Markov Chain Monte Carlo

Mark Gales

Lent 2019

Stochastic Processes: Handout 3

IIA Module 3M1: Mathematical Methods

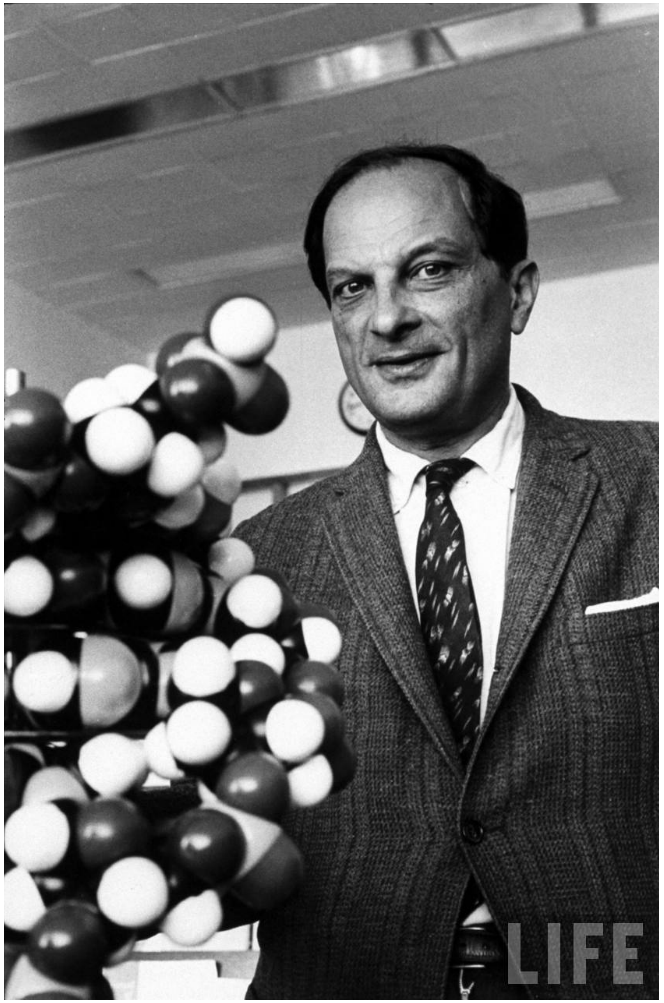
Overview

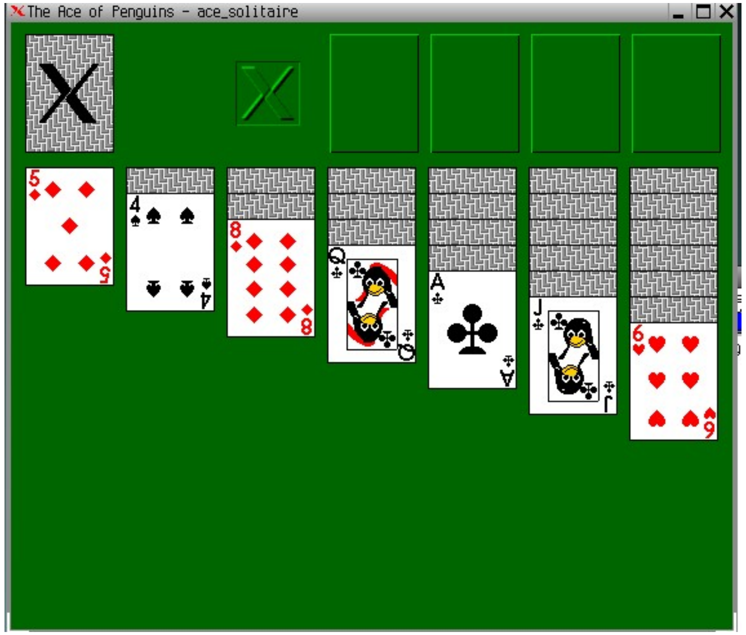
- Previous lectures examined discrete/continuous Markov chains
  - discrete and finite space Markov chains
  - stochastic processes (yielding partial differential equations)

- In this lecture we will look another use of Markov processes
  - schemes are all based on Monte Carlo methods
    basically we're going to use random samples!

In this lecture we will look at how to generate/use samples from distributions

Stanislaw Ulam





Monte-Carlo Methods

- General class of approaches
  - named after casino location
- Rely on random samples
  - often used when analytical approaches fail
  - estimate complicated integrals

$$\int h(\mathbf{x})d\mathbf{x}$$

  - generate random samples, $\mathbf{x}^{(i)}$
- Link with Markov chains
  - Markov Chain Monte Carlo (MCMC)

Numerical Integration

- Consider a highly complicated, multi-dimensional (d-dimensional), integration
  - use histogram approach (in d-dimensions!)

$$\int h(\mathbf{x})d\mathbf{x} \approx \sum_{i=1}^{N} h(\mathbf{x}^{(i)})\delta x_1 \delta x_2 \dots \delta x_d$$

  - uniformly spread N histogram centers $\mathbf{x}^{(i)}$ - highly inefficient
  - histograms in regions with very small (or no) contribution to the integral

- Can we improve on this basic histogram style approach?
  - range of approaches available in literature ...

Adopt a Monte-Carlo style approach

Monte-Carlo Integration

- Rather than histograms, sample from a weighting function $w(\mathbf{x})$
  - normalised form (note $w(\mathbf{x}) > 0$) is a valid PDF:

$$p(\mathbf{x}) = \frac{w(\mathbf{x})}{\int w(\mathbf{x})d\mathbf{x}}$$

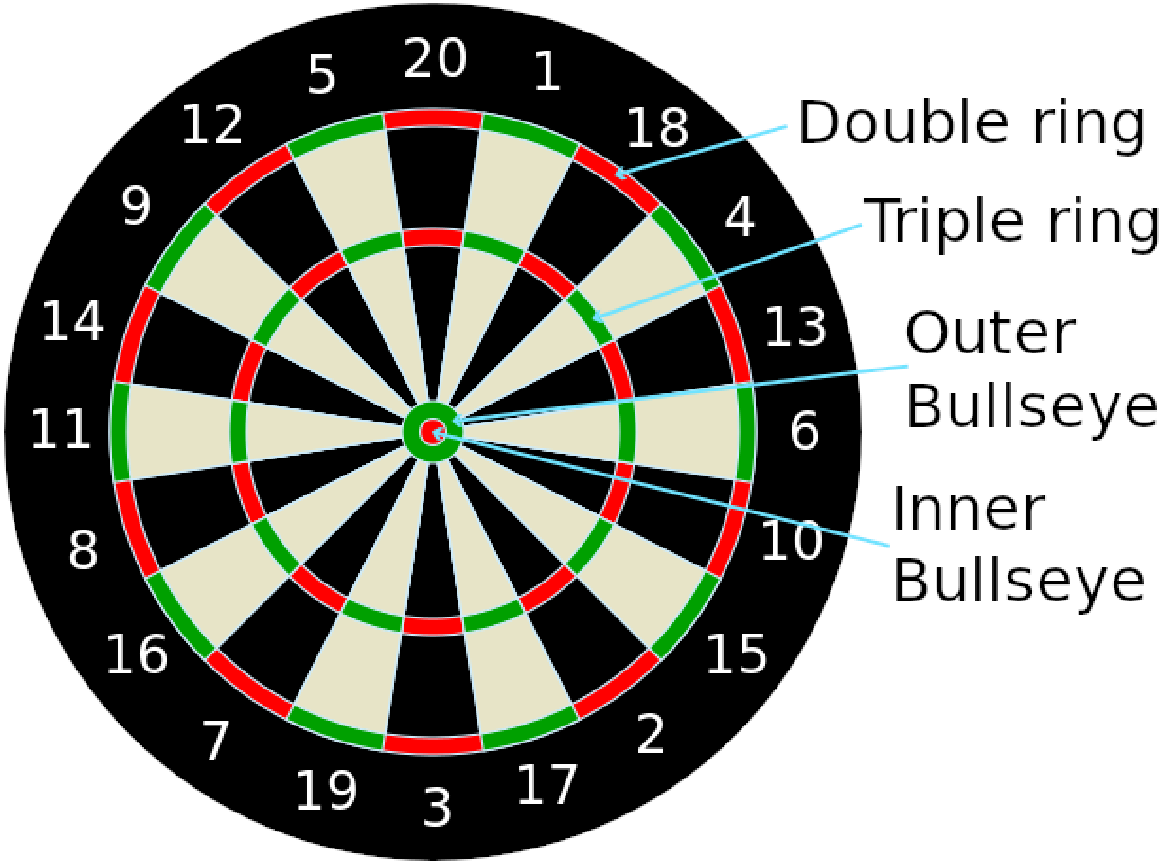
  - use this to model the distribution of the points
  - initially consider uniform distribution over "volume" V, $p(\mathbf{x}) = 1/V$

- Integration can then be rewritten as

$$\int h(\mathbf{x})d\mathbf{x} = \int \frac{h(\mathbf{x})}{p(\mathbf{x})}p(\mathbf{x})d\mathbf{x} \approx \frac{1}{N}\sum_{i=1}^{N}\frac{h(\mathbf{x}^{(i)})}{p(\mathbf{x}^{(i)})} = \frac{V}{N}\sum_{i=1}^{N}h(\mathbf{x}^{(i)})$$

  - samples, $\mathbf{x}^{(i)}$ drawn from $p(\mathbf{x})$ (or $w(\mathbf{x})$)
  - selecting appropriate "volume" for $p(\mathbf{x})$ will avoid "wasted" samples

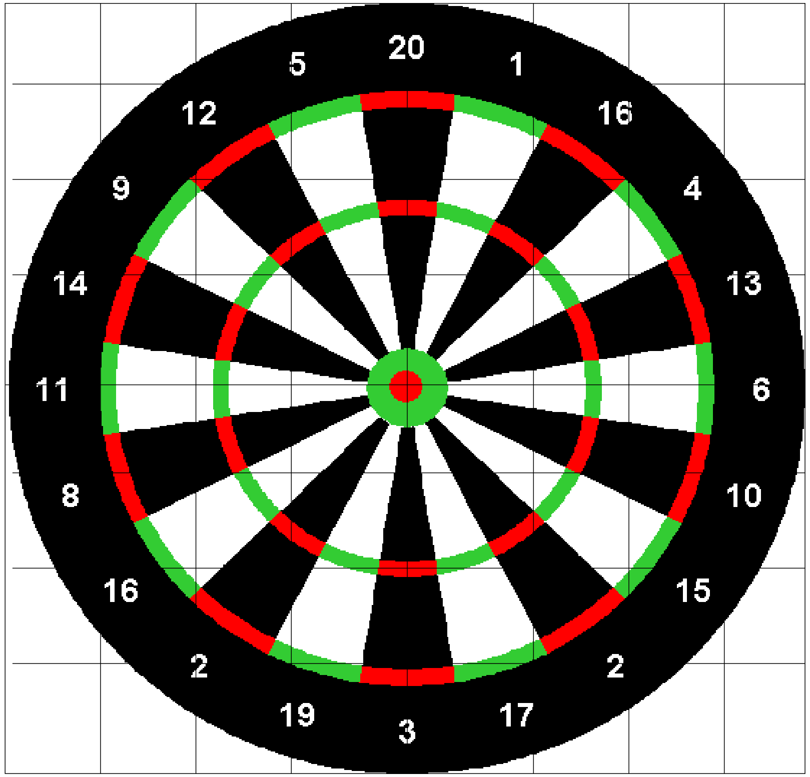
Dartboard

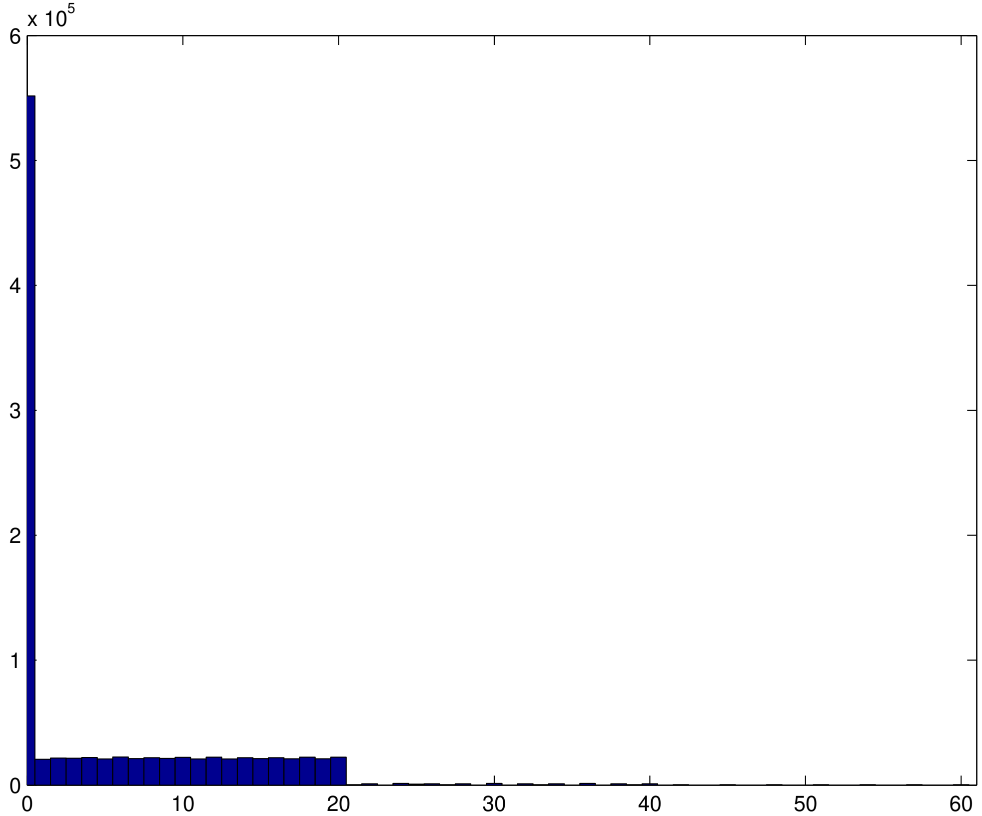


"Dartboard diagram" by Tijmen Stam

Interested in the "score" volume for a dartboard $(4.1138 s(core) m^2)$

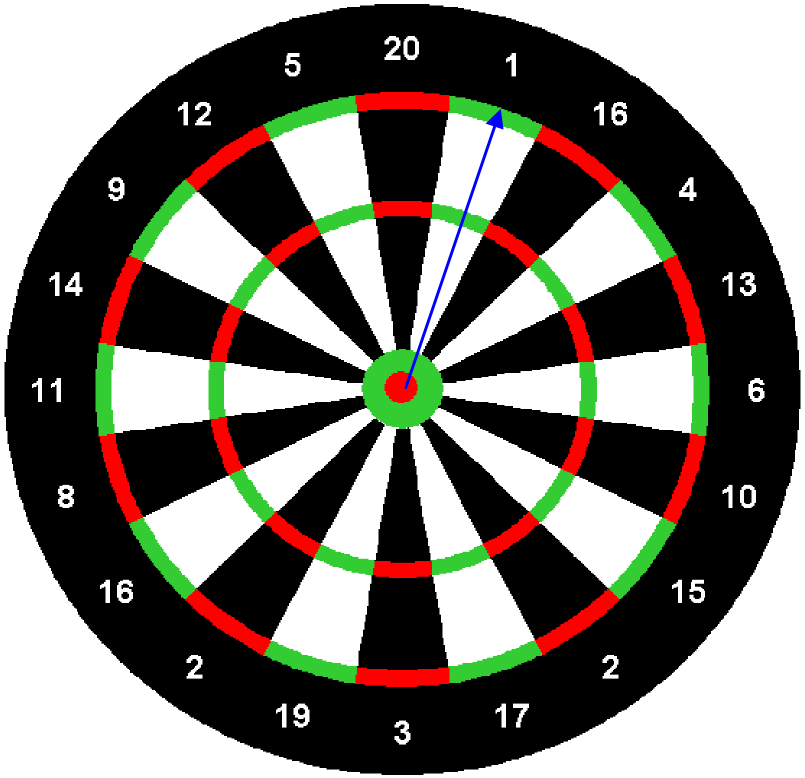
Dartboard - Uniform Samples (Square)

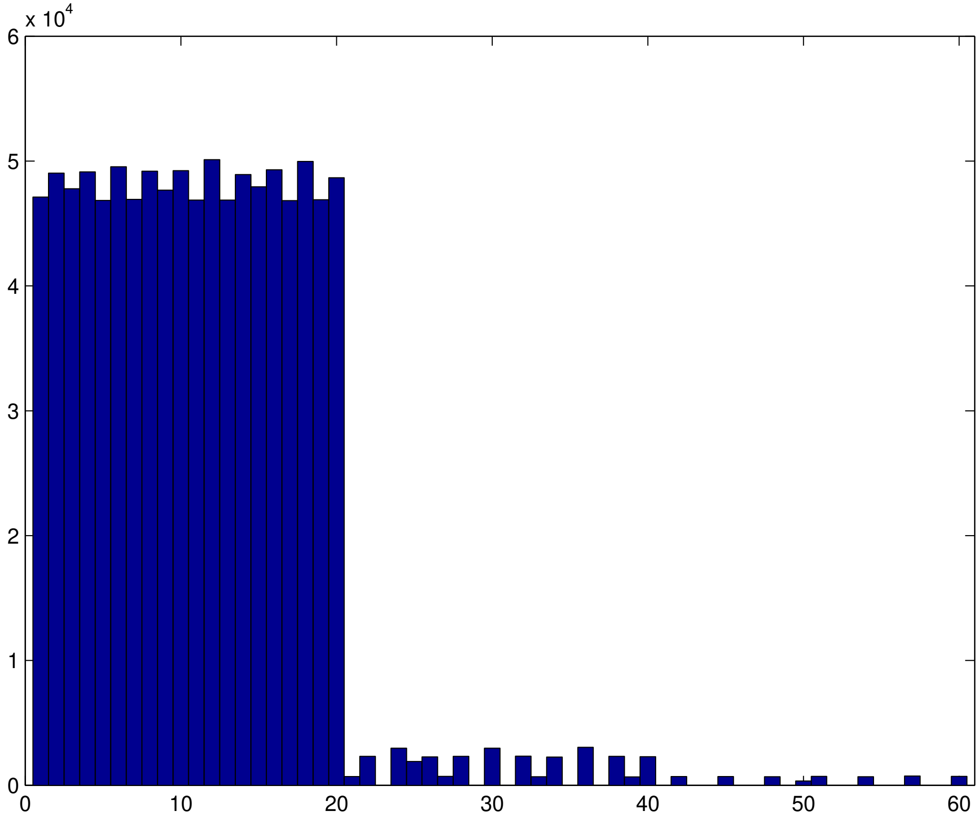




- Uniform samples from the space around the dartboard (1 million samples)
  - lots of misses, but correct answer (in the limit)

Dartboard - - Uniform Scoring Samples





- Uniform samples within the dartboard
  - no misses, but distribution of scores is correct

Mean and Variance of Estimate

- For Monte-Carlo methods would like to minimise number of samples
  - reducing the number of samples increases the variance of the estimate

Interested in the "score" volume for a dartboard ($4.1138s(core)m^2$)

- Take the dartboard - 1 million samples/run, 100 runs
  - exact solution (algebraic): mean 4.1138, standard deviation 0.0
  - uniform (square) sampling: mean 4.1132, standard deviation 0.0067
  - uniform (scoring) sampling: mean 4.1140, standard deviation 0.0026
- Smaller variance (Standard deviation) the better

Importance Sampling

- Current approaches yields correct answers
  - but can we select a better set of samples rather than uniform?
- Consider computing the expected value of a (multi-dimensional) function

$$\mathcal{E}\left\{f(\mathbf{x})\right\} = \int f(\mathbf{x})p(\mathbf{x})d\mathbf{x} \approx \frac{1}{N}\sum_{i=1}^{N} f(\mathbf{x}^{(i)})$$

  - where $\mathbf{x}^{(i)}$ is a sample from $p(\mathbf{x})$
  - but what if we can't sample from $p(\mathbf{x})$ but only $q(\mathbf{x})$

Use importance sampling

Importance Sampling

- Now draw samples from the distribution $q(\mathbf{x})$

$$\mathcal{E}\left\{f(\mathbf{x})\right\} = \int f(\mathbf{x})\frac{p(\mathbf{x})}{q(\mathbf{x})}q(\mathbf{x})d\mathbf{x} \approx \frac{1}{N}\sum_{i=1}^{N} f(\mathbf{x}^{(i)})\frac{p(\mathbf{x}^{(i)})}{q(\mathbf{x}^{(i)})}$$

  - the ratio $\frac{p(\mathbf{x}^{(i)})}{q(\mathbf{x}^{(i)})}$ is the “importance” of the drawn sample
  - some regions *over-represented* in samples $q(\mathbf{x}) > p(\mathbf{x})$
  - some regions *under-represented* in samples $p(\mathbf{x}) > q(\mathbf{x})$

- It is possible to run importance sampling without normalised distributions
  - can draw samples from distributions even when normalisation term can’t be computed
  - see the examples paper

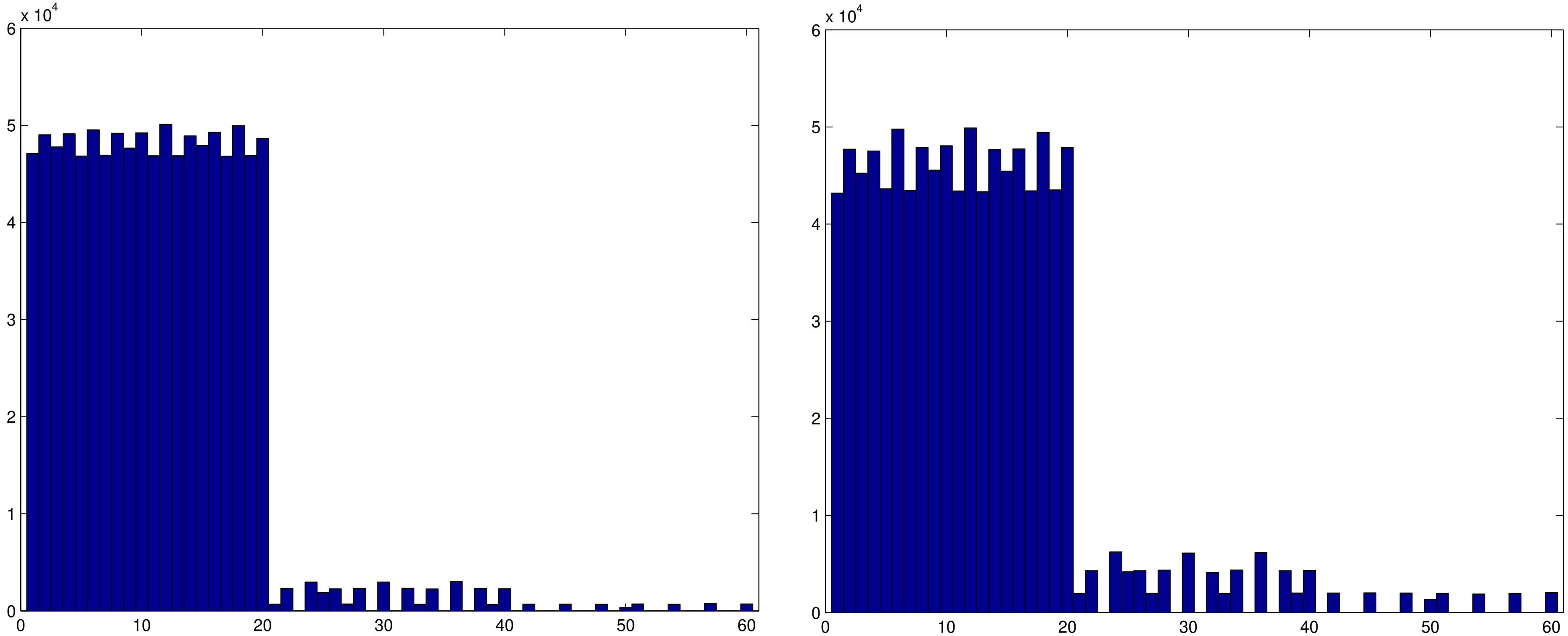
Sampling and Numerical Integration

- Return to the dartboard example with non-uniform sampling
- Similar form to importance sampling
  - assume possible to sample from a distribution $p(\mathbf{x})$ then

$$\int h(\mathbf{x})d\mathbf{x} = \int \frac{h(\mathbf{x})}{p(\mathbf{x})}p(\mathbf{x})d\mathbf{x} \approx \frac{1}{N}\sum_{i=1}^{N}\frac{h(\mathbf{x}^{(i)})}{p(\mathbf{x}^{(i)})}$$

  - the ratio $\frac{h(\mathbf{x}^{(i)})}{p(\mathbf{x}^{(i)})}$ is the “importance” of the drawn sample
- The closer that (scaled) $p(\mathbf{x})$ is to $h(\mathbf{x})$ the better!
  - feels intuitively right e.g. don’t sample from areas of zero “score”
  - if $p(\mathbf{x}) \propto h(\mathbf{x})$ get the answer in one sample ...

Dartboard - Score Biased Samples



- Focus on doubles (x2), triples (x3), bull - inner (x4.76) and outer ring (x2.38)
  - no misses, distribution of scores is not correct, but integral correct

Mean and Variance of Estimate

- For Monte-Carlo methods would like to minimise number of samples
  - reducing the number of samples increases the variance of the estimate
- Take the dartboard - 1 million samples/run, 100 runs
  - uniform (square) sampling: mean 4.1132, standard deviation 0.0067
  - uniform (scoring) sampling: mean 4.1140, standard deviation 0.0026
  - biased (score) sampling: mean 4.1140, standard deviation 0.0021
- Smaller variance (Standard deviation) the better

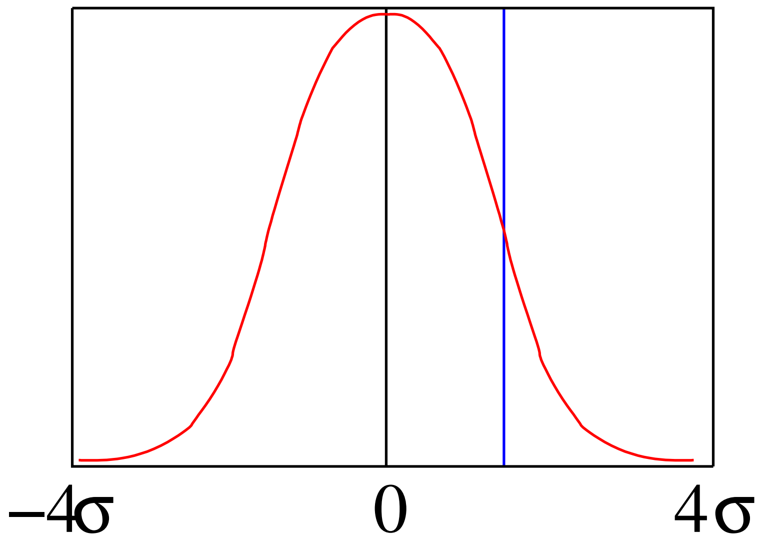
Generating Random Samples

- Modified problem to how to draw samples from a distribution
  - for "dartboard example" simple to modify samples to "scores"
  - generally awkward to generate samples from arbitrary distributions
- Packages (e.g. `matlab`/`octave`) support some standard distributions
  - Uniform distributions
  - Gaussian distributions

Is it possible to come up with general schemes

Rejection Sampling

- Alternative method to draw a sample is rejection sampling e.g. for a Gaussian



  - the peak value of a Gaussian is $1/\sqrt{2\pi\sigma^2}$
  - enclose in a box at, for example, $\pm 4\sigma$

- Draw samples uniformly from from the 2-D box,

  - accept those under the curve - take x value only
  - reject those above the line

Rejection Rate

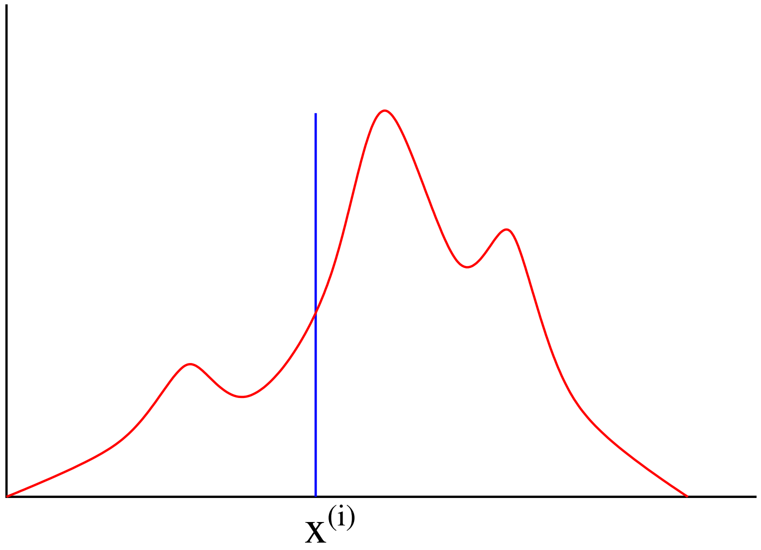
- Possible to compute percentage of samples rejected:
  - area of box is $8/\sqrt{2\pi}$
  - area under the curve is 1 (small fraction under of course!)

- Fraction of samples accepted is

$$\frac{\sqrt{2\pi}}{8} = 0.313$$

- What about two dimensions (assume Gaussians independent) - $0.313^2 = 0.093$
  - very rapidly becomes highly wasteful!
  - curse of dimensionality (again)

Metropolis-Hastings Algorithm



- Want to draw samples from the distribution above, $p(\mathbf{x})$

- Current sample is $\mathbf{x}^{(i)}$, generate another sample, $\mathbf{x}^{(\star)}$, from $p(\mathbf{x}|\mathbf{x}^{(i)})$

$$\mathbf{x}^{(\star)} = \mathbf{x}^{(i)} + \mathbf{z}, \quad \mathbf{z} \sim \mathcal{N}(0, \mathbf{I})$$

  – this would just yield Gaussian distributed variables ...

Metropolis-Hastings Algorithm (cont)

- Let's bias the samples we use so that we prefer "good" samples
  - accept the sample $\mathbf{x}^{(i+1)} = \mathbf{x}^{(\star)}$ with probability α where

$$\alpha = \min\left\{\frac{p(\mathbf{x}^{(\star)})}{p(\mathbf{x}^{(i)})}, 1\right\}$$

  - else reject the sample $\mathbf{x}^{(i+1)} = \mathbf{x}^{(i)}$

- Seems sensible, but what do we know about

$$\mathbf{x}^{(1)}, \mathbf{x}^{(2)}, \mathbf{x}^{(3)}, \ldots, \mathbf{x}^{(n)}$$

  - what is their distribution (as $n \to \infty$)?

They have the same distribution as $p(\mathbf{x})$

Metropolis-Hastings Algorithm

- Overall samples drawn from $p(\mathbf{x})$ BUT
  - initial samples (clearly) depend on $\mathbf{x}^{(0)}$
    often ignored - burn-in phase
  - samples correlated with "neighbouring" samples
    sometimes thinning is performed - only take every n^{th} sample
- Also need to decide proposal distribution
  - in the example the distribution of $\mathbf{z}$
  - for this example the proposal distribution is symmetric
- The general form of the Metropolis-Hastings algorithm uses

$$\alpha = \text{min}\left\{\frac{p(\mathbf{x}^{(\star)})p(\mathbf{x}^{(i)}|\mathbf{x}^{(\star)})}{p(\mathbf{x}^{(i)})p(\mathbf{x}^{(\star)}|\mathbf{x}^{(i)})}, 1\right\}$$

  - no assumption of symmetry required $p(\mathbf{x}^{(i)}|\mathbf{x}^{(\star)}) \neq p(\mathbf{x}^{(\star)}|\mathbf{x}^{(i)})$

Detailed Balance - Revisited

- When discussing discrete Markov Chains introduced detailed balance

$$\pi_j p_{j,k} = \pi_k p_{k,j}$$

  - what can we say when a process satisfies detailed balance?

- If a finite process (transition matrix $\mathbf{P}$) is in detailed balance then

$$(\boldsymbol{\pi}\mathbf{P})_k = \sum_j \pi_j p_{j,k} = \sum_j \pi_k p_{k,j} = \pi_k$$

  - hence $\boldsymbol{\pi}$ is a stationary distribution of $\mathbf{P}$

- If a distribution $\boldsymbol{\pi}$ satisfies detailed balance with transition matrix $\mathbf{P}$
  - then $\boldsymbol{\pi}$ is a stationary distribution for the transition matrix $\mathbf{P}$
  - a stationary distribution of $\mathbf{P}$ does not necessarily satisfy detailed balance

Samples from a Finite State-Space Markov Chain

- Given a limiting distribution $\boldsymbol{\pi}$
  - how to generate samples from a process with that limiting distribution?
- In the same fashion as the continuous distribution, have proposal function
  - in this case a transition matrix $\mathbf{R}$ with $r_{j,j} = 0$, $r_{j,k} > 0 \quad (j \neq k)$
  - true transition matrix is not known
- Process is:
  1. Choose an arbitrary starting state X_0, $i = 0$
  2. Given X_i, select $\hat{X}_{i+1}$ by sampling from the i^{th} row of $\mathbf{R}$
  3. Accept $(X_{i+1} = \hat{X}_{i+1})$ this sample with probability α

$$\alpha = \min\left\{\frac{\pi_{\hat{X}_{i+1}} r_{\hat{X}_{i+1},X_i}}{\pi_{X_i} r_{X_i,\hat{X}_{i+1}}}, 1\right\}$$

  4. else reject sample $(X_{i+1} = X_i)$, $i = i + 1$, goto (2)

Stationary Distribution of Metropolis-Hastings

- The above process is the equivalent of a transition matrix $\mathbf{P}$ with:

$$p_{j,k} = r_{j,k}\text{min}\left\{\frac{\pi_k r_{k,j}}{\pi_j r_{j,k}}, 1\right\} \quad \text{if } j \neq k, \quad p_{j,j} = 1 - \sum_{k\neq j} r_{j,k}\text{min}\left\{\frac{\pi_k r_{k,j}}{\pi_j r_{j,k}}, 1\right\}$$

- Need to show that $\mathbf{P}$ and $\boldsymbol{\pi}$ are in detailed balance

$$\begin{aligned}
\pi_j p_{j,k} &= \pi_j r_{j,k}\text{min}\left\{\frac{\pi_k r_{k,j}}{\pi_j r_{j,k}}, 1\right\} \\
&= \text{min}\left\{\pi_k r_{k,j}, \pi_j r_{j,k}\right\} \\
&= \pi_k r_{k,j}\text{min}\left\{1, \frac{\pi_j r_{j,k}}{\pi_k r_{k,j}}\right\} = \pi_k p_{k,j}
\end{aligned}$$

  – the process is also irreducible and aperiodic - so regular ergodic
  [Reference: in general for $\mathbf{R}$ all states must communicate, and $\mathbf{R}$ aperiodic]

- The above criteria are sufficient to show that $\boldsymbol{\pi}$ is the limiting distribution

Gibbs Sampling (Reference)

- Another MCMC approach is Gibbs sampling
  - draw samples from a (complicated) multivariate distribution, e.g. $p(\mathbf{x})$
- Assume that the current state of the system is given by $\mathbf{x}^{(i)}$
  - want to draw sample element $x_1^{(i+1)}$ from

$$p(x_1|x_2^{(i)}, x_3^{(i)}, \ldots, x_d^{(i)})$$

  - possible to then draw sample element $x_j^{(i+1)}$ from

$$p(x_j|x_1^{(i+1)}, \ldots, x_{j-1}^{(i+1)}, x_{j+1}^{(i)}, \ldots, x_d^{(i)})$$

  - eventually yield $\mathbf{x}^{(i+1)}$, then possible to repeat
- Yields a sequence of samples $\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(n)}$, similar to Metropolis-Hastings:
  - samples correlated: possible to have burn-in, and use thinning

Summary

- Stochastic approaches for optimisation/integration (examples)
  - Monte-Carlo approaches based on generating random samples
- Need to generate samples similar to function
  - affects variance of numerical integration approximation
  - speed of stochastic optimisation
- General approaches for generating samples:
  - Markov Chain Monte Carlo (MCMC)
  - Gibbs Sampling