where $B(\alpha_1, \ldots, \alpha_K)$ is the natural generalization of the beta function to K variables:

$$B(\boldsymbol{\alpha}) \triangleq \frac{\prod_{k=1}^{K} \Gamma(\alpha_k)}{\Gamma(\alpha_0)} \tag{2.76}$$

where $\alpha_0 \triangleq \sum_{k=1}^{K} \alpha_k$.

Figure 2.14 shows some plots of the Dirichlet when $K = 3$, and Figure 2.15 for some sampled probability vectors. We see that $\alpha_0 = \sum_{k=1}^{K} \alpha_k$ controls the strength of the distribution (how peaked it is), and the α_k control where the peak occurs. For example, $\mathrm{Dir}(1, 1, 1)$ is a uniform distribution, $\mathrm{Dir}(2, 2, 2)$ is a broad distribution centered at $(1/3, 1/3, 1/3)$, and $\mathrm{Dir}(20, 20, 20)$ is a narrow distribution centered at $(1/3, 1/3, 1/3)$. If $\alpha_k < 1$ for all k, we get "spikes" at the corner of the simplex.

For future reference, the distribution has these properties

$$\mathbb{E}[x_k] = \frac{\alpha_k}{\alpha_0}, \text{ mode}[x_k] = \frac{\alpha_k - 1}{\alpha_0 - K}, \text{ var}[x_k] = \frac{\alpha_k(\alpha_0 - \alpha_k)}{\alpha_0^2(\alpha_0 + 1)} \tag{2.77}$$

where $\alpha_0 = \sum_k \alpha_k$. Often we use a symmetric Dirichlet prior of the form $\alpha_k = \alpha/K$. In this case, the mean becomes $1/K$, and the variance becomes $\mathrm{var}[x_k] = \frac{K-1}{K^2(\alpha+1)}$. So increasing α increases the precision (decreases the variance) of the distribution.

2.6 Transformations of random variables

If $\mathbf{x} \sim p()$ is some random variable, and $\mathbf{y} = f(\mathbf{x})$, what is the distribution of $\mathbf{y}$? This is the question we address in this section.

2.6.1 Linear transformations

Suppose $f()$ is a linear function:

$$\mathbf{y} = f(\mathbf{x}) = \mathbf{A}\mathbf{x} + \mathbf{b} \tag{2.78}$$

In this case, we can easily derive the mean and covariance of $\mathbf{y}$ as follows. First, for the mean, we have

$$\mathbb{E}[\mathbf{y}] = \mathbb{E}[\mathbf{A}\mathbf{x} + \mathbf{b}] = \mathbf{A}\boldsymbol{\mu} + \mathbf{b} \tag{2.79}$$

where $\boldsymbol{\mu} = \mathbb{E}[\mathbf{x}]$. This is called the **linearity of expectation**. If $f()$ is a scalar-valued function, $f(\mathbf{x}) = \mathbf{a}^T\mathbf{x} + b$, the corresponding result is

$$\mathbb{E}\left[\mathbf{a}^T\mathbf{x} + b\right] = \mathbf{a}^T\boldsymbol{\mu} + b \tag{2.80}$$

For the covariance, we have

$$\mathrm{cov}[\mathbf{y}] = \mathrm{cov}[\mathbf{A}\mathbf{x} + \mathbf{b}] = \mathbf{A}\boldsymbol{\Sigma}\mathbf{A}^T \tag{2.81}$$

where $\boldsymbol{\Sigma} = \mathrm{cov}[\mathbf{x}]$. We leave the proof of this as an exercise. If $f()$ is scalar valued, the result becomes

$$\mathrm{var}[y] = \mathrm{var}\left[\mathbf{a}^T\mathbf{x} + b\right] = \mathbf{a}^T\boldsymbol{\Sigma}\mathbf{a} \tag{2.82}$$

We will use both of these results extensively in later chapters. Note, however, that the mean and covariance only completely define the distribution of $\mathbf{y}$ if $\mathbf{x}$ is Gaussian. In general we must use the techniques described below to derive the full distribution of $\mathbf{y}$, as opposed to just its first two moments.

2.6.2 General transformations

If X is a discrete rv, we can derive the pmf for y by simply summing up the probability mass for all the x's such that $f(x) = y$:

$$p_y(y) = \sum_{x:f(x)=y} p_x(x) \tag{2.83}$$

For example, if $f(X) = 1$ if X is even and $f(X) = 0$ otherwise, and $p_x(X)$ is uniform on the set $\{1, \ldots, 10\}$, then $p_y(1) = \sum_{x \in \{2,4,6,8,10\}} p_x(x) = 0.5$, and $p_y(0) = 0.5$ similarly. Note that in this example, f is a many-to-one function.

If X is continuous, we cannot use Equation 2.83 since $p_x(x)$ is a density, not a pmf, and we cannot sum up densities. Instead, we work with cdf's, and write

$$P_y(y) \triangleq P(Y \leq y) = P(f(X) \leq y) = P(X \in \{x | f(x) \leq y\}) \tag{2.84}$$

We can derive the pdf of y by differentiating the cdf.

In the case of monotonic and hence invertible functions, we can write

$$P_y(y) = P(f(X) \leq y) = P(X \leq f^{-1}(y)) = P_x(f^{-1}(y)) \tag{2.85}$$

Taking derivatives we get

$$p_y(y) \triangleq \frac{d}{dy} P_y(y) = \frac{d}{dy} P_x(f^{-1}(y)) = \frac{dx}{dy} \frac{d}{dx} P_x(x) = \frac{dx}{dy} p_x(x) \tag{2.86}$$

where $x = f^{-1}(y)$. We can think of dx as a measure of volume in the x-space; similarly dy measures volume in y space. Thus $\frac{dx}{dy}$ measures the change in volume. Since the sign of this change is not important, we take the absolute value to get the general expression:

$$p_y(y) = p_x(x) \left| \frac{dx}{dy} \right| \tag{2.87}$$

This is called **change of variables** formula. We can understand this result more intuitively as follows. 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 $p_y(y) \approx p_x(x) |\frac{\delta x}{\delta y}|$. For example, suppose $X \sim U(-1, 1)$, and $Y = X^2$. Then $p_y(y) = \frac{1}{2} y^{-\frac{1}{2}}$. See also Exercise 2.10.

2.6.2.1 Multivariate change of variables *

We can extend the previous results to multivariate distributions as follows. Let f be a function that maps $\mathbb{R}^n$ to $\mathbb{R}^n$, and let $\mathbf{y} = f(\mathbf{x})$. Then its **Jacobian matrix J** is given by

$$\mathbf{J}_{\mathbf{x} \to \mathbf{y}} \triangleq \frac{\partial(y_1, \ldots, y_n)}{\partial(x_1, \ldots, x_n)} \triangleq \begin{pmatrix} \frac{\partial y_1}{\partial x_1} & \cdots & \frac{\partial y_1}{\partial x_n} \\ \vdots & \ddots & \vdots \\ \frac{\partial y_n}{\partial x_1} & \cdots & \frac{\partial y_n}{\partial x_n} \end{pmatrix} \tag{2.88}$$

$|\det \mathbf{J}|$ measures how much a unit cube changes in volume when we apply f.

If f is an invertible mapping, we can define the pdf of the transformed variables using the Jacobian of the inverse mapping $\mathbf{y} \rightarrow \mathbf{x}$:

$$p_y(\mathbf{y}) = p_x(\mathbf{x}) \Big| \det \left(\frac{\partial \mathbf{x}}{\partial \mathbf{y}} \right) \Big| = p_x(\mathbf{x}) |\det \mathbf{J}_{\mathbf{y} \rightarrow \mathbf{x}}| \tag{2.89}$$

In Exercise 4.5 you will use this formula to derive the normalization constant for a multivariate Gaussian.

As a simple example, consider transforming a density from Cartesian coordinates $\mathbf{x} = (x_1, x_2)$ to polar coordinates $\mathbf{y} = (r, \theta)$, where $x_1 = r\cos\theta$ and $x_2 = r\sin\theta$. Then

$$\mathbf{J}_{\mathbf{y} \rightarrow \mathbf{x}} = \begin{pmatrix} \frac{\partial x_1}{\partial r} & \frac{\partial x_1}{\partial \theta} \\ \frac{\partial x_2}{\partial r} & \frac{\partial x_2}{\partial \theta} \end{pmatrix} = \begin{pmatrix} \cos\theta & -r\sin\theta \\ \sin\theta & r\cos\theta \end{pmatrix} \tag{2.90}$$

and

$$|\det \mathbf{J}| = |r\cos^2\theta + r\sin^2\theta| = |r| \tag{2.91}$$

Hence

$$p_{\mathbf{y}}(\mathbf{y}) = p_{\mathbf{x}}(\mathbf{x}) |\det \mathbf{J}| \tag{2.92}$$

$$p_{r,\theta}(r, \theta) = p_{x_1,x_2}(x_1, x_2) r = p_{x_1,x_2}(r\cos\theta, r\sin\theta) r \tag{2.93}$$

To see this geometrically, notice that the area of the shaded patch in Figure 2.16 is given by

$$P(r \le R \le r + dr, \theta \le \Theta \le \theta + d\theta) = p_{r,\theta}(r, \theta) dr d\theta \tag{2.94}$$

In the limit, this is equal to the density at the center of the patch, $p(r, \theta)$, times the size of the patch, $r\ dr\ d\theta$. Hence

$$p_{r,\theta}(r, \theta) dr d\theta = p_{x_1,x_2}(r\cos\theta, r\sin\theta) r\ dr\ d\theta \tag{2.95}$$

2.6.3 Central limit theorem

Now consider N random variables with pdf's (not necessarily Gaussian) $p(x_i)$, each with mean μ and variance σ^2. We assume each variable is **independent and identically distributed** or **iid** for short. Let $S_N = \sum_{i=1}^N X_i$ be the sum of the rv's. This is a simple but widely used transformation of rv's. One can show that, as N increases, the distribution of this sum approaches

$$p(S_N = s) = \frac{1}{\sqrt{2\pi N \sigma^2}} \exp\left(-\frac{(s - N\mu)^2}{2N\sigma^2} \right) \tag{2.96}$$

Hence the distribution of the quantity

$$Z_N \triangleq \frac{S_N - N\mu}{\sigma\sqrt{N}} = \frac{\overline{X} - \mu}{\sigma/\sqrt{N}} \tag{2.97}$$

converges to the standard normal, where $\overline{X} = \frac{1}{N}\sum_{i=1}^N x_i$ is the sample mean. This is called the **central limit theorem**. See e.g., (Jaynes 2003, p222) or (Rice 1995, p169) for a proof.

In Figure 2.17 we give an example in which we compute the mean of rv's drawn from a beta distribution. We see that the sampling distribution of the mean value rapidly converges to a Gaussian distribution.