# Scientific Computing CS660 Fall '11

Angjoo Kanazawa

September 26, 2011

## 1 September 1st Class 1

### 1.1 Logistics

Prof. Howard Elman CSI 2120 TR 2pm-3:15pm class url:http://www.cs.umd.edu/ elman/660.11/syl.html

- Scientific Computing puts heavier emphasis on computing, Numerical Analysis is more about proofs/theories.
- 4-6 hw asg: **35%** Penalty on late assignments (-15% after 24 hrs, -30% after 48 hrs.
- in-class midterm: **25%**
- final project: **40%**

### 1.2 Content

**Newton's Method**: Root finding. Objective: Fine $x$ s.t. $f(x) = 0$, where $f$ is a scalar, $f : \mathbf{R} \Rightarrow \mathbf{R}$ function. Where does the function cross 0 (x-axis)?

Given $x_n$, some guess, find where the line through $(x_n, f(x_n))$ tangent to the solution curve intersects the $x$-axis. Call that pt of intersection $x_{n+1}$

The equation of the tangent line:

$$\frac{y - f(x_n)}{x - x_n} = f'(x_n)$$

Set $y = 0$: then

$$\begin{aligned}
\frac{0 - f(x_n)}{x_{n+1} - x_n} &= f'(x_n) \\
\frac{x_{n+1} - x_n}{-f(x_n)} &= 1/f'(x_n) \\
x_{n+1} &= x_n - f(x_n)/f'(x_n)
\end{aligned}$$

**Another Derivation** Consider the taylor series $f(x_n + (x - x_n)) = f(x_n) + f'(x_n)(x - x_n) + 1/2f''(x_n)(x - x_n)^2 +$ etc. This is a function of some variable $x$. Approximate it just by using the first two terms (linear approximation). So

above becomes a new function $f(x_n+(x-x_n)) = f(x_n)+f'(x_n)(x-x_n) = l(x)$. Find where $l(x) = 0$. That is: $x_{n+1} = x_n - f(x_n)/f'(x_n)$

This is not guaranteed to work, i.e. when the tangent doesn't cross the $x$-axis.

**Problem**: given $\alpha \in \mathbf{R} > 0$ Find $1/\alpha$ without doing any division. First thing you need to do if identify (concoct) a $f(x)$ whose root is $1/\alpha$. Naive: try $f(x) = x - 1/\alpha$. But this won't work, because this requires division. Howabout: $f(x) = \alpha x - 1$. $f'(x) = \alpha$, so in the newton iteration..

$$\begin{aligned} x_{n+1} &= x_n - \frac{\alpha x - 1}{\alpha} \\ &= x_n - x_n + 1/\alpha \\ &= 1/\alpha \end{aligned}$$

No! because you need to divide.

Answer: $f(x) = \alpha - 1/x$. Not transparent, because intuitively it looks like there's a divide into it. $f'(x) = --1/x^2 = 1/x^2$. Then, $f/f' = \alpha x^2 - x$. So given $x_n$, the iteration goes $x_{n+1} = x_n - (\alpha x_n^2 - x_n) = 2x_n - \alpha x_n^2 = x_n(2 - \alpha x_n)$.

Numerical example in matlab: let $\alpha = 2$, solve with this method. Use $x_0 = 0.1$. Notice that $err(i)/err(i-1)^2$ is constant and *is* $\alpha$.

$$\frac{|x - x_{n+1}|^2}{x - x_n} \approx 1/2 \frac{f''(x_n)}{f'(x_n)}$$

Notice the $err(i)$ decreases faster as iteration moves on, this is the super linear convergence property of Newton's method. With $\alpha = 0.25$, same thing.

**Analysis of the Newton's Method**: $|x_{x_{n+1}}|/|x-x_n|^2 \approx 1/2|f''(x_n)/f'(x_n)| \approx 1/x$ So @ $x = 1/\alpha$, this turns out to be $\alpha$, just for this example.

The ratio was derived from the idea to find a patter in the errors s.t. $e_{n+1}$ $ce_n^2$, for some $c \in \mathbf{R}$

The question is to find trends in data, and this relationship $e_{n+1}$ $ce_n^p$ is useful to tell us the rate of convergence as the solution approches the optimal one. (I think $p$ goes down to the golden ratio). Our goal is to find what $p$ is for a specific problem.

# 2 September 6th class 2

## 2.1 Process of Scientific Computing

- Start with a mathematical model. (In general we don't have an analytic solution, so we get insight from numerical computation)
- Example with heat conduction in a bar:
  - A 1-D object $\in [0,1]$, $u(x)$ =temperature in the bar, with $u(0) = 0, u(1) = 0$. $q$=heat flow induced by a heater of intensity $f$.
  - We want what the temperature will be given $q$.
  - get some models: *Fourrier's Law*: $q = -ku'$, $k$=conductivity coefficient, transfer of heat in direction of decreasing temperature (hence the -). *Conservation of energy*: $q' = f$.
  - **Goal**: find $u$. **Equation of interest**: $-(ku')' = f$, the 1D diffusion equation.
  - Typical strategy: lay down a grid $x_o = 0, x_1, \ldots, x_n, x_n + 1 = 1$. Compute a discrete solution, $\bar{u}$, vector of size $n$. $\bar{u} = [u_1, \cdot, u_n]^T$, where $u_i \approx u(x_i)$
  - **Claim**: we can find the discrete sol, $\bar{u}$ by solving an algebraic system of equations. For this example, this system is a matrix equation

$$A\bar{u} = \bar{b}$$

- No matter how hard I try, we're never going to get the exact solution. This process $A\bar{u} = \bar{b}$ leads to errors
- **Sources of error**:
  1. Modeling error: we may not know $k$ exactly.
  2. Discretization error/Truncation error: difference between the discrete and the continuous values (from the approximation on a discrete set of points)
  3. Representation error: we don't have the entire $\mathbf{R}$, we only have a finite set, in floating point format. ($A$ and $b$ may have error)
  4. Additional error: from the computation of $\tilde{u}$, will get something else $\hat{\tilde{u}} \neq \tilde{u}$
     **we can show**:
     $$\frac{||\hat{\tilde{u}} - \tilde{u}||}{\hat{\tilde{u}}||} \leq K(A)\mu$$
     That is if we solve the problem appropriately (like being careful about pivoting etc).
- We're really trying to solve $\tilde{A}\tilde{u} = \tilde{b}.$, where $\tilde{A} \approx A$, $\tilde{u} = \bar{u}$. Typically $\approx$ is machine precision, $10^{-16}$
- In the end, we want $u(x)$, and would be happy with $u(x_j)$, $j = 1, \ldots, n$. That will get $\tilde{u_j}$
- The moral is that when we do this stuff, we're not just doing mathematics.

## 2.2 Floating Point Arithmetic

decimal numbers (base 10). Consider the example 6522 and 10.31

- $6522 = 6(10^3) + 5(10^2) + 2(10^1) + 2(10^0)$
- $10.31 = 1(10^1) + 0 + 3(10^-1) + 1(10^-2)$
- Normalize the numbers! then,,
- $6522 = 6.522 \times 10^3$ so we can express numbers with one digit to the left o fthe decimal point.
- $10.31 = 1.031 \times 10^1$
- For any number but 0, it has a fomr $z \times 10^p$, where $z \in [1, 10)$.

Computers use binary representation. Example: $3_{10} = 1(2^1) + 1(2^0) = 11$ or $1.1000 \times 2^1$.

$23_{10} = 16 + 4 + 2 + 1 = 1(2^4) + 0(2^3) + 1(2^2) + 1(2^1) = 10111_2$ or $1.0111 \times 2^4$ *normalized*. Here, normalized means $z \times 2^p$, where $z \in [1, 2)$

# 3 September 8th Class 3

## 3.1 Floating Point Operations

Example: addition using d=5 binary digits

Add $3 + 23$.

Normalized 5-digit binary expressions:

$3 = 2^1 + 2^0 = 1.1000 \times 2^1$

$23 = 16 + 4 + 2 + 1 = 1.0111 \times 2^4$

To add these, shift the smaller number so that the exponents agree.

$$1.1000 \times 2^1 = 0.11000 \times 2^2 \quad (1)$$
$$= 0.01100 \times 2^3 \quad (2)$$
$$= 0.00110 \times 2^4 \quad (3)$$

May need to round the result: 3+34 32+2 =

Shift smaller number

$$3 ==> 0.000110 \times 2^5$$
$$34 ==> 1.00010 \times 2^5$$
$$- - - - - - - - - - - - - - - - - -$$
$$37 ==> 1.00101$$

The way the arithmetic is done: take 2 floating point numbers. The computer hardware will take the correct result (stored temporarily), but will always round to $d$ digits. In our examples we are using only $d = 5$. in the real world we are usually dealing with $d = 27$ or more. number will always be rounded to closest representation– can be either up or down.

so in this case, the stored result is rounded to $1.0011 \times 2^5$ which is $32+4+2 = 38$.

Exercise: what two number would have the property whose sum's **correct** result is 1.001001? Answer: 32+4+0.5. (Note: 2.5 in binary is $1.0100 \times 2^1 = 2 + 1/2$).

In general, given the real $x$, $fl(x)$ = closest floating point number to $x$. $|x = fl(x)|$ is called the rounding error. For operations op $= \{+, -, \times, \div\}$, if $x$ and $y$ are floating point numbers, then $fl(x \text{ op } y)$ = floating point number closest to $x$ *op* $y$.

It is possible that both inputs are perfectly good FP numbers, but that the result of the operation is not able to be represented in the floating point system (eg. d is too small). For example, in our single precision example, $d = 23$, $m = -126 \leq p \leq M = 127$. Then, if $x = 1 \times 2^{100}$ and $y = 1 \times 2^{100}$ then $x * y = 1 \times 2^{200}$ which results in an overflow.

**Definition** *Machine Precision* or *Machine Epsilon* is the smallest floating point number $\mu$ such that $fl(1 + \mu) \neq 1$ $(> 1)$

In IEEE arithmetic, for any real $x \neq 0$:

$$\frac{|x - fl(x)|}{|x|} \leq \mu$$

Example: $d = 5$. Find $\mu$.

$$\begin{aligned} 1 &= 1.0000 \times 2^0 \\ \text{Add } z &= 0.00001 \times 2^0 \\ &\text{-----------------} \\ &= 1.00001 \end{aligned}$$

=¿ which gets rounded to 1.0001.

$fl(1 + z) \neq 1$

Next smallest number to $z$. Normalized, $z = 1.0000 \times 2^{-5}$. Next smaller number is $z = 1.1111 \times 2^{-6}$.

$$\begin{aligned} 1 &= 1.00000 \\ +z &= 0.0000011111 \\ &\text{----------------------} \\ &= 1.0000011111 \end{aligned}$$

which gets rounded to 1.

In general, in $d$-digit binary arithmetic, the machine precision $\mu$ is $2^{-d}$.

For IEEE single precision, this is $2^{-23} \approx 1.2x10^{-7}$ For IEEE double precision, this is $2^{-53} \approx 1.1x10^{-16}$

Does IEEE arithmetic support the existence of numbers like $2^m$ or $2^M$?

## 3.2 Relative Error

We have a number $x$, and a representation of $x \approx \hat{x}$.

The relative error is given by

$$\frac{||x - \hat{x}||}{||x||}$$

And the absolute error is given by $||x - \hat{x}||$.

The relative error is more important than the absolute error. Consider for example a census, x% absolute error in the number of ppl in class compared to the same x% absolute error in the population of nyc does not mean the same thing.

## 3.3 Forward vs. Backward Error Anlysis

Generically speaking: we are seeking $y = f(x)$, and get $\hat{(y)} \neq y$. We want insight into $\frac{||y-\hat{y}||}{||y||}$.

**Definition:** *Forward error analysis* tries to keep track of errors as the computation proceeds.

Example:

Solve $Ax = b$. We get $\hat{x}$ instead. And we are interested in $\frac{||x-\hat{x}||}{||x||}$. Forward error analysis (egf. for gaussian elimination) is not possible. (Well, it is possible, but the estimates tend to be wildly inaccurate).

**Definition:***Backward error analysis* makes the claim that $\hat{x}$ is the solution of a perturbed problem, $\hat{A}\hat{x} = b$ such that (if the solution is done right):

$$\frac{||x - \hat{x}||}{||x||} \lessapprox \mu(p(n)\mu)$$

where $p(n)$ is a slowly growing function of $n$ = problem size.
We use this observation: $Ax = b$ $\hat{A}\hat{x} = b$
and

$$\begin{aligned} A(x - \hat{x}) &= Ax - A\hat{x} && (4)\\ &= Ax - \hat{A}\hat{x} + \hat{A}\hat{x} - A\hat{x} && (5)\\ &= b - b + (\hat{A} - A)\hat{x} && (6)\\ &= 0 + E && (7) \end{aligned}$$

where $E$ is our error.

$$\begin{aligned} &\rightarrow x - \hat{x} = A^{-1}E\hat{x} && (8)\\ &||x - \hat{x}|| \leq ||A^{-1}||||E||||\hat{x}|| && (9)\\ &||A^{-1}||||A||||E||/||A||||\hat{x}|| && (10) \end{aligned}$$

$||A^{-1}||||A||$ is called $\kappa(A)$, and $||E||/||A|| \lessapprox \mu$
this $\rightarrow ||x - \hat{x}||/||\hat{x}|| \lessapprox \kappa(A)\mu$
$\mu \approx 10^{-16}$
$\kappa(A)$ depends on $A$, the statement of the problem..
We claim, that with a bit more work, we could put $||x||$ in the denominator.

# 4 September 13rd Class 4

## 4.1 Intro to Probability

**Discrete models**: Consider a game with a finite or countable number of outcomes, $\Omega$, the *sample space*. **Definition:** *Probability* for each $\omega_j \in \Omega$, we have a number $p_j(\omega_j) = p_j$ s.t. $p_j \in [0,1]$, and $\sum_{j=1}^{\infty} p_j = 1$

Examples:

1. Roll a fair die outcomes: $\Omega = \{1,2,3,4,5,6\}$, $p_j = 1/6$, $j = 1,2,\ldots,6$
2. Roll two fair dice, outcomes: $\Omega = \{(1,1),(1,2),\cdots,(1,6),(2,1),\cdots,(6,6)\}$, $p_j = 1/36$, $j = 1,\ldots,36$.
3. put $n$ distinct balls into $N$ urns $N > n$. 1st ball has $N$ choices, so does the 2nd ball, etc, so the total number of outcomes is $N^n$.
4. A die is rolled until a six appears. possible outcomes:
   (a) Get a 6 on 1st trial: 6, $p(\omega_1) = 1/6$
   (b) something other than 6 on 1st trial: 16,26,36,46,56 (5 ways), $p(\omega_2) = 5/6(1/6) = 5/36$
   (c) etc $p(\omega_i) = (1 - p(\omega_1))^{i-1})p(\omega_1) = (5/6)^{i-1}(1/6)$

   Here $\Omega$ is countably infinite. The sum:

$$\begin{aligned}\sum_{j=1}^{\infty} p_j &= 1/6 + (5/6)(1/6) + (5/6)^2(1/6) + \cdots \\ &= 1/6 \sum_{l=0}^{\infty} (5/6)^l \\ &= 1/6(\frac{1}{1-5/6}) = 1/6(\frac{1}{1/6}) = 1\end{aligned}$$

**Definition:** An *event* is a subset $A \subseteq \Omega$. Notation $P(A) = \sum_{\omega_J \in A} p(\omega_j)$. Properties: ¶$(\Omega) = 1, P() = 0, P(A^c) = 1 - P(A)$. If $A_1, A_2, \ldots$ are pairwise disjoint, then $P(\cup_i A_i) = \sum_i P(A_i)$.

**Definition:** A *random variable* $X$ is a function $X : \Omega \to \mathbf{R}$. $\forall E \in \mathbf{R}$, define an *event* $\{X \in E\} = \{\omega \in \Omega | X(\omega) \in E\}$ Can talk about probability of such an event as $P(X \in E)$

**Definition:** A *distribution* of a discrete random variable $X$ is a collection of distinct real numbers, or a set of values $\{m_j\} \in [0,1]$, s.t. $\sum m_j = 1$, and $P(X = x_j) = m_j$.

Example: *Binomial distribution.* Consider a random experiment with two possible outcomes. Psobability of success $= p$, failure $q = 1 - p$. With coins, $p = 1/2$. Perform this experiment $n$ times, $X =$# of successes. $X$ has a $Binom(p,n)$. The probability of exactly $l$ sucesses is $P(X = l) = \binom{n}{k} p^k q^{n-k}$

# 5 September 15th Class 5

## 5.1 Probability Cont.

Experiment: three tosses of a coin $|\Omega| = 2^3$

**Definitino:** The distribution of discrete random variables is a collection of real positives $\{x_j\}_{j=1}^{\infty}$, s.t. $P(X = x_j) = m_j$

**Definition:** A *mass function* of a discrete random variable $X$ is $m : \mathbf{R} \to [0, 1]$ s.t. $m(x) = P(X = x) \forall x \in \mathbf{R}$.

**Definition:** The binomial distriubtion is a random experiment with two outcomes where $p$ is the probability of success, $1 - p$ is the probability of failure. It performs the experiment $n$ times, and $X$ denotes the number of total successes.

Q: what is the probability of exactly $k$ successes? $\binom{n}{k} p^k (1-p)^{n-k}$, let's check $\sum_{j=1}^{\infty} m_j = 1$. This is $\sum_{k=0}^{n} p^k (1-p)^{n-k} = (p + (1-p))^n = 1^n = 1$ so yes.

**Definition:** Continuous random variables and distributions. ex: Average weight of 100 randomly selected , where $\Omega$=$\{$100 tuples of weights $\}$, $X(\omega) = 1/100$ (sum entries of $\omega$). Random number uniformly chosen from $[a, b]$.

**Definition:** A function $f$ on $\mathbf{R}$ is called a *density function* if it's non-negative and integrable and $\int_{-\infty}^{\infty} f(x)dx = 1$.

A random variable with a density function $f$ means

$$P(\{\omega \in \Omega | X(\omega) \in A\}) = \int_A f(x)dx$$

for $A \subseteq \mathbf{R}$.

example: we say $X$ is uniformly distributed on $[a, b]$ if it has density function

$$f(x) = \begin{cases} \frac{1}{b-a} & \text{for } x \in [a, b] \\ 0 & \text{otherwise} \end{cases}$$

Alternatively, $X$ is normally distributed with mean $\mu$ and variance $\sigma^2$ if

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

Exercise: check $\int_{-\infty}^{\infty} f(x)dx = 1$

**Defintion:** *mean*, $\mathbf{E}(x) = \mu$, of a random variable. Discrete case: $\mu = \sum_{x \in \mathbf{R}} x \times m(x)$, continuous case: $\mu = \int_{-\infty}^{\infty} x \times f(x)dx$

$$\begin{cases} \sum_{x \in \mathbf{R}} x \times m(x) & \text{Discrete} \\ \int_{-\infty}^{\infty} x \times f(x) & \text{Continuous} \end{cases}$$

**Definition:** *Distribution of a continuous r.v.*: $F(x) = P(X(\omega) \leq x) = \int_{-\infty}^{x} f(\xi)d\xi$. So if $F(x)$ is normal with $\mathcal{N}(\mu, \sigma)$, $\mu$ is at the tip of the gaussian mountain, and $\mu$ has val 0.5 in the CDF, $F(x)$.

Let $\Phi : \mathbf{R} \to \mathbf{R}$. *Claim:*

$$E(\Phi(x)) = \begin{cases} \sum_{x \in \mathbf{R}} \Phi(x) m(x) & \text{Discrete} \\ \int_{-\infty}^{\infty} \Phi(x) f(x) dx & \text{Continuous} \end{cases}$$

This is not obvious.

Now how to define R.V, $X$, formally. $\Omega \to \mathbf{R}$. Using the example of 3 coin tosses, $\Omega = \{000, 001, \ldots, 111\}$. $X(\omega) =$ decimal version of binary value of $\omega$. Then, $X(\omega) = [1, 7] \in \mathbf{N}$. Now, we can do: $\forall x \in \mathbf{R}$, define $m(x)$ by

$$m(x) = \begin{cases} 1/8 & \text{if } 1 \geq x \leq 7 \\ 0 & \text{otherwise} \end{cases}$$

Let us call 1 or more consecu, tive tosses of same type as *run*s. This is a new R.V. where $\Omega$ is the same, and $Y(\omega) =$number of runs.

i.e. $Y(000) = 1 = \Phi(0), Y(001) = 2 = \Phi(1), Y(010) = 3 = \Phi(2), Y(011) = 2, \ldots$. Note $Y = \Phi(X)$. So $m_y(y) = P(Y = y)$, and $m_y(1) = 1/4, m_y(2) = 1/2, m_y(3) = 1/4, m_y(\text{else}) = 0$. Then,

$$\begin{aligned} \mathbf{E}(Y) &= \sum_{y \in R} y m_y(y) \\ &= 1/4 + 2(1/2) + 3(1/4) = 2 \end{aligned}$$

This is the same as

$$\begin{aligned} \sum_{x \in R} \Phi(x) m(x) &= (1 + 2 + 3 + 2 + 3 + 2 + 1)(1/8) \\ &= 16/8 = 2 \end{aligned}$$

# 6 September 20th Class 5

## 6.1 Continue on Probability

Given random variable $X$ and $E(x) = \mu = \int_{-\infty}^{\infty} xf(x)dx$, the variance is $var(x) = \sigma = E[(X-\mu)^2] = E(X^2) - E(X)^2$ If $x$ has a desity function $f(x)$, then $E(\phi(x)) = \int_{-\infty}^{\infty} \phi(x)f(x)dx$

## 6.2 Monte Carlo Integeration

Let $X$ be a uniformly distributed random variable on $[0,1]$. If $\phi : [0,1] \to \mathbf{R}$, suppose we want an approximation to $\int_0^1 \phi(x)dx = I(\phi)$.

Sample $X$ from $[0,1]$, and get $x_1, x_2, \ldots, x_n$. We'll approximate $I(\phi)$ by the average $\frac{1}{N}\sum_{i=1}^{N} \phi(x_i)$. This is the quadrature rule $Q_{MC,N}(\phi)$.

Notice: $\phi(x)$ is a random variable. The expected value of $\phi(x)$ is $E(\phi(x)) = \int_0^1 \phi(x)dx = I(\phi)$, because $X$ is uniform. (so $f(x) = \begin{cases} 1 & \text{if } x \in [0,1] \\ 0 & \text{otherwise} \end{cases}$)

The approximation $Q_{MC,N}(\phi) = Q_N(\phi)$ is called the sample mean of $\phi(x)$. Approximates the mean of $\phi$, which is $I(\phi)$ or $\mu(\phi)$. i.e. a trivial example if $\phi(x) = x$, then $I(\phi) = \int_0^1 xdx = \frac{1}{2}x^2|_0^1 = 1/2$

Contrast this with standard ways to do quadrature. *examples:*

- the trapezoidal rule: area under the curve is approximated by the trapezoid bewteen $f(a)$ and $f(b)$. $\int_a^b \phi(x)dx \approx 1/2(b-a)(\phi(a)+\phi(b))$
- the simpson's rule: $\int_a^b \phi(x)dx \approx 1/6(\phi(a) + 4\phi(\frac{a+b}{2}) + \phi(b))(b-a)$
- Composite trapezoidal rule on $[0,\ 1]$: $\approx\ h/2\phi(x_0) + h\sum_i^{N-1} f(x_i) + h/2\phi(x_N)$

### 6.2.1 Monte Carlo history

From high energy physics by John von Newmann and Stanislaw Ulam. Monte Carlo, an island in the meditteranian where people gamble.

## 6.3 Error Analysis

Simpson's rule is better. The error analysis of sympson's rule is: $I(\phi) - Q_s(\phi) \leq ch^4 = c1/N^4 = O(1/N^4)$

With trapezoid it's $O(1/N^2)$. Is this better than monte carlo?
In $1D$, monte carlo does worse.

*Claim:* error $|I(\phi) - Q_{MN,N}(\phi)|$" $\leq$ "$O(1/\sqrt{N})$. Suppose $N = 10000$, so $1/100$, but in simpson's rule's error, $O(1/10^{16})$. Simpson's does way better.

From probability analysis, using the Law of Large Numbers, if we let $S_N = \phi(x_1) + \cdots + \phi(x_N)$, and the sample mean $S_N/N = \mu_N$, (which is our approximated integral), by LNN, $\lim_{n\to\infty} P(|\frac{S_N}{N} - I(\phi)| < \epsilon) = 1$, where $I(\phi)$ is the real $\mu$. That is $P(|\frac{S_N}{N} - I(\phi)| > \epsilon) \to 0$ as $N \to \infty$. In words, it's "very unlikely" that $Q_{MC,N}(\phi)$ is different from $I(\phi)$ when $N$ is very large.

**Chebyshev's inequality**: Given $c > o$,

$$P(|\frac{S_N}{N} - \mu| \geq c) \leq \frac{\sigma^2}{Nc^2}$$

Commentary: suppose we want this probability to correspond to 95% confidence, that is the probability on the left to be .05, then we require $.05 \leq \frac{\sigma^2}{Nc^2}$, solve for $c$ and we get $c^2 \leq \frac{\sigma^2}{.05N}$ or $c \leq \frac{\sigma}{\sqrt{.05N}}$. So with 95% confidence, $error \leq \frac{\sigma}{\sqrt{.05}}\frac{1}{\sqrt{N}}$ That's what we can say about the error. If we want 99% confidence, then $\sqrt{.05}$ becomes $\sqrt{.01}$

Two cons: convergence is quite a bit slower, we're not as confident as the other quadrature methods (the others give us a guarantee)

Suppose instead of uniform, say we had two box, $N$ samples of 2-tuples, then with double integral and think about monte carlo. The procedure is the same, sample, evaluate the value at that sample, then take the average. Suppose we have a cube, same thing. Nothing of monte carlo is tied to the underlying domain. The analysis using LLN and chebyshev is always the same.

Consider 2D. Simpsons rule on $[a, b] \times [c, d]$.

$$\int_0^\phi \int_a^b \phi(x, y)dxdy \approx \text{let } \Phi(y) = \frac{b-a}{6}(\phi(a, y) + 4\phi(\frac{a+b}{2}, y) + \phi(b, y))$$
$$\approx \frac{d-c}{6}(\Phi(c) + 4\Phi(\frac{c+d}{2}) + \Phi(d))$$

, where $\Phi(\frac{c+d}{2}) = \phi(\frac{a+b}{2}, \frac{c+d}{2})$ Total number of points $N = n^2$, $h = 1/n = 1/\sqrt{N}$. The error is $O(h^4) = O(1/n^4) = O(1/N^2)$

In 3D, the error on simpsons is still $O(h^4)$, but $h = 1/n, N = n^3$, so this is $O(1/N^{4/3})$, in $D$ dimensions, it's $O(1/N^{4/d})$)

In monte carlo, the error analysis is the same in any dimension. Has nothing to do with the integral. so the big pro is that we can never be certain but we're always working with the same error around $\frac{\sigma}{\sqrt{a}}\frac{1}{\sqrt{N}}$.

Now, which technique is better? If we ignore the uncertainty, we're asking when is $\frac{1}{N^{4/d}} < \frac{1}{N^{1/2}}$?

$$\frac{1}{N^{4/d}} < \frac{1}{N^{1/2}}?$$
$$N^{1/2} < N^{4/d}$$
$$\frac{1}{2} < \frac{4}{d}$$
$$d < \frac{4}{1/2} = 8$$

So when $d < 8$, simpson is better, else, MC is better. (Although we're ignoring the constant and the uncertainty factor so it won't be exact like this but still around there)

In model of particle physics, the number of dimensions is basically is equal to the number of particles, which could be hundreds. It's not unusual to have large $d$.

*Example:* of high dimensional integral (in the text/wiki). In particle physics, a *partition function* $Z(\beta)$ describes the statistical properties a system of $d$ particles, thermodynamic equilibrium,
$Z(\beta) = \int exp(-\beta H(p_1, \cdots, p_d, x_1, \cdots, x_d) d^3p_1 \cdots d^3p_d d^3x_1 \cdots d^3x_d$,
where this integral is a $3d$ integral. where $d$ = number of particles, and $\beta = \frac{1}{\alpha\tau}$, a Boltzmann constant, $\tau$ is temperature, and $H$ is hte hamiltonian function. And the point is that people care about this. This is about a 3 million dimensional integral.

# 7 September 22 Class 6

## 7.1 Continue on Monte Carlo

Monte Carlo Integration: Given a r.v. $Y$, $P(Y \subseteq A) = E(I_A(Y))$, where $I_A(Y) = \begin{cases} 1 & \text{if } Y \in A \\ 0 & \text{otherwise} \end{cases}$. And $E(I_A(Y)) = \int_A I_a(y)dy$, where the integral can be multi-dimensional. One way to deal with high-dimensional integral is to use Monte Carlo Integration.

*Review*: Monte Carlo gives us the approximation $\int_{\mathbf{R^D}} \phi(x)dx = I(\phi)$, where $I(\phi) \approx 1/N \sum_{n=1}^{N} Q(x_n) = Q_{MC,N}(\phi)$ where $\{x_n\}$ are samples of randome variable $X$.
We showed that the error was $|I(\phi) - Q_{MC,N}(\phi)|$" $\leq$ " $\frac{\sigma}{c\sqrt{N}}$
If the goal is to make this bound $\leq \tau$ tolerance, i.e. $\frac{\sigma}{c\sqrt{N}} \leq \tau$, then we need

$$\sqrt{N} \leq \frac{\sigma}{c\tau}$$
$$N \leq \frac{\sigma^2}{c^2\tau^2}$$

many samples.

## 7.2 Variance Reduction Methods

The aim is to reduce $\sigma$ 2 examples of idea to do this

### 7.2.1 Antithetical variables

Assume $\phi \in [0,1]$, and $I(\phi) = \int_0^1 \phi(x)dx = E(\phi(x))$ (remember !!) where $X$ is a r.v. from a uniform distribution on $[0,1]$. ($E(x) = \int_0^1 xdx = 1/2x^2|_0^1 = 1/2$, and we approximate $I(\phi)$ with $1/N \sum_n = 1^N Q(x_n)$).
Antithetical variables are $x$, and $x^c$ s.t. $x + x^c = 1$ or $x^c = 1 - x$, $x = \mu + d$, $x^c = \mu - d$

Assume $\sigma$ small. Write $d = \sigma\hat{d}$ ($d$ will not grow too much because they're from a bounded distribution..) Let $D = X - \mu$, another r.v., and let $\hat{D} = D/\sigma$ also a r.v. Then,

$$\begin{aligned} E(\hat{D}) &= \frac{1}{\sigma}E(D) \\ &= \frac{1}{\sigma}E(x - \mu) \\ &= \frac{1}{\sigma}[E(x) - E(\mu)] \\ &= \frac{1}{\sigma}[\mu - \mu] = 0 \end{aligned}$$

Now, consider $\phi(x)$. (he used $1/2$ as $\mu$ following the example on board)

$$\begin{aligned}\phi(x) &= \phi(\mu + d) = \phi(\mu + \sigma\hat{d}) \\ &= \phi(\mu) + \phi'(\mu)\sigma\hat{d} + \phi''(\xi)\sigma^2\hat{d}^2 \text{ by taylor series} \\ &\approx \phi(\mu) + \phi'(\mu)\sigma\hat{d} + O(\sigma^2)\end{aligned}$$

And $\phi(x^c)$

$$\begin{aligned}\phi(x^c) &= \phi(\mu - d) = \phi(\mu - \sigma\hat{)} \\ &\approx \phi(\mu) - \phi'(\mu)\sigma\hat{d} + O(\sigma^2)\end{aligned}$$

SO,

$$\frac{\phi(x) + \phi(x^c)}{2} = \phi(\mu) + O(\sigma^2)$$

Now look at $I(\phi) = E(\phi(x)) = \int_0^1 \phi(x)dx$. This is

$$\begin{aligned}&= E[\phi(\mu) + \phi'(\mu)\sigma\hat{D} + O(\sigma^2)] \\ &= \phi(\mu) + \phi'(\mu)\sigma E[\hat{D}] + O(\sigma^2) \\ &= \phi(\mu) + 0 + O(\sigma^2)\end{aligned}$$

Now a new integration method. Smaple $X$ $N$ times and get $x_1, \ldots, x_N$, also let $x_1^c, \ldots, x_N^c$. Let the complementary quantities:

$$Q_{MC,N}^c(\phi) = \frac{1}{2N}(\phi(x_1) + \phi(x_1^c) + \cdots + \phi(x_n) + \phi(x_n^c)$$

For each $n$,

$$\begin{aligned}\phi(x_n) = \phi(\mu + \sigma\hat{d}_n) &= \phi(\mu) + \phi'(\mu)\sigma\hat{d}_n + O(\sigma^2) \\ \phi(x_n^c) = \phi(\mu - \sigma\hat{d}_n) &= \phi(\mu) - \phi'(\mu)\sigma\hat{d}_n + O(\sigma^2) \\ 1/2(\phi(x_n) + \phi(x_n^c)) &= \phi(\mu) + O(\sigma^2) \\ \Rightarrow Q_{MC,N}^c(\phi) = \frac{1}{N}(N\phi(\mu) + N(\phi(\sigma^2)) &= \phi(\mu) + O(\sigma^2)\end{aligned}$$

Now the difference: $I(\phi) - Q_{M,N}^c(\phi) = \phi(\mu) - \phi(\mu) + O(\sigma^2)$ So **the error is 1/4th the size** or $\sigma^2$ instead of $\sigma$. Smmmary:

Method 1 Take $N$ samples, error $\frac{\sigma}{\sqrt{N}}$

Method 1 Take $2N$ samples, error $\frac{\sigma}{\sqrt{2N}}$

Method 2 Take $N + N_{\text{antithetical}}$ samples, error $\frac{\sigma^2}{\sqrt{N}}$

### 7.2.2 Importance Sampling

$I(\phi) = \int_0^1 \phi(x)dx$, written in a different way is

$$\int_0^1 \frac{\phi(x)}{p(x)} p(x)dx$$

Where $p$ is a positive function on $[0,1]$ s.t. $\int_0^1 p(y)dy = 1$ . Notice $p(x)$ is a density function.

Suppose we can randomly sample variable $x$ from the distribution defined by $p$. This means $P(X \in A) = \int_A p(x)dx$.
**Rule:**

1. Sample $x_1, x_2, \ldots, x_n$
2. Form the MC cost estimate, $I_N(\phi) = \frac{1}{N} \sum_{n=1}^{N} \frac{\phi(x_n)}{p(x_n)}$

The variance of this process is

$$\hat{\sigma}^2 = \int_0^1 (\frac{\phi(x)}{p(x)} - I(f))^2 p(x)dx$$

Error is proportional to $\frac{\hat{\sigma}}{\sqrt{N}}$ We want $p$ s.t. $\hat{\sigma} < \sigma_\phi$ $I(f)$ is just a number. We would like $p$ close to $\frac{\phi}{I(\phi)}$. But we don't know what $I(\phi)$ is.