A.3 Probability Distributions

A probability is a measure over a set of events that satisfies three axioms:

1. The measure of each event is between 0 and 1. We write this as $0 \le P(X = x_i) \le 1$, where X is a random variable representing an event and x_i are the possible values of X. In general, random variables are denoted by uppercase letters and their values by lowercase letters.
2. The measure of the whole set is 1; that is, $\sum_{i=1}^{n} P(X = x_i) = 1$.
3. The probability of a union of disjoint events is the sum of the probabilities of the individual events; that is, $P(X = x_1 \vee X = x_2) = P(X = x_1) + P(X = x_2)$, in the case where x_1 and x_2 are disjoint.

A **probabilistic model** consists of a sample space of mutually exclusive possible outcomes, together with a probability measure for each outcome. For example, in a model of the weather tomorrow, the outcomes might be *sun, cloud, rain*, and *snow*. A subset of these outcomes constitutes an event. For example, the event of precipitation is the subset consisting of $\{$*rain, snow*$\}$.

We use $\mathbf{P}(X)$ to denote the vector of values $\langle P(X = x_1), \ldots, P(X = x_n)\rangle$. We also use $P(x_i)$ as an abbreviation for $P(X = x_i)$ and $\sum_x P(x)$ for $\sum_{i=1}^{n} P(X = x_i)$.

The conditional probability $P(B \mid A)$ is defined as $P(B \cap A)/P(A)$. A and B are conditionally independent if $P(B \mid A) = P(B)$ (or equivalently, $P(A \mid B) = P(A)$).

For continuous variables, there are an infinite number of values, and unless there are point spikes, the probability of any one exact value is 0. So it makes more sense to talk about the value being within a range. We do that with a **probability density function**, which has a slightly different meaning from the discrete probability function. Since $P(X = x)$—the probability that X has the value x exactly—is zero, we instead measure how likely it is that X falls into an interval around x, compared to the width of the interval, and take the limit as the interval width goes to zero:

Probability density function

$$P(x) = \lim_{dx \to 0} P(x \le X \le x + dx)/dx\,.$$

The density function must be nonnegative for all x and must have

$$\int_{-\infty}^{\infty} P(x)\,dx = 1\,.$$

We can also define the **cumulative distribution** $F_X(x)$, which is the probability of a random variable being less than x:

Cumulative distribution

$$F_X(x) = P(X \le x) = \int_{-\infty}^{x} P(u)\,du\,.$$

Note that the probability density function has units, whereas the discrete probability function is unitless. For example, if values of X are measured in seconds, then the density is measured in Hz (i.e., 1/sec). If values of $\mathbf{X}$ are points in three-dimensional space measured in meters, then density is measured in $1/m^3$.

One of the most important probability distributions is the **Gaussian distribution**, also known as the **normal distribution**. We use the notation $\mathcal{N}(x; \mu, \sigma^2)$ for the normal distribu-

Gaussian distribution