Refresher on probability

Kevin P. Murphy

Last updated September 18, 2006

Probability theory is nothing but common sense reduced to calculation. — Pierre Laplace, 1812

1 Joint, marginal and conditional distributions

Let X and Y be **discrete random variables (rv's)**, each with K possible values. (We consider continuous rv's in Section 3.)[1] The **sum rule** specifies how to compute a **marginal distribution** from a **joint distribution**:

$$p(X=i) = \sum_{j=1}^{K} p(X=i, Y=j) \tag{1}$$

Since the variables are discrete, we can represent $p(X,Y)$ as a $K \times K$ table, and we can represent $p(X=i)$ as a $K \times 1$ vector. See Figure 1 for an example.

The **product rule** specifies how to compute a joint distribution from the product of a marginal and a **conditional distribution**:

$$p(X=i, Y=j) = p(X=i|Y=j)p(Y=j) \tag{2}$$

We can represent $p(X|Y)$ as a $K \times K$ matrix $M(Y,X)$, where each row sums to one (this is called a **stochastic matrix**):

$$\sum_{i=1}^{K} M(j,i) = 1 \tag{3}$$

By symmetry we can write

$$p(X=i, Y=j) = p(Y=j|X=i)p(X=i) \tag{4}$$

We can use this definition to compute a conditional probability

$$p(X=i|Y=j) = \frac{p(X=i, Y=j)}{p(Y=j)} \propto p(Y=j|X=i)p(X=i) \tag{5}$$

This is called **Bayes' rule** and is often written

$$\text{posterior} \propto \text{ likelihood} \times \text{ prior} \tag{6}$$

where $p(X=i)$ is the prior probability that $X=i$, $p(Y=j|X=i)$ is the likelihood that Y has value j given that X has value i, and $p(X=i|Y=j)$ is the posterior probability that $X=i$ given that we observe that $Y=j$. The constant of proportionality is $1/p(Y=j)$, where $p(Y=j)$ is the **marginal likelihood** of the data (marginalized over X).

[1] We use lower-case p to denote either a probability density function (for continuous rv's) or a probability mass function (for discrete rv's). Also, we follow the standard convention that random variables are denoted by upper case letters, and values of random variables are denoted by lower case letters. However, when we start treating parameters as random variables (see Section **??**), we will use usually use lower-case greek letters for both the variable and its value.

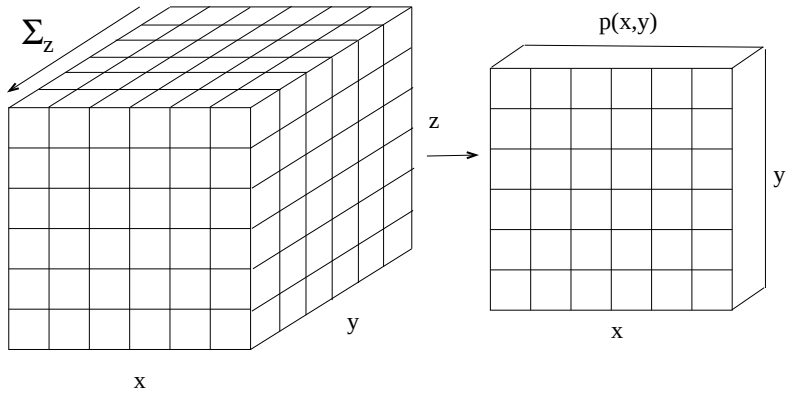

Figure 1: Computing $p(x, y) = \sum_z p(x, y, z)$ by marginalizing over dimension Z. Source: Sam Roweis.

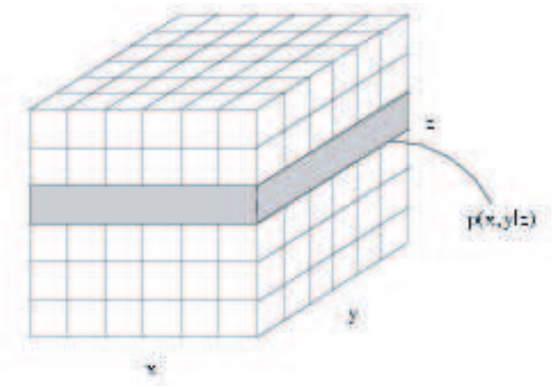

Figure 2: Computing $p(x, y|z)$ by extracting the slice from $p(x, y, z)$ corresponding to $Z = z$ and then renormalizing. Source: Sam Roweis.

We can compute the **joint posterior** $p(X, Y|Z)$ similarly; the result is a 2D "slice" of the matrix (see Figure 2) which satisfies $\sum_{x,y} p(x, y|z) = 1$.

The product rule can be applied multiple times to yield the **chain rule of probability**:

$$p(X_1, \ldots, X_n) = p(X_1)p(X_2|X_1)p(X_3|X_2, X_1) \ldots p(X_n|X_{1:n-1}) \tag{7}$$

where we introduce the notation $1 : n - 1$ to denote $\{1, 2, \ldots, n - 1\}$.

2 Conditional independence

We can simplify the chain rule by making **conditional independence assumptions**. We say Z and Y are conditionally independent given X, written as $Z \perp Y|X$, **if and only if (iff)** $p(Z, Y|X) = p(Z|X)p(Y|X)$. Suppose we make a **Markov assumption** that the "future" is independent of the "past" given the "present":

$$X_{i+1} \perp X_{1:i-1}|X_i \tag{8}$$

Then the chain rule simplifies to

$$p(X_1, \ldots, X_n) = p(X_1)p(X_2|X_1)p(X_3|X_2) \ldots p(X_n|X_{n-1}) = p(X_1) \prod_{i=2}^{n} p(X_i|X_{i-1}) \tag{9}$$

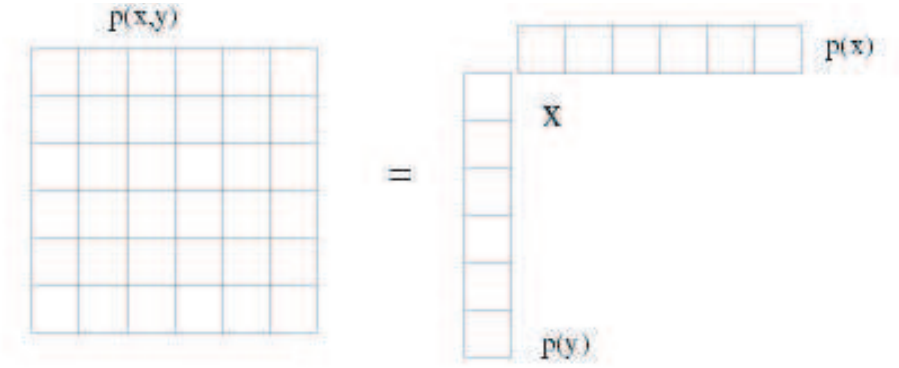

Figure 3: Computing $p(x, y) = p(x)p(y)$, where $X \perp Y$. Source: Sam Roweis.

Thus the joint is decomposed into a product of small terms. We shall see other examples of this kind of simplification in Section **??** below. If X and Y are unconditionally or **marginally independent**, $X \perp Y$, we can write $p(X, Y) = p(X)p(Y)$: see Figure 3.

3 Continuous random variables

In the examples above, $p(X = i)$ is the probability that X takes on value i; this **probability mass function (pmf)** satisfies $\sum_i p(X = i) = 1$. If X is a continuous random variable, e.g., $X \in \mathbb{R}$ or $X \in \mathbb{R}^+$, then we use a **probability density function (pdf)** which satisfifes $\int_S p(X = x)dx = 1$, where we integrate over the **support** S of the distribution. It is called a density because we can multiply it by an interval of size dx to find the probability of being in that interval:

$$p(x)dx \approx P(x \leq X \leq x + dx) \tag{10}$$

Note that it is possible for $p(x) > 1$ for any given x, so long as the density integrates to 1.

We will see many examples of pdf's in this book, but here we introduce the most famous distributon, namely the Gaussian or Normal distribution. For one-dimensional variables, this is defined as

$$\mathcal{N}(x|\mu, \sigma) \stackrel{\text{def}}{=} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(x-\mu)^2} \tag{11}$$

where μ is the mean and σ is the standard deviation, which we explain below. See Figure 4 for an example. and see Chapter **??** for more details on the Gaussian distribution (such as how to estimate μ and σ from data).

If $Z \sim \mathcal{N}(0, 1)$, we say Z follows a **standard normal** distribution. Its **cumulative distribution function (cdf)** is defined as

$$\Phi(x) = \int_{-\infty}^{x} p(z)dz \tag{12}$$

which is called the **probit distribution**. This has no closed form expression, but is built in to most software packages (eg. **normcdf** in the matlab statistics toolbox). In particular, we can compute it in terms of the **error (erf) function**

$$\Phi(x; \mu, \sigma) = \frac{1}{2}[1 + erf(z/\sqrt{2})] \tag{13}$$

where $z = (x - \mu)/\sigma$ and

$$erf(x) \stackrel{\text{def}}{=} \frac{2}{\sqrt{\pi}} \int_0^x e^{-t^2} dt \tag{14}$$

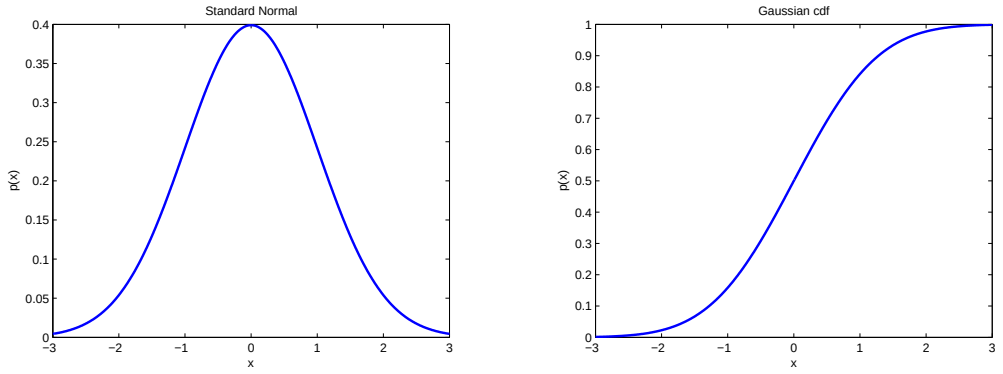

Figure 4: A standard normal pdf and cdf. The matlab code used to produce these plots is `xs=-3:0.01:3; plot(xs,normpdf(xs,mu,sigma)); plot(xs,normcdf(xs,mu,sigma));` , where $xs = [-3, -2.99, -2.98, \ldots, 2.99, 3.0]$ is a vector of points at which the density is evaluated.

Let us see how we can use the cdf to compute how much probability mass is contained in the interval $\mu \pm 2\sigma$. If $X \sim \mathcal{N}(\mu, \sigma^2)$, then $Z = (X - \mu)/\sigma \sim \mathcal{N}(0, 1)$. The amount of mass contained inside the 2σ interval is given by

$$p(a < X < b) = p(\frac{a-\mu}{\sigma} < Z < \frac{b-\mu}{\sigma}) \tag{15}$$

$$= \Phi(\frac{b-\mu}{\sigma}) - \Phi(\frac{a-\mu}{\sigma}) \tag{16}$$

Since

$$p(Z \leq -1.96) = \text{normcdf}(-1.96) = 0.025 \tag{17}$$

we have

$$p(-1.96\sigma < X - \mu < 1.96\sigma) = 1 - 2 \times 0.025 = 0.95 \tag{18}$$

Often we approximate this by replacing 1.96 with 2, and saying that the interval $\mu \pm 2\sigma$ contains 0.95 mass.

It is also useful to compute quantiles of a distribution. A **α-quantile** is the value $f_\alpha = x$ s.t., $f(X \leq x) = \alpha$, where f is the pdf. For example, the median is the 50%-quantile. Also, if $Z \sim \mathcal{N}(0, 1)$, then the 2.5% quantile is $N_{0.025} = \Phi^{-1}(0.025) = -1.96$, where Φ^{-1} is the inverse of the Gaussian cdf:

$$z = \text{norminv}(0.025) = -1.96 \tag{19}$$

$$p(Z \leq z) = \text{normcdf}(z) = 0.025 \tag{20}$$

By symmetry of the Gaussian, $\Phi^{-1}(0.025) = -\Phi^{-1}(1 - 0.025) = \Phi^{-1}(0.975)$.

4 Expectation

We define the **expected value** of an RV X to be

$$\mu_X \stackrel{\text{def}}{=} E\,X \stackrel{\text{def}}{=} \sum_x x p(X = x) \tag{21}$$

We replace the sum by an integral if X is continuous. By **linearity of expectation**, we can push E inside $\sum$:

$$E\left(\sum_i a_i X_i\right) = \sum_i a_i E\left(X_i\right) \tag{22}$$

where the a_i are constants. If X is a **random vector**,

$$X = \begin{pmatrix} X_1 \\ \vdots \\ X_p \end{pmatrix} \tag{23}$$

then its mean is denoted by

$$\vec{\mu} = \begin{pmatrix} \mu_1 \\ \vdots \\ \mu_p \end{pmatrix} \tag{24}$$

Note that we will often just write μ instead of $\vec{\mu}$. If a is a vector and A a matrix, we have the following two important results (which follow from linearity of expectation):

$$E\left(a^T X\right) = a^T \mu \tag{25}$$

$$E\left(AX\right) = A\mu \tag{26}$$

In particular, for any two random variables X,Y, whether independent or not, we have

$$E\left[aX + bY + c\right] = aE\ X + bE\ Y + c \tag{27}$$

Also, if X,Y are independent,

$$E\left[XY\right] = \left[E\ X\right]\left[E\ Y\right] \tag{28}$$

The **conditional expectation** is defined as

$$E\left(X|Y = y\right) \stackrel{\text{def}}{=} \sum_x x\ p(x|y) \tag{29}$$

Note that whereas $E\left(X\right)$ is a number, $E\left(X|Y\right)$ is a function of Y. The important **rule of iterated expectations** is

$$E\left[E\left(Y|X\right)\right] = E\left(Y\right) \tag{30}$$

This is easy to prove:

$$E\left[E\left(Y|X\right)\right] = \sum_x E\left(Y|X = x\right)p(X = x) \tag{31}$$

$$= \sum_x \sum_y yp(Y = y|X = x)p(X = x) \tag{32}$$

$$= \sum_y y\left[\sum_x p(Y = y, X = x)\right] \tag{33}$$

$$= \sum_y yp(Y = y) \tag{34}$$

$$= E\ Y \tag{35}$$

5 Variance

The **variance** is a measure of spread:

$$\sigma^2 \stackrel{\text{def}}{=} \text{Var}X = E(X - \mu)^2 \tag{36}$$

$$= \int (x - \mu)^2 p(x)dx \tag{37}$$

$$= \int x^2 p(x)dx + \mu^2 \int p(x)dx - 2\mu \int xp(x)dx \tag{38}$$

$$= E[X^2] - \mu^2 \tag{39}$$

from which we infer the useful result $E[X^2] = \mu^2 + \sigma^2$. The **standard deviation** is defined as

$$\sigma_X \stackrel{\text{def}}{=} \sqrt{\text{Var } X} \tag{40}$$

It is easy to show

$$\text{Var } (aX + b) = a^2 \text{Var } (X) \tag{41}$$

where a and b are constants.

The variance of a sum is

$$\text{Var } [X + Y] = \text{Var } X + \text{Var } Y + 2\text{Cov}(X, Y) \tag{42}$$

The **conditional variance** is defined as

$$\text{Var } (Y|X = x) \stackrel{\text{def}}{=} \sum_y (y - E\ (Y|x))^2\ p(y|x) \tag{43}$$

The **rule of iterated variance** is

$$\text{Var } (Y) = E\ \text{Var } (Y|X) + \text{Var } E\ (Y|X) \tag{44}$$

This can be proved as follows. Let $\mu = E[Y|X]$. Then

$$\begin{aligned} E\ \text{Var } (Y|X) + \text{Var } E\ (Y|X) &= E\left[E(Y^2|X) - \mu^2\right] + E[\mu^2] - [E^2\mu] && (45)\\ &= E(Y^2) - E[\mu^2] + E[\mu^2] - E^2(\mu) && (46)\\ &= E(Y^2) - (E(E[Y|X]))^2 && (47)\\ &= E(Y^2) - (E^2 Y) && (48)\\ &= \text{Var } Y && (49) \end{aligned}$$

6 Covariance

The **covariance** between two RVs X and Y is defined as

$$\begin{aligned} \text{Cov}(X, Y) &\stackrel{\text{def}}{=} E\ ((X - \mu_X)(Y - \mu_Y)) && (50)\\ &= E\ (XY) - E\ (X)E\ (Y) && (51) \end{aligned}$$

and the correlation is defined as

$$\rho(X, Y) \stackrel{\text{def}}{=} \frac{\text{Cov}(X, Y)}{\sigma_X \sigma_Y} \tag{52}$$

We can show $-1 \leq \rho(X, Y) \leq 1$ as follows.

$$\begin{aligned} 0 &\leq \text{Var } (\frac{X}{\sigma_X} + \frac{Y}{\sigma_Y}) && (53)\\ &= \text{Var } (\frac{X}{\sigma_X}) + \text{Var } (\frac{Y}{\sigma_Y}) + 2\text{Cov}(\frac{X}{\sigma_X}, \frac{Y}{\sigma_Y}) && (54)\\ &= \frac{\text{Var } X}{\sigma_X} + \frac{\text{Var } Y}{\sigma_Y} + 2\text{Cov}(\frac{X}{\sigma_X}, \frac{Y}{\sigma_Y}) && (55)\\ &= 1 + 1 + 2\rho && (56) \end{aligned}$$

Hence $\rho \geq -1$. Similarly,

$$0 \leq \text{Var } (\frac{X}{\sigma_X} - \frac{Y}{\sigma_Y}) = 2(1 - \rho) \tag{57}$$

so $\rho \leq 1$.

If $Y = aX + b$, then $\rho(X, Y) = 1$ if $a > 0$ and $\rho(X, Y) = -1$ if $a < 0$. Thus **correlation only measures linear relationships** between RVs. If X and Y are independent, then $\text{Cov}(X, Y) = \rho = 0$; however, the converse is not true, as we see below.

The **partial correlation coefficient** is defined as

$$r_{XY|Z} \stackrel{\text{def}}{=} \frac{r_{XY} - r_{XZ} r_{YZ}}{\sqrt{(1 - r_{XZ}^2)(1 - r_{YZ}^2)}} \tag{58}$$

and measures the linear dependence of X and Y when Z is fixed.

If X is a random vector, its **covariance matrix** is defined to be

$$\text{Var}\,(X) = \Sigma \stackrel{\text{def}}{=} E\left[(X - E\,X)(X - E\,X)'\right] = \begin{pmatrix} \text{Var}\,(X_1) & \text{Cov}(X_1, X_2) & \cdots & \text{Cov}(X_p, X_p) \\ \text{Cov}(X_2, X_1) & \text{Var}\,(X_2) & \cdots & \text{Cov}(X_2, X_p) \\ \vdots & \vdots & \ddots & \vdots \\ \text{Cov}(X_p, X_1) & \text{Cov}(X_p, X_2) & \cdots & \text{Var}\,(X_p) \end{pmatrix} \tag{59}$$

If a is a vector and A a matrix, we have the following two important results:

$$\text{Var}\,(a^T X) = a^T \Sigma a \tag{60}$$

$$\text{Var}\,(AX) = A \Sigma A^T \tag{61}$$

The **conditional covariance** is defined as

$$\text{Cov}(X, Y|Z = z) \stackrel{\text{def}}{=} \sum_{x,y} p(x, y|z)(x - E\,(x|z))(y - E\,(Y|z)) \tag{62}$$

which is a function of Z.

7 Uncorrelated does not necessarily imply independent

Consider two RVs $X, Y \in \{-1, 0, 1\}$ with the following joint distribution:

$$p(X, Y) = \left(\begin{array}{c|ccc} X/Y| & 0 & -1 & 1 \\ \hline -1| & 0.25 & 0 & 0 \\ 0| & 0 & 0.25 & 0.25 \\ 1| & 0.25 & 0 & 0 \end{array}\right) \tag{63}$$

The marginal distributions are clearly $p(X) = p(Y) = (0.25, 0.5, 0.25)$. We will first show that X and Y are uncorrelated. We have

$$E\,(X, Y) = \sum_{x \in \{-1,0,1\}} \sum_{y \in \{-1,0,1\}} x\, y\, p(x, y) \tag{64}$$

$$= -1 \cdot 0 \cdot 0.25 + 0 \cdot -1 \cdot 0.25 + 0 \cdot 1 \cdot 0.25 + 1 \cdot 0 \cdot 0.25 = 0 \tag{65}$$

and

$$E\,X = \sum_{x \in \{-1,0,1\}} x p(x) = -1 \cdot 0.25 + 0 \cdot 0.5 + 1 \cdot 0.25 = 0 \tag{66}$$

Similarly $E\,Y = 0$. Hence

$$\text{Cov}(X, Y) = E\,(X, Y) - E\,(X) E\,(Y) = 0 - 0 \tag{67}$$

However, it is easy to see that X and Y are not independent: i.e., $p(X, Y) \neq p(X)p(Y)$. We can simply multiply out the two marginals, c.f., Figure 3.

$$\begin{pmatrix} 0.25 \\ 0.5 \\ 0.25 \end{pmatrix} \begin{pmatrix} 0.25 & 0.5 & 0.25 \end{pmatrix} = \begin{pmatrix} 0.0625 & 0.1250 & 0.0625 \\ 0.1250 & 0.2500 & 0.1250 \\ 0.0625 & 0.1250 & 0.0625 \end{pmatrix} \tag{68}$$

8 Change of variables

Let X be an rv with pdf $p_x(x)$. Let $Y = g(X)$ where g is differentiable and strictly monotonic (so $x = g^{-1}(y)$ exists). What is $p_y(y)$? Observations falling in the range $(x, x + \delta x)$ will get transformed into $(y, y + \delta y)$, where $p_x(x)\delta x \approx p_y(y)\delta_y$. Hence

$$p_y(y) = p_x(x)|\frac{dx}{dy}| \tag{69}$$

Now let x_1, x_2 have joint distribution $p_x(x_1, x_2)$ and let $(y_1, y_2) = g(x_1, x_2)$, where g is an invertible transform. Then

$$p_y(y_1, y_2) = p_x(x_1, x_2)|J_{x/y}| = p_x(x_1, x_2)|J_{y/x}^{-1}| \tag{70}$$

where J is the **Jacobian** (how much the unit volume changes), defined as

$$J_{x/y} = \frac{\partial(x_1, x_2)}{\partial(y_1, y_2)} \stackrel{\text{def}}{=} det \begin{pmatrix} \frac{\partial x_1}{\partial y_1} & \frac{\partial x_1}{\partial y_2} \\ \frac{\partial x_2}{\partial y_1} & \frac{\partial x_2}{\partial y_2} \end{pmatrix} \tag{71}$$

where det is the determinant (since we use $|J|$ to denote absolute value). More mnemonically , we can write this as

$$p_{new} = p_{old}|J_{old/new}| = p_{old}|J_{new/old}^{-1}| \tag{72}$$

As an example, consider transforming a density from polar (r, θ) to Cartesian (x, y) coordinates:

$$(r, \theta) \rightarrow (x = r\cos\theta, y = r\sin\theta) \tag{73}$$

Then

$$\begin{aligned} J_{new/old} &= \frac{\partial(x, y)}{\partial(r, \theta)} && (74)\\ &= det \begin{pmatrix} \frac{\partial x}{\partial r} & \frac{\partial x}{\partial \theta} \\ \frac{\partial y}{\partial r} & \frac{\partial y}{\partial \theta} \end{pmatrix} && (75)\\ &= det \begin{pmatrix} \cos\theta & -r\sin\theta \\ \sin\theta & r\cos\theta \end{pmatrix} && (76)\\ &= -r\sin^2\theta - r\cos^2\theta && (77)\\ &= = -r && (78) \end{aligned}$$

Hence

$$p_{X,Y}(x, y) = p_{R,\Theta}(r, \theta)|J_{new/old}^{-1}| = p_{R,\Theta}(r, \theta)\frac{1}{r} \tag{79}$$

To see this geometrically, notice that

$$p_{R,\Theta}(r, \theta)drd\theta = P(r \leq R \leq r + dr, \theta \leq \Theta \leq \theta + d\theta) \tag{80}$$

is the area of the shaded patch in Figure 5, which is clearly $rdrd\theta$, times the density at the center of the patch. Hence

$$P(r \leq R \leq r + dr, \theta \leq \Theta \leq \theta + d\theta) = p_{X,Y}(r\cos\theta, r\sin\theta)r\ dr\ d\theta \tag{81}$$

Hence

$$p_{R,\Theta}(r, \theta) = p_{X,Y}(r\cos\theta, r\sin\theta)r \tag{82}$$

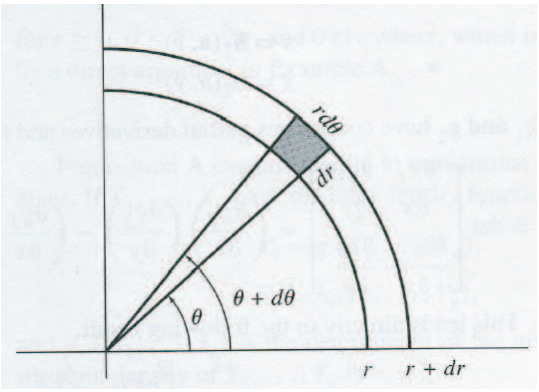

Figure 5: Change of variables from polar to Cartesian. The area of the shaded patch is $rdrd\theta$. Source: [Ric95] Figure 3.16.

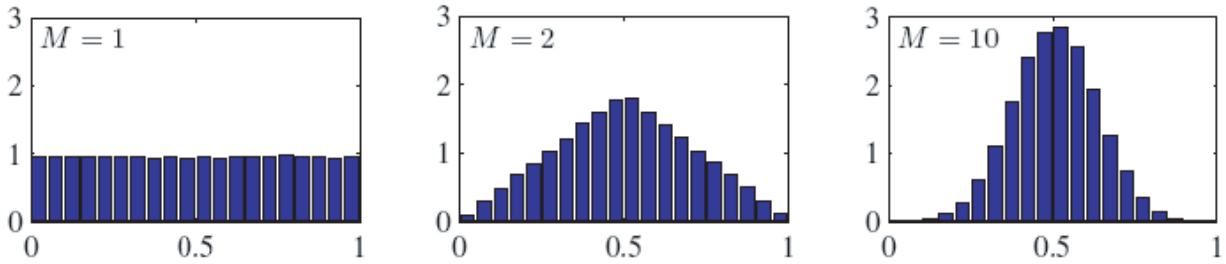

Figure 6: The central limit theorem in pictures. We plot a histogram of $\frac{1}{M}\sum_{i=1}^{M} x_i$, where $x_i \sim U(0,1)$. As $M \to \infty$, the distribution tends towards a Gaussian. Source: [Bis06] Fig 2.6.

9 Central limit theorem

One reason for the widespread use of Gaussian distributions is because of the **central limit theorem**, which says the following. Let $X_1, \ldots, X_n$ be iid with mean μ and variance σ^2. Let $\overline{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$. Then

$$Z_n = \frac{\sqrt{n}(\overline{X}_n - \mu)}{\sigma} \to N(0,1) \tag{83}$$

i.e., sums of iid **random variables (rv's)** converge (in distribution) to a Gaussian. See Figure 6 for an example. Hence if there are many random factors that exert an additive effect on the output, then rather than modeling each factor separately, we can model their net effect, which is to add Gaussian noise to the output. Thus we use a Gaussian to summarize our **ignorance** of the true causes of the output. (The Gaussian can also be motivated by the fact that it is the unique distribution which maximizes the **entropy** subject to first and second moment constraints. We discuss this later.)

References

[Bis06] C. Bishop. *Pattern recognition and machine learning.* Springer, 2006. Draft version 1.21.

[Ric95] J. Rice. *Mathematical Statistics and Data Analysis.* Duxbury, second edition, 1995.