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

$$
\begin{aligned}
I = \int_a^b h(x) &= (b-a) \int h(x) p(x) dx \\
p(x) &= U(x|a,b) = \frac{1}{(b-a)} I(a < x < b) \\
I &\approx \frac{1}{S} \sum_{s=1}^{S} h(x^s)
\end{aligned}
$$

- Thus the method can also be applied in non-statistical settings.

Estimating π

- 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 $h(x, y) = I(x^2 + y^2 \leq r^2)$

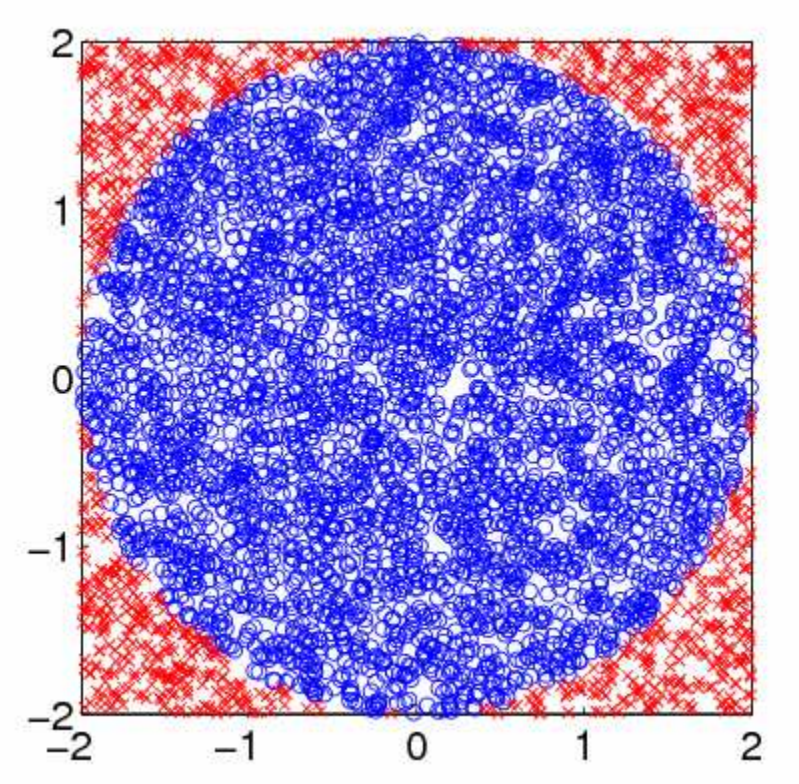
$$\begin{aligned} I &= (b_x - a_x)(b_y - a_y) \int \int h(x, y) p(x) p(y) dx dy \\ &= (2r)(2r) \int \int h(x, y) p(x) p(y) dx dy \\ &= 4r^2 \int \int h(x, y) p(x) p(y) dx dy \\ &\approx 4r^2 \frac{1}{S} \sum_s h(x^s, y^s) \end{aligned}$$

Estimating π

- 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)
```

$$\hat{\pi} = 3.1416, \quad se = 0.09$$

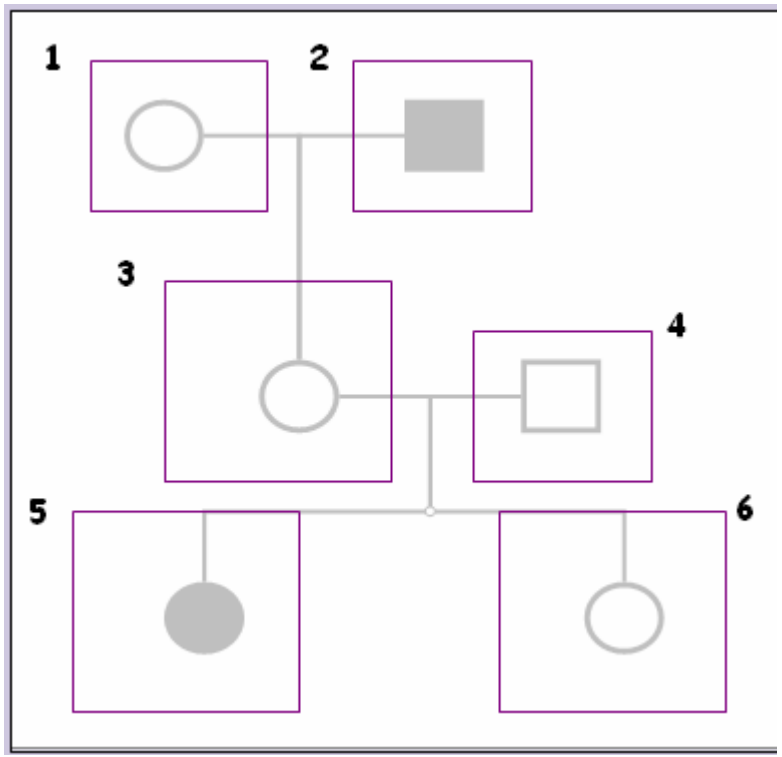


Outline

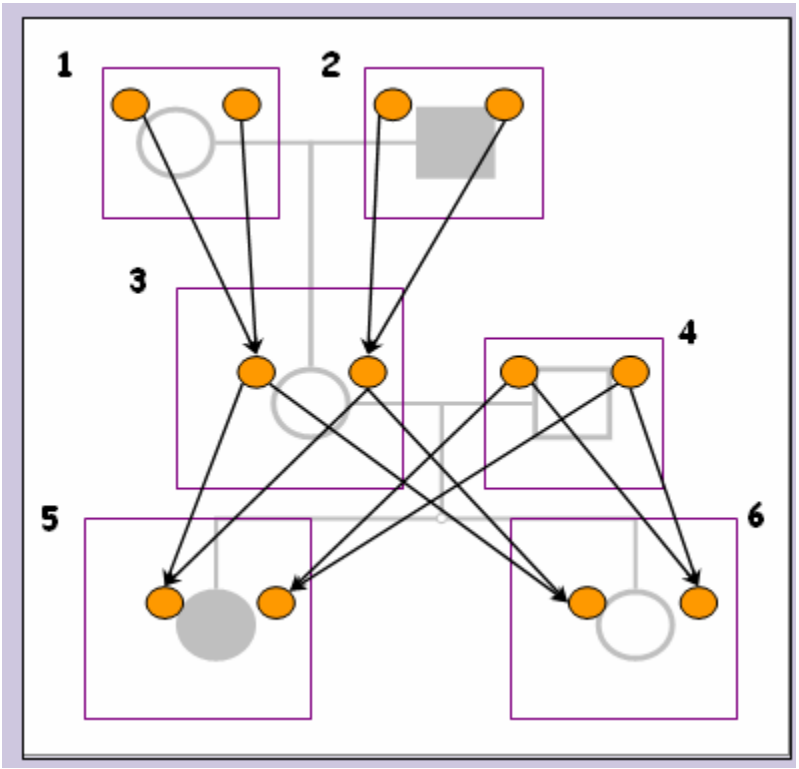
- Monte Carlo integration
- Genetic linkage analysis

Pedigree analysis

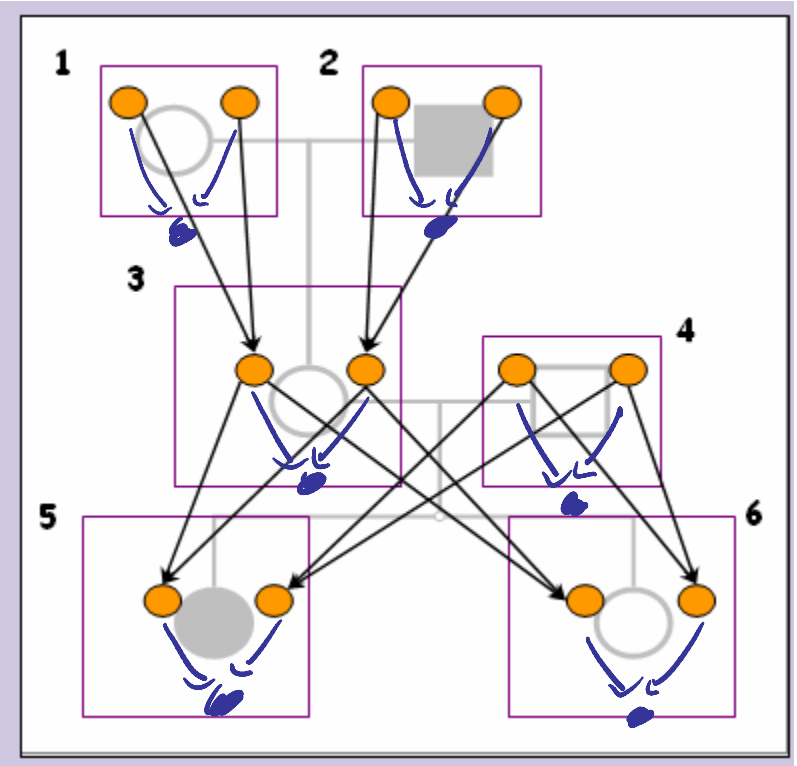
Family tree



Maternal and paternal copies of each gene



Genotype, phenotype



Penetrance model

- Phenotype = bloodtype, genotype = alleles

G_p	G_m	$p(P=a)$	$p(P=b)$	$p(P=o)$	$p(P=ab)$
a	a	1	0	0	0
a	b	0	0	0	1
a	o	1	0	0	0
b	a	0	0	0	1
b	b	0	1	0	0
b	o	0	1	0	0
o	a	1	0	0	0
o	b	0	1	0	0
o	o	0	0	1	0

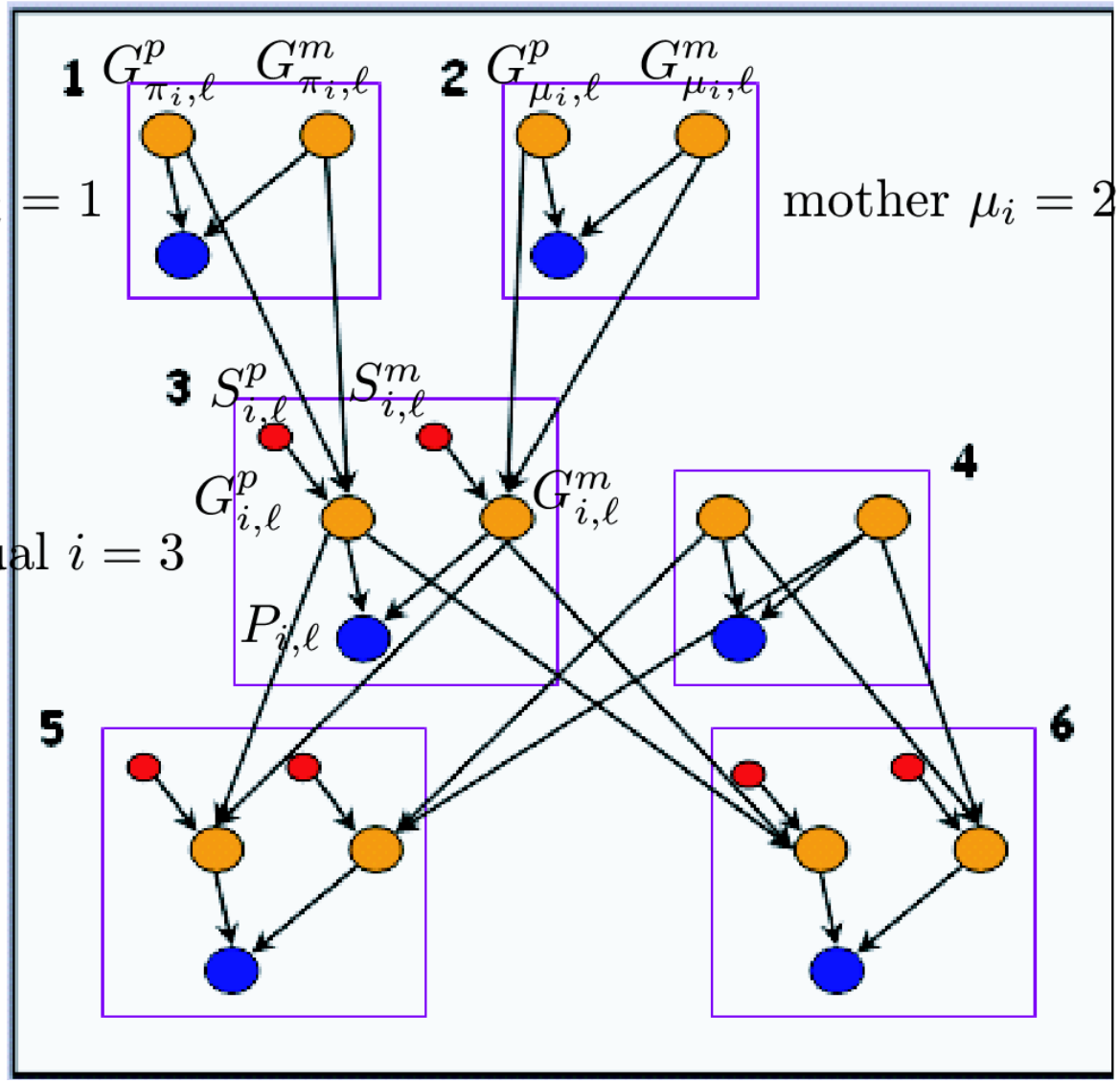
Haplotypes

locus $\ell = 1$

father $\pi_i = 1$

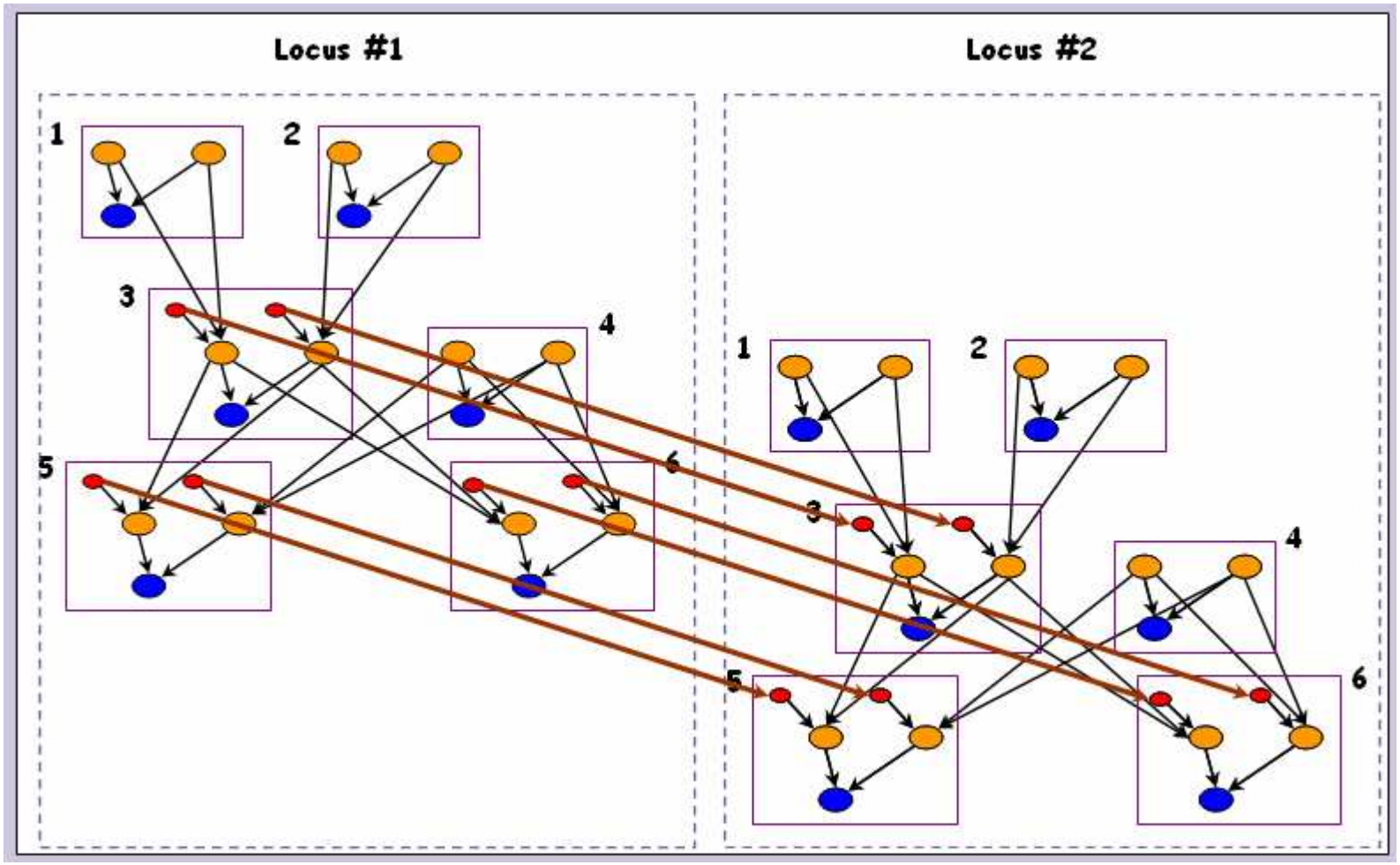
mother $\mu_i = 2$

individual $i = 3$



$$p(G^p_{i,\ell}|G^p_{\pi_i,\ell}, G^m_{\pi_i,\ell}, S^p_{i,\ell}) = \begin{cases} \delta(G^p_{i,\ell} - G^p_{\pi_i,\ell}) & \text{if } S^p_{i,\ell} = 0 \\ \delta(G^p_{i,\ell} - G^m_{\pi_i,\ell}) & \text{if } S^p_{i,\ell} = 1 \end{cases}$$

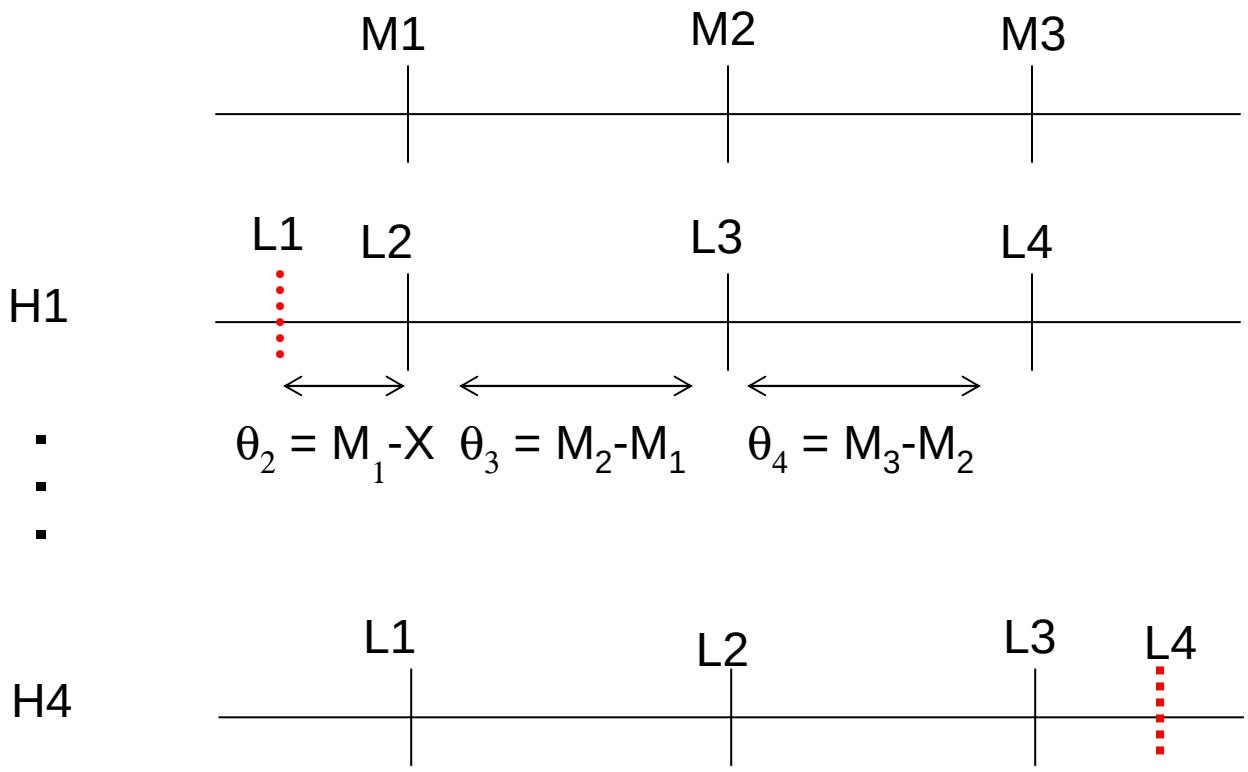
Cross-over



$S_{i,\ell-1}$	$p(S_{i,\ell}=0)$	$p(S_{i,\ell}=1)$
0	$1-\theta_\ell$	θ_ℓ
1	θ_ℓ	$1-\theta_\ell$

Linkage analysis

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



$$p(D|\boldsymbol{\theta}(X)) = \sum_{G^p_{1:n,1:L}} \sum_{G^m_{1:n,1:L}} \sum_{S^p_{1:n,1:L}} \sum_{S^m_{1:n,1:L}} p(G^p_{1:n,1:L}, G^m_{1:n,1:L}, S^p_{1:n,1:L}, S^m_{1:n,1:L}, D|\boldsymbol{\theta}(X))$$