Last Week ...

- Introduction to R
  - Programming Language

- Introduction to R functions
  - Standard Statistical Analyses
  - Useful Graphics Toolkit

- Contrast to C, which provides very basic built-in functionality
For 815 Students

- Group Project
  - 3 or 4 students

- An exemplar anonymized data set will be provided to each group
  - Should not be used outside class

- Groups will be assigned to ensure a balance of computing experience
Project I
Correlated Quantitative Measures

- Consider a sample of individuals measured for a set of correlated continuous variables
  - E.g., blood pressure, weight, hip diameter...

- Model multivariate normal data
  - Estimate the number of mixture components
  - Estimate the properties of mixture components
Project I - Tasks

- Evaluate Multivariate Normal Likelihood

- Likelihood optimization
  - Likely to require MCMC methods

- Evaluate Neighbor Joining Trees as an Alternative
Consider a sample of individuals measured for a set of correlated discrete variables
- E.g. DNA bases at specific genes / locations

Model data using a mixture of multinomials
- Estimate the number of mixture components
- Estimate the properties of mixture components
- Identify variables that provide the most (and the least) discrimination between components
Consider pairs of relatives measured for a specific trait, with some observations partially censored
  - E.g. blood pressure when some individuals are taking medication

Assume that a measure of the degree of genetic similarity between individuals is also available

Estimate the contributions of genes and environment to the original, uncensored, trait
Project III - Tasks

- Read in complex data set
  - Sorting, Searching

- Evaluate Bivariate Normal Likelihood
  - Use quadrature for integration

- Likelihood optimization
815 Students

- If you prefer one of these projects...

- E-mail me today
  - I’ll try to accommodate your preference

- Group assignments available on Wednesday
  - Should plan to meet with me by October 6
Today

- Strategies for comparing algorithms
- Common relationships between algorithm complexity and input data
- Compare two simple search algorithms
Objectives

- Framework for
  - Empirical Testing
  - Theoretical Analysis

- Highlight performance characteristics of algorithms
Specific Questions

- Compare two algorithms for one task
- Predict performance in a new environment
  - If we had a computer that was 10x faster and could handle 10x more data, how would approach perform?
- Set values of algorithm parameters
Two Common Mistakes

- Ignore performance of algorithm
  - Shun faster algorithms to avoid complexity in program
  - Instead, wait for simple $N^2$ algorithms, when $N \log N$ alternatives exist of modest complexity available

- Too much weight on performance of algorithm
  - Improving a very fast program is not worth it
  - Spending too much time tinkering with code is rarely good use of time
Empirical analysis

- Given two algorithms … which is better?

- Run both
  - Say, algorithm A takes 3 seconds
  - Say, algorithm B takes 30 seconds

- Empirical studies may not always be practical
  - Some algorithms may take too long to run!
Choices of Input Data

- Actual data
  - Measures performance in use

- Random data
  - Generic approach, may not be representative

- Perverse data
  - Attempt worst case analysis
Limitations of Empirical Analysis

- Quality of implementation
  - Is our favored implementation coded more carefully than another?

- Extraneous factors
  - Compiler
  - Machine
  - Computer system
Limitations of Empirical Analysis

- Requires a working program

- Theoretical analysis is an alternative
  - Estimate potential gains

- Predict effectiveness relative to new algorithms or computers (that may not yet exist)
Theoretical Analysis

- Predict performance of algorithm based on theoretical properties
- “Independent” of actual implementation
- Several constructs occur frequently in algorithm analysis
Limitations of Theoretical Analysis

- Efficiency can depend on compiler
- Efficiency may fluctuate with input data
- Some algorithms are not well understood
The idea...

- Given a code fragment

  ```
  #Find parent of node i
  i = a[i];
  ```

- Consider how many times it is executed
- But not how long each execution takes
Two typical analyses

- Average-case for random input
- Worst-case

- Are these representative of real world problems?
  - Check with empirical predictions…
The Primary Parameter N

- Examples
  - Degree of polynomial
  - Number of characters in a string
  - Size of file to be sorted
  - Number of input data items
  - Some other abstract measure of problem size

- With multiple parameters, we can often hold one of them constant
## Running time as a function of $N$

<table>
<thead>
<tr>
<th>$f(N)$</th>
<th>Description</th>
<th>Running time when $N$ doubles…</th>
</tr>
</thead>
<tbody>
<tr>
<td>$1$</td>
<td>constant</td>
<td>-</td>
</tr>
<tr>
<td>$\log N$</td>
<td>logarithmic</td>
<td>constant increase</td>
</tr>
<tr>
<td>$N$</td>
<td>linear</td>
<td>doubles</td>
</tr>
<tr>
<td>$N \log N$</td>
<td>log-linear</td>
<td>more than doubles</td>
</tr>
<tr>
<td>$N^2$</td>
<td>quadratic</td>
<td>increases fourfold</td>
</tr>
<tr>
<td>$N^3$</td>
<td>cubic</td>
<td>increases eightfold</td>
</tr>
<tr>
<td>$2^N$</td>
<td>exponential</td>
<td>running time squares</td>
</tr>
</tbody>
</table>
Running time as a function of $N$

- Multiple terms may be involved
  - e.g. $N + N \log N$

- Typically, we ignore
  - Smaller terms
  - Constant coefficient
  - Focus on inner loop

- In rare cases, smaller terms and constant coefficient will be important
### Time to Solve Large Problem

<table>
<thead>
<tr>
<th>operations per second</th>
<th>Problem Size N = 1,000,000</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>N</td>
</tr>
<tr>
<td>10^6</td>
<td>seconds</td>
</tr>
<tr>
<td>10^9</td>
<td>instant</td>
</tr>
<tr>
<td>10^{12}</td>
<td>instant</td>
</tr>
</tbody>
</table>
## Time to Solve Huge Problem

<table>
<thead>
<tr>
<th>operations per second</th>
<th>Problem Size $N = 1,000,000,000$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$N$</td>
</tr>
<tr>
<td>$10^6$</td>
<td>hours</td>
</tr>
<tr>
<td>$10^9$</td>
<td>seconds</td>
</tr>
<tr>
<td>$10^{12}$</td>
<td>instant</td>
</tr>
</tbody>
</table>
Big-Oh Notation

- Algorithm is $O(N)$ or $O(N \log N)$
  - Common statement
  - What does it mean?

- Summarizes performance for large $N$

- Focuses on leading terms of expression describing running time
**Big-Oh Notation**

- Consider function \( g(N) \)

- It is said to be \( O(f(N)) \)

- If there exist \( c_0 \) and \( N_0 \) such that:
  - \( N > N_0 \) implies \( c_0f(N) > g(N) \)
From N to Running Time...

- Common relationships
  - \( N^2 \)
  - \( \log N \)
  - \( N \log N \)
  - \( N \)
  - \( CN \)

- Describe examples of how these arise
- Cost of running program is \( C_N \)
O(N²)

- Loop through input successively, eliminate one item at a time

\[
C_N = C_{N-1} + N \quad \text{for } N \geq 2, \quad C_1 = 1
\]

\[
= C_{N-2} + (N - 1) + N
\]

\[
= 1 + 2 + \ldots + (N - 1) + N
\]

\[
= \frac{N(N + 1)}{2}
\]
**O(log N)**

- Recursive program, halves input in one step

\[
C_{2^n} = C_{2^{n-1}} + 1 \quad \text{for } N \geq 2, \quad C_1 = 1
\]

\[
= C_{2^{n-2}} + 1 + 1
\]

\[
= C_{2^{n-3}} + 3
\]

\[
\vdots
\]

\[
= C_{2^0} + n
\]

\[
= n + 1
\]

\[
N = 2^n
\]
O(N log N)

- Recursive program, processes each item, splits input into two halves, examines each one...

\[
C_N = 2C_{N/2} + N \quad \text{for } N \geq 2, \quad C_1 = 0
\]

\[
C_{2^n} = 2C_{2^{n-1}} + 2^n
\]

\[
\frac{C_{2^n}}{2^n} = \frac{2C_{2^{n-1}} + 2^n}{2^n}
\]

\[
= \frac{C_{2^{n-1}}}{2^{n-1}} + 1
\]

\[
= \frac{C_{2^{n-2}}}{2^{n-2}} + 1 + 1
\]

\[
\ldots
\]

\[
= n
\]
O(2N)

- Halves input, must examine each item...

\[ C_N = C_{N/2} + N \quad \text{for } N \geq 2, \quad C_1 = 1 \]

\[ = N + \frac{N}{2} + \frac{N}{4} + \frac{N}{8} + \ldots \]

\[ \approx 2N \]
Application

- Analysis of two search algorithms
- Consider a set of items
  - Evaluate functions to decide whether a particular item is present…
int search(int a[], int value, int start, int stop) {
    // Variable declarations
    int i;

    // Search through each item
    for (i = start; i <= stop; i++)
        if (value == a[i])
            return i;

    // Search failed
    return -1;
}
Sequential Search Properties

- **Algorithm:**
  - Look through array sequentially, until we find a match

- **Average cost**
  - If match found: \( \frac{N}{2} \)
  - If match not found: \( N \)

- **Actual cost depends on fraction of successful searches**
Better Sequential Search

- If items are sorted...

- Stop unsuccessful search early, when we reach item with higher value
  - Cost for unsuccessful searches is now $N/2$

- Overall, algorithm is still $O(N)$
Binary Search

```c
int search(int a[], int value, int start, int stop)
{
    while (stop >= start)
    {
        // Find midpoint
        int mid = (start + stop) / 2;

        // Compare midpoint to value
        if (value == a[mid])
            return mid;

        // Reduce input in half!!!
        if (value > a[mid])
            { start = mid + 1; }
        else
            { stop = mid - 1; }
    }

    // Search failed
    return -1;
}
```
Binary Search Properties

- Algorithm:
  - Halve number of items to consider with each comparison

- Worst-case cost
  - Maximum cost is never greater than \( \log_2 N \)

- Much better than sequential search, but even better methods exist!
## Sequential vs. Binary Search

<table>
<thead>
<tr>
<th>N</th>
<th>S (1,000)</th>
<th>B (1,000)</th>
<th>S (10,000)</th>
<th>B (10,000)</th>
<th>S (100,000)</th>
<th>B (100,000)</th>
</tr>
</thead>
<tbody>
<tr>
<td>125</td>
<td>1</td>
<td>1</td>
<td>13</td>
<td>2</td>
<td>130</td>
<td>20</td>
</tr>
<tr>
<td>250</td>
<td>3</td>
<td>0</td>
<td>25</td>
<td>2</td>
<td>251</td>
<td>22</td>
</tr>
<tr>
<td>500</td>
<td>5</td>
<td>0</td>
<td>49</td>
<td>3</td>
<td>492</td>
<td>23</td>
</tr>
<tr>
<td>1250</td>
<td>13</td>
<td>0</td>
<td>128</td>
<td>3</td>
<td>1276</td>
<td>25</td>
</tr>
<tr>
<td>2500</td>
<td>26</td>
<td>1</td>
<td>267</td>
<td>3</td>
<td>*</td>
<td>28</td>
</tr>
</tbody>
</table>

Timings in seconds, for M searches in table of N elements
Summary

- Outline principles for analysis of algorithms
- Introduced some common relationships between N and running time
- Described two simple search algorithms
Further Reading

- Read chapter 2 of Sedgewick
Tip of the Day: Defensive Programming

- Document code
  - Indicate intended purpose
  - Specify required inputs
  - Always indicate author

- Check for error conditions