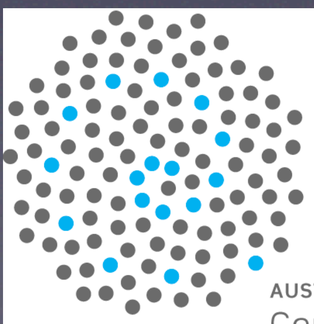
Importance sampling strategies for forwards and backwards processes in population genetics

Martin O'Hely
MASCOS
University of Queensland



AUSTRALIAN RESEARCH COUNCIL
Centre of Excellence for Mathematics and Statistics of Complex Systems

THE UNIVERSITY OF QUEENSLAND
AUSTRALIA

ARC Centre for Complex Systems
School of ITEE | The University of Queensland | ST LUCIA QLD 4069 | AUSTRALIA

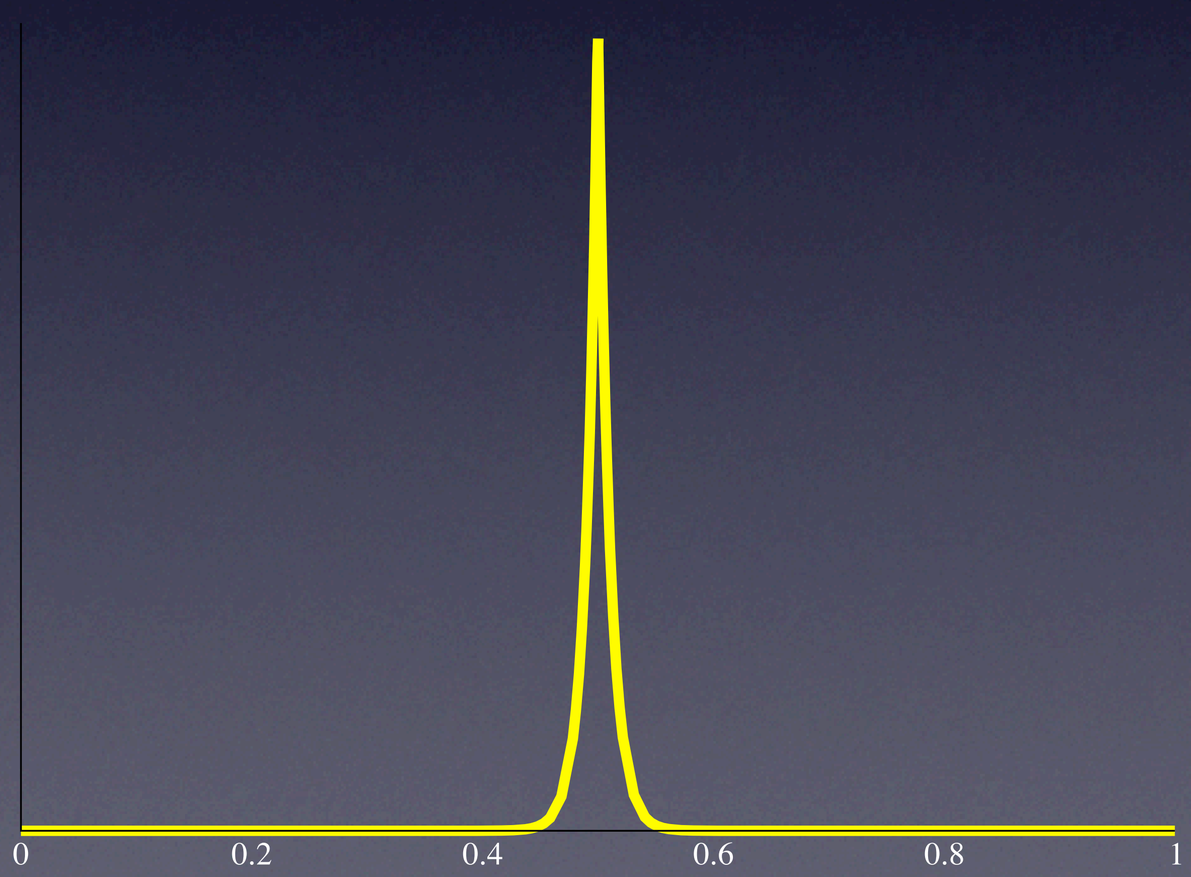
T: +61 7 3365 1003 | F: +61 7 3365 1533 | E: outreach@accs.edu.au

Importance sampling?

$$\int_A f(\mathbf{x})\, d\mathbf{x}$$

Estimate by uniformly-at-random choosing points from *A* then averaging

- Can be inefficient: *f* may only be large at a few points of *A*

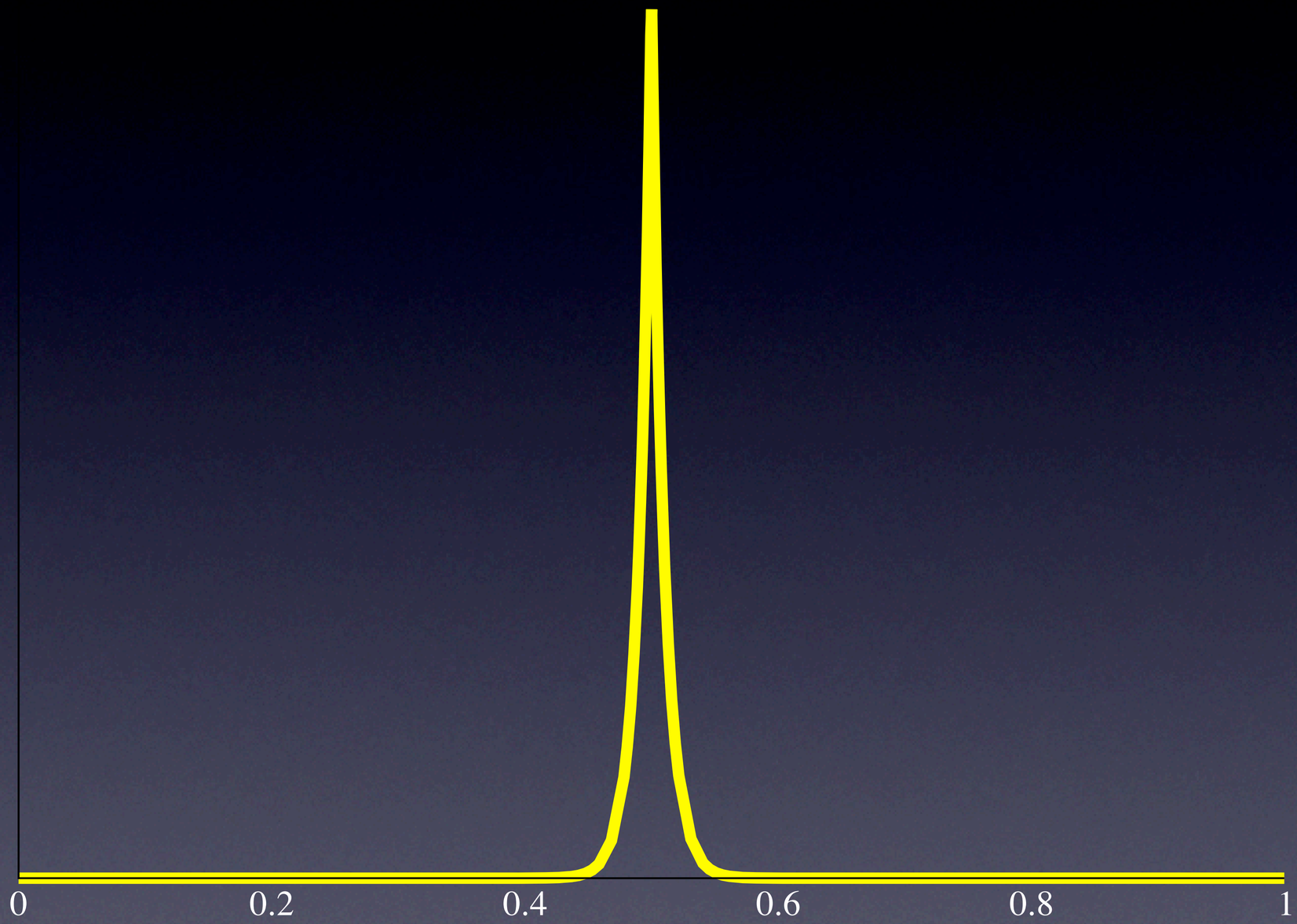


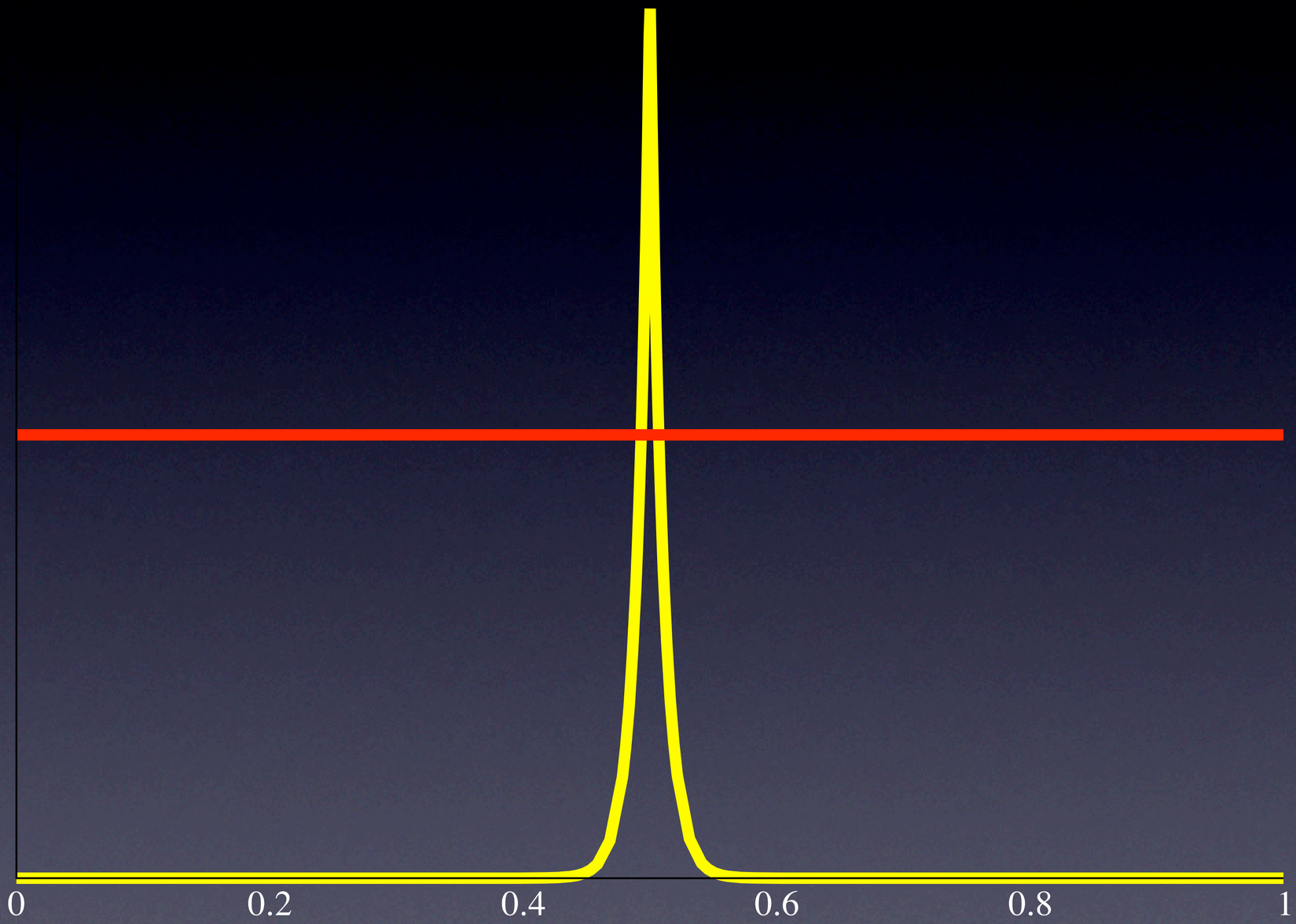
Importance sampling (IS)

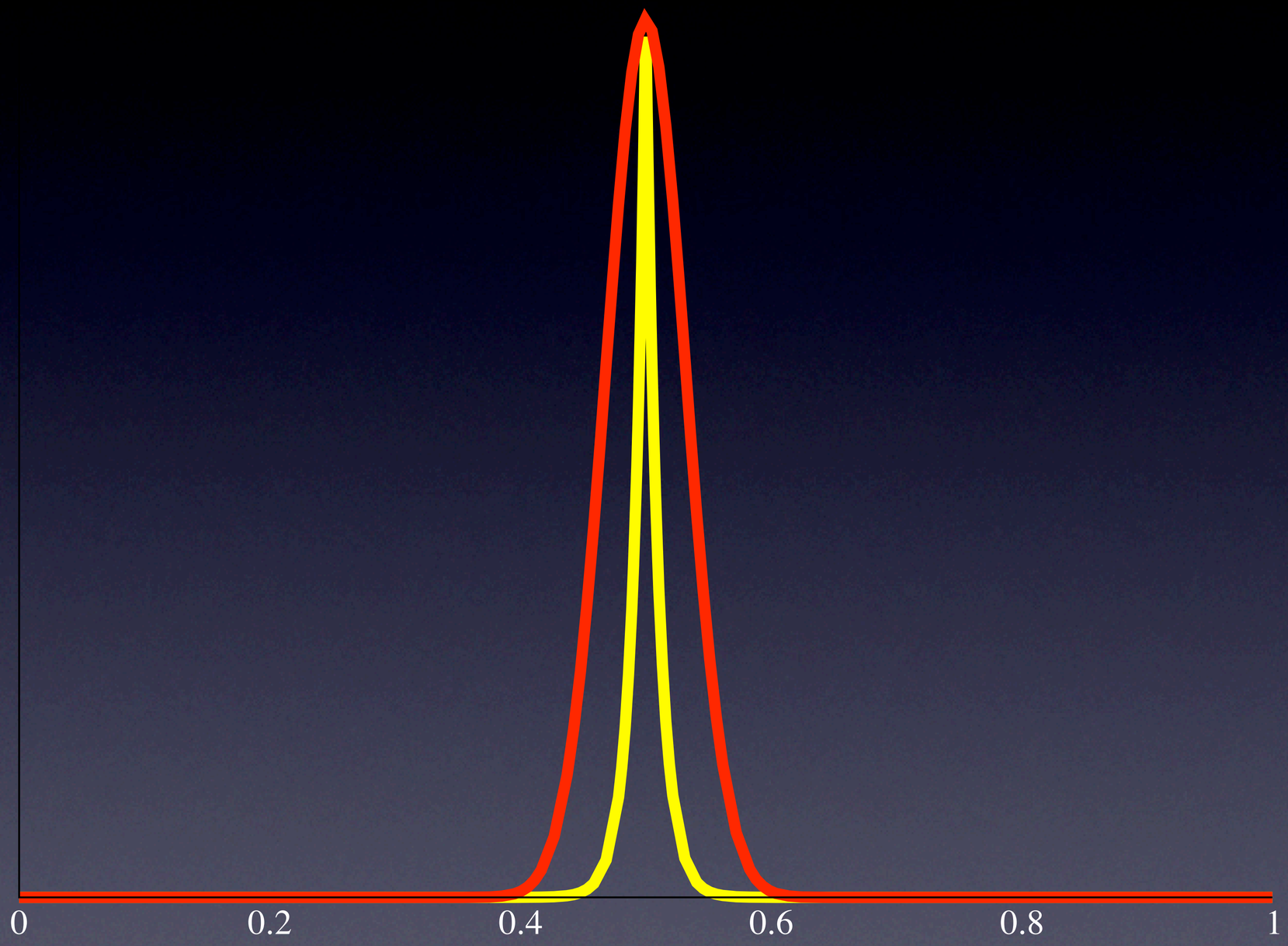
$$\int_A \frac{f(\mathbf{x})}{g(\mathbf{x})} g(\mathbf{x})\, d\mathbf{x}$$

Estimate by choosing points from A under the distribution with density g then averaging the values of $f(\mathbf{x})/g(\mathbf{x})$

- $1/g(\mathbf{x})$ is called the IS weight


0
0.2
0.4
0.6
0.8
1


0
0.2
0.4
0.6
0.8
1


0
0.2
0.4
0.6
0.8
1

Toy problem

Simple symmetric random walk on {0, 1, ... , *N*}

- Estimate: the probability that, starting at 1, the walk will hit *N* before 0.
  - It's well-known that this is 1/*N*
  - In fact, starting at *i* the answer is *i*/*N*
  - For illustrative purposes, estimate by

$$\int_{\text{paths } \pi} I_{\text{hit } N \text{ before } 0}(\pi)\, d\pi$$

(i.e. score 1 for an SSRW which hits *N*, 0 for one which hits 0)

Population genetics?

Simple symmetric random walk on $\{0, 1, \ldots, N\}$ models the spread of a neutral mutant gene in a population of size N.

- Starting at 1 reflects a mutation that arises in a single individual
- Hitting 0 is equivalent to the loss of the mutant gene from the population
- Hitting N represents the fixation of the gene in the population

Toy problem & crude strategy

Simple symmetric random walk on {0, 1, ... , *N*}

- Expected number of steps to hit 0 or *N* from 1 is

$$N - 1$$

- Variance of one run's probability estimate is

$$\frac{N - 1}{N^2}$$

- Standard deviation of *M* runs' estimate is (about)

$$\frac{1}{\sqrt{NM}}$$

An importance sampling strategy

When at 1, *always* go straight to 2

- Expected number of steps to hit **0** or ***N*** is

$$(N-1)(N-2)$$

- Variance of one run's probability estimate is

$$\frac{N^2-3N+2}{N^2(3N-2)}$$

- Standard deviation of ***M*** runs' estimate is about

$$\frac{1}{\sqrt{3NM}}$$

- Better estimate in the same time by doing (*N*-2)*M* crude runs

An importance sampling strategy which doesn't help

When at 1, *always* go straight to 2

- Expected number of steps to hit 0 or *N* is

$$(N-1)(N-2)$$

- Variance of one run's probability estimate is

$$\frac{N^2-3N+2}{N^2(3N-2)}$$

- Standard deviation of *M* runs' estimate is about

$$\frac{1}{\sqrt{3NM}}$$

- Better estimate in the same time by doing (*N*-2)*M* crude runs

An importance sampling strategy that helps but has flaws

When at i go to

$i+1$ with probability $(i+1)/(2i)$

$i-1$ with probability $(i-1)/(2i)$

- Variance of one run's probability estimate is 0, requires on average $(N+1)(N-1)/3$ steps

Flaw: this only works because these step probabilities are based on the hitting probabilities: path $i_0, i_1, \ldots, i_T=N$ has IS weight

$$\frac{i_0}{i_1}\frac{i_1}{i_2}\cdots\frac{i_{T-1}}{i_T} = \frac{i_0}{N}$$

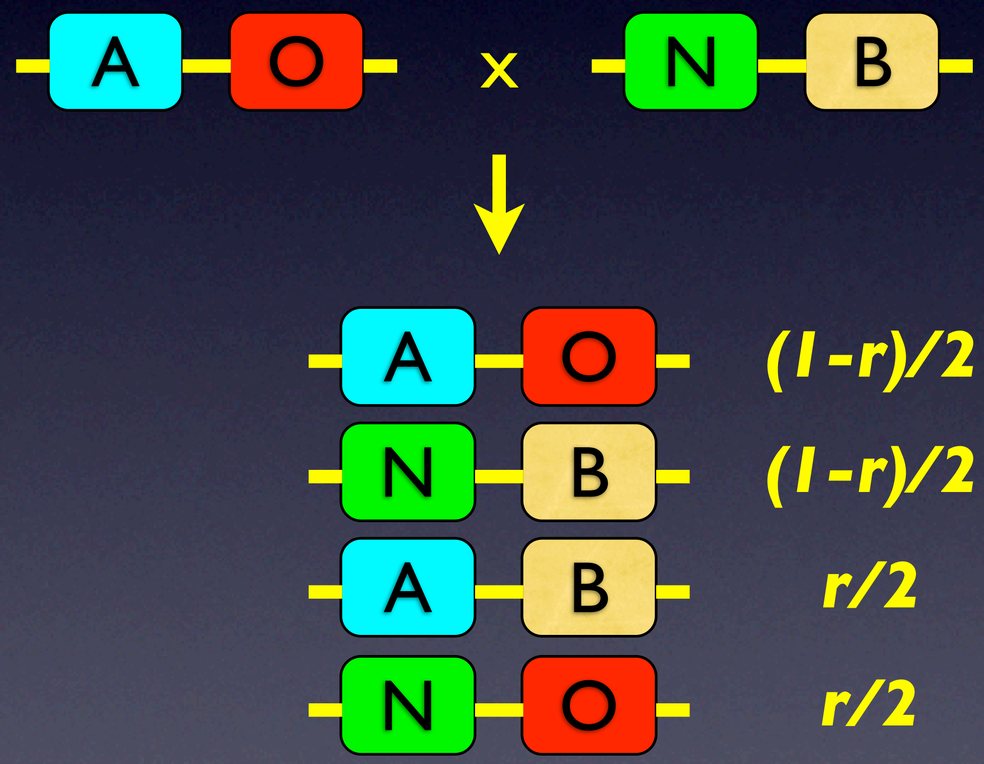
Can we get anything useful out of this zero-variance IS?

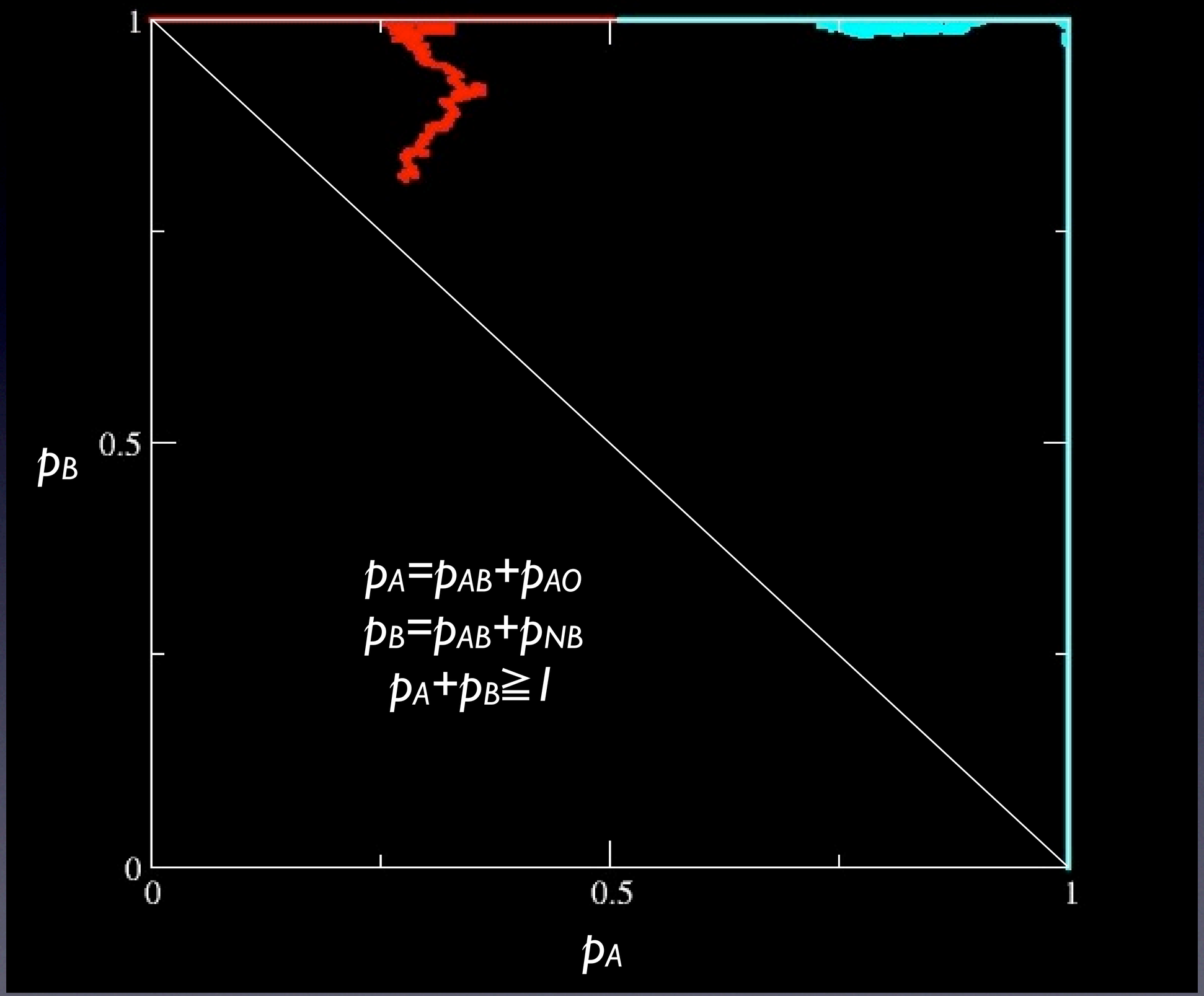
- Use a one-dimensional process as an approximation of a higher-dimensional process
- "Discovery" of the right IS scheme

Considered by C. James (UQ Maths Honours project 2004) & L. Wockner (UQ Maths Honours project 2006–7)

Approximating higher-dimensional processes

- Individuals have two parents and may be either a copy of one or a mix of the two (and maybe with mutation)
- Functional alleles A, B; non-functional N, O
- AO, NB and AB all OK
- NO can't survive
- Biologically less-than-reasonable




1
0.5
0
0
0.5
1
p_B
p_A
$p_A=p_{AB}+p_{AO}$
$p_B=p_{AB}+p_{NB}$
$p_A+p_B \geqq 1$

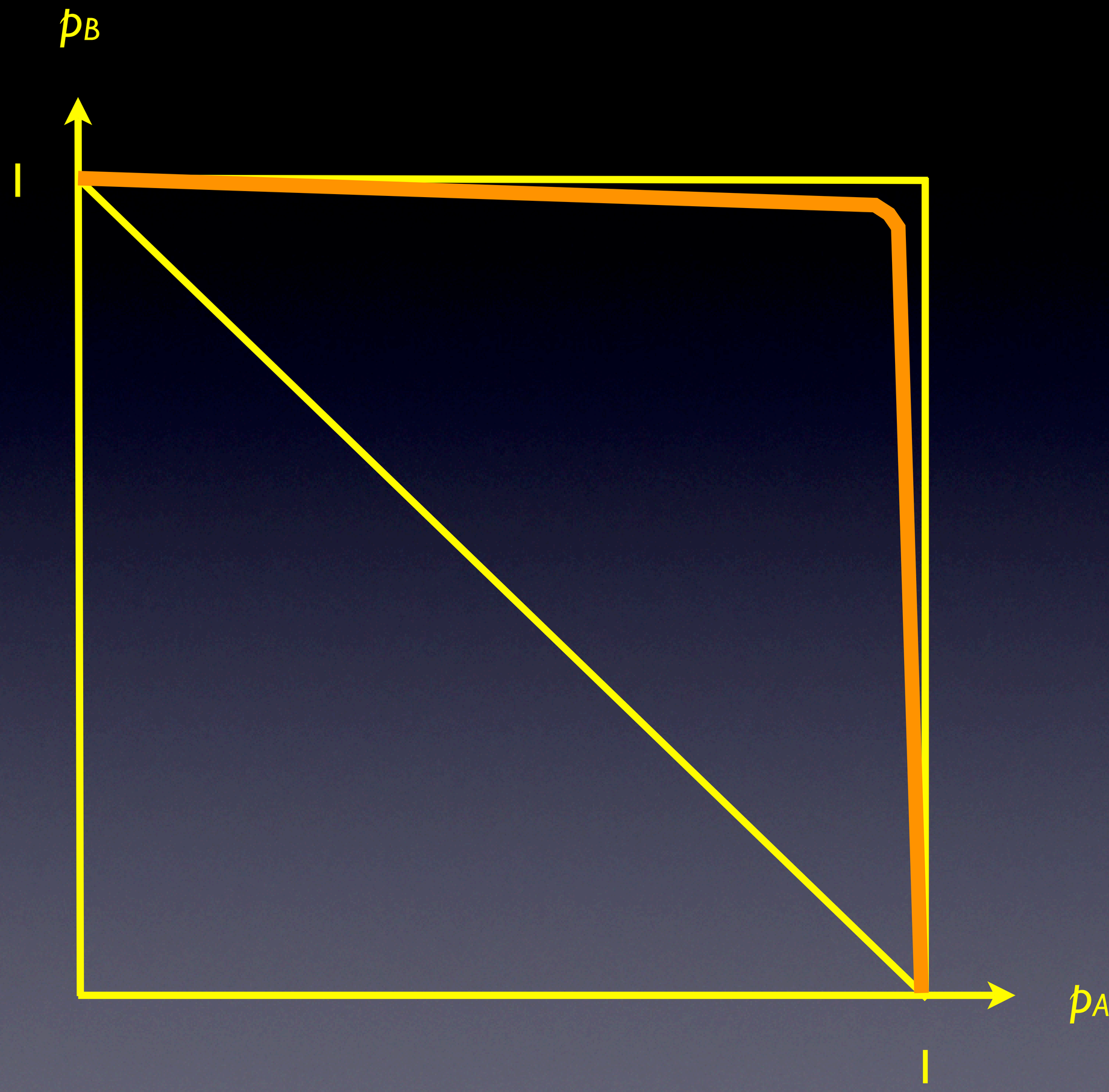
Unlinked haploid case: observations

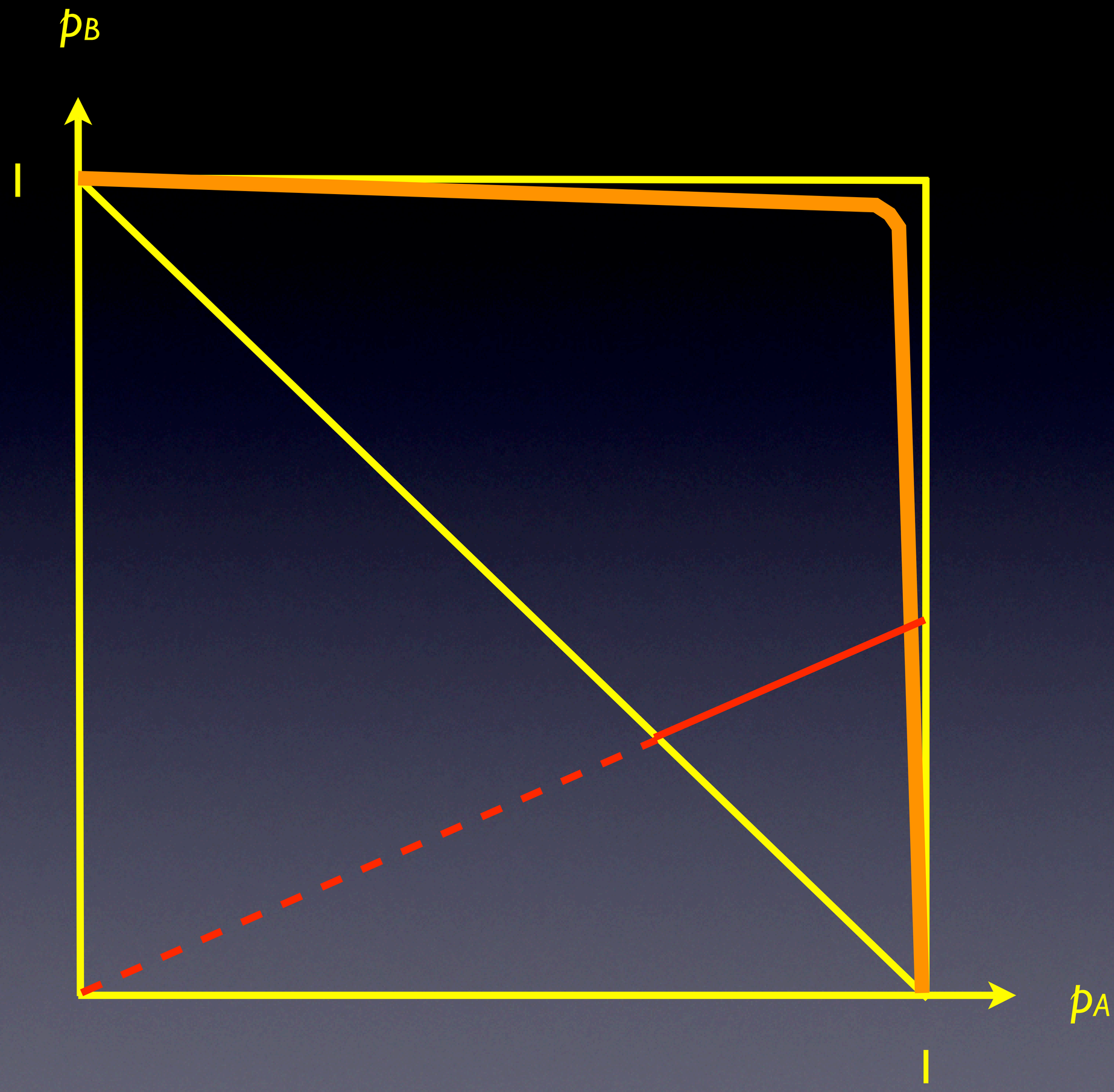
- Describe by frequencies of functional haplotypes: p_{AB}, p_{AO}, p_{NB}.
- Two dimensional but observed to collapse quickly to

$$\mu(p_A + p_B - 1) = r(1-\mu)(1-p_A)(1-p_B)$$

- Overall behaviour described by

$$\frac{p_A}{p_A + p_B}$$


p_B
I
p_A
I


p_B
I
p_A
I

Unlinked haploid case: analysis

- In a large population limit, the process remains deterministically on the hyperbola and follows a diffusion along it (via Ethier & Nagylaki 1980)
- Analysis of the diffusion leads to hitting probabilities which we can use to arrive at optimal IS
- For a finite population size, this is just an approximation
- We found that the map change probability was estimated with a relative standard deviation of 0.01% over 100 replicate runs

Diffusion approximations of Markov chains with applications to population genetics, Advances in Applied Probability **12**

Further applications: diploid case

- For a diploid version of the model we can't currently show that the requirements of Ethier & Nagylaki are fulfilled
- However we *can* say what the limiting diffusion would be *if* we could prove the technical bits, and we *can* analyze this diffusion
  - Thus we can implement an IS scheme as before
- Work in progress

Other applications

Non-nearest neighbour interactions / non-overlapping generations

- Much better efficiency for crude simulation but optimal IS may ruin this
  - Work in progress

Another IS application

- Three populations 1, 2, *H*.
- Allele frequencies

$$\mathbf{f}^k = (f_1^k, \ldots, f_I^k)$$

in population *k*=1,2;

$$\mathbf{f}^H = p\mathbf{f}^1 + (1-p)\mathbf{f}^2$$

- Samples (populations *k*=1,2,*H*)

$$\mathbf{n}^k = (n_1^k, \ldots, n_I^k)$$

- Goal: estimate likelihood of a particular *p* by assigning hybrid samples to parental populations sequentially

Crude & better solutions

- Pick a hybrid allele uniformly-at-random, assign to parental population via p
- Calculating the likelihood involves an expression

$$p^{\text{target population}}(\text{new number of chosen allele in target population})$$
$$\text{new number of all alleles in chosen population}$$

- Tempting to choose target population and allele simultaneously using the above expression as a weight.
- But simulation is a *lot* more efficient selecting the allele uniformly-at-random, then using this weight to pick the target population: in some test cases hundreds of times more efficient

Conclusions

- Importance sampling can make simulation methods vastly more efficient
- It is possible to be *too* clever: the fanciest looking method is not always the best
- It is also possible to be not clever enough: a step "in the right direction" methodologically might be a step backwards in efficiency

Acknowledgements

Phil Pollett

Leesa Wockner

Robert Cope

Lounès Chikhi

Jean-Marie Cornuet

Mark Beaumont

MASCOS

CNRS