APPENDIX

Sampling from distributions D

Sampling from uniform distributions can readily be achieved by using pseudo-random numbers generated by standard library functions. It is less obvious how this can be applied to more general distributions such as Gaussians. However, if the cumulative distribution function (CDF) is calculated, the overall process immediately becomes clear. At that stage the real problem is that it is not possible to deal with the general case analytically, so it is necessary to resort to numerical methods, which tend to be slow. This means that for widely used distributions such as the Gaussian, special methods have to be devised: the Box−Muller method and its polar variant have proved highly useful in this situation.

Look out for:

- the CDF approach
- the problems that occur when even the simple Gaussian distribution is to be used
- how these are tackled by the Box−Muller approach
- the gain achieved using the polar variant of the Box−Muller method.

Sampling from distributions is used in Chapter 14, Machine Learning: Probabilistic Methods for quickly generating 2-D datasets; Chapter 22, Surveillance outlines how it is used as part of the particle filter approach to object tracking.

D.1 INTRODUCTION

In many areas of statistics including machine learning, it is useful to be able to represent probability distributions by samples taken from them. Modeling distributions in this way have the effect of greatly simplifying subsequent calculations and vastly reducing computational load. In particular, the Gaussian distribution crops up very frequently in practical situations, and it is often necessary to sample it. However, even for such a simple, well-defined function, it is not altogether obvious how to draw a suitable set of samples. Indeed, the only immediately obvious case is that of the uniform distribution, which is readily sampled using pseudo-random numbers generated by standard library functions (e.g., "rand").

Fortunately, it is not a big step to proceed from the uniform distribution to the Gaussian distribution. All we need to do is to imagine the M original measurements that would have composed the discrete histogram $f(x)$ of the distribution, place all M of them in numerical order—thereby obtaining a continuous uniform

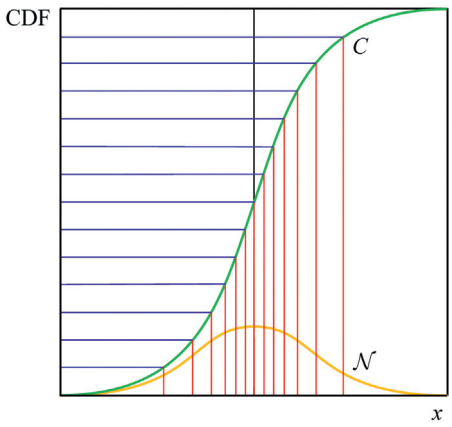

FIGURE D.1

Formation of the cumulative distribution function (CDF). The CDF of the approximately Gaussian distribution function $\mathcal{N}$ (orange, near the bottom of the figure) is shown higher up in green. Regular, uniformly generated sample positions are selected (blue, on the left ordinate axis): these are carried across to the CDF curve C and sample locations are marked in red on the original (orange) distribution. High CDF gradient points give high densities of samples along the x-axis: these are at positions where the original distribution probabilities are greatest.

distribution—and take samples from them using a suitable process. By examining the corresponding values of the original distribution, we obtain the desired set of sample values. All this may be seen from Fig. D.1, which shows the original distribution $\mathcal{N}$ and its CDF. The ordinate on the left is the continuous uniform distribution forming the set of all possible samples taken from the CDF, and Fig. D.1 shows these being sampled at regular intervals. Intuitively, it is clear that the samples lead to densities (of sample values) along the x-axis that are proportional to the gradient of the CDF curve; this would also be so for random samples, though in that case the sample locations would have to be smoothed out to get accurate density values. Note that the case of sampling at regular intervals has been used for clarity, but it is usual in statistical and machine learning applications to apply random sampling.

To understand the process more fully, we define the CDF according to the formula:

$$C(x) = \int_{-\infty}^{x} f(x)\mathrm{d}x \tag{D.1}$$

Differentiating immediately leads to the result:

$$f(x) = \frac{\mathrm{d}C(x)}{\mathrm{d}x} \tag{D.2}$$

This proves our earlier intuitive result—that the density of sampling along x is proportional to the gradient $\frac{dC(x)}{dx}$.

Finally, we should normalize the starting histogram f to obtain a true probability distribution p: in that case the CDF will range from 0 to 1, and will take the form:

$$c(x) = \int_{-\infty}^{x} p(x)dx \tag{D.3}$$

We then have

$$p(x) = \frac{dc(x)}{dx} \tag{D.4}$$

D.2 THE BOX–MULLER AND RELATED METHODS

While the CDF procedure seems sound, it is seldom easy to apply, because of the need to obtain and invert the CDF. In fact, only a limited number of possible probability functions lead to closed form solutions—e.g., the exponential distribution and the Cauchy distribution: in most other cases solutions need to be determined numerically, by methods such as piecewise-linear or polynomial approximation. However, the Gaussian distribution is needed for so many applications that special methods have been devised for sampling it. One such method is to calculate it in 2-D using polar coordinates. To carry this out, we define the Gaussian distribution over a suitable range, such as -1 to $+1$ in each of the x and y directions—though we also restrict it to a distance 0 to 1 radially. Thus we will be sampling within the unit circle: $0 \le r \le 1,\ \ 0 \le \theta \le 2\pi$. The r samples can be obtained from the uniform distribution Unif (0, 1) and the theta samples from Unif (0, 2π). Finally, x and y values can be obtained from the formulae

$$x = r\cos\theta \tag{D.5}$$

$$y = r\sin\theta \tag{D.6}$$

In fact, this approach is useful but excessively computation intensive. However, the method known as the Box–Muller transform overcomes this problem. This first expresses the distribution as

$$\begin{aligned} p(x,y) &= p(x)p(y) \\ &= \frac{1}{(2\pi)^{1/2}}\exp\left(-\frac{x^2}{2}\right) \times \frac{1}{(2\pi)^{1/2}}\exp\left(-\frac{y^2}{2}\right) \\ &= \frac{1}{2\pi}\exp\left(-\frac{x^2+y^2}{2}\right) \\ &= \frac{1}{2\pi}\exp\left(-\frac{r^2}{2}\right) \end{aligned} \tag{D.7}$$

Note that this equation has converted from Cartesians to polar coordinates, though no variation arises from the θ coordinate: this means that the θ distribution is subject to a uniform distribution over angles. The Box−Muller distribution takes one further step, using two independent standard uniform random variables u_1 and u_2, it produces independent standard x and y distributions using the following formulae:

$$R^2 = -2 \ln u_1 \tag{D.8}$$

$$\theta = 2\pi u_2 \tag{D.9}$$

This works because R^2 is the square of the norm of the standard 2-D Gaussian variable (x, y) and has the chi-squared distribution with two degrees of freedom, which means that it is identical to the Gaussian distribution. Using Eqs. (D.8) and (D.9) now permits the values of x and y to be deduced using the standard polar coordinate Eqs. (D.5) and (D.6). Note that the method produces two independent Gaussian distributions, of which one can be ignored if only a 1-D solution is required: it is of course useful for sampling from 2-D Gaussian distributions. Refer to the original paper by Box and Muller (1958) and also Pike (1965) for the subtleties of the method and proofs of its validity.

Finally, note that a separate "polar" form of the Box−Muller method is commonly used to eliminate the need to calculate the cos and sine functions, thereby speeding up the algorithm by a further significant factor: this version of the algorithm was developed by Knop (1969) and is perhaps the most widely used as it was included in Press et al. *Numerical Recipes in C* (1997).

In this book, sampling from distributions is implemented in Chapter 14, Machine Learning: Probabilistic Methods, where it is used to generate the starting data in Fig. 14.3 from multiple 2-D Gaussians. It is also relevant for the particle filter calculation presented in Section 22.4: see especially the top parts of Fig. 22.8.

D.3 BIBLIOGRAPHICAL AND HISTORICAL NOTES

Space prevents a more detailed study of sampling from distributions. Bishop (2006) provides useful theory, leading up to the Box−Muller method. For further work it is useful to refer to the original paper by Box and Muller (1958), to Pike (1965), and the quite recent paper by Martino et al. (2012); and for the polar form of the method to Knop (1969) and Press et al. (1997). Rubinstein and Kroese (2007) cover more complex—and often approximate—cases that apply for empirically obtained distributions.