

*Maximum Likelihood Estimation  
for Allele Frequencies*

**Biostatistics 666**

## Previous Series of Lectures: Introduction to Coalescent Models

---

- **Computationally efficient framework**
  - Alternative to forward simulations
  - Amenable to analytical solutions
- **Predictions about sequence variation**
  - Number of polymorphisms
  - Frequency of polymorphisms
  - Distribution of polymorphisms across haplotypes

# Coalescent Models: Key Ideas

---

- Proceed backwards in time
- Genealogies shaped by
  - Population size
  - Population structure
  - Recombination rates
- Given a particular genealogy ...
  - Mutation rate predicts variation

## Next Series of Lectures

---

- Estimating allele and haplotype frequencies from genotype data
  - Maximum likelihood approach
  - Application of an E-M algorithm
- Challenges
  - Using information from related individuals
  - Allowing for non-codominant genotypes
  - Allowing for ambiguity in haplotype assignments

# Objective: Parameter Estimation

---

- Learn about **population** characteristics
  - E.g. allele frequencies, population size
- Using a specific **sample**
  - E.g. a set sequences, unrelated individuals, or even families

# Maximum Likelihood

---

- A general framework for estimating model parameters
- Find the set of parameter values that maximize the probability of the observed data
- Applicable to many different problems

## Example: Allele Frequencies

---

- Consider...
  - A sample of  $n$  chromosomes
  - $X$  of these are of type “a”
  - Parameter of interest is allele frequency...

$$L(p | n, X) = \binom{n}{X} p^X (1-p)^{n-X}$$

## Evaluate for various parameters

---

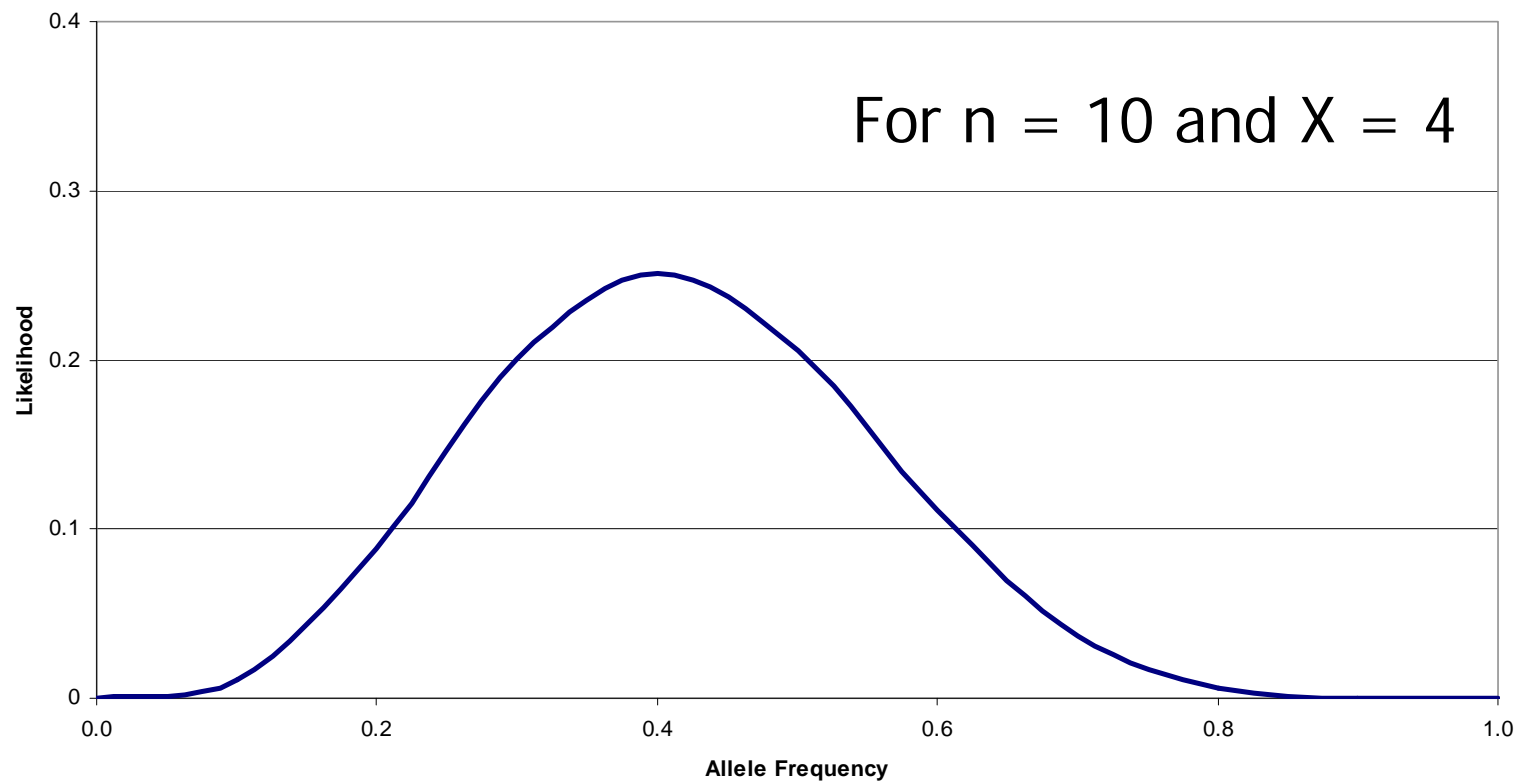
<b>p</b>	<b>1-p</b>	<b>L</b>
0.0	1.0	0.000
0.2	0.8	0.088
0.4	0.6	0.251
0.6	0.4	0.111
0.8	0.2	0.006
1.0	0.0	0.000

For  $n = 10$  and  $X = 4$



# Likelihood Plot

---



## In this case

---

- The likelihood tells us the data is most probable if  $p = 0.4$
- The likelihood curve allows us to evaluate alternatives...
  - Is  $p = 0.8$  a possibility?
  - Is  $p = 0.2$  a possibility?

## Example: Estimating $4N\mu$

---

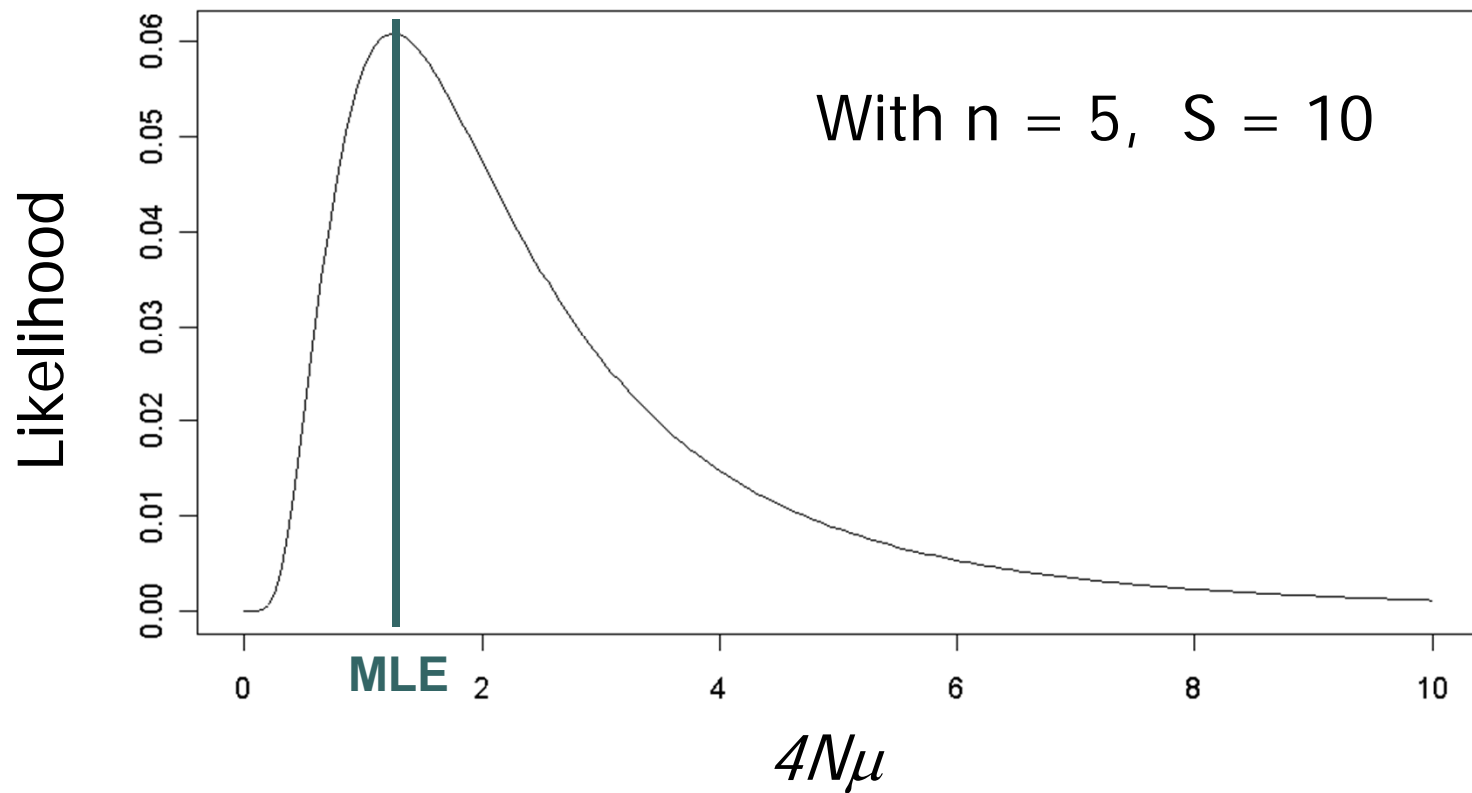
- Consider  $S$  polymorphisms in sample of  $n$  sequences...

$$L(\theta | n, S) = P_n(S | \theta)$$

- Where  $P_n$  is calculated using the  $Q_n$  and  $P_2$  functions defined previously

# Likelihood Plot

---



## Maximum Likelihood Estimation

---

- Two basic steps...
  - a) Write down likelihood function
$$L(\theta | x) \propto f(x | \theta)$$
  - b) Find value of  $\hat{\theta}$  that maximizes  $L(\theta | x)$
- In principle, applicable to any problem where a likelihood function exists

# MLEs

---

- Parameter values that maximize likelihood
  - $\theta$  where observations have maximum probability
- Finding MLEs is an optimization problem
- How do MLEs compare to other estimators?

# Comparing Estimators

---

- How do MLEs rate in terms of ...
  - **Unbiasedness**
  - **Consistency**
  - **Efficiency**
- For a review, see Garthwaite, Jolliffe, Jones (1995) *Statistical Inference*, Prentice Hall

# Analytical Solutions

---

- Write out log-likelihood ...

$$\ell(\theta | data) = \ln L(\theta | data)$$

- Calculate derivative of likelihood

$$\frac{d\ell(\theta | data)}{d\theta}$$

- Find zeros for derivative function



# Information

---

- The second derivative is also extremely useful

$$I_{\theta} = -E \left[ \frac{d^2 \ell(\theta | data)}{d\theta^2} \right]$$

$$V_{\hat{\theta}} = \frac{1}{I_{\theta}}$$

- The speed at which log-likelihood decreases
- Provides an asymptotic variance for estimates

## Allele Frequency Estimation ...

---

- When individual chromosomes are observed this does not seem tricky...
- What about with genotypes?
- What about with parent-offspring pairs?

## Coming up ...

---

- We will walk through allele frequency estimation in three distinct settings:
  - Samples single chromosomes ...
  - Samples of unrelated Individuals ...
  - Samples of parents and offspring ...

# I. Single Alleles Observed

---

- Consider...
  - A sample of  $n$  chromosomes
  - $X$  of these are of type “a”
  - Parameter of interest is allele frequency...

$$L(p | n, X) = \binom{n}{X} p^X (1-p)^{n-X}$$

## Some Notes

---

- The following two likelihoods are just as good:

$$L(p; X, n) = \binom{n}{X} p^X (1-p)^{n-X}$$

$$L(p; x_1, x_2, \dots, x_n, n) = \prod_{i=1}^n p^{x_i} (1-p)^{1-x_i}$$

- For ML estimation, constant factors in likelihood don't matter

# Analytic Solution

---

- The log-likelihood

$$\ln L(p | n, X) = \ln \binom{n}{X} + X \ln p + (n - X) \ln(1 - p)$$

- The derivative

$$\frac{d \ln L(p | X)}{dp} = \frac{X}{p} - \frac{n - X}{1 - p}$$

- Find zero ...

## Samples of Individual Chromosomes

---

- The natural estimator (where we count the proportion of sequences of a particular type) and the MLE give identical solutions
- Maximum likelihood provides a justification for using the “natural” estimator

## II. Genotypes Observed

---

- Use notation  $n_{ij}$  to denote the number of individuals with genotype  $i / j$
- Sample of  $n$  individuals

Genotype Counts				
Genotype	$A_1A_1$	$A_1A_2$	$A_2A_2$	Total
Observed Counts	$n_{11}$	$n_{12}$	$n_{22}$	$n=n_{11}+n_{12}+n_{22}$
Frequency	$p_{11}$	$p_{12}$	$p_{22}$	1.0



# Allele Frequencies by Counting...

---

- A natural estimate for allele frequencies is to calculate the proportion of individuals carrying each allele

Allele Counts			
Genotype	A <sub>1</sub>	A <sub>2</sub>	Total
Observed Counts	$n_1 = 2n_{11} + n_{12}$	$n_2 = 2n_{22} + n_{12}$	$2n = n_1 + n_2$
Frequency	$p_1 = n_1 / 2n$	$p_2 = n_2 / 2n$	1.0

## MLE using genotype data...

---

- Consider a sample such as ...

Genotype Counts				
Genotype	A <sub>1</sub> A <sub>1</sub>	A <sub>1</sub> A <sub>2</sub>	A <sub>2</sub> A <sub>2</sub>	Total
Observed Counts	n <sub>11</sub>	n <sub>12</sub>	n <sub>22</sub>	n=n <sub>11</sub> +n <sub>12</sub> +n <sub>22</sub>
Frequency	p <sub>11</sub>	p <sub>12</sub>	p <sub>22</sub>	1.0

- The likelihood as a function of allele frequencies is ...

$$L(p;n) = \frac{n!}{n_{11}!n_{12}!n_{22}!} (p^2)^{n_{11}} (2pq)^{n_{12}} (q^2)^{n_{22}}$$

## Which gives...

---

- Log-likelihood and its derivative

$$\ell = \ln L = (2n_{11} + n_{12}) \ln p_1 + (2n_{22} + n_{12}) \ln(1 - p_1) + C$$

$$\frac{d\ell}{dp_1} = \frac{2n_{11} + n_{12}}{p_1} - \frac{2n_{22} + n_{12}}{(1 - p_1)}$$

- Giving the MLE as ...

$$\hat{p}_1 = \frac{(2n_{11} + n_{12})}{2(n_{11} + n_{12} + n_{22})}$$

## Samples of Unrelated Individuals

---

- Again, natural estimator (where we count the proportion of alleles of a particular type) and the MLE give identical solutions
- Maximum likelihood provides a justification for using the “natural” estimator

### III. Parent-Offspring Pairs

---

Parent	Child			
	$A_1A_1$	$A_1A_2$	$A_2A_2$	
$A_1A_1$	$a_1$	$a_2$	0	$a_1+a_2$
$A_1A_2$	$a_3$	$a_4$	$a_5$	$a_3+a_4+a_5$
$A_2A_2$	0	$a_6$	$a_7$	$a_6+a_7$
	$a_1+a_3$	$a_2+a_4+a_6$	$a_5+a_7$	N pairs

# Probability for Each Observation

---

Parent	Child			
	$A_1A_1$	$A_1A_2$	$A_2A_2$	
$A_1A_1$				
$A_1A_2$				
$A_2A_2$				
				1.0

# Which gives...

---

$$\ln L =$$

$$p_2 = 1 - p_1$$

$$B = 3a_1 + 2(a_2 + a_3) + a_4 + (a_5 + a_6)$$

$$C = (a_2 + a_3) + a_4 + 2(a_5 + a_6) + 3a_7$$

$$\hat{p}_1 = \frac{B}{(B + C)}$$

## Samples of Parent Offspring-Pairs

---

- The natural estimator (where we count the proportion of alleles of a particular type) and the MLE no longer give identical solutions
- In this case, we expect the MLE to be more accurate



# Comparing Sampling Strategies

---

- We can compare sampling strategies by calculating the information for each one

$$I_{\theta} = -E \left[ \frac{d^2 \ell(\theta | data)}{d\theta^2} \right]$$

$$V_{\hat{\theta}} = \frac{1}{I_{\theta}}$$

- Which one to you expect to be most informative?

## How informative is each setting?

---

- Single chromosomes  $Var(p) = \frac{pq}{N}$
- Unrelated individuals  $Var(p) = \frac{pq}{2N}$
- Parent offspring pairs  $Var(p) = \frac{pq}{3N - a_4}$

## Other Likelihoods

---

- Allele frequencies when individuals are...
  - Diagnosed for Mendelian disorder
  - Genotyped at two neighboring loci
  - Phenotyped for the ABO blood groups
- Many other interesting problems...
- ... but some have no analytical solution

# Today's Summary

---

- Examples of Maximum Likelihood
- Allele Frequency Estimation
  - Allele counts
  - Genotype counts
  - Pairs of Individuals

## Take home reading

---

- Excoffier and Slatkin (1995)
  - *Mol Biol Evol* **12**:921-927
- Introduces the E-M algorithm
- Widely used for maximizing likelihoods in genetic problems

# *Properties of Estimators*

For Review

## Unbiasedness

---

- An estimator is unbiased if

$$E(\hat{\theta}) = \theta$$

$$\text{bias}(\hat{\theta}) = E(\hat{\theta}) - \theta$$

- Multiple unbiased estimators may exist
- Other properties may be desirable

# Consistency

---

- An estimator is consistent if

$$P(|\hat{\theta} - \theta| > \varepsilon) \rightarrow 0 \text{ as } n \rightarrow \infty$$

- for any  $\varepsilon$
- Estimate converges to true value in probability with increasing sample size



## Mean Squared Error

---

- MSE is defined as

$$\begin{aligned}MSE(\hat{\theta}) &= E\left(\{(\hat{\theta} - \bar{\theta}) + (\bar{\theta} - \theta)\}^2\right) \\ &= \text{var}(\hat{\theta}) + \text{bias}(\hat{\theta})^2\end{aligned}$$

- If  $MSE \rightarrow 0$  as  $n \rightarrow \infty$  then the estimator must be consistent
  - The reverse is not true

## Efficiency

---

- The relative efficiency of two estimators is the ratio of their variances

$$\text{if } \frac{\text{var}(\hat{\theta}_2)}{\text{var}(\hat{\theta}_1)} > 1 \text{ then } \hat{\theta}_1 \text{ is more efficient}$$

- Comparison only meaningful for estimators with equal biases

# Sufficiency

---

- Consider...
  - Observations  $X_1, X_2, \dots, X_n$
  - Statistic  $T(X_1, X_2, \dots, X_n)$
- $T$  is a sufficient statistic if it includes all information about parameter  $\theta$  in the sample
  - Distribution of  $X_i$  conditional on  $T$  is independent of  $\theta$
  - Posterior distribution of  $\theta$  conditional on  $T$  is independent of  $X_i$

## Minimal Sufficient Statistic

---

- There can be many alternative sufficient statistics.
- A statistic is a minimal sufficient statistic if it can be expressed as a function of every other sufficient statistic.

# Typical Properties of MLEs

---

- **Bias**
  - Can be biased or unbiased
- **Consistency**
  - Subject to regularity conditions, MLEs are consistent
- **Efficiency**
  - Typically, MLEs are asymptotically efficient estimators
- **Sufficiency**
  - Often, but not always
- Cox and Hinkley, 1974

*Strategies for Likelihood  
Optimization*

For Review

## Generic Approaches

---

- Suitable for when analytical solutions are impractical
- Bracketing
- Simplex Method
- Newton-Rhapson

# Bracketing

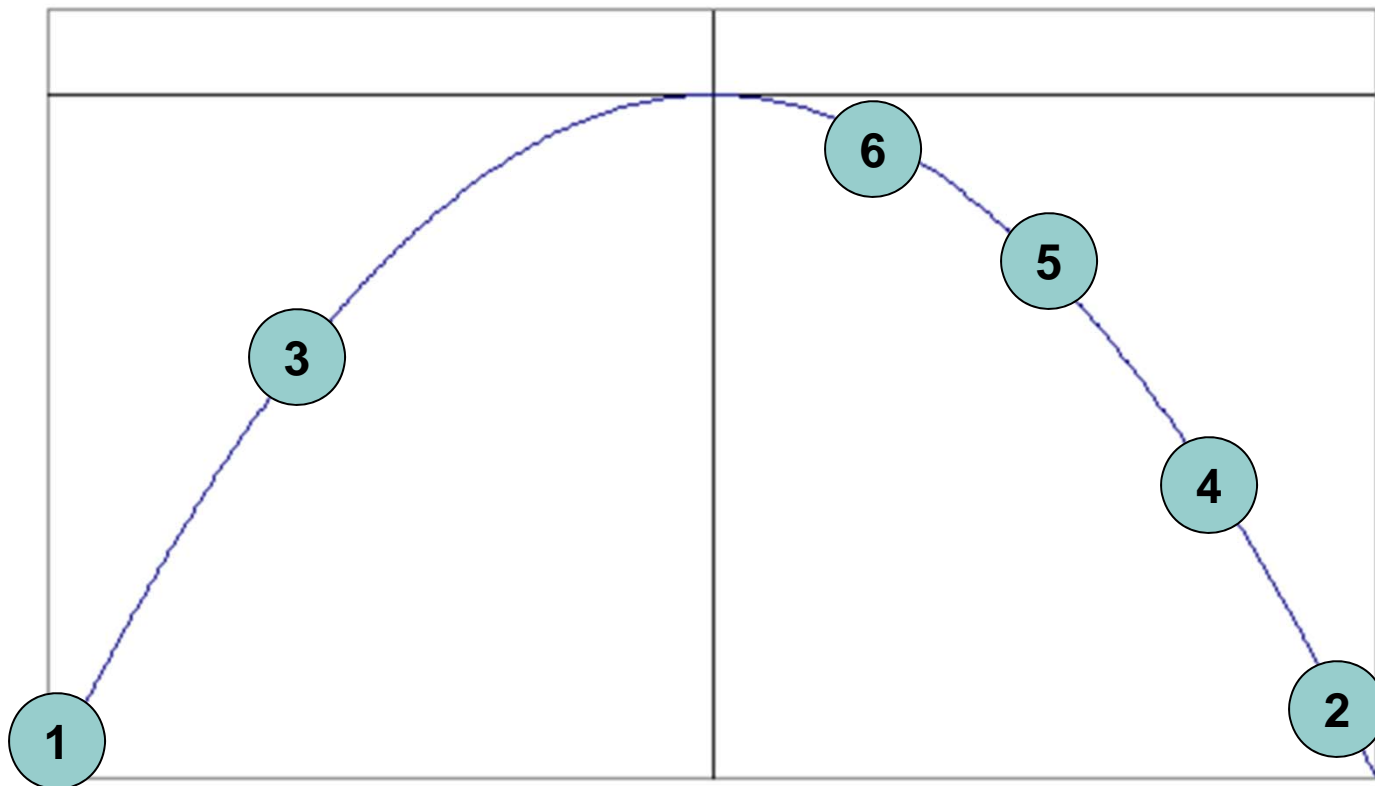
---

- Find 3 points such that
  - $\theta_a < \theta_b < \theta_c$
  - $L(\theta_b) > L(\theta_a)$  and  $L(\theta_b) > L(\theta_c)$
- Search for maximum by
  - Select trial point in interval
  - Keep maximum and flanking points



# Bracketing

---



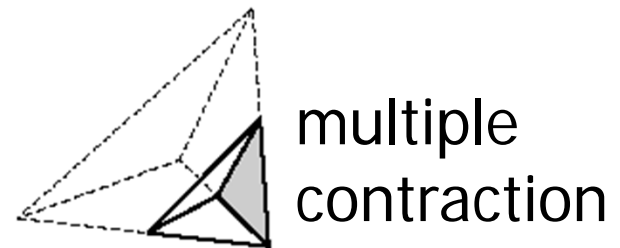
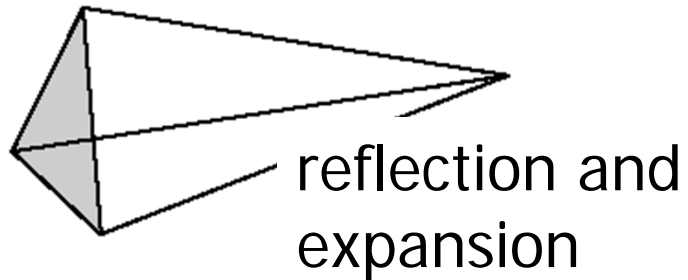
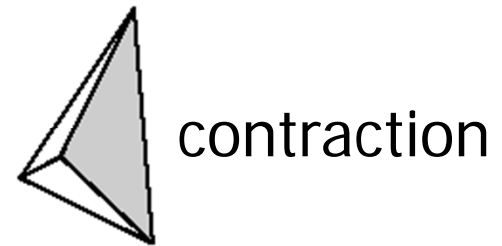
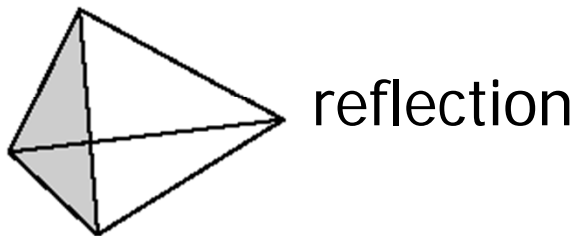
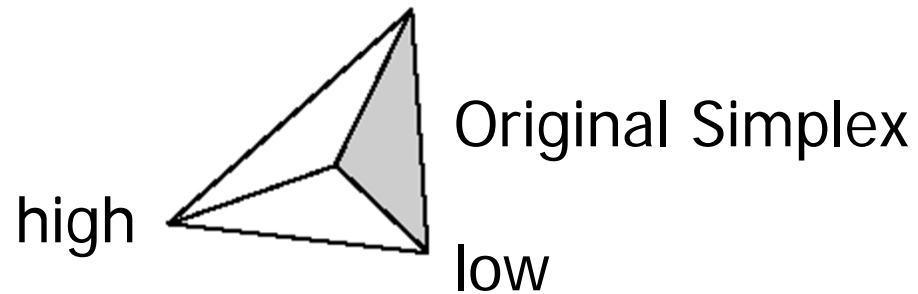
# The Simplex Method

---

- Calculate likelihoods at simplex vertices
  - Geometric shape with  $k+1$  corners
  - E.g. a triangle in  $k = 2$  dimensions
- At each step, move the high vertex in the direction of lower points

# The Simplex Method II

---



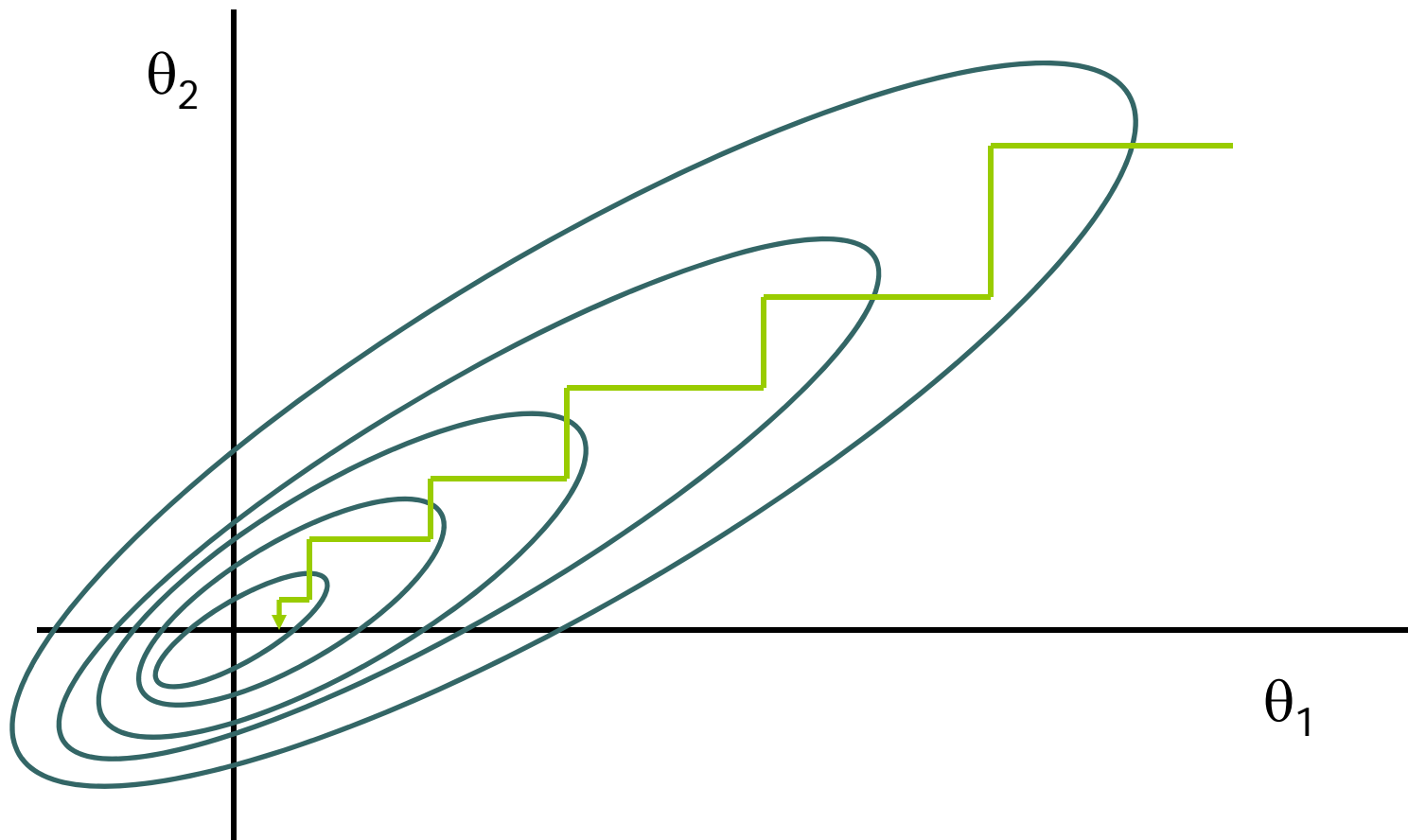
## One parameter maximization

---

- Simple but inefficient approach
- Consider
  - Parameters  $\theta = (\theta_1, \theta_2, \dots, \theta_k)$
  - Likelihood function  $L(\theta; x)$
- Maximize  $\theta$  with respect to each  $\theta_i$  in turn
  - Cycle through parameters

# The Inefficiency...

---



# Steepest Descent

---

- Consider
  - Parameters  $\theta = (\theta_1, \theta_2, \dots, \theta_k)$
  - Likelihood function  $L(\theta; \mathbf{x})$

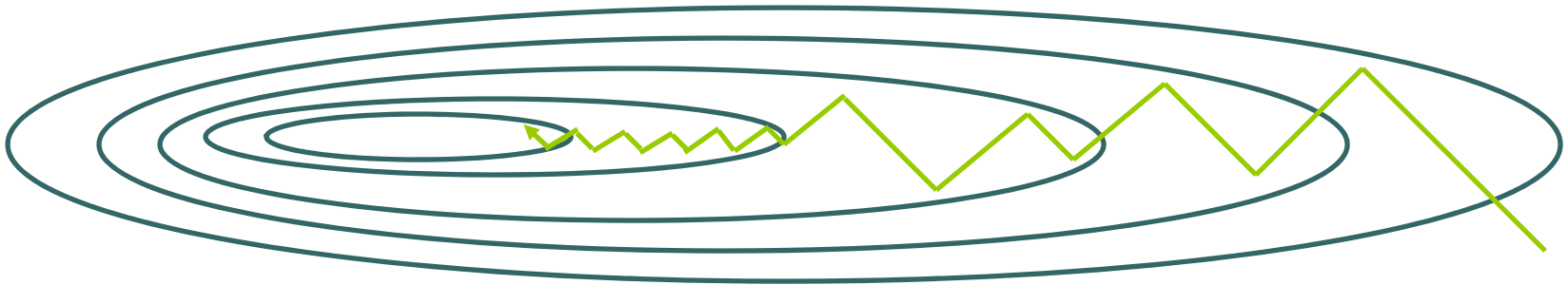
- Score vector

$$S = \frac{d \ln(L)}{d\theta} = \left( \frac{d \ln(L)}{d\theta_1}, \dots, \frac{d \ln(L)}{d\theta_k} \right)$$

- Find maximum along  $\theta + \delta S$

Still inefficient...

---



Consecutive steps are perpendicular!

# Local Approximations to Log-Likelihood Function

---

In the neighborhood of  $\boldsymbol{\theta}_i$

$$\ell(\boldsymbol{\theta}) \approx \ell(\boldsymbol{\theta}_i) + S(\boldsymbol{\theta} - \boldsymbol{\theta}_i) - \frac{1}{2} (\boldsymbol{\theta} - \boldsymbol{\theta}_i)^t \mathbf{I}_\theta (\boldsymbol{\theta} - \boldsymbol{\theta}_i)$$

where

$\ell(\boldsymbol{\theta}) = \ln L(\boldsymbol{\theta})$  is the loglikelihood function

$\mathbf{S} = d\ell(\boldsymbol{\theta}_i)$  is the score vector

$\mathbf{I}_\theta = -d^2\ell(\boldsymbol{\theta}_i)$  is the observed information matrix



## Newton's Method

---

Maximize the approximation

$$\ell(\boldsymbol{\theta}) \approx \ell(\boldsymbol{\theta}_i) + \mathbf{S}(\boldsymbol{\theta} - \boldsymbol{\theta}_i) - \frac{1}{2} (\boldsymbol{\theta} - \boldsymbol{\theta}_i)^t \mathbf{I}(\boldsymbol{\theta} - \boldsymbol{\theta}_i)$$

by setting its derivative to zero...

$$\mathbf{S} - \mathbf{I}(\boldsymbol{\theta} - \boldsymbol{\theta}_i) = \mathbf{0}$$

and get a new trial point

$$\boldsymbol{\theta}_{i+1} = \boldsymbol{\theta}_i + \mathbf{I}^{-1}\mathbf{S}$$

# Fisher Scoring

---

- Use expected information matrix instead of observed information:

$$E\left[-\frac{d^2\ell(\theta)}{d\theta^2}\right]$$

instead of

$$-\frac{d^2\ell(\theta | data)}{d\theta^2}$$

## Compared to Newton-Rhapson:

Converges faster when estimates are poor.

Converges slower when close to MLE.