CS340 Machine learning
BN3
Outline

• Monte Carlo integration
• Genetic linkage analysis
Monte Carlo integration

• Suppose we want to evaluate the integral
  \[ E[h(X)] = I = \int h(x)p(x)dx \]

• In low dimensions, we can use numerical integration (eg. quadrature: in matlab, quad, dblquad, triplequad).

• In higher dimensions, a better approach is to sample \( S \) values \( x^s \) from \( p(x) \) and then use the law of large numbers
  \[
  \hat{I} = \frac{1}{S} \sum_{s=1}^{S} h(x^s)
  \]
  which has standard error
  \[
  se = \sqrt{\frac{\hat{\sigma}^2}{S}}, \quad \hat{\sigma}^2 = \frac{1}{S-1} \sum_{s=1}^{S} (h(x^s) - \hat{I})^2
  \]
Definite integrals

• We can evaluate a definite integral by sampling uniformly within the range

\[
I = \int_{a}^{b} h(x) = (b - a) \int h(x)p(x)dx
\]

\[
p(x) = U(x|a,b) = \frac{1}{(b-a)}I(a < x < b)
\]

\[
I \approx \frac{1}{S} \sum_{s=1}^{S} h(x^s)
\]

• Thus the method can also be applied in non-statistical settings.
Estimating $\pi$

• Area of circle is

$$I = \int_{-r}^{r} \int_{-r}^{r} I(x^2 + y^2 \leq r^2) \, dx \, dy$$

so $\pi = I/r^2$. Let $h(x, y) = I(x^2 + y^2 \leq r^2)$

$$I = (b_x - a_x)(b_y - a_y) \int \int h(x, y)p(x)p(y) \, dx \, dy$$

$$= (2r)(2r) \int \int h(x, y)p(x)p(y) \, dx \, dy$$

$$= 4r^2 \int \int h(x, y)p(x)p(y) \, dx \, dy$$

$$\approx 4r^2 \frac{1}{S} \sum_s h(x^s, y^s)$$
Estimating $\pi$

- **Matlab**

```matlab
r=2;
S=5000;
xs = unifrnd(-r,r,S,1);
ys = unifrnd(-r,r,S,1);
rs = xs.^2 + ys.^2;
inside = (rs <= r^2);
samples = 4*(r^2)*inside;
Ihat = mean(samples);
piHat = Ihat/(r^2);
se = sqrt(var(samples)/S)

$\hat{\pi} = 3.1416, \quad se = 0.09$
```
• Monte Carlo integration
• Genetic linkage analysis
Pedigree analysis

Family tree

Maternal and paternal copies of each gene
Genotype, phenotype
**Penetrance model**

- Phenotype = bloodtype, genotype = alleles

<table>
<thead>
<tr>
<th>$G_p$</th>
<th>$G_m$</th>
<th>$p(P = a)$</th>
<th>$p(P = b)$</th>
<th>$p(P = o)$</th>
<th>$p(P = ab)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>a</td>
<td>a</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>a</td>
<td>b</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>a</td>
<td>o</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>b</td>
<td>a</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>b</td>
<td>b</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>b</td>
<td>o</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>o</td>
<td>a</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>o</td>
<td>b</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>o</td>
<td>o</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
</tbody>
</table>
Haplotypes

locus \( \ell = 1 \)

father \( \pi_i = 1 \)

mother \( \mu_i = 2 \)

individual \( i = 3 \)

\[
p(G_{i,\ell}^p | G_{\pi_i,\ell}^p, G_{\pi_i,\ell}^m, S_{i,\ell}^p) = \begin{cases} 
\delta(G_{i,\ell}^p - G_{\pi_i,\ell}^p) & \text{if } S_{i,\ell}^p = 0 \\
\delta(G_{i,\ell}^p - G_{\pi_i,\ell}^m) & \text{if } S_{i,\ell}^p = 1
\end{cases}
\]
Cross-over

\[
\begin{array}{c|cc}
S_{i,\ell-1} & p(S_{i,\ell} = 0) & p(S_{i,\ell} = 1) \\
0 & 1 - \theta_{\ell} & \theta_{\ell} \\
1 & \theta_{\ell} & 1 - \theta_{\ell}
\end{array}
\]
### Linkage analysis

- Compute likelihood of data as a function of unknown gene location $X$

\[
p(D|\theta(X)) = \sum G^p_{1:n,1:L} G^m_{1:n,1:L} S^p_{1:n,1:L} S^m_{1:n,1:L} p(G^p_{1:n,1:L}, G^m_{1:n,1:L}, S^p_{1:n,1:L}, S^m_{1:n,1:L}, D|\theta(X))
\]