2.4 Conditional PDFs

Treated as a regular PDF. Given random variables x, y, z (which can themselves be collections of random variables):

Marginalization: $f_{x|z}(x|z) := \sum_{y\in\mathcal{Y}} f_{xy|z}(x,y|z)$ (or integral if CRV)

Conditioning: $f_{x|yz}(x|y,z) := \dfrac{f_{xy|z}(x,y|z)}{f_{y|z}(y|z)}$

This is intuitive, z just parametrizes the PDF.

Independence

- Random variables x and y are said to be **independent** if $f(x|y) = f(x)$.
  - Combining this with the conditioning axiom, we have an equivalent definition: $f(x,y) = f(x)\,f(y)$. The probability of two independent events is the product of individual probabilities.
  - It is easy to show that $f(x|y) = f(x) \Leftrightarrow f(y|x) = f(y)$, as one would expect. **PSET 1: P4**
- **Conditional independence:** Random variables x and y are said to be **conditionally independent** on z if $f(x|y,z) = f(x|z)$. Knowledge of z makes x and y independent.
  - It is easy to show that the above implies $f(x,y|z) = f(x|z)\,f(y|z)$. **PSET 1: P5**
- Can also come up with examples where $f(x,y|z) = f(x|z)\,f(y|z)$, *but* $f(x,y) \neq f(x)\,f(y)$. **PSET 1: P6+7**
- Independence greatly simplifies algorithms and allows us to decouple information and processes.

2.5 Expected Value

- $\mathrm{E}[x] := \sum_{x\in\mathcal{X}} x f_x(x)$ (or integral if CRV)
  Intuitive definition: weighted sum of values, with probability being the weight.
- Applies to conditional PDFs:

  $$\mathrm{E}[x|y] := \sum_{x\in\mathcal{X}} x f_{x|y}(x|y)$$

  Just treat y as a parameter.

Law of the unconscious statistician

Let x be a random variable, $x \in \mathcal{X}$. Let $y = g(x)$ with $y \in \mathcal{Y}$ and $\mathcal{Y} = g(\mathcal{X})$; that is,
$\mathcal{Y} = \{y \mid y = g(x), x \in \mathcal{X}\}$ (the set $\mathcal{Y}$ consists of all possible values of $g(x)$ when $x \in \mathcal{X}$).
Then y is a random variable and

$$\mathrm{E}[y] = \sum_{y\in\mathcal{Y}} y f_y(y) = \sum_{x\in\mathcal{X}} g(x) f_x(x) \text{ (or integral if CRV),}$$ **PSET 1: P9**

or simply $\mathrm{E}[g(x)]$. Why this is useful: to calculate $\mathrm{E}[g(x)]$ we don't have to first calculate the PDF of y.

Date compiled: March 4, 2013

Mean and Variance

- $\mathrm{E}[x]$ is called the **mean**, generally a vector.
- $\mathrm{Var}[x] := \mathrm{E}[(x - \mathrm{E}[x])(x - \mathrm{E}[x])^T]$ is called the **variance**, generally a matrix. Often called the **co-variance** when a matrix, but we will simply use the term variance.

2.6 Sampling a Distribution

Most math libraries have functions that can generate **uniformly distributed**, random real numbers[1] in the range $[0, 1]$. The function `rand` in Matlab for example. That is, we have a way to generate a sample of a CRV u with PDF

$$f_u(u) = \begin{cases} 1 & u \in [0,1] \\ 0 & \text{otherwise.} \end{cases}$$

Furthermore, repeated calls to the random number generator are independent. That is, the joint PDF associated with two calls to the `rand` function is:

$$f_{u_1,u_2}(u_1, u_2) = \begin{cases} 1 & u_1, u_2 \in [0,1] \\ 0 & \text{otherwise,} \end{cases}$$

from which it follows that $f_{u_1,u_2}(u_1, u_2) = f_{u_1}(u_1)\, f_{u_2}(u_2)$.

We now present four algorithms for generating random variables with arbitrary PDFs using samples of CRVs with uniform distributions.

A1. One DRV: Given a *desired* PDF $\hat{f}_x(x)$ for a DRV x, can we come up with a procedure that generates x from u? Without loss of generality, let $\mathcal{X} = \{\dots, -1, 0, 1, \dots\}$ be the set of integers.

- Let $\hat{F}_x(x) := \sum_{\bar{x}=-\infty}^{x} \hat{f}_x(\bar{x})$, often called the **cumulative distribution function** (CDF). Clearly, $\hat{F}_x(-\infty) = 0$, $\hat{F}_x(\infty) = 1$, and it is a non-decreasing function.
- Let u be generated from $f_u(u)$, the uniform distribution. Solve for x such that $\hat{F}_x(x-1) < u$ and $\hat{F}_x(x) \geq u$ (see Figure 2.1). Then we claim that this defines a random variable x with PDF $\hat{f}_x(x)$. That is, $f_x(x) = \hat{f}_x(x)$.

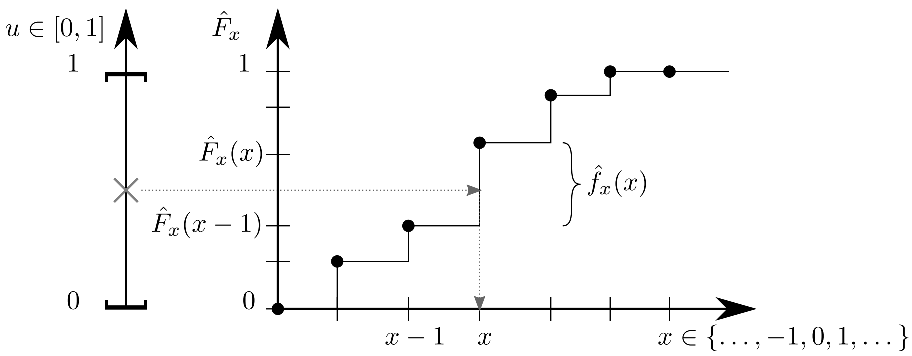

Figure 2.1: Sampling a discrete distribution.

[1] Strictly speaking, a computer does not generate real numbers since it uses finitely many bits to represent a number. For most applications however, the approximation of the random number generator output as real numbers is valid.

Proof:
First, notice that we can always solve for such an x; potentially there are problems with $u = 0$ and $u = 1$, but the probability of this happening is 0, since u is a CRV (although in practice we would check for these two cases, and simply re-sample if we obtained them).

For a fixed x, determine the values of u for which $\hat{F}_x(x-1) < u$ and $\hat{F}_x(x) \geq u$: We have

$$u \leq \hat{F}_x(x) = \hat{F}_x(x-1) + \hat{f}_x(x) \quad \text{and, thus,} \quad \hat{F}_x(x-1) < u \leq \hat{F}_x(x-1) + \hat{f}_x(x)\,.$$

The probability of u being in $(\hat{F}_x(x-1)\,, \hat{F}_x(x)]$ is

$$\int\limits_{\hat{F}_x(x-1)}^{\hat{F}_x(x)} 1\; du = \int\limits_{\hat{F}_x(x-1)}^{\hat{F}_x(x-1)+\hat{f}_x(x)} 1\; du = \hat{f}_x(x)$$

as desired. That is, the random variable $x \in \mathcal{X}$ has PDF $f_x(x) = \hat{f}_x(x)$.

A2. More than one DRV: What if we have a desired joint PDF $\hat{f}_{xy}(x, y)$, where x and y are single (scalar-valued) DRVs?

Option 1: This only works if $\mathcal{X}$ and $\mathcal{Y}$ are finite. Let N_x and N_y be the number of elements in $\mathcal{X}$ and $\mathcal{Y}$, respectively. Define $\mathcal{Z} = \{1, 2, ..., N_x N_y\}$, which has a total of $N_x N_y$ elements. Define

$$\hat{f}_z(1) = \hat{f}_{xy}(1, 1)\,,\ \hat{f}_z(2) = \hat{f}_{xy}(1, 2)\,, \dots, \hat{f}_z(N_x N_y) = \hat{f}_{xy}(N_x, N_y)$$

You can then apply the results of Algorithm A1 to $\hat{f}_z(z)$.

Option 2 This works for both finite and infinite number of elements. Decompose $\hat{f}_{xy}(x, y) = \hat{f}_{x|y}(x|y)\,\hat{f}_y(y)$. Apply A1 to first get a value for y via $\hat{f}_y(y)$, then with y fixed, apply A1 again to get a value for x via $\hat{f}_{x|y}(x|y)$. Note that the independence of the uniform number generator between successive calls is key.

Both options have been described with two DRVs, but they clearly generalize to an arbitrary number of DRVs.

A3. One CRV: Given a desired PDF $\hat{f}_x(x)$ for a continuous random variable x, $\mathcal{X} = (-\infty, \infty)$.

- Let $\hat{F}_x(x) := \int\limits_{-\infty}^{x} \hat{f}_x(\bar{x})\, d\bar{x}$, the CDF of x.
- Let u be uniform on $[0, 1]$. Let $x = \hat{F}_x^{-1}(u)$. That is, x is *any* solution to $u = \hat{F}_x(x)$. Then x has PDF $f_x(x) = \hat{f}_x(x)$.

Proof:
Assume that $\hat{F}_x(\cdot)$ is strictly increasing (proof can be generalized). Let a be arbitrary.

$$\begin{aligned} F_x(a) &= \Pr(x \leq a) \quad \text{(by definition)} \\ &= \Pr(\hat{F}_x^{-1}(u) \leq a) \end{aligned}$$

Since $\hat{F}_x(\cdot)$ is strictly increasing,

$$\hat{F}_x^{-1}(u) \leq a \Leftrightarrow u \leq \hat{F}_x(a)\,,$$

and, therefore,

$$F_x(a) = \Pr(u \leq \hat{F}_x(a)) = \hat{F}_x(a) \quad \text{(since } u \text{ is uniform)}$$

Since a is arbitrary, $F_x(x) = \hat{F}_x(x)\ \forall\, x$, which implies $f_x(x) = \hat{f}_x(x)$.

A4. More than one CRV: This is analogous to Option 2 in Algorithm 2. Decompose $\hat{f}_{xy}(x, y) = \hat{f}_{x|y}(x|y)\,\hat{f}_y(y)$. Apply A3 to first get a value for y via $\hat{f}_y(y)$, then with y fixed, apply A3 again to get a value for x via $\hat{f}_{x|y}(x|y)$.

2.7 Change of Variables

We often encounter situations where we have to calculate the PDFs of functions of random variables. We will focus on one random variable; similar formulas can be derived for joint random variables.

Discrete Random Variable

Let $f_y(y)$ be given. Consider function $x = g(y)$. What is $f_x(x)$? Let $\mathcal{X} = g(\mathcal{Y})$. For each $x_j \in \mathcal{X}$, let $\mathcal{Y}_j = \{y_{j,i}\}$ be the set of all $y \in \mathcal{Y}$ such that $g(y_{j,i}) = x_j$. We claim that

$$f_x(x_j) = \sum_{y_{j,i} \in \mathcal{Y}_j} f_y(y_{j,i})$$

Proof:

- Since $\mathcal{Y}_j \cap \mathcal{Y}_k = \emptyset$ when $j \neq k$ and $\bigcup_j \mathcal{Y}_j = \mathcal{Y}$, it is clear that $\sum_{x \in \mathcal{X}} f_x(x) = 1$.
- $\Pr(x = x_j) = \Pr(y \in \mathcal{Y}_j) = f_x(x_j)$ as required.

Continuous Random Variable

This is a bit more complicated. We will assume that $g(y)$ is continuously differentiable and strictly monotonic, and that $f_y(y)$ is continuous. Strictly monotonic means that $\frac{dg}{dy}(y) > 0$ for all y or that $\frac{dg}{dy}(y) < 0$ for all y. Assume that $\frac{dg}{dy}(y) > 0$ for all y, the other case is similar (**PSET 1: P8**). We claim that

$$f_x(x) = \frac{f_y(y)}{\frac{dg}{dy}(y)}.$$

Proof:

- $\Pr(y \in [\bar{y}, \bar{y} + \Delta y]) = \int_{\bar{y}}^{\bar{y}+\Delta y} f_y(y)\, dy \to f_y(\bar{y})\, \Delta y$ as $\Delta y \to 0$
- Let $\bar{x} = g(\bar{y})$. Then
$$\begin{aligned} g(\bar{y} + \Delta y) &\to g(\bar{y}) + \frac{dg}{dy}(\bar{y})\Delta y \text{ as } \Delta y \to 0 \\ &= \bar{x} + \Delta x \text{ where } \Delta x := \frac{dg}{dy}(\bar{y})\Delta y \end{aligned}$$
- $\Pr(x \in [\bar{x}, \bar{x} + \Delta x]) = \Pr(y \in [\bar{y}, \bar{y} + \Delta y]) \to f_x(\bar{x})\, \Delta x$ as $\Delta x \to 0$.
Therefore, by letting $\Delta y \to 0$, and hence $\Delta x \to 0$:
$$f_x(\bar{x}) = \frac{f_y(\bar{y})}{\frac{dg}{dy}(\bar{y})}.$$

This gives us another way for generating samples of a PDF:

- Given a desired PDF $\hat{f}_x(x)$, and a method for sampling $f_y(y)$, find a function $x = g(y)$ such that
$$\frac{dg}{dy}(y) = \frac{f_y(y)}{\hat{f}_x(g(y))}.$$
Equivalently, solve $\frac{dx}{dy} = \frac{f_y(y)}{\hat{f}_x(x)}$, a differential equation.
- *Example:* Let y be uniformly distributed on $[0, 1]$. Then $f_y(y) = 1$. Solve
$$\frac{dx}{dy} = 1/\hat{f}_x(x) \quad \Rightarrow \quad dy = dx\, \hat{f}_x(x) \quad \Rightarrow \quad y = \int_{-\infty}^{x} \hat{f}_x(\bar{x}) d\bar{x} = \hat{F}_x(x)\,.$$ This is algorithm **A3**!