Equational reasoning for conditioning as disintegration

Chung-chieh Shan
ccshan@post.harvard.edu

Dylan Thurston
dpt@bostoncoop.net

Conditional distributions are widely used for practical inference, even when the condition has zero probability (such as setting a continuous variable to an observed value). This popularity contrasts with the scary pitfalls (such as Borel's paradox) that beset rigorous treatments of conditioning. In general, conditional expectations may arise that do not correspond to any conditional distribution at all. This gap between theory and practice has made it difficult to automate conditioning calculations soundly. In particular, given a generative model, we want to produce its conditional expectations mechanically, so as to optimize them numerically, simplify them symbolically, and so on.

Disintegration [1] is a rigorous approach to conditioning that covers a wide range of applications yet admits intuitive conditional distributions and allows their 'guilt-free manipulation'. In the present work, we mechanize this approach by adding a 'Lebesgue measure' operation to the usual monadic representation of stochastic experiments [2]. We show how to compute conditional distributions by equating expressions in this representation.

By 'compute', we mean producing a symbolic algebraic expression usually and giving up in failure occasionally, not approximating a real number to arbitrary precision. In fact, the latter task is impossible for conditional probabilities in general [3]. As for the former task, a system for computing conditional distributions that guarantees any expression it produces to be correct in some sense (explained below) is useful because we can feed a mathematical expression to existing software such as computer algebra systems, numerical solvers, and optimized code generators.

We use the following stochastic experiment as a running example. It flips a fair coin c to decide whether to scale a uniform random number x by 1 or by 3.

$$\begin{aligned} c &\sim \text{Uniform}[0,1] \\ s &= \text{if } c < 1/2 \text{ then 1 else 3} \\ x &\sim \text{Uniform}[-1,1] \\ y &= sx \end{aligned} \tag{1}$$

Suppose we observe y to be 0. The propositions $y = 0$ and $x = 0$ are logically equivalent, so they are one and the same event. When we condition on $x = 0$, the posterior marginal distribution of s is of course concentrated on $s = 1$ and $s = 3$, each with probability $1/2$. However, when we condition on $y = 0$, the posterior should be $s = 1$ with probability $3/4$ and $s = 3$ with probability $1/4$. This example is thus a simplified version of Borel's paradox.

We can think of this experiment as defining a measure μ. We can then compute the expectation of a 'test function' f from c,s,x,y to reals:

$$E(f) = \int f(c,s,x,y)\,\mu = \int_0^1 \left(\left(\int_{-1}^1 \frac{1}{2} \big(f(c,s,x,y)|_{y=sx} \big) dx \right) \Big|_{s=\text{if } c < 1/2 \text{ then 1 else 3}} \right) dc. \tag{2}$$

Despite its clumsy notation, this formula corresponds directly to (1): it is an integral over c of a (Dirac delta) integral over s of an integral over x of a (Dirac delta again) integral over y.

If there are any distributions conditional on y for us to speak of, they had better be a family μ_y of measures over c,s,x, parametrized by y. We do not even require these measures to be normalized. The basic idea of disintegration is that we want μ_y to satisfy

$$E(f) = \int \Big(\int f(c,s,x,y) \mu_y \Big) \nu \tag{3}$$

for all measurable f. The outer integral is with respect to some measure ν over y, which may be normalized like the marginal distribution over y or unnormalized like the standard Lebesgue measure over y. The inner integral is with respect to the measure μ_y over c,s,x, which again may be normalized or not. Note that we want to find μ_y independently of any particular test function f, so that we may call μ_y a conditional distribution, and not just call $\int f(c,s,x,y)\mu_y$ a conditional expectation. That is one reason disintegration is a stronger notion of conditioning than Kolmogorov's, and one reason μ_y does not always exist (and when it exists, is unique only up to an almost-sure equivalence).

The requirement (3) can be understood intuitively in several different ways:

1. If the original probability measure μ can be chopped into thin slices—one for each y—then the slices can be used as unnormalized conditional distributions.
2. If the original stochastic experiment can be rearranged so that y is chosen first (by sampling from ν), without affecting the ultimate measure μ, then we can describe a 'conditional stochastic experiment' by replacing that first step to choose y with a step that clamps y at the given value.
3. A conditional probability density can be defined as the ratio of a joint density to a marginal density, even when the marginal density is not with respect to the Lebesgue measure over y, but with respect to any measure ν. For instance, if ν is the marginal distribution over y, then the marginal density is simply 1.

The second intuition above is especially useful, because it turns out that the 'monadic' representation exemplified by (1) is a convenient format in which to reason about equality of measures. To represent unnormalized measures in the same format, we add two kinds of operations that go beyond describing steps in a stochastic experiment. First, we write '$x \sim$ Lebesgue' to denote the Lebesgue measure over x (which corresponds to the integration $\int_{-\infty}^{\infty} \cdots dx$). Second, if δ is a scalar expression then we write 'weight δ' to scale the integral by the density δ. For example, the expressions

$$\begin{array}{lll} x \sim \text{Uniform}[-1,1] & \text{and} & x \sim \text{Lebesgue} \\ & & \text{weight } (\text{if } -1 < x < 1 \text{ then } 1/2 \text{ else } 0) \end{array} \tag{4}$$

are equivalent, because the integrals

$$\int_{-1}^{1} \frac{1}{2} f(x)\,dx \qquad \text{and} \qquad \int_{-\infty}^{\infty} (\text{if } -1 < x < 1 \text{ then } 1/2 \text{ else } 0)\, f(x)\,dx \tag{5}$$

are equal. Also useful is the fact that the expressions

$$\begin{array}{lll} y \sim \text{Lebesgue} & \text{and} & z \sim \text{Lebesgue} \\ z = x + y & & y = z - x \end{array} \tag{6}$$

are equivalent. Finally, if $s > 0$ then the expressions

$$\begin{array}{lll} x \sim \text{Lebesgue} & \text{and} & y \sim \text{Lebesgue} \\ y = sx & & x = y/s \\ & & \text{weight } 1/s \end{array} \tag{7}$$

are equivalent.

Now, to condition our running example (1) on y, we reason equationally as follows:

$$\begin{array}{lll} (1) & = & c \sim \text{Uniform}[0,1] \\ & & s = \text{if } c < 1/2 \text{ then } 1 \text{ else } 3 \\ & & x \sim \text{Lebesgue} \\ & & \text{weight } (\text{if } -1 < x < 1 \text{ then } 1/2 \text{ else } 0) \\ & & y = sx \end{array} \Big\} \text{by (4)}$$

$$\begin{array}{lll} & = & c \sim \text{Uniform}[0,1] \\ & & s = \text{if } c < 1/2 \text{ then } 1 \text{ else } 3 \\ & & x \sim \text{Lebesgue} \\ & & y = sx \\ & & \text{weight } (\text{if } -1 < x < 1 \text{ then } 1/2 \text{ else } 0) \end{array} \Big\} \text{exchange scalar with (Dirac delta) integral}$$

$$
\begin{aligned}
= \quad & c \sim \text{Uniform}[0,1] \\
& s = \text{if } c < 1/2 \text{ then } 1 \text{ else } 3 \\
& \left.\begin{array}{l} y \sim \text{Lebesgue} \\ x = y/s \\ \text{weight } 1/s \end{array}\right\} \text{by (7)} \\
& \text{weight } (\text{if } -1 < x < 1 \text{ then } 1/2 \text{ else } 0) \\
= \quad & \left.\begin{array}{l} y \sim \text{Lebesgue} \\ c \sim \text{Uniform}[0,1] \\ s = \text{if } c < 1/2 \text{ then } 1 \text{ else } 3 \end{array}\right\} \text{exchange integrals} \\
& x = y/s \\
& \text{weight } 1/s \\
& \text{weight } (\text{if } -1 < x < 1 \text{ then } 1/2 \text{ else } 0)
\end{aligned}
$$

To compute the distribution of s given $y = 0$, we replace the first step '$y \sim$ Lebesgue' in the result above by '$y = 0$', then simplify using the fact that $-1 < 0/s < 1$ and we do not care about x (or about c, for that matter) to get

$$
\begin{aligned}
& y = 0 \\
& c \sim \text{Uniform}[0,1] \\
& s = \text{if } c < 1/2 \text{ then } 1 \text{ else } 3 \\
& \text{weight } 1/2s
\end{aligned}
\tag{8}
$$

as desired.

We emphasize that these equivalences follow purely symbolic rewriting patterns, whose side conditions ($s > 0$; the order of integrals can be changed, etc.) are easy to discharge. In this sense, we have mechanized the disintegration approach to conditioning, in a way that could fit into a rewrite system for statistical modeling [4]. Another reason for the search for a derivation to be easy is that equivalences are typically used in the direction that sweeps the conditioning variable and its dependencies 'up', that is, earlier in the generative process.

Moreover, our representation and calculation of conditional distributions generalizes Bhat et al.'s [5] representation and calculation of probability densities. For example, Bhat et al. demonstrate their procedure on the hierarchical model

$$
\begin{aligned}
& x \sim \text{Uniform}[0,1] \\
& y \sim \text{Uniform}[0,x].
\end{aligned}
\tag{9}
$$

We can easily calculate the marginal density over y of this model: apply a generalization of (4) and exchange integrals to get

$$
\begin{aligned}
& y \sim \text{Lebesgue} \\
& x \sim \text{Uniform}[0,1] \\
& \text{weight if } 0 < y < x \text{ then } 1/x \text{ else } 0,
\end{aligned}
\tag{10}
$$

so the density at y is $\int_0^1 (\text{if } 0 < y < x \text{ then } 1/x \text{ else } 0)\, dx$, essentially same as Bhat et al.'s result. It is also trivial for us to incorporate Bhat et al.'s rules for log, exp, etc. as equivalences and check their validity in a modular way.

References

[1] Joseph T. Chang and David Pollard. Conditioning as disintegration. *Statistica Neerlandica*, 51 (3):287–317, November 1997.

[2] Norman Ramsey and Avi Pfeffer. Stochastic lambda calculus and monads of probability distributions. In *POPL '02: Symposium on Principles of Programming Languages*, pages 154–165, January 2002.

[3] Nathanael L. Ackerman, Cameron E. Freer, and Daniel M. Roy. Noncomputable conditional distributions. In *LICS 2011: Symposium on Logic in Computer Science*, pages 107–116, June 2011.

[4] Sooraj Bhat, Ashish Agarwal, Alexander Gray, and Richard Vuduc. Toward interactive statistical modeling. *Procedia Computer Science*, 1(1):1829–1838, 2010. International Conference on Computational Science 2010.

[5] Sooraj Bhat, Ashish Agarwal, Richard Vuduc, and Alexander Gray. A type theory for probability density functions. In *POPL '12: Symposium on Principles of Programming Languages*, January 2012.