Lecture 5

Multi-Dimensional Distributions

Contents

1 Goals and Outline

For the most part, we shall focus on 2D distributions – the simplest "multi"-dimensional distribution. Besides ease of visualization, this will allow us to introduce some terminology, including:

- Joint PDF
- Marginal PDF
- Conditional PDF

These concepts are connected through a famous application of MCMC: **Bayes Theorem**. For most practical problems, Bayes theorem usually leads to a probability distribution function,[1] which for non-trivial cases, is explored using MC or MCMC.

Unlike the transformation method, rejection sampling (direct Monte Carlo method) is perfectly sufficient for simple low-dimensional PDFs, but is inadequate for complicated high-dimensional PDFs. For these cases, we often need to rely on MCMC.

2 Joint, Marginal and Conditional PDF

Let's anchor our discussion to an example.[2]

Consider a general **2D normal** or **Gaussian** distribution,[3]

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

It describes how variables x and y covary. A PDF which describes the variation of multiple random variables together is called a **joint probability distribution function**.

The joint PDF $p_{X,Y}(x,y)$ has 5 parameters: it is centered at (μ_x, μ_y); σ_x and σ_y are the standard deviation in the x and y direction; and $-1 < \rho < 1$ is called the correlation coefficient.

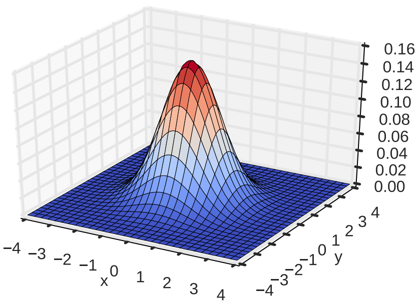
Lets visualize it for different values of these parameters to get a feel for how the PDF works.

First, let's set $\sigma_x = \sigma_y = 1.0$; $\rho = \mu_x = \mu_y = 0$. Next, let us increase $\sigma_y = 4.0$, leaving other quantities the same.

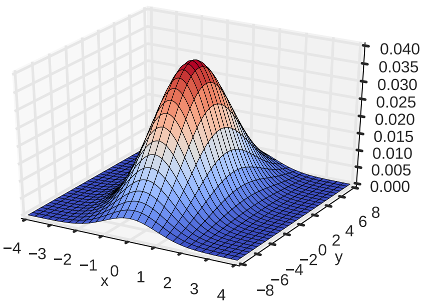
[1] Bayesian inference is a sampling problem $X \sim \pi(x)$

[2] "show, then tell!"

[3] sometimes also called bivariate normal, or binormal



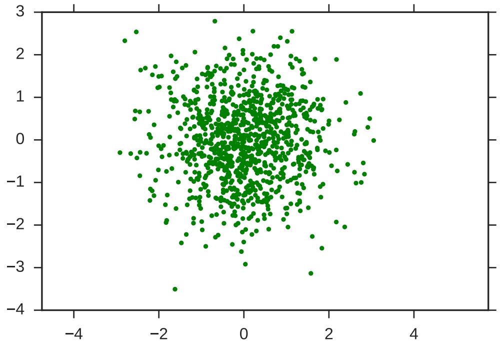
$\sigma_x = \sigma_y = 1.0$

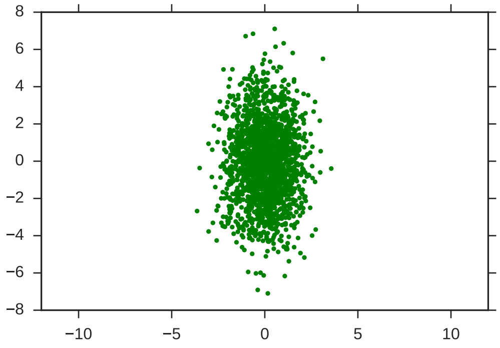


$\sigma_x = \sigma_y = 4.0$

As σ_y increases we obtain an increased spread in the y-direction.

Let us draw samples from such “uncorrelated” ($\rho = 0$) distributions.[4]


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

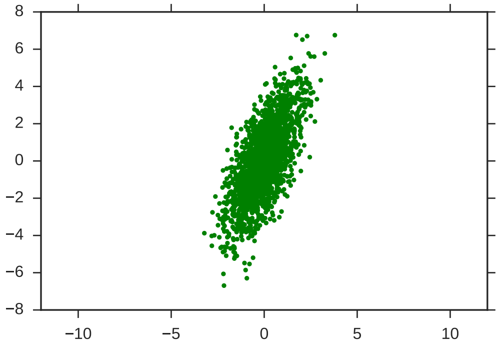

$\sigma_x = 1.0; \sigma_y = 2.0; \rho = 0$

Note that the scale for the two plots is different; σ_x and σ_y control the scale or dispersion. We set $\mu_x = \mu_y = 0$, in these visualizations; changing them simply shifts the location of the “center” or the “peak”.

2.1 Independent and Correlated Random Variables

Finally, let us change ρ to something non-zero. For 2D normal distributions, ρ controls “tilt”.

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



[4] In python, this can be done using `numpy.random.multivariate_normal(mean, cov, size)`. For 2D Gaussian the covariance matrix (`cov`) is related to the correlation coefficient via,

$$\Sigma = \begin{bmatrix} \sigma_x^2 & \rho\sigma_x\sigma_y \\ \rho\sigma_x\sigma_y & \sigma_y^2 \end{bmatrix}.$$

Practically, such a PDF may correspond to something like the joint PDF of height (x) and weight (y). Generally tall people are heavier than short people, and there is a positive correlation between height and weight ($\rho = 0.7$).

If x and y are **correlated**, knowing the value of one *tells you something* about the other.[5]

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

Two random variables X and Y are **uncorrelated** if,

$$E[XY] = E[X]E[Y], \tag{2}$$

and, thus their covariance,

$$\text{Cov}[X, Y] = E[XY] - E[X]E[Y] = 0. \tag{3}$$

Two random variables X and Y are independent if,

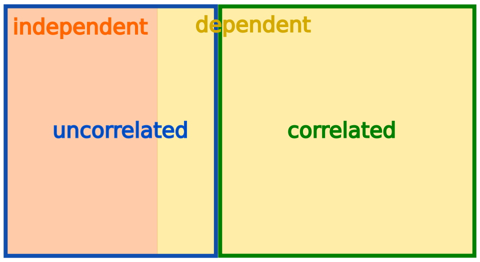
$$p_{X,Y}(x, y) = p_X(x)p_Y(y). \tag{4}$$

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.

The following Venn diagram illustrates their relationship.



Exercise: Consider $X \sim U[-1, 1]$ and $Y = |X|$.

By definition, Y and X are dependent variables. Show that the domain of $Y \sim U[0, 1]$, and $E[XY] = 0 = E[X]E[Y]$. Thus, X and Y are uncorrelated, even though they are dependent.

Exercise: Consider $X \sim \mathcal{N}(0, 1)$ and $Y = X^2$.

2.2 Joint Distribution

Consider two random variables X_1 and X_2 (think of height and weight, if you like). If they are continuous, then a PDF $p_{X_1,X_2}(x_1, x_2)$,[6] which describes their collective variation is called a joint probability distribution function.

[5] The relationship between independent and uncorrelated variables is discussed in this handout from CMU Advanced Data Analysis. Correlation quantifies linear dependence between two variables. It does not capture more complicated behavior, which can sometimes be obvious on a scatterplot.

[6] like the 2D normal distribution we just saw

The interpretation of a 2D PDF is similar to that of a 1D PDF.

$p_{X_1,X_2}(x_1,x_2)\,dx_1\,dx_2$ represents the probability of a state (X_1,X_2) such that $x_1 \leq X_1 < x_1+dx_1$ and $x_2 \leq X_2 < x_2+dx_2$.

Normalization implies,

$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} p_{X_1,X_2}(x_1,x_2)\,dx_1\,dx_2 = 1 \tag{5}$$

If the random variables are *independent*, $p_{X_1,X_2}(x_1,x_2) = p_{X_1}(x_1)p_{X_2}(x_2)$, where $p_{X_1}(x_1)$ and $p_{X_2}(x_2)$ are normalized 1D PDFs.

2.3 Marginal Distribution

The marginal distribution of a particular random variable represented by $p_{X_i}(x_i)$, is obtained from a joint PDF by integrating out all the other random variables.

Thus, $p_{X_i}(x_i)$ can be obtained from an n-dimensional joint PDF $p_{X_1,...,X_n}(x_1,...,x_n)$,

$$p_{X_i}(x_i) = \int p_{X_1,...,X_n}(x_1,...,x_n)\,dx_1...\,dx_{i-1}\,dx_{i+1}...\,dx_n. \tag{6}$$

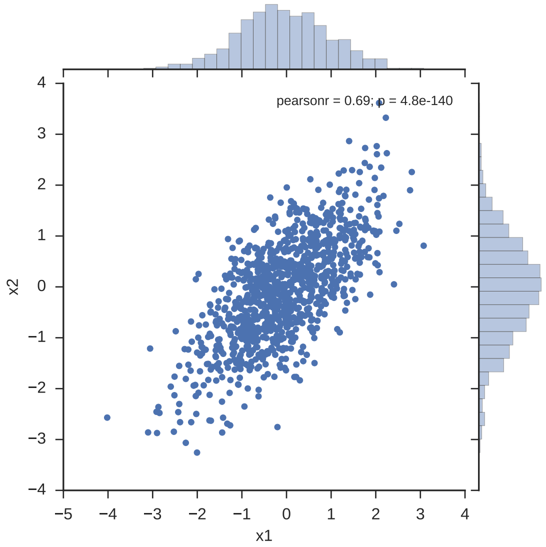
Note that we integrate over all dimensions except x_i.[7]

In particular for a 2D distribution, dropping the subscripts,

$$p(x_1) = \int p(x_1,x_2)\,dx_2 \tag{7}$$

$$p(x_2) = \int p(x_1,x_2)\,dx_1. \tag{8}$$

For the 2D normal distribution with $\sigma_x = 1.0$, $\sigma_y = 2.0$, and $\rho = 0.7$ that we saw earlier, this can be visualized graphically as the distributions on the top ($p(x_1)$) and right ($p(x_2)$) of the panel below.



Some properties of marginal distributions are summarized below:

[7] If we integrate over x_i as well, the integral equals 1, due to the normalization condition!

(i) Marginal distributions are 1D PDFs. They offer a particular summary of the joint PDF.[8]

(ii) 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.[9]

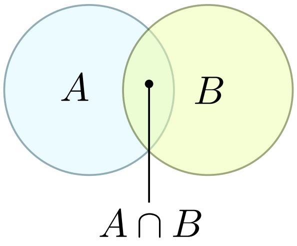
(iii) In the previous plot, notice that the marginal distributions looked Gaussian. They did not reflect the correlation between the quantities. This information was lost.

2.4 Conditional Distribution

Conditional probability of event A given B is represented as Pr$(A|B)$. It is given by,

$$\text{Pr}(A|B) = \frac{\text{Pr}(A \cap B)}{\text{Pr}(B)}, \tag{9}$$

where Pr$(A \cap B)$ can be read as the "probability of A and B". The following Venn diagram may be helpful in seeing this relationship.



Or perhaps, an example will help.

Example 1: If you roll a fair die, what is the probability of observing a number (n) less than 3, given n is even?
Solution: We want to compute Pr($n < 3$ | n is even).

$$\text{Pr}(n < 3 \,|\, n \text{ is even}) = \frac{\text{Pr}(n < 3 \text{ and even})}{\text{Pr}(n \text{ is even})} = \frac{(1/6)}{(3/6)} = 1/3.$$

You may explicitly enumerate the cases to convince yourself of the answer.

Example 2: Life expectancy tables
The life expectancy of a newborn in the US is around 75 years. However, the "conditional" life expectancy for a 70 year old, rises to more than 80 (link). Thus, if x is a person's current age, and y is life expectancy, that person may care about the distribution Pr$(y|x)$, rather than the marginal Pr(y).

Suppose we want to describe the distribution of x_1 at a particular value of x_2, given a joint PDF $p(x_1, x_2)$, . Dropping subscripts, this conditional PDF is written as $p(x_1|x_2)$. You may read the "|" sign as "given".

Eqn. 9 can be generalized to:

$$p(x_1|x_2) = \frac{p(x_1, x_2)}{p(x_2)}, \quad \text{conditional} = \frac{\text{joint}}{\text{marginal}}, \tag{10}$$

[8] An analogy might be 2D shadows (marginal PDF) of 3D (joint PDF) objects.

[9] Said differently, for a 2D distribution you cannot uniquely reconstruct $p(x_1, x_2)$ from $p(x_1)$ and $p(x_2)$.

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

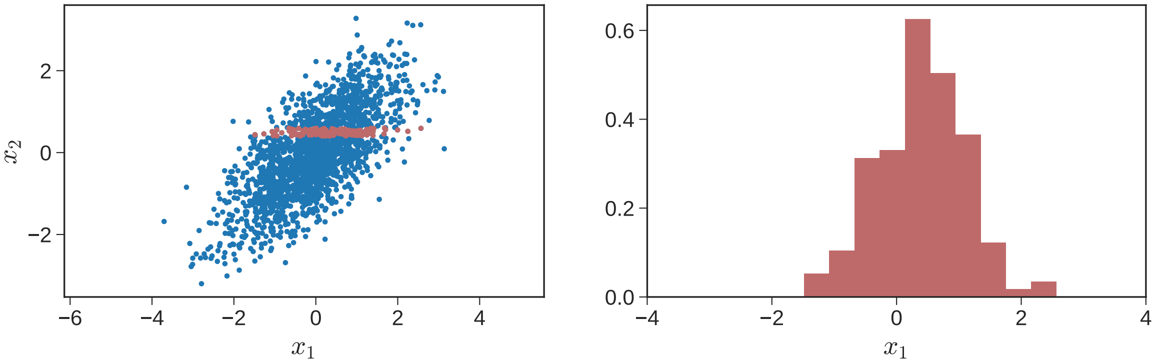
Thus,

$$p(x_1|x_2) = \frac{p(x_1, x_2)}{\int p(x_1, x_2) dx_1} \tag{11}$$

Conditional PDFs can also be obtained from the joint PDF.

Lets look at some pictures again. We return to our familiar 2D normal joint PDF with $\rho = 0.7$, Suppose we want to consider the distribution $p(x_1|x_2 = 0.5)$.

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



The 1D conditional PDF on the right is not normalized. Note that it is not centered at zero (see exercise iii).

3 Bayes Theorem

Bayesian analysis is an increasingly important application of MCMC. It is useful to introduce it here, because it relies on joint, marginal, and conditional PDFs. Let's first derive Bayes rule using a mathematical approach. Later, we shall interpret it using "physical reasoning" to gain some intuition on what it implies, and how it relates to MCMC.

3.1 Derivation

We saw that,

$$\begin{aligned} p(x_1|x_2) &= \frac{p(x_1, x_2)}{p(x_2)}, \\ p(x_2|x_1) &= \frac{p(x_1, x_2)}{p(x_1)}. \end{aligned} \tag{12}$$

Thus, it is easy to combine them as:

$$p(x_1|x_2)p(x_2) = p(x_2|x_1)p(x_1), \tag{13}$$

or

$$p(x_1|x_2) = \frac{p(x_2|x_1)p(x_1)}{p(x_2)}. \tag{14}$$

Using eqns. 8 and 12, the $p(x_2)$ in the denominator can be written as,

$$\begin{aligned} p(x_2) &= \int p(x_1, x_2) dx_1, & (15) \\ &= \int p(x_2|x_1) p(x_1) dx_1. & (16) \end{aligned}$$

Using eqn. 14, this allows us to write Bayes theorem in a compact symmetric form:

$$\boxed{p(x_1|x_2) = \frac{p(x_2|x_1)p(x_1)}{\int p(x_2|x_1)p(x_1)dx_1}} \tag{17}$$

The denominator of eqn. 17 can be thought of as a normalizing factor; it does not depend on x_1,[10] which ensures that the PDF $p(x_1|x_2)$ is properly normalized. Often, it is not easy to compute for large multidimensional systems.[11] As it turns out, for many important problems, we don't have to explicitly evaluate the denominator.

3.2 Interpretation

Thus, the form:

$$p(x_1|x_2) \propto p(x_2|x_1)p(x_1) \tag{18}$$

is often used in computational implementations of Bayes Rule.

A 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, and color-code it:

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

where m=model, and d = data.

We start with a prior belief $p(m)$ of the correct model. This is called the **prior probability**. $p(d|m)$ is the probability of observing the data d, if the model prior model m is correct. It is called the **likelihood**. Finally, $p(m|d)$ is our updated belief in the model m based on the strength of the evidence d. It is called the **posterior probability**.

Let's do an example to illustrate these ideas.

3.3 Example

A (potentially unfair) coin is tossed $N = 10$ times. We observe n_H = 7 heads and n_T = 3 tails. Let θ be the (unknown) probability of tossing heads. Thus, $1 - \theta$ is the probability of tossing tails.[12] For a fair coin, $\theta = 0.5$. For a coin that is "heads-biased", $\theta > 0.5$.

Suppose, we think the coin is probably fair, but want to allow for other possibilities. We may

[10] since it is integrated out!

[11] If you are familiar with statistical thermodynamics, the denominator can be thought of as a partition function, which is related to the concepts like entropy or free-energy.

[12] The range of θ is 0 to 1.

encapsulate this prior belief about θ by a triangular distribution peaked at $\theta = 0.5$,

$$p(\theta) = \begin{cases} 4\theta & 0 \leq \theta < 0.5, \\ 4(1-\theta) & 0.5 \leq \theta < 1. \end{cases} \tag{19}$$

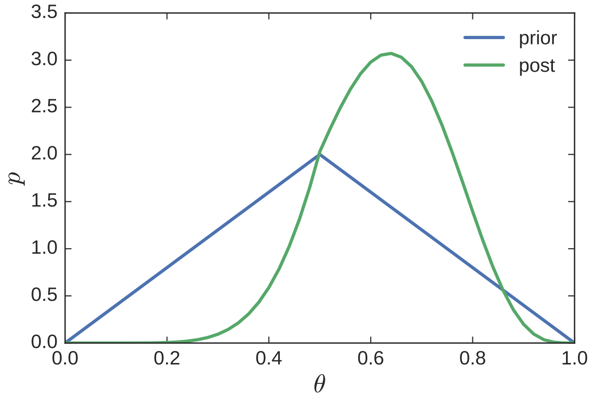
Next we turn to the likelihood, i.e., given θ, what is the probability of observing n_H heads in N trials? It turns out that we have seen this set up before: the answer turns out to be the binomial distribution.

$$p(d|\theta) \propto {}^N C_{n_H} \theta^{n_H} (1-\theta)^{N-n_H}. \tag{20}$$

Thus, the posterior distribution is given by $p(\theta|d) = p(d|\theta)p(\theta)$. It is useful to take "log" on both sides, to keep numbers manageable. Terms that do not depend on θ contribute only a multiplicative constant; they can be dropped since are subsumed into the normalization constraint anyway. Thus,

$$\log p(\theta|d) \propto n_H \log \theta + (N - n_H) \log(1-\theta) + \log p(\theta).$$

This is answer. To visualize, we can select different values of θ between 0 and 1, evaluate the RHS to compute $p(\theta|d)$ at that θ. Once $p(\theta|d)$ is computed at all the different θ, it can be normalized.



See the python program A.2.1 used to plot this. Play with the program by varying n_H and N.[13]

4 Sampling 2D distributions

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:

(i) transformation method

(ii) rejection sampling (accept-reject)

[13] Conjugate priors: A conjugate prior is a probability distribution that, when multiplied by the likelihood and divided by the normalizing constant, yields a posterior probability distribution that is in the same family of distributions as the prior. In this example, we could have used a Beta distribution as a prior, which would have made the problem completely analytical. This would have allowed us to express the posterior as a Beta distribution.

4.1 Transformation Method

Consider two 2D distributions $f_{X,Y}(x,y)$ and $g_{U,V}(u,v)$, where the functional relationships $x(u,v)$ and $y(u,v)$ are known.

Then the relationship between the two PDFs is given by (see appendix A.1 for an example),

$$g(u,v) = f(x,y)\,|\text{det}(J)|, \tag{21}$$

where the Jacobian J is,

$$J = \frac{\partial(x,y)}{\partial(u,v)} = \begin{bmatrix} \partial x/\partial u & \partial x/\partial v \\ \partial y/\partial u & \partial y/\partial v \end{bmatrix}. \tag{22}$$

How does this sit in relation to the transformation method for 1D distributions? There, we used $x = F^{-1}(u).$, where $u \sim U[0,1]$. Is the older recipe, consistent with eqn. 21?

For 1D, the general formula specializes to,

$$g(u) = f(x)\,|dx/du| \tag{23}$$

For simplicity, assume that $x(u)$ is a monotonously increasing function, so that $|dx/du| = dx/du$. Thus, $g(u) = f(x)\,dx/du$, which can be rephrased as:

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

which is consistent with our older recipe. We can think of eqn. 21 as a generalization of the 1D transformation rule. There, 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 transformation**.

Let us illustrate this with a direct example.[14]

Example: Given the joint PDF of (u_1, u_2),

$$g(u_1,u_2) = \begin{cases} 2e^{-u_1-2u_2} & u_1 > 0, u_2 > 0 \\ 0 & \text{elsewhere.} \end{cases}$$

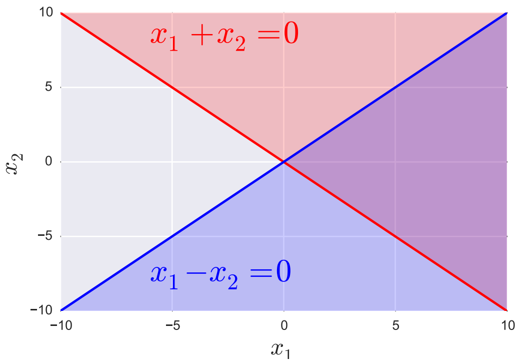
Find the PDF $f(x_1,x_2)$ of $x_1 = u_1 + u_2$, and $x_2 = u_1 - u_2$.

Solution:

We note that $u_1 = (x_1 + x_2)/2$ and $u_2 = (x_1 - x_2)/2$. The first task is to find the domain of x_1 and x_2. We are given that g is non-zero only for positive u_1 and u_2.

Observe that $2u_1 = x_1 + x_2 > 0$, and $2u_2 = x_1 - x_2 > 0$. Let's plot the two lines $x_1 + x_2 = 0$, and $x_1 - x_2 = 0$, and figure out what side the inequality lies on.

[14] Example 2 in A.1 is another such example



The domain of x_1 and x_2 is given by $x_1 > 0$ and $-x_1 < x_2 < x_1$.

The determinant of the Jacobian is given by,

$$|\det(J)| = \frac{\partial(x_1, x_2)}{\partial(u_1, u_2)} = \left|\det \begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix}\right| = |-2| = 2.$$

Thus,

$$\begin{aligned} g(u_1, u_2) &= f(x_1, x_2)\, |\det(J)| \\ 2e^{-(x_1+x_2)/2-(x_1-x_2)} &= 2f(x_1, x_2) \\ f(x_1, x_2) &= e^{-3x_1/2+x_2/2} \end{aligned}$$

Check the normalization condition as an excercise,

$$\int_0^\infty dx_1 \int_{-x_1}^{x_1} dx_2\, e^{-3x_1/2+x_2/2} = 1.$$

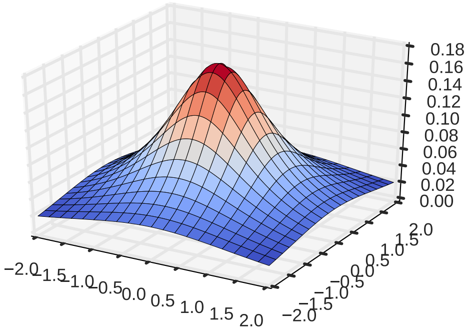
4.2 Rejection Sampling

Unlike the transformation method, which has limited use as a sampling method in higher dimensions, rejection sampling generalizes quite well. Let us see what we mean by an example.

Example: Sample,

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

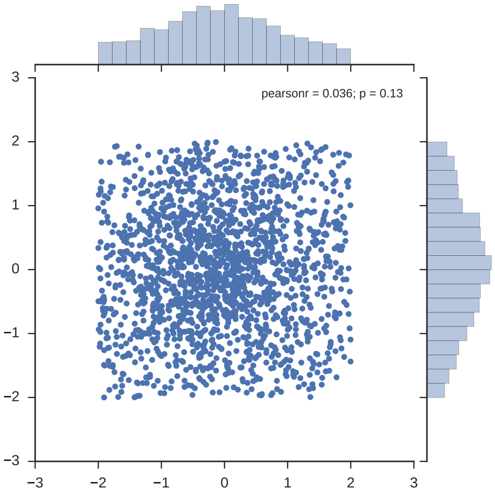
with $-2 \leq x_1 \leq 2$ and $-2 \leq x_2 \leq 2$.



Here, we simply throw three darts: $X_1 \sim U[-2, 2]$, $X_2 \sim U[-2, 2]$, and finally a third dart $u \sim U[0, f_{\max}]$, where $f_{\max}$ is the largest value of the PDF $f(x_1, x_2)$ on the domain (see code A.2.2).

We accept the "dart" (X_1, X_2), if $u \le f(X_1, X_2)$. That is, we return (X_1, X_2) as a sample point. If not, we just throw a fresh set of three darts.

For this particular problem, our acceptance ratio was ≈ 0.36, which means only about a third of the darts thrown "landed". The samples and the marginal distributions are shown below.



4.2.1 Limitations of Rejection Sampling

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

- **High number of dimensions**

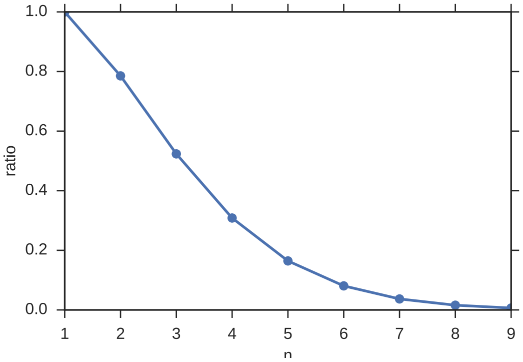
  Example: Find the ratio of the volumes of an n-dimensional sphere and the n-dimensional cube that circumscribes it.

  Solution: The volume of a n-dimensional hypercube is $V_C = L^n$ (think line → square → cube). The volume of a n-dimensional hypersphere of radius R is,

$$V_S = \frac{\pi^{n/2}}{\Gamma(\frac{n}{2} + 1)} R^n.$$

Consider finding the volume of a hypersphere of radius R by throwing darts inside a hypercube of edge length $L = 2R$. The fraction of darts that land in the sphere is quite small as the number of dimensions $n > 10$.

$$\frac{V_S}{V_C} = \frac{\pi^{n/2}}{2^n \Gamma(\frac{n}{2}+1)}.$$



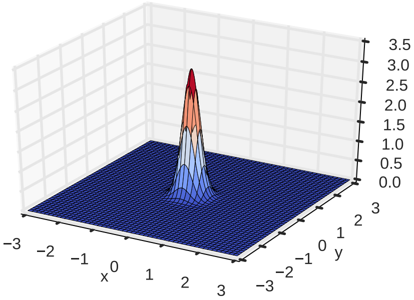
- **Sharply Peaked PDFs**

Consider a sharply peaked PDF with an *infinite domain*,

$$f(x_1, x_2) = \frac{a}{\pi} e^{-a(x_1^2+x_2^2)},$$

with $a = 10$.

Unless one does work to enclose the peak in a box, a lot of dart throws will be wasted.



- **Many Physical Problems**

Later, we might look at physical problems like randomly dispersing coins in a shoe box, and computing the displacement of coupled harmonic oscillators, which present problems to direct MC methods.

5 Problems

5.1 Thought Questions

(i) If x_1 and x_2 are independent, using the "conditional = joint/marginal" formula, show that the conditional probability $p(x_1|x_2) = p(x_1)$

(ii) Let X and Y be two random variables. State whether the following statements are true or false:

- X and Y independent $\implies$ X and Y uncorrelated
- X and Y uncorrelated $\implies$ X and Y independent
- X and Y dependent $\implies$ X and Y correlated
- X and Y uncorrelated $\implies$ X and Y dependent
- X and Y correlated $\implies$ X and Y dependent

(iii) For a 2D normal distribution given by eqn. 1, show that:

$$p(x_1|x_2) \sim \mathcal{N}\left(\mu_1 + \frac{\sigma_1}{\sigma_2}\rho(x_2-\mu_2),\, (1-\rho^2)\sigma_1^2\right).$$

(iv) Approximately 1/125 of all births are fraternal twins, and 1/300 are identical twins. Elvis Presley had a twin brother (who died at birth). What is the probability that Elvis was an identical twin. Assume that the probability of girl/boy is 1/2.

(v) In the coin tossing example for Bayesian inference, how would the relationship between the posterior and likelihood change, if the prior distibution $p(\theta)$ were uniform (noninformative prior).

(vi) Check the normalization condition for the 2D transformation example:

$$\int_0^\infty dx_1 \int_{-x_1}^{x_1} dx_2 \, e^{-3x_1/2+x_2/2} = 1.$$

(vii) Consider a 2D uniform distrbution defined on the unit square $[0,1)^2$,

$$g(u_1, u_2) = U[0,1)U[0,1) = 1,$$

where u_1 and u_2 are *independent* uniform random numbers between 0 and 1.

Suppose we define:

$$\begin{aligned} x_1 &= \sqrt{-2\ln u_1} \cos 2\pi u_2 \\ x_2 &= \sqrt{-2\ln u_1} \sin 2\pi u_2 \end{aligned}$$

(i) What is the range of values that x_1 and x_2 can take? Explain. (10 points)

(ii) Find the distribution $f(x_1, x_2)$ using the 2D transformation rule. (25 points)

Note: This might not necessarily be required, but:

$$\begin{aligned} u_1 &= \exp\left(-\frac{x_1^2+x_2^2}{2}\right) \\ u_2 &= \frac{\tan^{-1}(x_2/x_1)}{2\pi} \end{aligned}$$

5.2 Numerical Questions

(i) Transformation of Variables

Let θ be the probability that a birth is female. Suppose $n_G = 241,945$ and $n_B = 251,527$ girl and boy births observed over a year, respectively.

From Bayes theorem:

$$p(\theta|n_G, n_B) \propto p(n_G, n_B|\theta)p(\theta)$$

Suppose we assume $p(\theta) = \mathcal{N}(0.5, 0.1)$,[15] and use a binomial distribution to characterize the likelihood:

$$p(n_G, n_B|\theta) =^{n_G+n_B} C_{n_G} \theta^{n_G}(1-\theta)^{n_B}.$$

Report the histogram of $p(\theta|n_G, n_B)$ using MC simulation to sample from the prior distribution.

[15] if you are particular, you can use a truncated normal distribution, to limit θ to $[0,1]$.

(ii) (Krauth) Consider an algorithm to generate samples of Z from $\pi_Z(z)$,

$$\begin{aligned} X &\sim U(-1,1) \\ Y &\sim U(-1,1) \\ Z &= X^2 + Y^2 \end{aligned}$$

(a) What is the domain of z?

(b) What is the distribution $\pi_Z(z)$? Is it true that z is uniformly distributed in the interval $z \in [0,1]$? **Hint**: Define a new variable $W = X$, find the joint distribution $\pi(z,w)$ using the 2D transformation rule, and marginalize w.

(c) Implement the algorithm, and generate a histogram to check your answers.

(iii) Bayesian Inference

You are given a coin, which may be fair or unfair. Let p be the probability that the coin shows "heads" (=1) on a given toss. Thus, the probability of "tails" (=0) is $1-p$.

The result of 1000 coin tosses is provided in the file headsQ2.txt. We want to use Bayesian inference to learn something about p from this data. Suppose we start with a uniform prior $\pi(p) = 1$ for $0 \leq p \leq 1$. Given a probability p of heads, the binomial distribution tells us that if we toss a coin n times, the likelihood of seeing k heads is given by the distribution,

$$\pi(k|n,p) = \binom{n}{k} p^k (1-p)^{n-k}.$$

Solve the following parts:

(a) The posterior distribution is proportional to the likelihood and the prior.

$$\pi(p|k,n) = \frac{\pi(k|n,p)\pi(p)}{\int_0^1 \pi(k|n,p)\pi(p)dp}.$$

Explain why the denominator is,

$$\int_0^1 \pi(k|n,p)\pi(p)dp, \text{ and not } \sum_{k=0}^{\infty} \pi(k|n,p)\pi(p).$$

(b) WolframAlpha suggests that the denominator symbolically evaluates to,

$$\int_0^1 \pi(k|n,p)\pi(p)dp = \int_0^1 \binom{n}{k} p^k (1-p)^{n-k} (1) dp = \frac{1}{n+1}.$$

For $n = 10$, vary $k = 0, 1, 2, ..., n$ and use symbolic/numerical integration (built-in function like "quad" or "integrate") and verify the general formula for this particular case.

(c) Plot the posterior distribution after the first 1, 10, 100, 500, and 1000 tosses.

(d) Consider the posterior distribution after 1000 tosses. What is the probability that $p > 0.5$? What is your point-estimate of p? Provide an error-bar to describe the uncertainty in the estimated p?

A Appendices

A.1 Transformation of Variables in 2D

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,

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

where the Jacobian,

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

The Jacobian tells you how a unit square in (u, v) co-ordinates gets stretched or squeezed, when transformed to (x, y) co-ordinates. Although the transformation can be nonlinear, the idea here is that we are looking at tiny areas, where a linear approximation (inherent in the first derivatives computed in the Jacobian) holds.

Some properties of the determinant of J are:

- The determinant remains same if we switch rows and columns.
- The sign of the Jacobian tells us something about the orientation; in particular a negative value for the determinant tells us that the sign of the cross-product of the vectors making up the “unit square” in (u, v) co-ordinates, gets reversed upon transformation.
- If,

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

  then

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

If you are unfamiliar, or rusty, it may be useful to review the intuition behind Jacobian determinants (Khan Academy or wikipedia).

When we transform variables, we want to assert

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

A.1.1 Example 1

Consider transformation between (x, y) and cylindrical coordinates (r, θ).

$$\begin{aligned} x &= r\cos\theta \\ y &= r\sin\theta \end{aligned}$$

Therefore,

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

where

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

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

A.1.2 Example 2

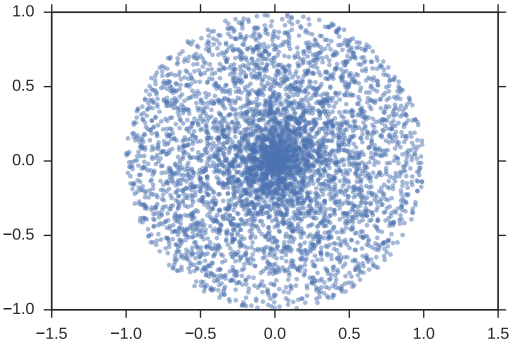
Using cylindrical coordinates (r, θ) sample points uniformly on a disc of radius a, centered at $(0, 0)$.

Incorrect Solution: It may be tempting to sample according to,

$$\begin{aligned} r &\sim U[0, a] \\ \theta &\sim U[0, 2\pi] \end{aligned}$$

Python Code

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



However, if we plot the samples, we note that the disc has too many points near the core. That is, the points are not distributed uniformly. We need to approach this problem more carefully.

We start by explicitly writing down the target distribution $f(x, y)$. It 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,

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

One way to approach this is to write $g(r, \theta) = R(r)T(\theta)$, where,

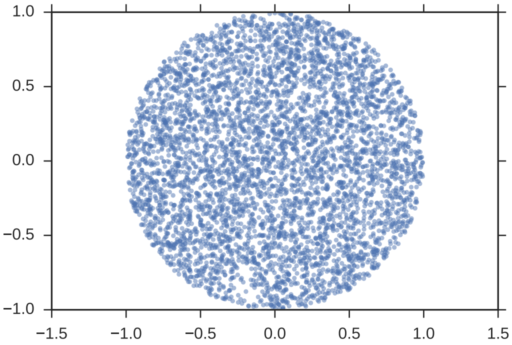
$$\begin{aligned} R(r) &= \frac{2r}{a^2}, \quad 0 < r \leq a \\ T(\theta) &= \frac{1}{2\pi}, \quad 0 < \theta \leq 2\pi. \end{aligned}$$

- both $R(r)$ and $T(\theta)$ are appropriately normalized
- unlike θ, r has to be sampled from a non-uniform distribution; $R(r)$ requires more points to be sampled at large r.
- From our knowledge of sampling 1D distributions, we may sample from $R(r)$ by setting $r \sim a\sqrt{u}$, with $u \sim U[0, 1]$.

Python Code

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

Points now appear uniformly distributed on the disc.



A.2 Python Programs

A.2.1 Bayesian Inference Example

```
import numpy as np
import matplotlib.pyplot as plt
from scipy.integrate import trapz

def T(theta):
        if theta < 0.5:
                T = 4*theta
        else:
                T = 4*(1-theta)
        return T

def plotPostDist(N, nH):

        # get prior
        thetav   = np.linspace(1e-6, 1.-1e-6, 51)
        priorD   = Tv(thetav)

        # get unnormalized posterior
        LogPostD = nH * np.log(thetav) + (N - nH) * np.log(1.-thetav)
                   + np.log(priorD)
        postD   = np.exp(LogPostD)

        # normalize posterior
        normFac = trapz(postD, thetav)
        postD   = postD/normFac

        # plot
        plt.plot(thetav, priorD, label='prior')
        plt.plot(thetav, postD, label='post')

        plt.xlabel(r'$\theta$')
        plt.ylabel(r'$p$')
        plt.legend()
        plt.show()

        return

# this lets me send scalar or vector theta to T()
Tv = np.vectorize(T)

plotPostDist(N=10, nH=7)
```

A.2.2 Sampling 2D Distributions Example

```
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)

    constraint = u <= fmax/(X1**2 + X2**2 + 1.0)

    return X1[constraint], X2[constraint]
```