Conditioning and density, mathematically and computationally

Chung-chieh Shan (with Wazim Mohammed Ismail)

INDIANA UNIVERSITY

Mathematical Foundations of Programming Semantics

June 14, 2014

This work is supported by DARPA grant FA8750-14-2-0007.

Probabilistic programming

Alice beat Bob at a game. Is she better than him at it?

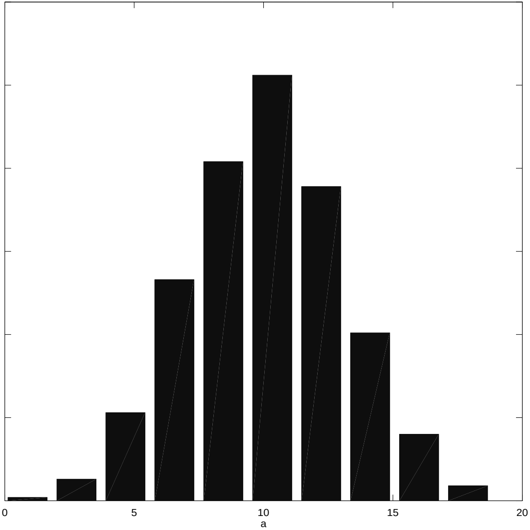
Generative story

Probabilistic programming

Alice beat Bob at a game. Is she better than him at it?

Generative story

```
a <- normal 10 3
```

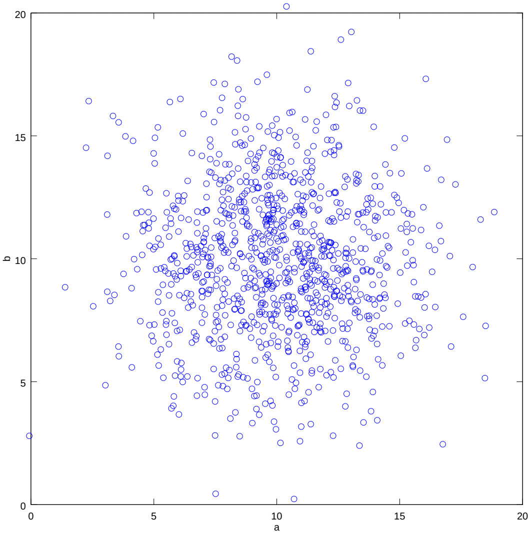


Probabilistic programming

Alice beat Bob at a game. Is she better than him at it?

Generative story

```
a <- normal 10 3
b <- normal 10 3
```

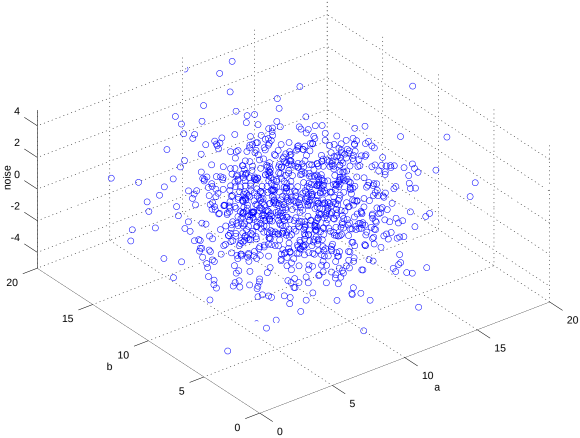


Probabilistic programming

Alice beat Bob at a game. Is she better than him at it?

Generative story

```
a <- normal 10 3
b <- normal 10 3
l <- normal  0 2
```



Probabilistic programming

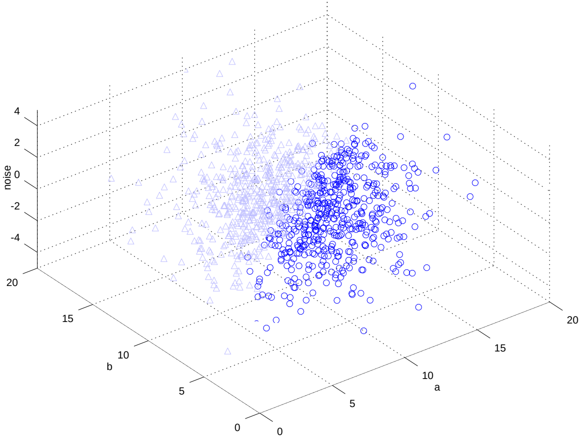
Alice beat Bob at a game. Is she better than him at it?

Generative story

```
a <- normal 10 3
b <- normal 10 3
l <- normal  0 2
```

Observed effect

```
condition (a-b > l)
```



Probabilistic programming

Alice beat Bob at a game. Is she better than him at it?

Generative story

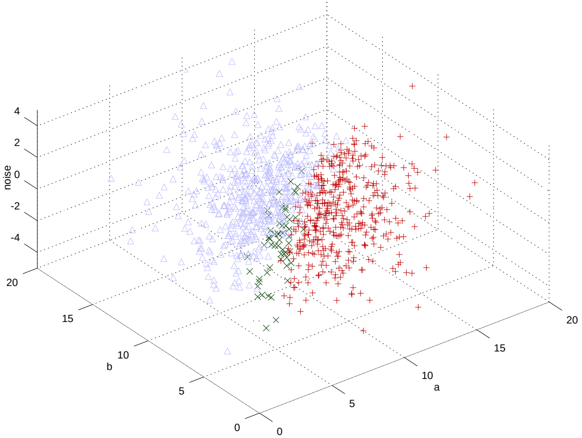
```
a <- normal 10 3
b <- normal 10 3
l <- normal  0 2
```

Observed effect

```
condition (a-b > l)
```

Hidden cause

```
return (a > b)
```



Probabilistic programming

Alice beat Bob at a game. Is she better than him at it?

Generative story

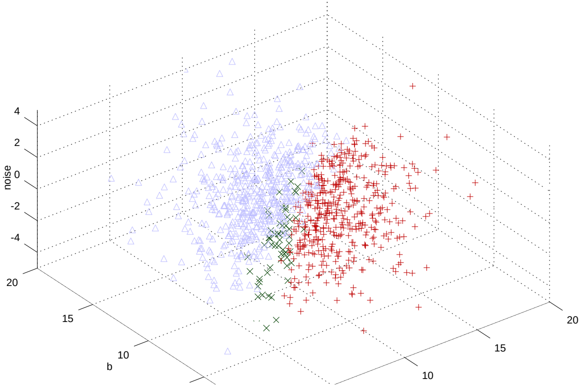
```
a <- normal 10 3
b <- normal 10 3
l <- normal  0 2
```

Observed effect

```
condition (a-b > l)
```

Hidden cause

```
return (a > b)
```



Denoted measure:

$$\lambda c.\ \int_{N(10,3)} da \int_{N(10,3)} db \int_{N(0,2)} dl\ \langle a - b > l \rangle\ c(a > b)$$

Importance sampling

Alice beat Bob at a game. Is she better than him at it?

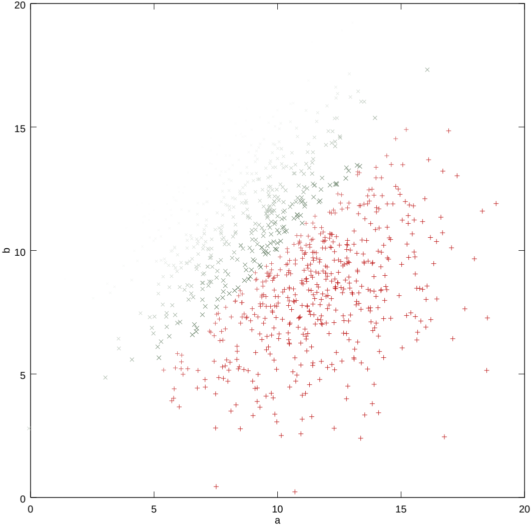
Generative story

```
a <- normal 10 3
b <- normal 10 3
```

weight $(1 + \mathrm{erf}(\frac{a-b}{2\sqrt{2}}))/2$

Hidden cause

```
return (a > b)
```

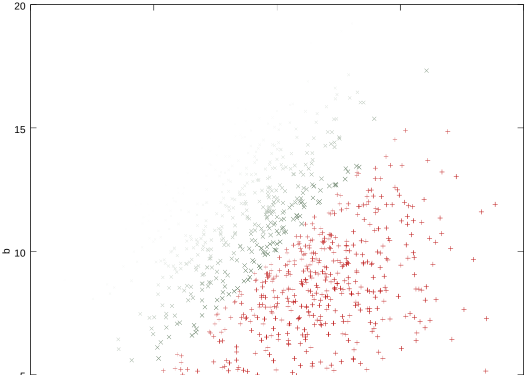


Alice beat Bob at a game. Is she better than him at it?

Generative story

```
a <- normal 10 3
b <- normal 10 3
```

weight $(1+\mathrm{erf}(\frac{a-b}{2\sqrt{2}}))/2$



Hidden cause

```
return (a > b)
```

Denoted measure:

$$\lambda c.\ \int_{N(10,3)} da \int_{N(10,3)} db\ \frac{1+\mathrm{erf}\left(\frac{a-b}{2\sqrt{2}}\right)}{2}\ c(a>b)$$

Conditioning on a real number

The bike was moving near location 9. Where's it now?

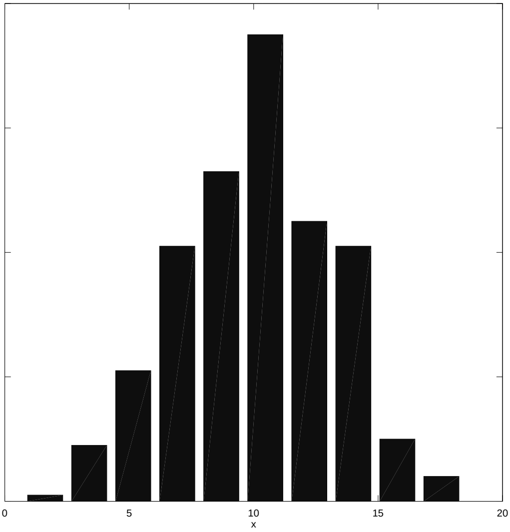
Generative story

Conditioning on a real number

The bike was moving near location 9. Where's it now?

Generative story

```
x  <- normal    10 3
```

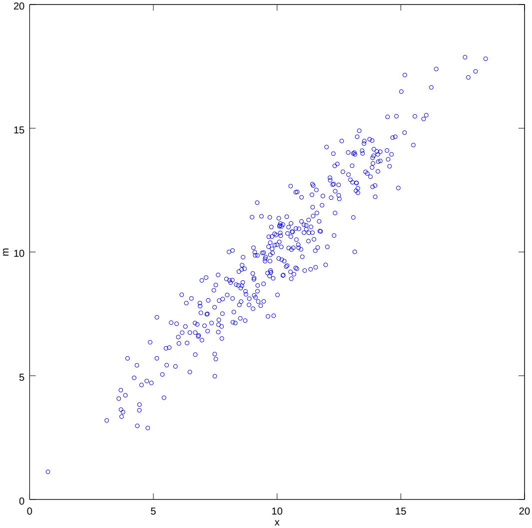


Conditioning on a real number

The bike was moving near location 9. Where's it now?

Generative story

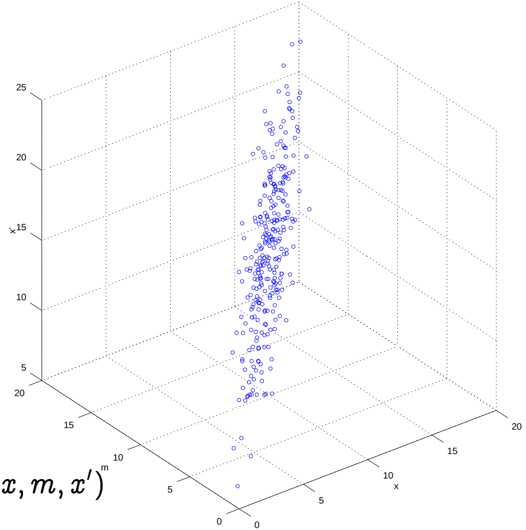
```
x  <- normal    10 3
m  <- normal     x 1
```



The bike was moving near location 9. Where's it now?

Generative story

```
x   <- normal    10 3
m   <- normal     x 1
x’ <- normal (x+5) 2
```



Denoted measure:

$$\lambda c. \int_{N(10,3)} dx \int_{N(x,1)} dm \int_{N(x+5,2)} dx'\, c(x, m, x')$$

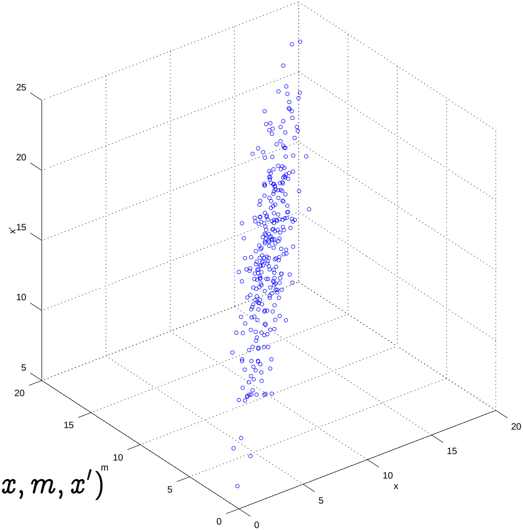
The bike was moving near location 9. Where's it now?

Generative story

```
x  <- normal    10 3
m  <- normal     x 1
x' <- normal (x+5) 2
```

Observed effect

```
condition (m = 9)
```



Denoted measure:

$$\lambda c. \int_{N(10,3)} dx \int_{N(x,1)} dm \int_{N(x+5,2)} dx'\, c(x, m, x')$$

Conditioning on a real number

The bike was moving near location 9. Where's it now?

Generative story

~~`x  <- normal    10 3`~~ `m <- normal 10` $\sqrt{10}$

~~`m  <- normal     x 1`~~ `x <- normal` $\left(\frac{9}{10}\mathtt{m} + \frac{1}{10}10\right)\ \sqrt{\frac{9}{10}}$

`x' <- normal (x+5) 2`

Observed effect

```
condition (m = 9)
```

Denoted measure:

$$\lambda c. \int_{N(10,\sqrt{10})} dm \int_{N(\frac{9}{10}m+\frac{1}{10}10,\sqrt{\frac{9}{10}})} dx \int_{N(x+5,2)} dx'\, c(x,m,x')$$

Conditioning on a real number

The bike was moving near location 9. Where's it now?

Generative story

~~x <- normal 10 3~~ ~~m <- normal 10 $\sqrt{10}$~~ let m = 9

~~m <- normal x 1~~ x <- normal $(\frac{9}{10}\mathtt{m} + \frac{1}{10}10)$ $\sqrt{\frac{9}{10}}$

x' <- normal (x+5) 2

Observed effect

~~condition (m = 9)~~

Denoted conditional measure:

$$\lambda m.\ \lambda c. \int_{N(\frac{9}{10}m+\frac{1}{10}10,\sqrt{\frac{9}{10}})} dx \int_{N(x+5,2)} dx'\ c(x,m,x')$$

Conditioning on a real number

The bike was moving near location 9. Where's it now?

Generative story

$$
\begin{array}{l}
\texttt{x \ \ <- normal } \tfrac{91}{10}\ \sqrt{\tfrac{9}{10}} \\
\texttt{x' <- normal (x+5) 2}
\end{array}
$$

Denoted marginal measure:

$$\lambda c. \int_{N(\frac{91}{10},\sqrt{\frac{9}{10}})} dx \int_{N(x+5,2)} dx'\, c(x,x')$$

Conditioning on a real number

The bike was moving near location 9. Where's it now?

Generative story

$$\texttt{x}\;\;\texttt{<- normal}\;\tfrac{91}{10}\;\sqrt{\tfrac{9}{10}}$$

```
x' <- normal (x+5) 2
```

Hidden cause

```
return x'
```

Denoted marginal measure:

$$\lambda c.\int_{N(\frac{91}{10},\sqrt{\frac{9}{10}})} dx \int_{N(x+5,2)} dx'\; c(x')$$

Conditioning on a real number

The bike was moving near location 9. Where's it now?

Generative story

```
x' <- normal
```
$\frac{141}{10}$ $\sqrt{\frac{49}{10}}$

Hidden cause

```
return x'
```

Denoted marginal measure:

$$\lambda c. \int_{N(\frac{141}{10}, \sqrt{\frac{49}{10}})} dx'\, c(x')$$

Importance sampling

Each sample has an *importance weight*

Importance sampling

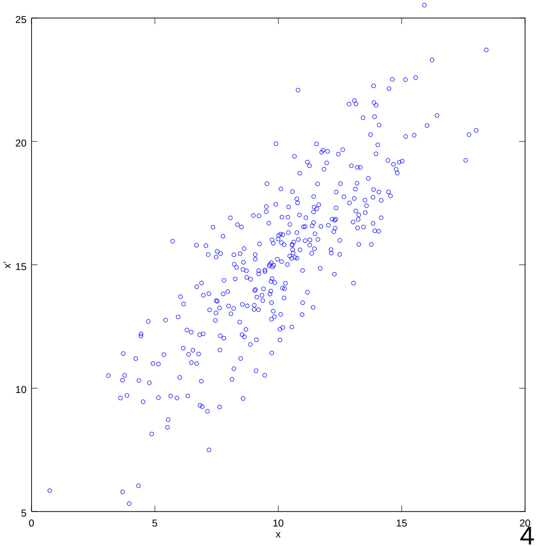
Each sample has an *importance weight*

Generative story

```
x  <- normal     10 3
m  <- normal      x 1
x' <- normal (x+5) 2
```

Observed effect

```
condition (m = 9)
```



Importance sampling

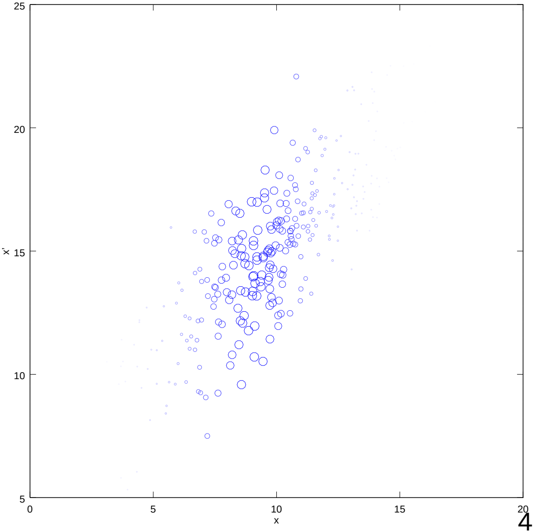
Each sample has an *importance weight*:
How much did we rig our random choices to avoid rejection?

Generative story

```
x  <- normal     10 3
weight e^{-(9-x)^2/2}/√(2π)
x' <- normal (x+5) 2
```

Denoted measure:

$$\lambda c. \int_{N(10,3)} dx \, \frac{e^{-(9-x)^2/2}}{\sqrt{2\pi}} \int_{N(x+5,2)} dx' \, c(x,m,x')$$



Probability or density?

Generative story

```
x  <-  flip True False
m  <-  (if x then flip 0 1
              else normal 0 1)
```

Observed effect

```
condition (m = 1)
```

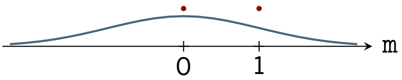
Hidden cause

```
return x
```

Probability or density?

Generative story

```
x  <-  flip True False
weight (if x then 1/2
             else e^{-1/2}/√(2π))
```



Hidden cause

```
return x
```

Ambient measures

Let μ be a measure over pairs (x, m).

```
μ = ...  <-  ...
    ...  <-  ...
    ...
    return (x,m)
```

$$= \lambda c. \iint \cdots\ c(x, m)$$

Condition μ on m = Express μ by choosing m first!
(easy when m is drawn from a fixed measure)

```
μ = m    <-  ...
    ...  <-  ...
    ...
    return (x,m)
```

$$= \lambda c. \int dm \int \cdots\ c(x, m)$$

It doesn't matter what the ambient measure τ is,
but there must be *one*, such as Lebesgue.

Ambient measures

Let μ be a measure over pairs (x, m).

```
μ = ... <- ...
    ... <- ...
    ...
    return (x,m)
```

$$= \lambda c. \int\!\!\int \cdots\; c(x, m)$$

Condition μ on m = Express μ by choosing m first!
(easy when m is drawn from a fixed measure)

```
μ = m   <- ...
    ... <- ...
    ...
    return (x,m)
```

$$= \lambda c. \int dm \int \cdots\; c(x, m)$$

It doesn't matter what the ambient measure τ is,
but there must be *one*, such as Lebesgue.

Ambient measures

Let μ be a measure over pairs (x, m).

$\mu =$

```
... <- ...
... <- ...
...
return (x,m)
```

$$= \lambda c. \int\!\!\int \cdots\ c(x, m)$$

Condition μ on m = Express μ by choosing m first!
(easy when m is drawn from a fixed measure)

$\nu(m) =$ ~~m <- τ~~

```
... <- ...
...
return (x,m)
```

$$= \lambda c. \cancel{\int_\tau dm} \int \cdots\ c(x, m)$$

It doesn't matter what the ambient measure τ is,
but there must be *one*, such as Lebesgue.

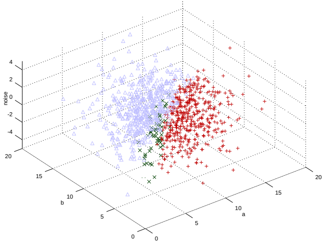
Discrete importance sampling revisited

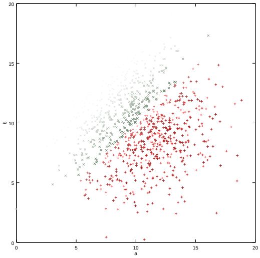
```
a <- normal 10 3
b <- normal 10 3
l <- normal 0 2
w <- dirac (a-b > l)

return (a > b)
```

=

```
a <- normal 10 3
b <- normal 10 3
l <- normal 0 2
w <- counting
weight (if a-b > l then 1
                    else 0)
return (a > b)
```





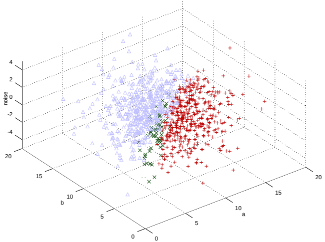
probability density (absolute continuity) w.r.t. `counting`

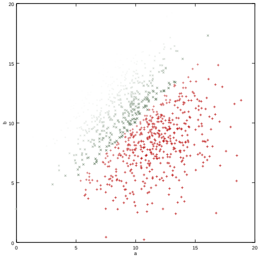
Discrete importance sampling revisited

```
a <- normal 10 3
b <- normal 10 3
l <- normal 0 2
w <- dirac (a-b > l)

return (a > b)
```

```
a <- normal 10 3
b <- normal 10 3
l <- normal 0 2
w <- counting
weight (if a-b > l then 1
                    else 0)
return (a > b)
```





probability density (absolute continuity) w.r.t. `counting`

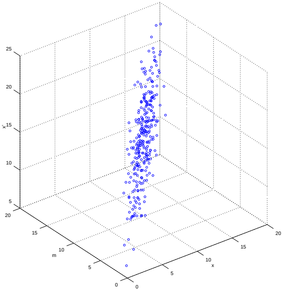
Continuous importance sampling revisited

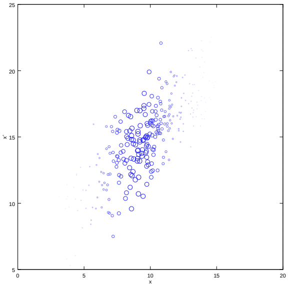
```
x   <- normal     10 3
m   <- normal      x 1

x' <- normal (x+5) 2
...
```

=

```
x   <- normal     10 3
m   <- lebesgue
```
weight $e^{-(m-x)^2/2} / \sqrt{2\pi}$
```
x' <- normal (x+5) 2
...
```





probability density (absolute continuity) w.r.t. `lebesgue`

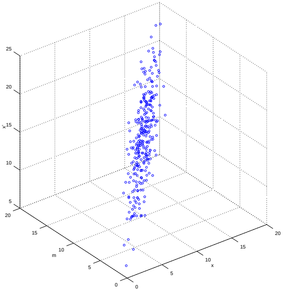
Continuous importance sampling revisited

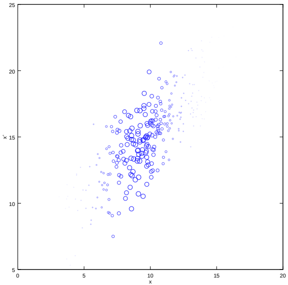
```
x  <- normal    10 3
m  <- normal     x 1

x' <- normal (x+5) 2
...
```

$=$

```
x  <- normal    10 3
m  <- lebesgue
```
weight $e^{-(m-x)^2/2}/\sqrt{2\pi}$
```
x' <- normal (x+5) 2
...
```



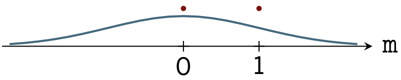


probability density (absolute continuity) w.r.t. `lebesgue`

Discrete+continuous importance sampling revisited

```
x <- flip True False
m <- (if x then flip 0 1 else normal 0 1)
...
```

```
x <- flip True False
m <- (x' <- flip True False
      if x' then flip 0 1 else normal 0 1)
weight (if x then if m==0 || m==1 then 2 else 0
             else if m==0 || m==1 then 0 else 2)
...
```

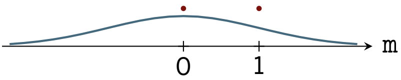


probability density (absolute continuity) w.r.t. marginal

Discrete+continuous importance sampling revisited

```
x <- flip True False
m <- (if x then flip 0 1 else normal 0 1)
...
```

```
x <- flip True False
m <- (x' <- flip True False
      if x' then flip 0 1 else normal 0 1)
weight (if x then if m==0 || m==1 then 2 else 0
             else if m==0 || m==1 then 0 else 2)
...
```



probability density (absolute continuity) w.r.t. marginal

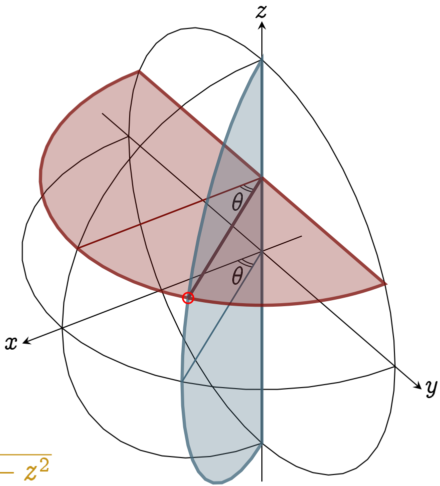
Borel's paradox

Requires change of variable:

z <- uniform -1 1
θ <- uniform $-\pi/2$ $\pi/2$
y <- dirac $\sqrt{1-z^2}\sin\theta$
...

$$\|$$

z <- uniform -1 1
y <- uniform $-\sqrt{1-z^2}$ $\sqrt{1-z^2}$
θ <- dirac $\arcsin(y/\sqrt{1-z^2})$
weight $2/\cos\theta$
...



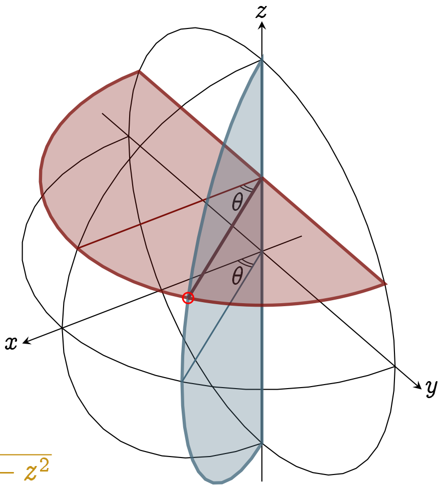
Borel's paradox

Requires change of variable:

z `<-` `uniform` -1 1
θ `<-` `uniform` $-\pi/2$ $\pi/2$
y `<-` `dirac` $\sqrt{1-z^2}\sin\theta$
...

$\|$

z `<-` `uniform` -1 1
y `<-` `uniform` $-\sqrt{1-z^2}$ $\sqrt{1-z^2}$
θ `<-` `dirac` $\arcsin(y/\sqrt{1-z^2})$
`weight` $2/\cos\theta$
...



Another desideratum

Generative story

```
x <- normal 0 1
m <- dirac (x * 2)
```

Observed effect

```
condition (m = 32)
```

Hidden cause

```
return (x == 16)
```

Should give `True` with 100% probability.

Another desideratum

Generative story

```
x <- normal 0 1
m <- dirac (x / 2)
```

Observed effect

```
condition (m = 32)
```

Hidden cause

```
return (x == 64)
```

Should give `True` with 100% probability.

Interim summary

Conditioning means disintegration.

(Chang & Pollard)

If the conditioned choice has a density with respect to an ambient measure,

then we can transform the program locally into importance sampling.

- Simple ambient measures: `counting`, `lebesgue`
- Complex ambient measures: marginal, …

How to infer an ambient measure?
How to calculate density with respect to it?

(Pfeffer, Bhat et al.)

Interim summary

Conditioning means disintegration.
(Chang & Pollard)

If the conditioned choice has a density
with respect to an ambient measure,
then we can transform the program locally
into importance sampling.

- Simple ambient measures:
`counting`, `lebesgue`
- Complex ambient measures:
marginal, …

How to infer an ambient measure?
How to calculate density with respect to it?
(Pfeffer, Bhat et al.)

Density calculator demo

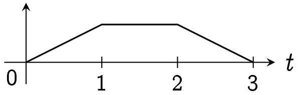
```
x <- uniform 0 1
x + x + uniform 0 1
```

$$\Downarrow$$

$$\int_0^1 \langle 0 < t - x - x < 1 \rangle \, dx$$

$$\Downarrow$$

$$\begin{cases} 0 & \text{if } t < 0 \\ t/2 & \text{if } t < 1 \\ 1/2 & \text{if } t < 2 \\ 3/2 - t/2 & \text{if } t < 3 \\ 0 & \text{if } 3 \le t \end{cases}$$



Density calculator demo

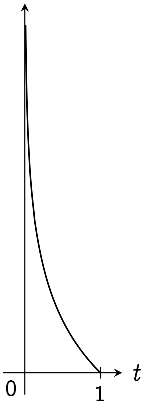
```
exp (log (uniform 0 1) + log (uniform 0 1))
```

$$\Downarrow$$

$$\begin{cases} \displaystyle\int_0^1 \langle 0 < e^{\ln t - \ln x} < 1 \rangle \, e^{\ln t - \ln x} \, dx \Big/ t & \text{if } 0 < t \\ 0 & \text{otherwise} \end{cases}$$

$$\Downarrow$$

$$\begin{cases} -\ln t & \text{if } 0 < t < 1 \\ 0 & \text{otherwise} \end{cases}$$



Density calculator demo

```
x <- uniform 0 1
(if x < .2 then 0 else uniform 0 1)
  + (if x > .7 then 1 else uniform 0 1)
```

Pfeffer's approximate algorithm randomly samples from one term then calculates density of the other term.

Mochastic vs stochastic lambda calculus

Mochastic/stochastic only	Stochastic
`x <- normal 0 1` `return (exp x + 1)`	`exp (normal 0 1) + 1`

`x <- normal 0 1`
`y <- normal 0 1`
`return (x + y)`
	`normal 0 1 + normal 0 1`

`x <- flip True False`
`if x`
`  then flip 0 1`
`  else normal 0 1`
	`if flip True False`
	`  then flip 0 1`
	`  else normal 0 1`

`x <- normal 0 1` `return (x + x)`	none

Stochastic lambda calculus: expectation

$$\llbracket \texttt{uniform 0 1} \rrbracket\ \rho\ c = \int_0^1 c(t)\, dt$$

$$\llbracket \texttt{exp}\ \varepsilon \rrbracket\ \rho\ c = \llbracket \varepsilon \rrbracket\ \rho\ (\lambda x.\, c(e^x))$$

Stochastic lambda calculus: density

$$[\![\texttt{uniform 0 1}]\!]\ \rho\ c = \int_0^1 c(t)\,dt$$

$$[\![\texttt{exp}\ \varepsilon]\!]\ \rho\ c = [\![\varepsilon]\!]\ \rho\ (\lambda x.\, c(e^x))$$

Goal: Given ε, find $D(\varepsilon)$ such that

$$[\![\varepsilon]\!]\ \rho\ c = \int D(\varepsilon)\ \rho\ t \times c(t)\,dt \qquad \text{for all } \rho, c.$$

Approach: Define $D(\varepsilon)$ by recursion on ε.

Stochastic lambda calculus: density

$$[\![\texttt{uniform 0 1}]\!]\ \rho\ c = \int_0^1 c(t)\,dt$$

$$[\![\texttt{exp}\ \varepsilon]\!]\ \rho\ c = [\![\varepsilon]\!]\ \rho\ (\lambda x.\, c(e^x))$$

Goal: Given ε, find $D(\varepsilon)$ such that

$$[\![\varepsilon]\!]\ \rho\ c = \int D(\varepsilon)\ \rho\ t \times c(t)\,dt \qquad \text{for all } \rho, c.$$

One base case:

$$D(\texttt{uniform 0 1})\ \rho\ t = \langle 0 < t < 1 \rangle$$

Stochastic lambda calculus: density

$$[\![\texttt{uniform 0 1}]\!]\ \rho\ c = \int_0^1 c(t)\, dt$$

$$[\![\mathrm{exp}\ \varepsilon]\!]\ \rho\ c = [\![\varepsilon]\!]\ \rho\ (\lambda x.\, c(e^x))$$

Goal: Given ε, find $D(\varepsilon)$ such that

$$[\![\varepsilon]\!]\ \rho\ c = \int D(\varepsilon)\ \rho\ t \times c(t)\, dt \qquad \text{for all } \rho, c.$$

One recursive case:

$$D(\mathrm{exp}\ \varepsilon)\ \rho\ t = ???$$

$$[\![\mathrm{exp}\ \varepsilon]\!]\ \rho\ c = \int_{-\infty}^{\infty} ??? \times c(t)\, dt$$

Stochastic lambda calculus: density

$$[\![\texttt{uniform 0 1}]\!]\ \rho\ c = \int_0^1 c(t)\,dt$$

$$[\![\mathrm{exp}\ \varepsilon]\!]\ \rho\ c = [\![\varepsilon]\!]\ \rho\ (\lambda x.\, c(e^x))$$

Goal: Given ε, find $D(\varepsilon)$ such that

$$[\![\varepsilon]\!]\ \rho\ c = \int D(\varepsilon)\ \rho\ t \times c(t)\,dt \qquad \text{for all } \rho, c.$$

One recursive case:

$$D(\mathrm{exp}\ \varepsilon)\ \rho\ t = {???}$$

$$\begin{array}{c}[\![\mathrm{exp}\ \varepsilon]\!]\ \rho\ c = \displaystyle\int_{-\infty}^{\infty} {???} \times c(t)\,dt \\ \| \\ [\![\varepsilon]\!]\ \rho\ (\lambda x.\, c(e^x)) = \displaystyle\int_{-\infty}^{\infty} D(\varepsilon)\ \rho\ x \times c(e^x)\,dx\end{array}$$

Stochastic lambda calculus: density

$$[\![\texttt{uniform 0 1}]\!]\ \rho\ c = \int_0^1 c(t)\,dt$$

$$[\![\texttt{exp}\ \varepsilon]\!]\ \rho\ c = [\![\varepsilon]\!]\ \rho\ (\lambda x.\, c(e^x))$$

Goal: Given ε, find $D(\varepsilon)$ such that

$$[\![\varepsilon]\!]\ \rho\ c = \int D(\varepsilon)\ \rho\ t \times c(t)\,dt \qquad \text{for all } \rho, c.$$

One recursive case:

$$D(\texttt{exp}\ \varepsilon)\ \rho\ t = \langle t > 0\rangle \frac{D(\varepsilon)\ \rho\ (\ln t)}{t}$$

$$\begin{array}{ccc} [\![\texttt{exp}\ \varepsilon]\!]\ \rho\ c & = & \displaystyle\int_{-\infty}^{\infty} \langle t > 0\rangle \frac{D(\varepsilon)\ \rho\ (\ln t)}{t} \times c(t)\,dt \\ \| & & \| \\ [\![\varepsilon]\!]\ \rho\ (\lambda x.\, c(e^x)) & = & \displaystyle\int_{-\infty}^{\infty} D(\varepsilon)\ \rho\ x \times c(e^x)\,dx \end{array}$$

Stochastic lambda calculus: density

$$
\begin{aligned}
[\![\varepsilon_1 + \varepsilon_2]\!]\ \rho\ c &= [\![\varepsilon_1]\!]\ \rho\ (\lambda x_1.\ [\![\varepsilon_2]\!]\ \rho\ (\lambda x_2.\ c(x_1 + x_2))) \\
&= [\![\varepsilon_2]\!]\ \rho\ (\lambda x_2.\ [\![\varepsilon_1]\!]\ \rho\ (\lambda x_1.\ c(x_1 + x_2)))
\end{aligned}
$$

Randomized algorithm for binary operators

$$
\begin{aligned}
D(\varepsilon_1 + \varepsilon_2)\ \rho\ t &= [\![\varepsilon_1]\!]\ \rho\ (\lambda x.\ D(\varepsilon_2)\ \rho\ (t - x)) \\
&= [\![\varepsilon_2]\!]\ \rho\ (\lambda x.\ D(\varepsilon_1)\ \rho\ (t - x))
\end{aligned}
$$

Stochastic lambda calculus: density

$$[\![42]\!]\ \rho\ c = c(42)$$

$$[\![\mathtt{var}]\!]\ \rho\ c = c(\rho\ \mathtt{var})$$

$$[\![\mathtt{var}\ \texttt{<-}\ \varepsilon_1\ \mathtt{;}\ \ \varepsilon_2]\!]\ \rho\ c = [\![\varepsilon_1]\!]\ \rho\ (\lambda x.\ [\![\varepsilon_2]\!]\ (\rho\{\mathtt{var} \mapsto x\})\ c)$$

Three strategies for bound variables

$$D(\mathtt{bool\_var})\ \rho\ t = \langle \rho\ \mathtt{bool\_var} = t \rangle$$

$$D(\mathtt{var}\ \texttt{<-}\ \varepsilon_1\ \mathtt{;}\ \ \varepsilon_2)\ \rho\ t = D(\varepsilon_2\{\mathtt{var} \mapsto \varepsilon_1\})\ \rho\ t$$

if ε_2 uses `var` at most once

$$D(\mathtt{var}\ \texttt{<-}\ \varepsilon_1\ \mathtt{;}\ \ \varepsilon_2)\ \rho\ t = [\![\varepsilon_1]\!]\ \rho\ (\lambda x.\ D(\varepsilon_2)\ (\rho\{\mathtt{var} \mapsto x\})\ (t))$$

Summary

Probabilistic programming

- Denote measure by generative story
- Run backwards to infer cause from effect

Mathematical reasoning

- Define conditioning as disintegration
- Perform importance sampling
- Derive density calculator

Need semantics and inference for loops

- Bag of words
- Brownian motion
- Probabilistic context-free grammars