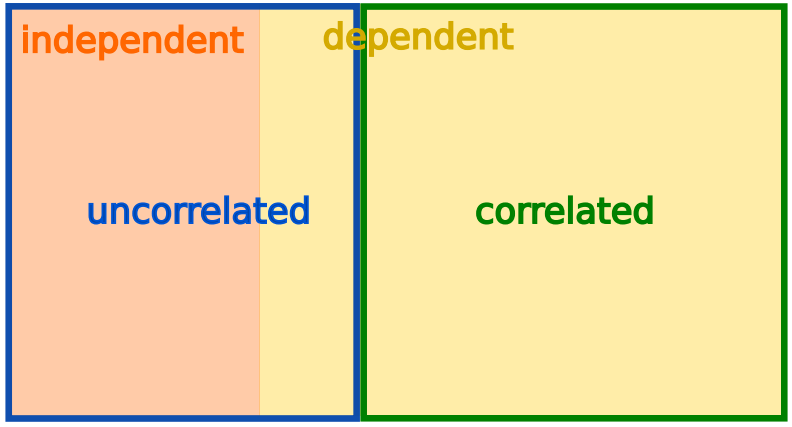
# Independent and Uncorrelated Variables

Two random variables $x$ and $y$ are independent if,

$$f(x, y) = f(x)f(y).$$

Multidimensional distributions can be explicitly written as product of 1D distributions, only when the random variables (corresponding to the "dimensions") are independent.

Independence is the stricter criterion. It implies no correlation. But uncorrelated variables need not be independent.



# Joint, Marginal and Conditional PDF

Consider two random variables $x_1$ and $x_2$ (think of height and weight, if you like).

If they are continuous, then a PDF $f(x_1, x_2)$, like the 2D normal distribution, which describes their collective variation is called a *joint* PDF.

The probability of a state $(X_1, X_2)$, with $x_1 \leq X_1 < x_1 + dx_1$ and $x_2 \leq X_2 < x_2 + dx_2$ is represented by $f(x_1, x_2)dx_1dx_2$.

Note that

$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(x_1, x_2)dx_1dx_2 = 1$$

If the random variables are **independent**,

$$f(x_1, x_2) = f(x_1)f(x_2).$$

# Marginal PDF

A marginal distribution of a particular random variable $x_i$, represented by $f(x_i)$ is obtained from a joint PDF by *integrating out the all the other random variables*.

Thus, for a $n$-dimensional joint PDF,

$$f(x_i) = \int f(x_1, x_2, ..., x_n) dx_1 dx_2 ... dx_{i-1} dx_{i+1} ... dx_n$$

In 2D

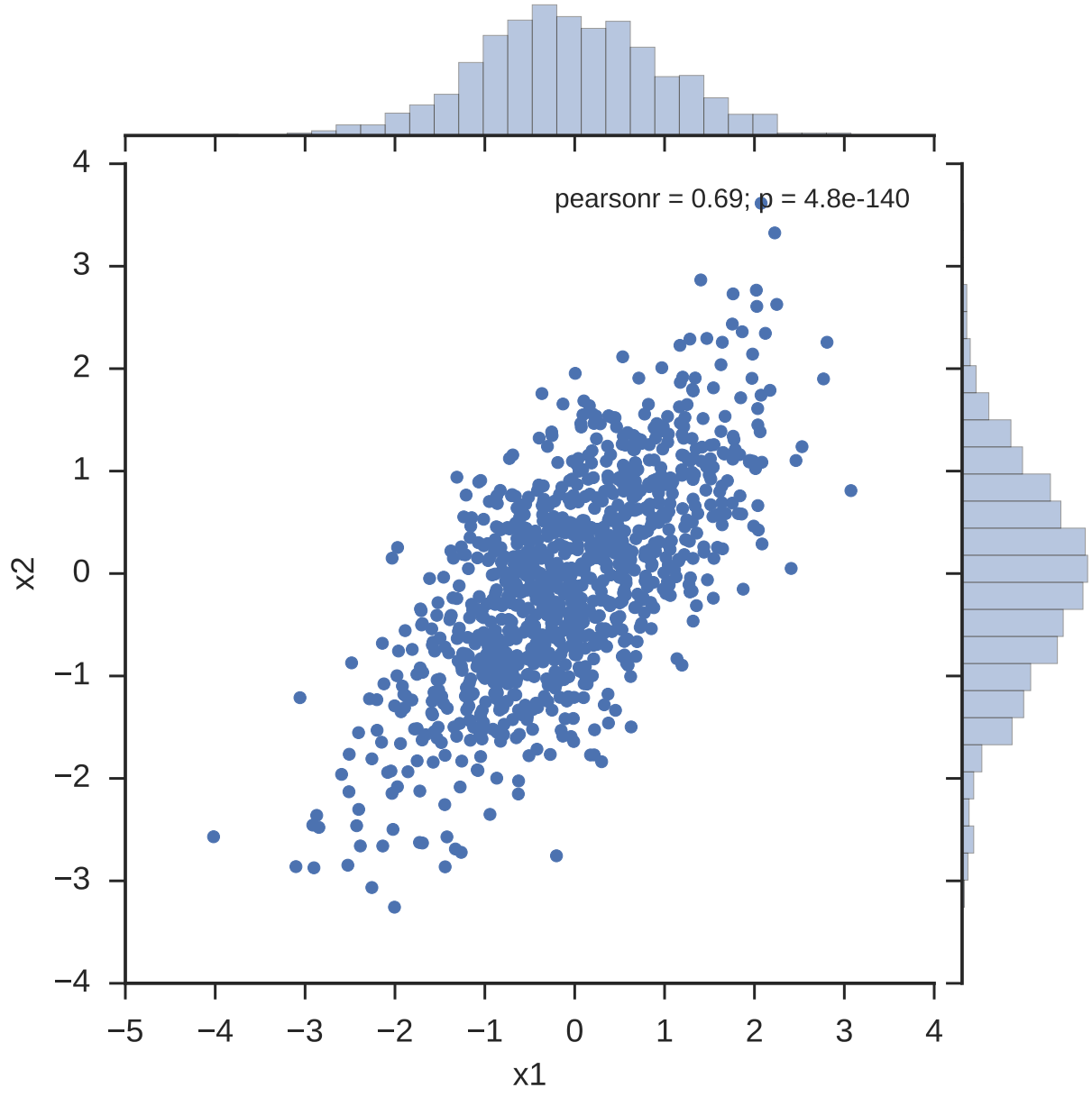
$$f(x_1) = \int f(x_1, x_2)\, dx_2,$$

and

$$f(x_2) = \int f(x_1, x_2)\, dx_1.$$

This can graphically be shown as:

# Marginal PDF



# Marginal PDF

Marginal distributions are 1D PDFs. They offer a particular summary of the joint PDF.

One can get marginal PDFs from joint PDFs, but not the other way around.

The joint PDF is the richest "lossless" description of the variation.

In the previous plot, notice that the marginal distributions looked Gaussian.

They did not reflect the correlation between the quantities.

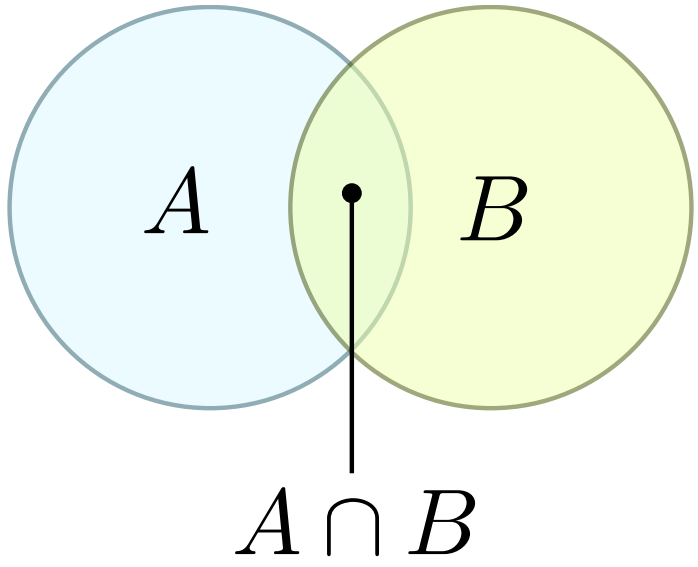
This information was lost.

Analogy: 2D shadows of 3D objects

# Conditional Probability

Conditional probability of event $A$ given $B$ is given by,

$$P(A|B) = \frac{P(A \cap B)}{P(B)}.$$




Example 1: If you roll a fair die, what is the probability of observing a number ($n$) less than 3, given $n$ is even?

# Conditional Probability Examples

We want to compute P($n < 3 \mid n$ is even).

$$= \frac{P(n < 3 \text{ and even})}{P(n \text{ is even})} = \frac{(1/6)}{(3/6)} = 1/3$$

Example 2: Life expectancy tables

The life expectancy of a newborn is 75 years.

The "conditional" life expectancy for a 70 year old is more than 80 (link).

Thus, if $y$ is life expectancy, what one may care about is $P(y|\text{age})$.

# Conditional PDF

In the previous joint PDF plot, suppose we wanted to describe the distribution of $x_1$ at a particular value of $x_2$.

This is written as $f(x_1|x_2)$. I read the "|" sign as "given".

It can be shown that:

$$f(x_1|x_2) = \frac{f(x_1, x_2)}{f(x_2)}, \qquad \text{conditional} = \frac{\text{joint}}{\text{marginal}}$$

where the $f(x_2)$ in the denominator is the marginal PDF.

Thus,

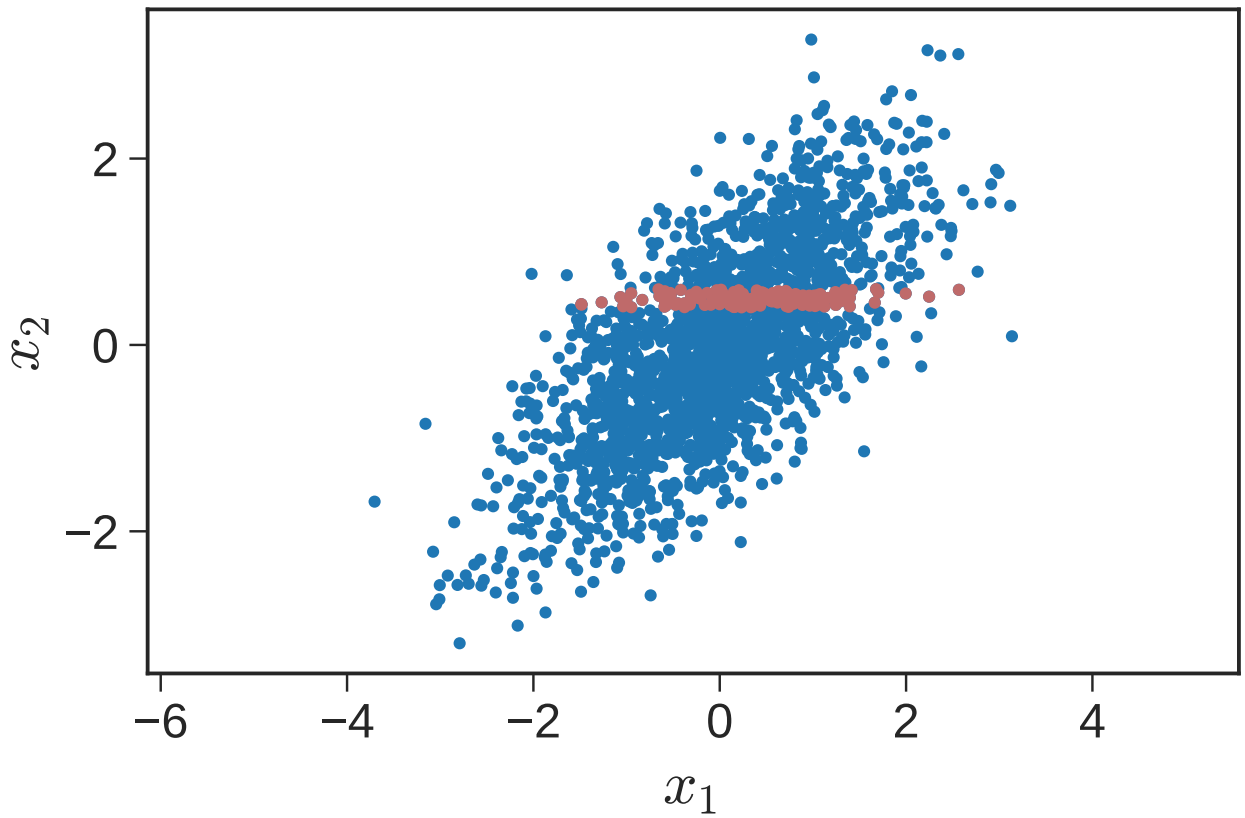
$$f(x_1|x_2) = \frac{f(x_1, x_2)}{\int f(x_1, x_2) dx_1}.$$

Conditional PDFs can be obtained from the joint PDF.

# Conditional PDF

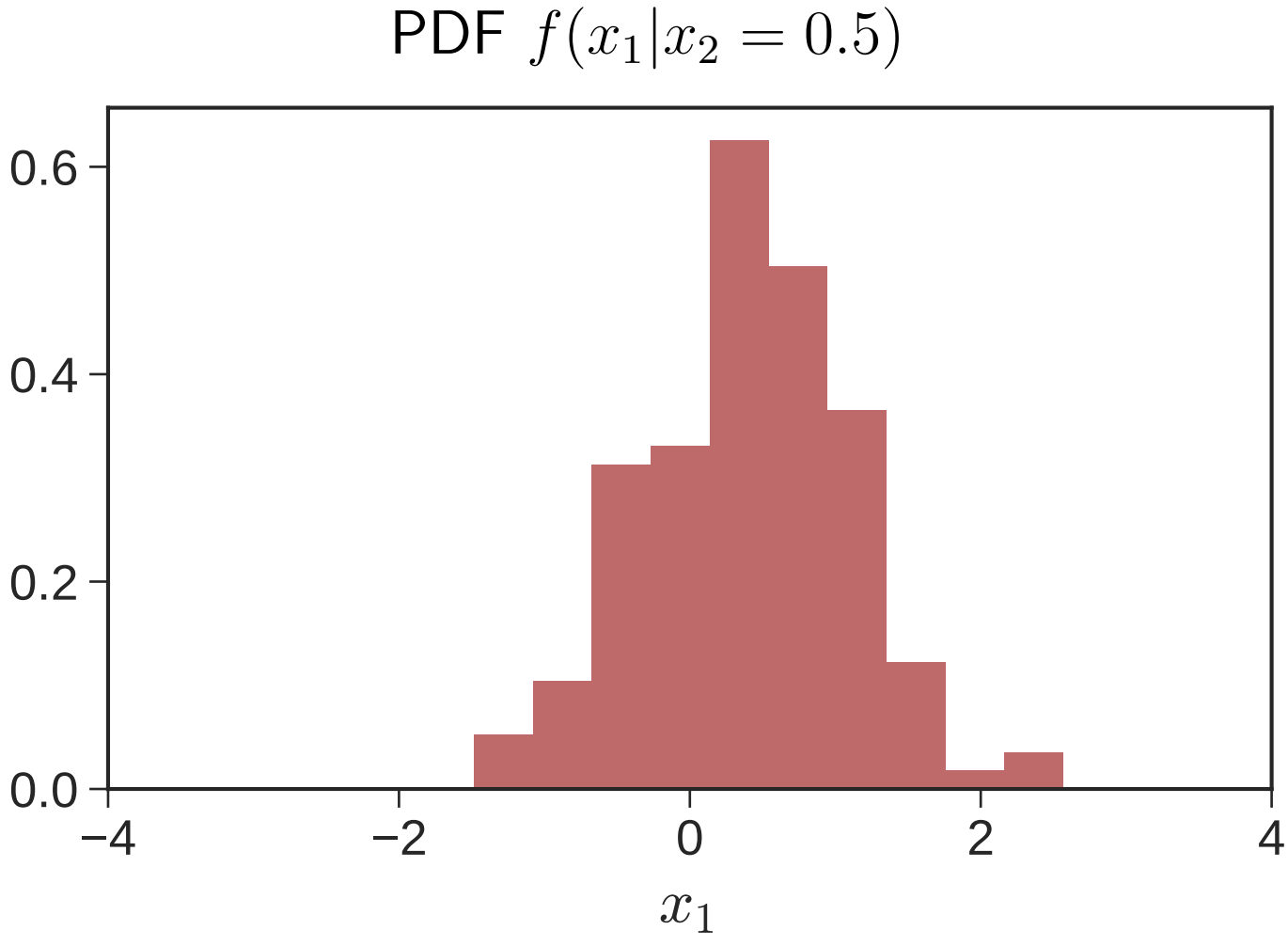
Lets look at some pictures again.

Suppose we want to consider the distribution $f(x_1|x_2 = 0.5)$




$f(x_1|x_2 = 0.5)$ is the 1D histogram/PDF of the red dots.

# Conditional PDF




note that it is not centered at zero

# Bayes Theorem

Bayesian analysis is an increasingly important application of MCMC.

It uses joint, marginal, and conditional PDFs.

Let's derive Bayes rule using a mathematical approach.

Later, we will interpret it using "physical reasoning" to gain some intuition on what it implies, and how it relates to MCMC.

We saw that

$$\begin{aligned} f(x_1|x_2) &= \frac{f(x_1,x_2)}{f(x_2)}, \\ f(x_2|x_1) &= \frac{f(x_1,x_2)}{f(x_1)}. \end{aligned}$$

Thus, it is easy to combine them as:

# Bayes Theorem

$$f(x_1|x_2)f(x_2) = f(x_2|x_1)f(x_1),$$

or

$$f(x_1|x_2) = \frac{f(x_2|x_1)f(x_1)}{f(x_2)},$$

The $f(x_2)$ in the denominator can be written as

$$f(x_2) = \int f(x_1, x_2)dx_1,$$

$$f(x_2) = \int f(x_2|x_1)f(x_1)dx_1$$

This allows us to write Bayes theorem in a compact symmetric form:

$$f(x_1|x_2) = \frac{f(x_2|x_1)f(x_1)}{\int f(x_2|x_1)f(x_1)dx_1}$$

# Bayes Theorem: Interpretation

$$f(x_1|x_2) = \frac{f(x_2|x_1)f(x_1)}{\int f(x_2|x_1)f(x_1)dx_1}$$

Denominator:

- It can be thought of as a normalizing factor; it is just a number, which ensures that the PDF $f(x_1|x_2)$ is normalized.
- Often it is not easy to compute for large multidimensional systems.
- If you have a physics/chemistry background, this can be thought of as a partition function, which is related to the concepts of entropy or free-energy.
- As it turns out, in many important problems, we don't need to explicitly evaluate the denominator.

# Bayes Theorem: Interpretation

Thus, the form:

$$f(x_1|x_2) \propto f(x_2|x_1)f(x_1)$$

is often used in computational implementations of Bayes Rule.

The typical use case of Bayes theorem is the statistical analysis of data for some type of inference.

Let us rewrite the theorem using more intuitive variables:

$$f(m|d) \propto f(d|m)f(m),$$

where $m$=model, and $d$ = data.

# Bayes Theorem: Data Analysis

$$f(m|d) \propto f(d|m)f(m),$$

**Prior Probability**

We start with a prior belief $f(m)$ of the correct model.

**Likelihood**

$f(d|m)$ is the probability of observing the data $d$, if the model $m$ were correct.

**Posterior probability**

$f(m|d)$ is our updated belief in the model $m$ based on the strength of the evidence $d$.

# Sampling 2D distribution

Some multidimensional problems can be approached with the methods developed for arbitrary 1D distributions.

However, their scope is fairly limited.

Let us consider two examples, which highlight the niches occupied by:

1. transformation method
2. accept-reject method

# Transformation method

Consider:

$$(X, Y) \sim f(x, y),$$

and

$$(U, V) \sim g(u, v).$$

Suppose the relationships $x(u, v)$ and $y(u, v)$ are known.

Then the relationship between the two PDFs is given by,

$$g(u, v) = f(x, y)|\mathsf{det}(J)|,$$

where the Jacobian $J$ is given by,

$$J = \frac{\partial(x, y)}{\partial(u, v)} = \begin{bmatrix} \partial x/\partial u & \partial x/\partial v \\ \partial y/\partial u & \partial y/\partial v \end{bmatrix}.$$

# Example: Origin of $J$

Consider transformation between $(x, y)$ and cylindrical coordinates $(r, \theta)$.

$$x = r\cos\theta$$
$$y = r\sin\theta$$

Therefore,

$$dA = dx\,dy = |\mathsf{det}(J)|\,dr\,d\theta,$$

where

$$J = \begin{bmatrix} \partial x/\partial r & \partial x/\partial\theta \\ \partial y/\partial r & \partial y/\partial\theta \end{bmatrix} = \begin{bmatrix} \cos\theta & -r\sin\theta \\ \sin\theta & r\cos\theta \end{bmatrix} = r$$

Therefore, $dx\,dy = r\,dr\,d\theta$.

# Consistency with 1D formula

In the 1D case, we assumed

$$u \sim U[0, 1],$$

which led to the method

$$x = F^{-1}(u).$$

Is the older recipe, consistent with this "new" formula?

For 1D, the general formula specializes to:

$$g(u) = f(x)|dx/du|.$$

Assume $x(u)$ is a monotonously increasing function, so that

$$|dx/du| = dx/du.$$

# Consistency with 1D formula

Thus, $g(u) = f(x)dx/du$, which can be rephrased as:

$$\int g(u)du = \int f(x)dx \implies u = F(x).$$

Thus, this is consistent with our older recipe.

The goal of solving $u = F(x)$ was to develop a transformation $x(u)$, which would enable us to sample from $f(x)$.

For multidimensional distributions, it is often hard to develop suitable transformations, given the distribution to sample from.

The utility of the transformation method (in such cases) is to figure out the distribution, that corresponds to a given the transformation.

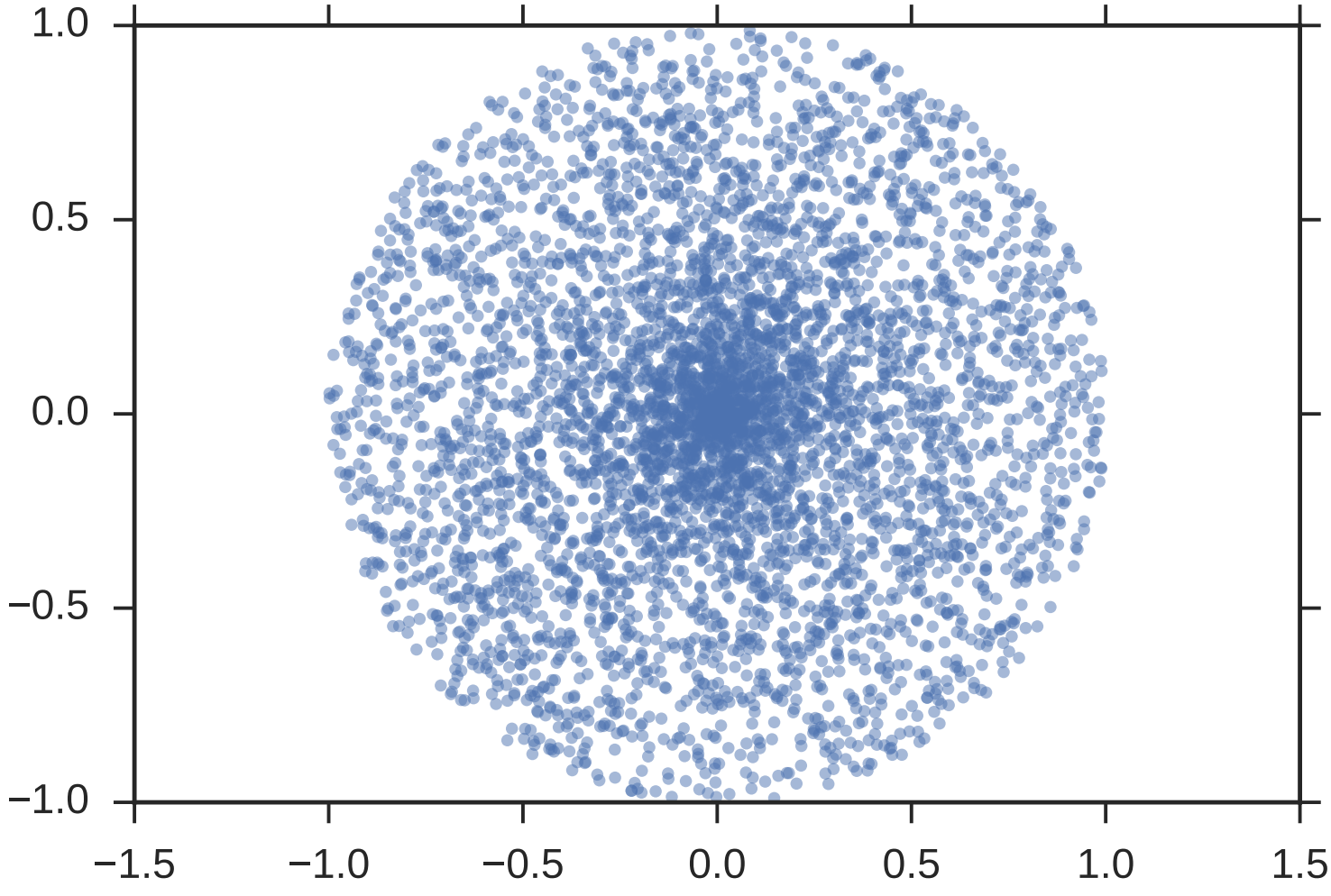
# Example

Using cylindrical coordinates $(r, \theta)$ sample points uniformly on a disc of radius $a$, centered at $(0, 0)$.

Wrong Solution: It may be tempting to sample:

$$r \sim U[0, a]$$
$$\theta \sim U[0, 2\pi]$$

```
a     = 1 # radius
npts  = 5000
r     = np.random.uniform(0, a, npts)
theta = np.random.uniform(0, 2*np.pi, npts)
x     = r * np.cos(theta)
y     = r * np.sin(theta)
```




too many points near the core

# Transformation of Variables

The target distribution is zero everywhere except,

$$f(x,y) = \frac{1}{\pi a^2}, \text{ for } x^2 + y^2 \leq a^2$$

Using the transformation rule,

$$\begin{aligned} g(r,\theta) &= \left|\frac{\partial(x,y)}{\partial(r,\theta)}\right| f(x,y) \\ &= r f(x,y) \\ &= \frac{r}{\pi a^2}, \text{ for } 0 < r \leq a \\ &= R(r)T(\theta) \end{aligned}$$

where,

$$R(r) = \frac{2r}{a^2}, 0 < r \leq a$$

$$T(\theta) = \frac{1}{2\pi}, 0 < \theta \leq 2\pi$$

# Example

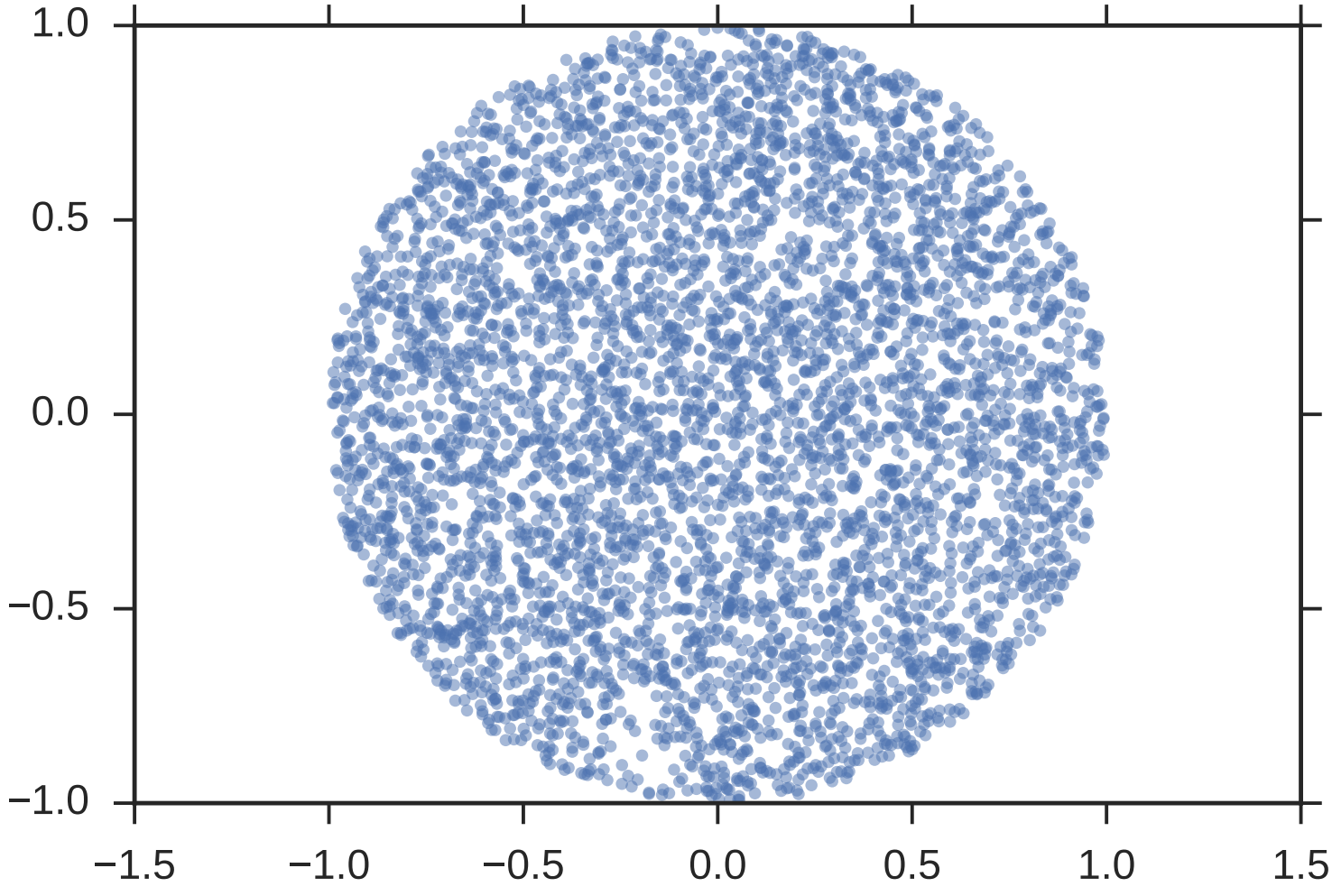
Note:

- unlike $\theta$, $r$ has to be sampled from a non-uniform distribution
- more points at large $r$.
- both $R(r)$ and $T(\theta)$ are appropriately normalized

One can sample from $R(r)$ by setting $r \sim a\sqrt{u}$, where $u \sim U[0, 1]$.

Code

```
r     = a * np.sqrt(np.random.uniform(0, 1, npts))
theta = np.random.uniform(0, 2*np.pi, npts)
x     = r * np.cos(theta)
y     = r * np.sin(theta)
```

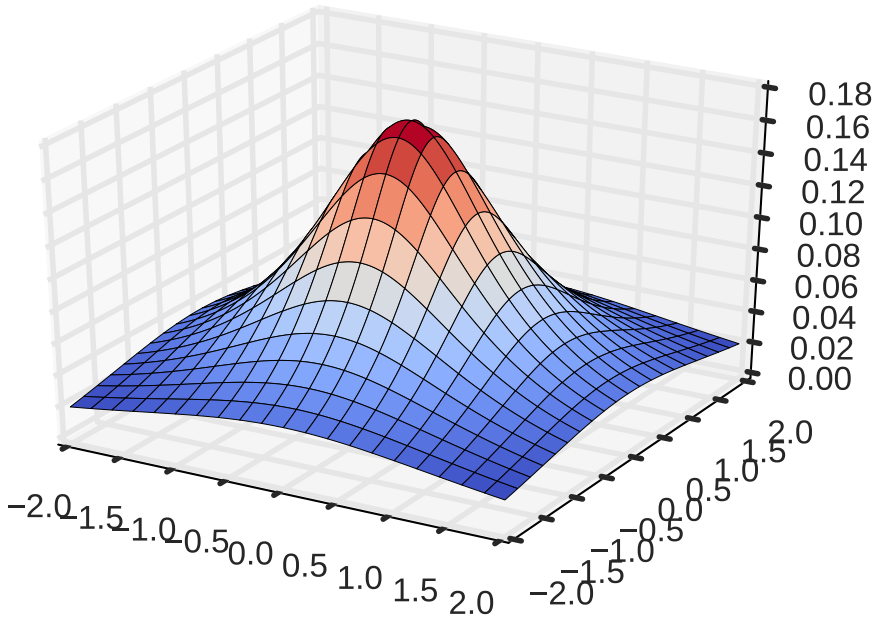

1.0
0.5
0.0
−0.5
−1.0
−1.5
−1.0
−0.5
0.0
0.5
1.0
1.5

# Accept-Reject Method

This method translates rather smoothly. Consider the following example; sample,

$$(x_1, x_2) \sim f(x_1, x_2) = \frac{0.1776}{x_1^2 + x_2^2 + 1},$$

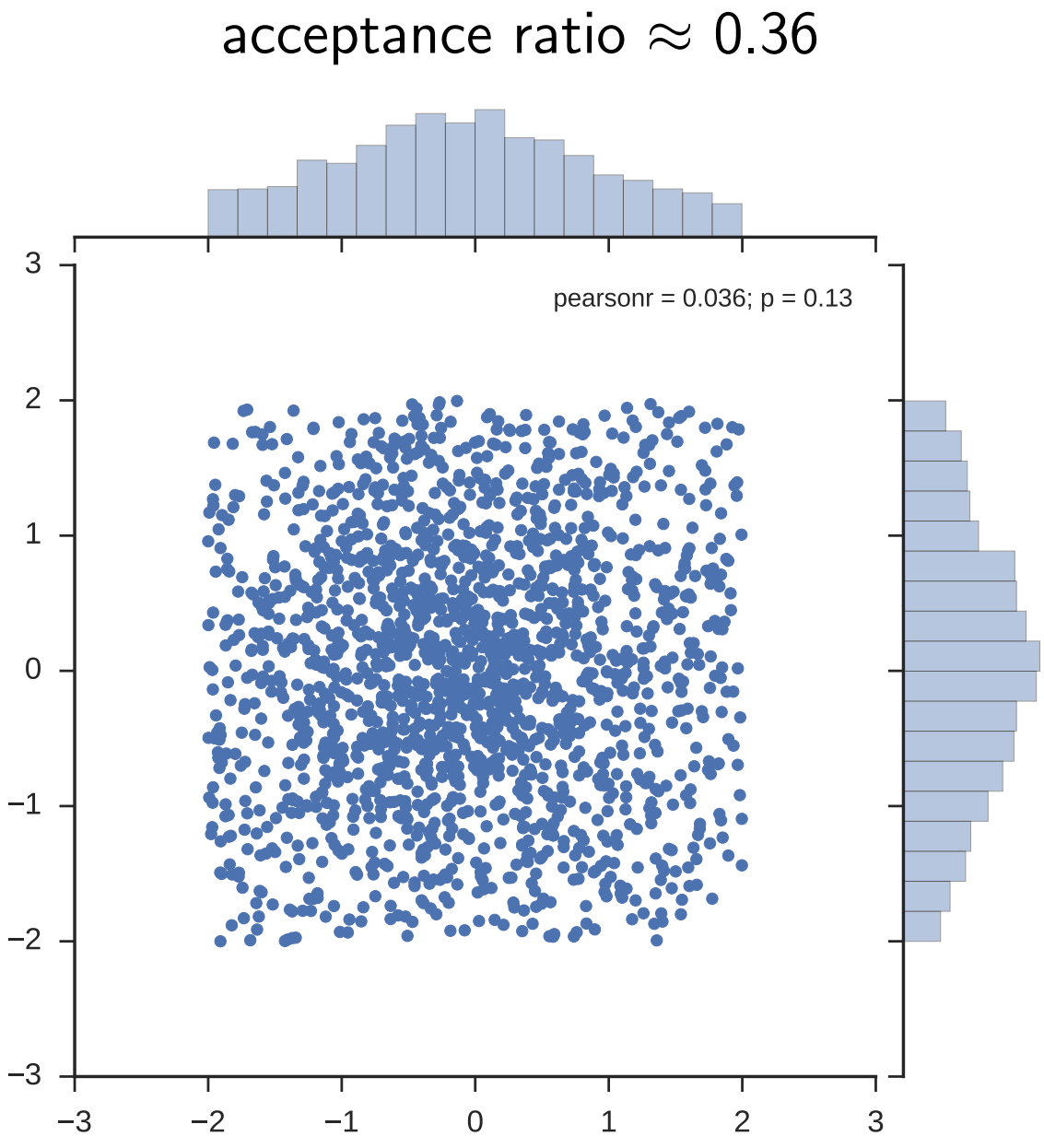
with $-2 \leq x_1 \leq 2$ and $-2 \leq x_2 \leq 2$.



# Solution: Throwing Darts!

```
def exampleSampler(ndarts):

    xmin = -2.
    xmax = 2.
    fmax = 0.1776

    x1 = np.random.uniform(xmin, xmax, ndarts)
    x2 = np.random.uniform(xmin, xmax, ndarts)
    u  = np.random.uniform(0., fmax, ndarts)

    const = u <= fmax/(x1**2 + x2**2 + 1.0)

    return x1[const], x2[const]
```

# Solution



# Limitations of Accept-Reject Method

Let us consider a few examples where the accept-reject method becomes impractical.

- High number of dimensions
- Peaked PDFs, “rough PDF landscapes”
- Many physical problems

# Appendix

# Appendix: Transformation of Variables

Consider a 1-1 mapping between $(x, y)$ and $(u, v)$ in $\Re^2$.

Suppose,

$$x = x(u, v)$$
$$y = y(u, v)$$

Then, it can be shown that:

$$dA = dx\ dy = \left|\det\left(\frac{\partial(x, y)}{\partial(u, v)}\right)\right| du\ dv = |\det(J)| du\ dv,$$

where the Jacobian,

$$J = \frac{\partial(x, y)}{\partial(u, v)} = \begin{bmatrix} \partial x/\partial u & \partial x/\partial v \\ \partial y/\partial u & \partial y/\partial v \end{bmatrix} = \begin{bmatrix} \partial x/\partial u & \partial y/\partial u \\ \partial x/\partial v & \partial y/\partial v \end{bmatrix}.$$

Note: determinant remains same if we switch rows and columns

# Example

Consider transformation between $(x, y)$ and cylindrical coordinates $(r, \theta)$.

$$x = r\cos\theta$$
$$y = r\sin\theta$$

Therefore,

$$dA = dx\,dy = |\mathsf{det}(J)|\,dr\,d\theta,$$

where

$$J = \begin{bmatrix} \partial x/\partial r & \partial x/\partial\theta \\ \partial y/\partial r & \partial y/\partial\theta \end{bmatrix} = \begin{bmatrix} \cos\theta & -r\sin\theta \\ \sin\theta & r\cos\theta \end{bmatrix} = r$$

Therefore, $dx\,dy = r\,dr\,d\theta$.

# Note

If,

$$J_1 = \frac{\partial(x,y)}{\partial(u,v)}, \text{ and } J_2 = \frac{\partial(u,v)}{\partial(x,y)},$$

then

$$\det(J_1) = \frac{1}{\det(J_2)}.$$

Note:

When we transform variables, we want to assert

$$f(x,y)dxdy = g(u,v)dudv$$

$$f(x,y)\left|\det\left(\frac{\partial(x,y)}{\partial(u,v)}\right)\right|dudv = g(u,v)dudv$$

$$f(x,y)|\det(J)| = g(u,v)$$