Introduction to Computational Statistics
(STAT-598G)

Jian Zhang
Department of Statistics

Purdue University
Course Information

Topics to be covered:

- C and R programming; data structures and algorithms
- Numerical methods; statistical applications

Prerequisites:

- STAT 516, STAT 517 or equivalent
- Some programming experience (not necessarily in C or R)

Grading: homework 40%, project 20%, final 40%

Website: http://www.stat.purdue.edu/~jianzhan/STAT598G/
Computing Environment

C programming

- GCC software (with gdb for debugging)
- Linux/Unix/Mac operating system
- Any editor (vi, emacs, pico, etc)

R programming

- R software package
- Available in Windows, Linux/Unix and Mac
- The R project [http://www.r-project.org/](http://www.r-project.org/)
Expected Outcomes

1. Should be able to write R and C code, and be able to modify/understand existing R and C code.

2. Understand basic data structures and algorithms in statistical applications.

3. Understand basic numerical methods such as optimization, sampling, etc.

4. Learn or review some statistics topics such as bootstrap, linear/logistic regression, EM, Bayesian inference.
Dynamic Programming

Example

Find the shortest path from A to J in the following graph:

![Graph Diagram]

Question: Is there a more efficient way than simply computing each possible path (can be exponentially many)?
Iterative Methods

Example
Consider the linear equation

\[ Ax = b \]

where \( A \in \mathbb{R}^{n \times n} \), \( x \in \mathbb{R}^{n \times 1} \) and \( b \in \mathbb{R}^{n \times 1} \). For simplicity we further assume \( A \) is non-singular, and in this case the solution is trivial: \( x = A^{-1}b \).

Question: What if \( A \) is sparse and \( n \) is so large that you cannot even load it into the memory? Also recall that matrix inverse takes \( O(n^3) \)!
Example
Consider the linear model \( y_i = x_i^T \theta + \epsilon_i \) where \( \epsilon_i \sim N(0, \sigma^2) \) iid, \( y_i \in \mathbb{R}, \ x_i \in \mathbb{R}^{p \times 1} \) and \( \theta \in \mathbb{R}^{p \times 1} \) is the unknown parameter we want to estimate. The least square estimation tries to find \( \theta \) using

\[
\arg\min_{\theta} \sum_{i=1}^{n} (y_i - x_i^T \theta)^2 = \arg\min \| Y - X \theta \|^2.
\]

Question: How to efficiently find \( \theta \) if \( X \) is huge and sparse? How to find a \( \theta \) which also satisfies some constraint such as \( \| \theta \| \leq t? \)
Example

Given an IID sample $X_1, \ldots, X_n \sim N(\mu, \sigma^2)$ with $\mu$ unknown. One obvious estimator of the $\mu$ is the sample mean

$$\bar{X} = \frac{1}{n} \sum_{i=1}^{n} X_i$$

whose expectation and variance can be easily computed as $\mathbb{E}[\bar{X}] = \mu$ and $\mathbb{V}[\bar{X}] = \frac{1}{n} \sigma^2$.

Question: We can also use the sample median $\tilde{X}$ as an estimator of $\mu$, and suppose we know $\mathbb{E}[\tilde{X}] = \mu$. Now, what is $\mathbb{V}[\tilde{X}]$?
Example

Given a function $g : \mathbb{R}^{p \times 1} \mapsto \mathbb{R}$ and a multivariate probability density function $f(x)$ for a random vector $X \in \mathbb{R}^{p \times 1}$. We want to compute

$$\mathbb{E}[g(X)] = \int f(x)g(x)dx,$$

which can be solved by multivariate calculus.

Question: If the above integration does not have a close form solution, what are the other choices? If we have $x_1, \ldots, x_n \sim f(x)$, how about $\frac{1}{n} \sum_{i=1}^{n} g(x_i)$?
Numerical Integration

Example

Given a function $g : \mathbb{R} \mapsto \mathbb{R}$ and a univariate probability density function $f(x)$ for a random variable $X$. We still want to compute

$$E[g(X)] = \int f(x)g(x) \, dx,$$

which can be solved by calculus or sampling.

Question: For this low dimensional integration, can we do better? Here “better” means to obtain the same accurate solution faster, or a more accurate solution with the same amount of time.
Expectation Maximization

Example

Given an IID sample \( x_1, \ldots, x_n \sim N(\mu, \Sigma) \) with both \( \mu \) and \( \Sigma \) unknown. We can use sample mean and sample covariance (or maximum likelihood) to estimate \( \mu \) and \( \Sigma \):

\[
\hat{\mu} = \frac{1}{n} \sum_{i=1}^{n} x_i
\]

and

\[
\hat{\Sigma} = \frac{1}{n} \sum_{i=1}^{n} (x_i - \hat{\mu})(x_i - \hat{\mu})^T.
\]

Question: If \( x_1, \ldots, x_n \sim iid N(\mu_1, \Sigma_1) \) and \( y_1, \ldots, y_m \sim iid N(\mu_2, \Sigma_2) \) and you only observe their union \( z_1, \ldots, z_{m+n} \) (without out which belongs to which group), can you still estimate \( \mu_1, \Sigma_1 \) and \( \mu_2, \Sigma_2 \)?
Bayesian Inference

Example
In Bayesian data analysis we often have a prior $\theta \sim \pi(\theta)$ and likelihood $p(D|\theta)$. Now the posterior can be computed as

$$p(\theta|D) = \frac{\pi(\theta)p(D|\theta)}{\int \pi(\theta)p(D|\theta)d\theta}. $$

The posterior distribution $p(\theta|D)$ is of central interest in Bayesian data analysis.

Question: How to obtain the posterior distribution when the integration cannot be easily computed? What if $\theta$ is of high dimension?