4F13: Machine Learning

http://learning.eng.cam.ac.uk/zoubin/ml06/

Department of Engineering, University of Cambridge
Michaelmas 2006

Lecture 8

Statistical sampling and Monte Carlo

Iain Murray

i.murray+ta@gatsby.ucl.ac.uk

Review: probabilistic modelling

Data $\mathcal{D}$, model $\mathcal{M}$; what do we know about x?

Bayesian prediction with unknown parameters θ:

$$P(x|\mathcal{D},\mathcal{M}) = \int P(x,\theta|\mathcal{D},\mathcal{M})\,\mathrm{d}\theta \quad \textbf{Marginalization}$$

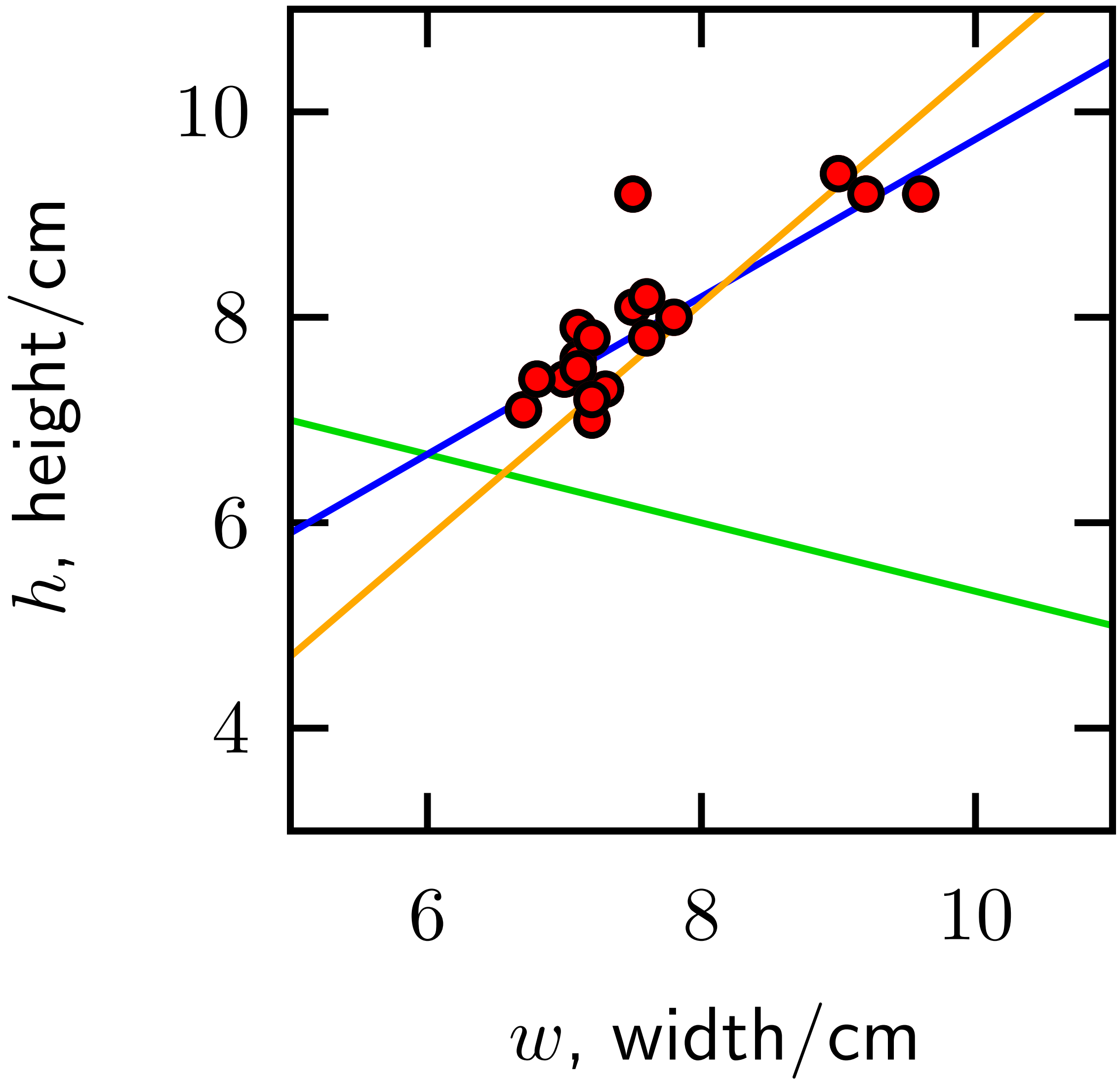
$$= \int P(x|\theta,\mathcal{D},\mathcal{M})\underbrace{P(\theta|\mathcal{D},\mathcal{M})}_{\text{from } \textbf{Bayes' rule}}\,\mathrm{d}\theta \quad \textbf{Product rule}$$

Also marginalize any hidden variables:

$$P(x|\mathcal{D},\mathcal{M}) = \int \sum_h P(x,h|\theta,\mathcal{D},\mathcal{M}) \sum_H P(\theta,H|\mathcal{D},\mathcal{M})\,\mathrm{d}\theta$$

$\ldots$

Regression example



Linear regression model

Three parameters:
$\theta = \{$slope m, shift c, noise $\sigma\}$

$$h = mw + c + \eta \sim \mathcal{N}(0, \sigma^2)$$

$$P(h|w, \mathcal{D}, \mathcal{M}) =$$
$$\int P(h|w, \theta, \mathcal{M}) P(\theta|\mathcal{D}, \mathcal{M})\, \mathrm{d}\theta$$

“Review” continued

The **EM algorithm** uses sums or integrals for its sufficient statistics:

$$E[s(h,v)|\theta] = \int s(h,v)P(h|v,\theta)\,\mathrm{d}h$$

Lecture 11 will be on **model comparison**:

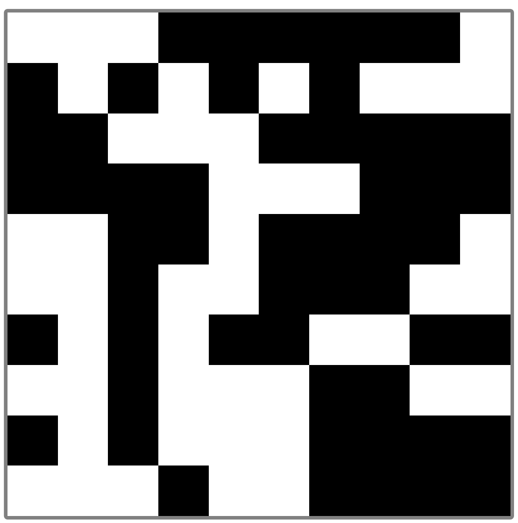
$$P(\mathcal{D}|\mathcal{M}) = \int P(\mathcal{D}|\theta,\mathcal{M})P(\theta|\mathcal{M})\,\mathrm{d}\theta$$

Inference only needs mechanical use of marginalization, the product rule and Bayes' rule. . . provided we can do all the sums and integrals

The trouble with sums

100 binary variables $x_i \in \{0,1\}$, could be:

– assignments of 100 data points in a mixture of 2 Gaussians
– a *tiny* patch of pixel labels in computer vision
– a *tiny* patch of idealized magnetic iron

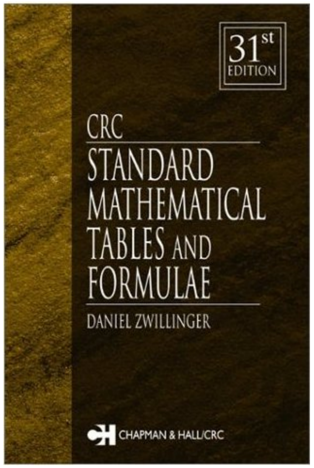


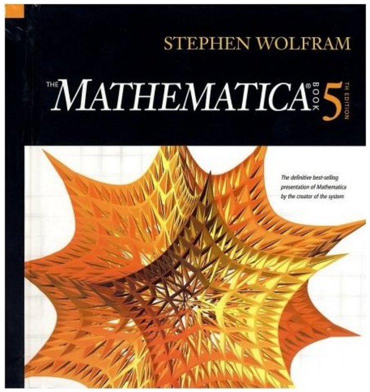
There are 2^{100} possible states

The age of the universe $\approx 2^{98}$ picoseconds

Sum might decompose (e.g. belief propagation)
. . . otherwise must approximate

The trouble with integrals





Only some integrals have analytic solutions

Numerical quadrature is feasible in low dimensions (1, 2 or 3)

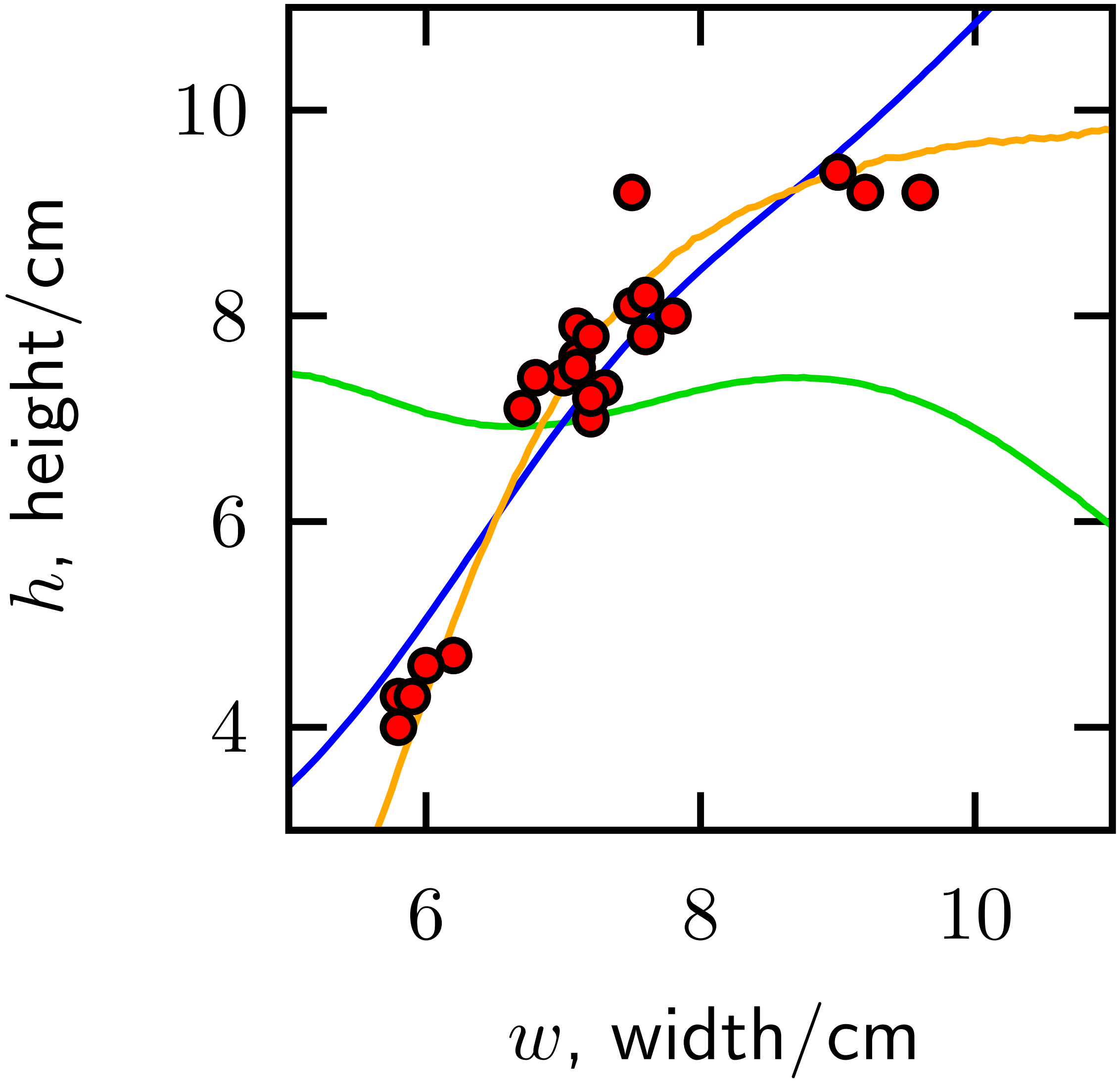
Multivariate integrals occasionally decompose into standard integrals:

– Linear Gaussian models
– Exponential family models with conjugate priors

Discretization or quadrature are infeasible in high dimensions:

– Only allow each variable to take on 2 settings
– There are 2^{100} possible joint settings. . .

Regression example (2)



Non-linear regression

Many parameters:
$\theta = \{\text{curve, noise}\}$

$$P(h|w, \mathcal{D}, \mathcal{M}) =$$
$$\int P(h|w, \theta, \mathcal{M}) P(\theta|\mathcal{D}, \mathcal{M})\, \mathrm{d}\theta$$

Statistical sampling

What is the average height h of people p in Cambridge $\mathcal{C}$?

$$E_{p\in\mathcal{C}}[h(p)] \equiv \frac{1}{|\mathcal{C}|}\sum_{p\in\mathcal{C}} h(p)$$

$$\approx \frac{1}{S}\sum_{s=1}^{S} h\big(p^{(s)}\big), \quad \text{for random survey of } S \text{ people } \{p^{(s)}\} \in \mathcal{C}$$

What is the distribution over unknown x?

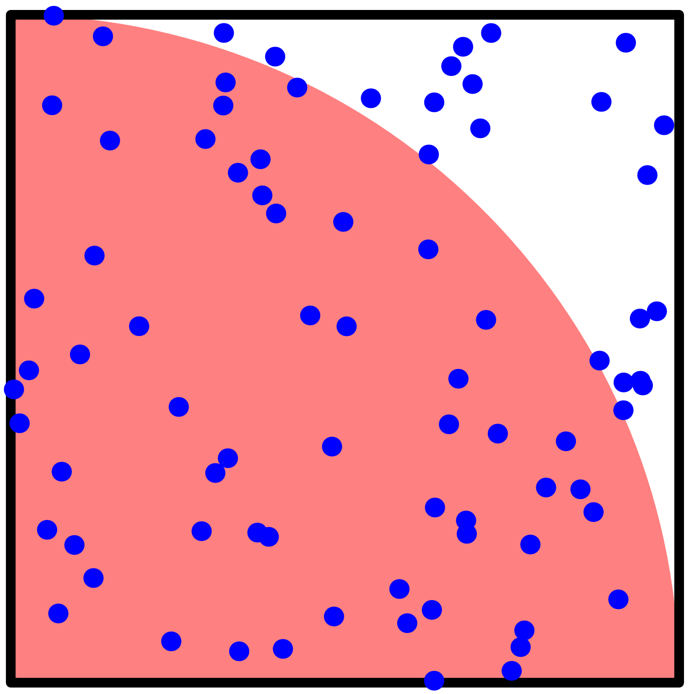
$$p(x|\mathcal{D}) = \int P(x|\theta,\mathcal{D}) P(\theta|\mathcal{D})\ \mathrm{d}\theta = E_{P(\theta|\mathcal{D})}[P(x|\theta,\mathcal{D})]$$

$$\approx \frac{1}{S}\sum_{s=1}^{S} P(x|\theta^{(s)},\mathcal{D}), \quad \theta^{(s)} \sim P(\theta|\mathcal{D})$$

This technique is also known as simple **Monte Carlo**

Estimates are unbiased, variance $\sim 1/S$ "independent of dimension"

A dumb approximation of π

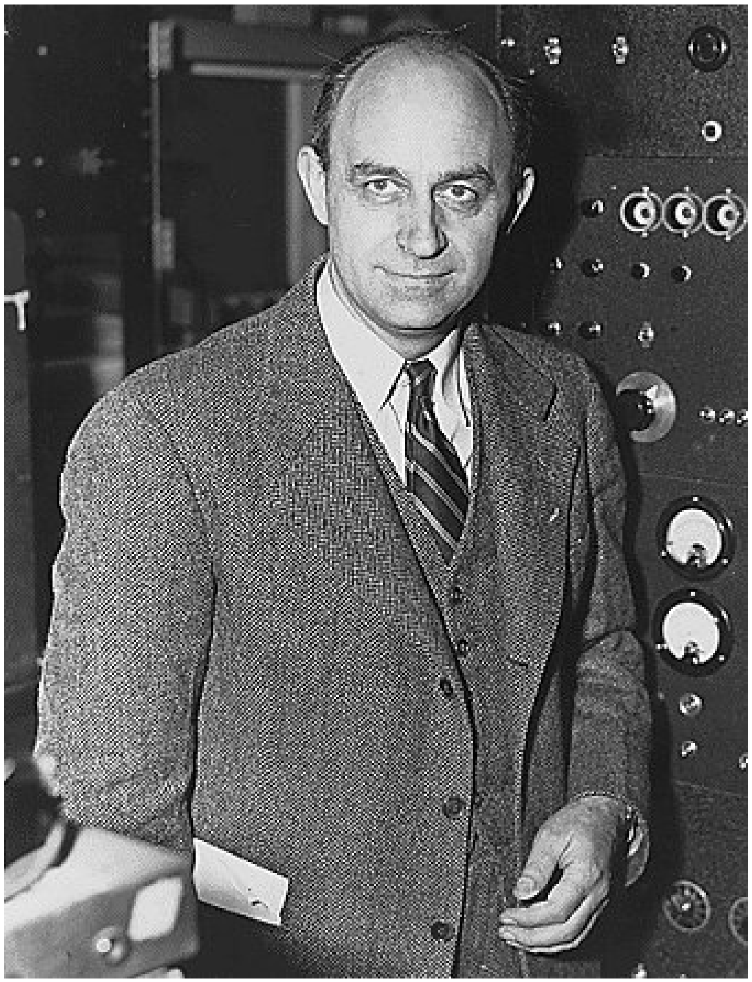


$$P(x,y) = \begin{cases} 1 & 0<x<1 \text{ and } 0<y<1 \\ 0 & \text{otherwise} \end{cases}$$

$$\pi = 4 \iint \mathbb{I}\left((x^2+y^2)<1\right) P(x,y)\,\mathrm{d}x\,\mathrm{d}y$$

```
octave:1> N=12; a=rand(N,2); 4*mean(sum(a.*a,2)<1)
ans = 3.3333
octave:2> N=1e7; a=rand(N,2); 4*mean(sum(a.*a,2)<1)
ans = 3.1418
```

Monte Carlo and Insomnia

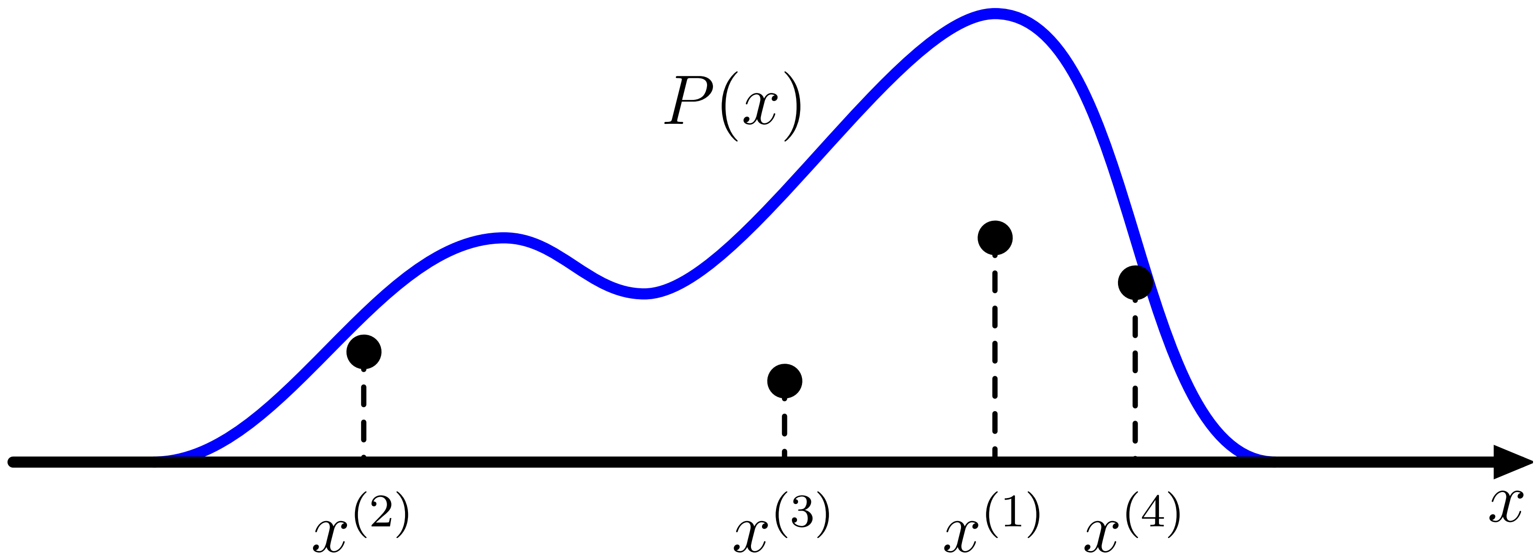


Enrico Fermi (1901–1954) took great delight in astonishing his colleagues with his remakably accurate predictions of experimental results. . . he revealed that his “guesses” were really derived from the statistical sampling techniques that he used to calculate with whenever insomnia struck in the wee morning hours!

—*The beginning of the Monte Carlo method*, N. Metropolis

Sampling from distributions

Draw points from the unit area under the curve



Draw probability mass to left of point, $u \sim$ Uniform[0,1]

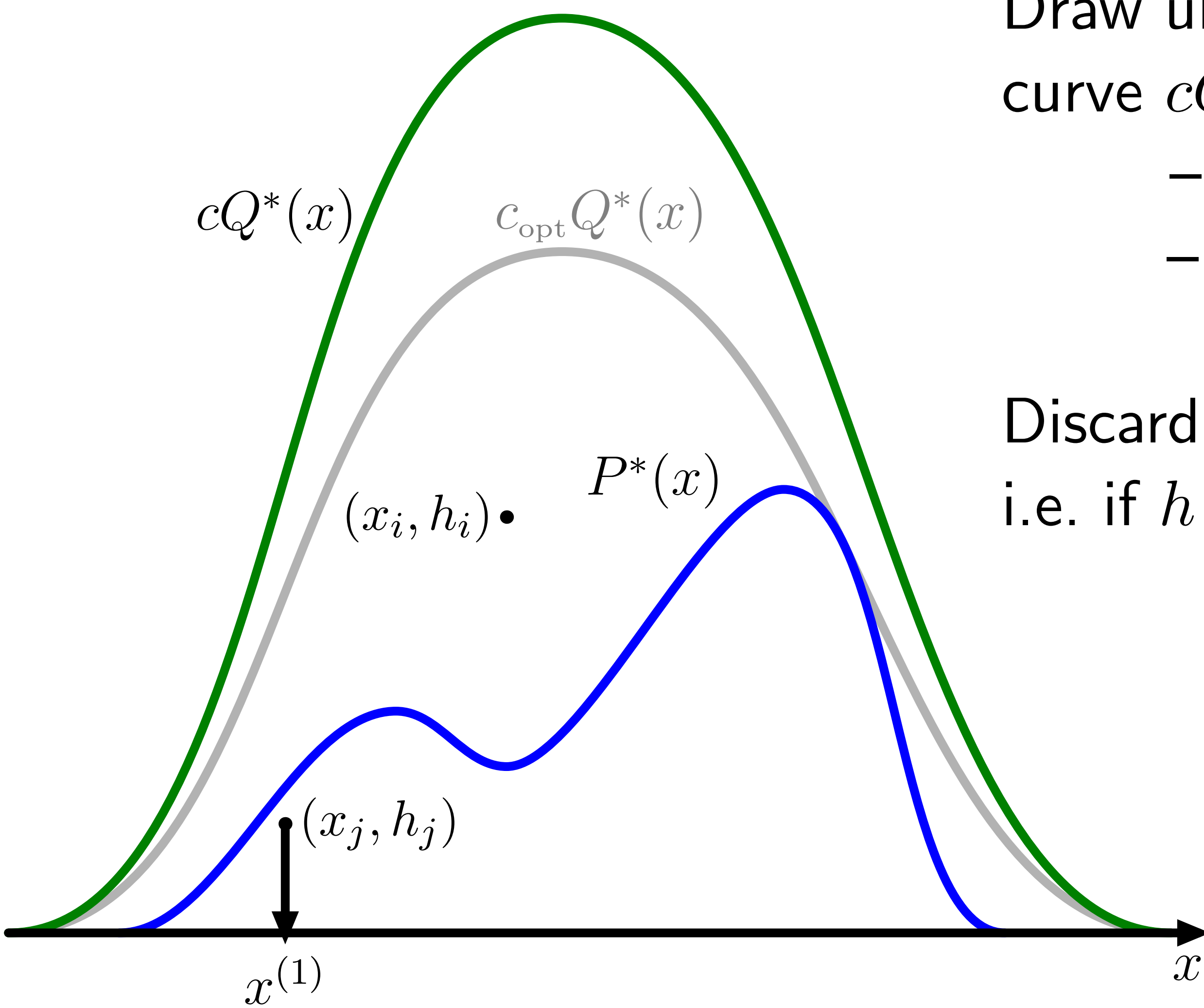
Sample $x(u) = c^{-1}(u)$, where $c(x) = \int_{-\infty}^{x} P(x')\, \mathrm{d}x'$

Problem: often can't even normalize P, e.g. $P(\theta|\mathcal{D}) \propto P(D|\theta)P(\theta)$

Will call unnormalized version P^* (as in MacKay's textbook).

Rejection sampling

Sampling underneath a $P^*(x) \propto P(x)$ curve is also valid



Draw underneath a simple curve $cQ^*(x) \geq P^*(x)$:

– Draw $x \sim Q(x)$

– height $h \sim \text{Uniform}[0, cQ^*(x)]$

Discard the point if above P^*, i.e. if $h > P^*(x)$

Importance sampling

Computing $P^*(x)$ and $Q^*(x)$, then *throwing* x *away* seems wasteful
Instead rewrite the integral as an expectation under Q:

$$\int f(x)P(x)\,\mathrm{d}x = \int f(x)\frac{P(x)}{Q(x)}Q(x)\,\mathrm{d}x, \qquad (Q(x)>0 \text{ if } P(x)>0)$$

$$\approx \frac{1}{S}\sum_{s=1}^{S} f(x^{(s)})\frac{P(x^{(s)})}{Q(x^{(s)})}, \quad x^{(s)} \sim Q(x)$$

Unbiased; but light-tailed $Q(x)$ can give the estimator infinite variance

. . . and you might not notice.

Importance sampling applies when the integral is not an expectation.

Importance sampling (2)

Previous slide assumed we could evaluate $P(x) = P^*(x)/\mathcal{Z}_P$

$$\int f(x)P(x)\,\mathrm{d}x \approx \frac{\mathcal{Z}_Q}{\mathcal{Z}_P}\,\frac{1}{S}\sum_{s=1}^{S} f(x^{(s)})\underbrace{\frac{P^*(x^{(s)})}{Q^*(x^{(s)})}}_{w^{(s)}}, \quad x^{(s)} \sim Q(x)$$

$$\approx \frac{\cancel{1}}{\cancel{S}}\sum_{s=1}^{S} f(x^{(s)})\frac{w^{(s)}}{\frac{\cancel{1}}{\cancel{S}}\sum_{s'} w^{(s')}}$$

This estimator is **consistent** but **biased**

Exercise: Prove that $\mathcal{Z}_P/\mathcal{Z}_Q \approx \frac{1}{S}\sum_s w^{(s)}$

Summary so far

- Sums and integrals, often expectations, occur frequently in statistics
- **Monte Carlo** approximates expectations with a sample average
- **Rejection sampling** draws samples from fiddly distributions
- **Importance sampling** applies Monte Carlo to any sum/integral
- If $Q(x)$ is a poor global fit:
  - rejection samplers almost always reject (large c needed)
  - importance sampling is dangerous (large or infinite variance)

In high dimensions finding a good $Q(x)$ is hard. What then?