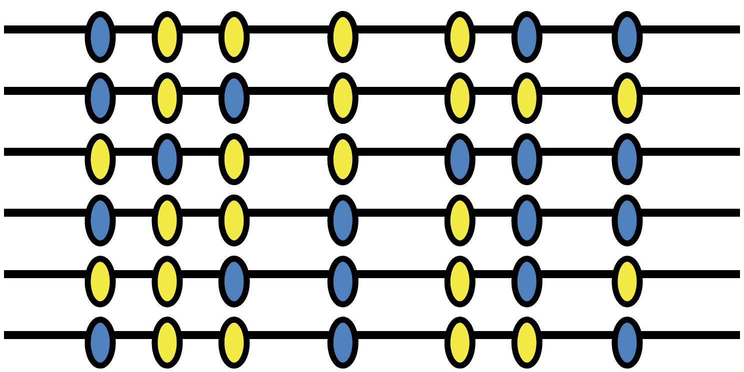
Computational Methods And The Coalescent

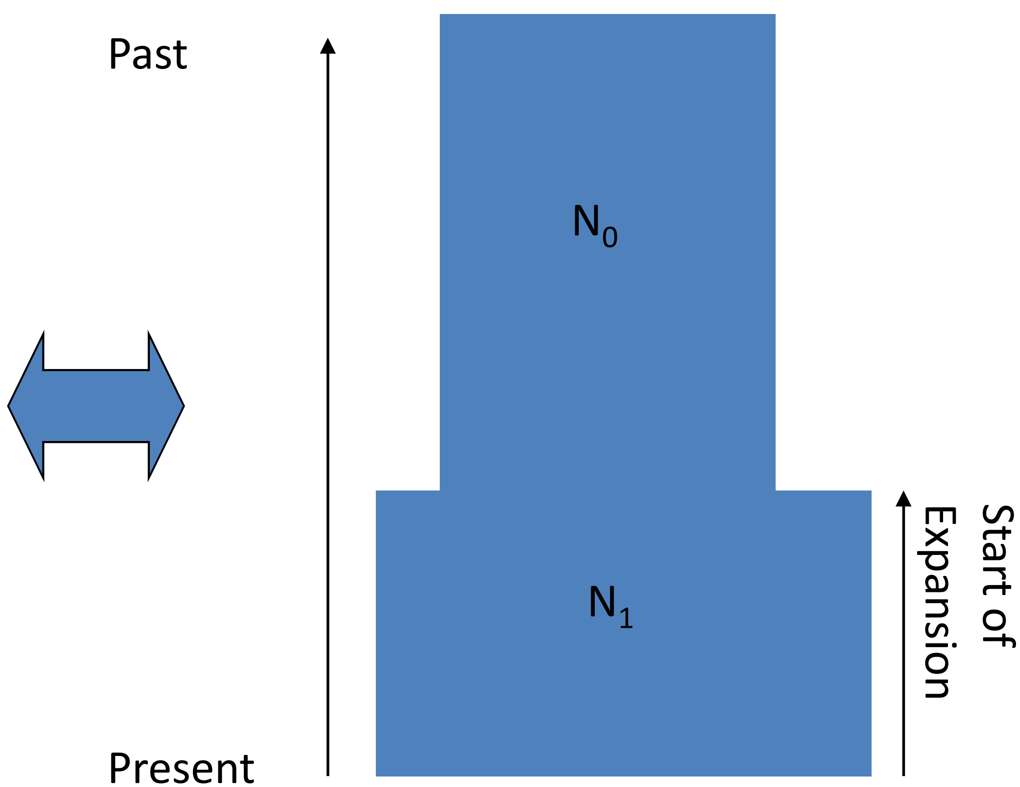
Biostatistics 666

Lecture by Guest Expert
Sebastian Zoellner

Example

We have a set of haplotypes and want to infer population growth (starting time and size increase).





General Case

Let G be the genotype data and C the set of parameters. We want to calculate P(C|G) (Bayesian) or L(C)=P(G|C) (Frequentist).

$$P(C|G) = \frac{P(G|C)P(C)}{P(G)} \propto P(G|C)P(C)$$

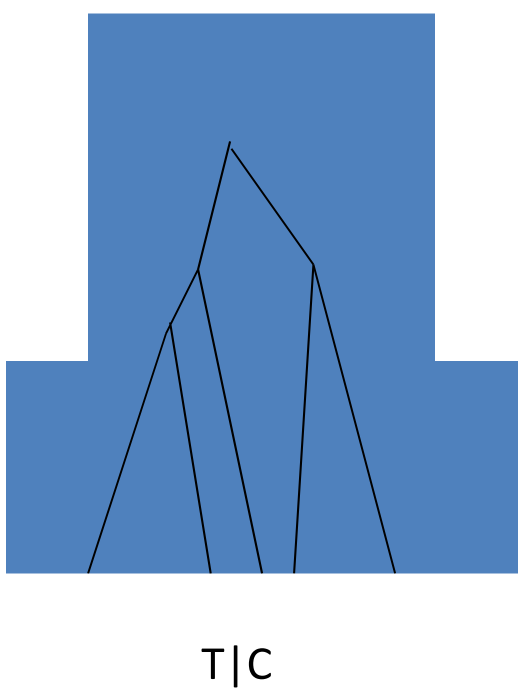
However, we can calculate neither.

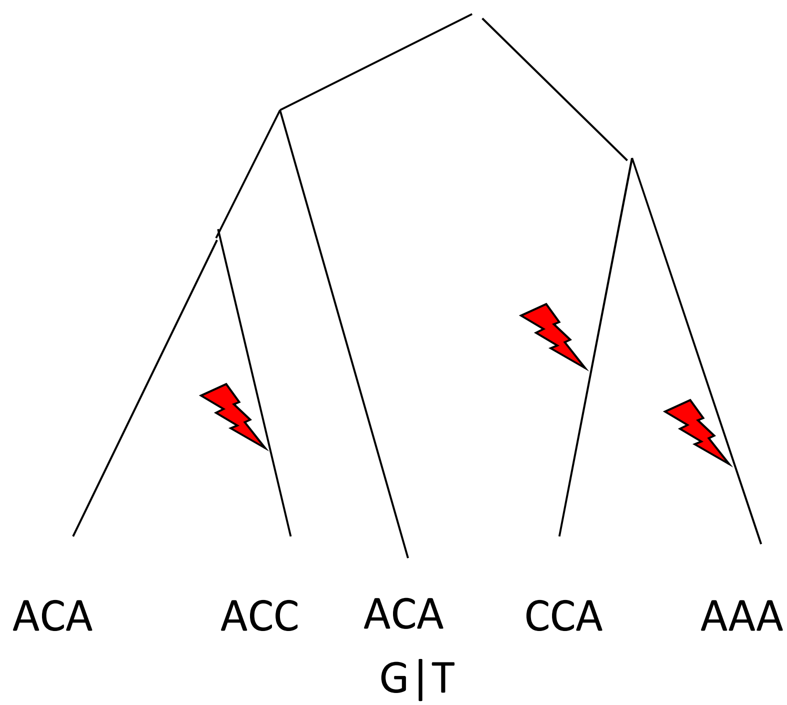
For a given coalescent tree T, we can calculate P(G|C,T) and P(T|C). Hence we calculate

$$P(G|C) = \int_T P(G|C,T)P(T|C)dT$$

Example

Consider the ancestry of the sample as the intermediate variable.





Monte Carlo Integration

The integral $P(G \mid C) = \int_T P(G \mid C, T) P(T \mid C) dT$

cannot be calculated.

To evaluate the integral, we can generate an iid sample $(x_1, \ldots, x_m)$

from $P(\mathrm{T} \mid \mathrm{C})$ and approximate the expectation by the empirical average

$$P(G \mid C) = \frac{1}{m} \sum_{j=1}^{m} P(G \mid C, x_j)$$

since this converges almost surely due to the Strong Law of Large Numbers.

Algorithm I

- Repeat n times:
  - Draw T from distribution P(T|C).
  - s+=P(G|T,C)
- Calculate P(G|C)≈s/n

Algorithm I.1

- Repeat n times:
  - Draw T from distribution $P(T|C)$.
  - Draw g from $P(g|T,C)$
  - If $g=G$: $s+=1$
- Calculate $P(G|C) \approx s/n$

Problem

In reality, dataset g has a very low probability of being identical with the initial data G. Hence the sum takes forever to converge.

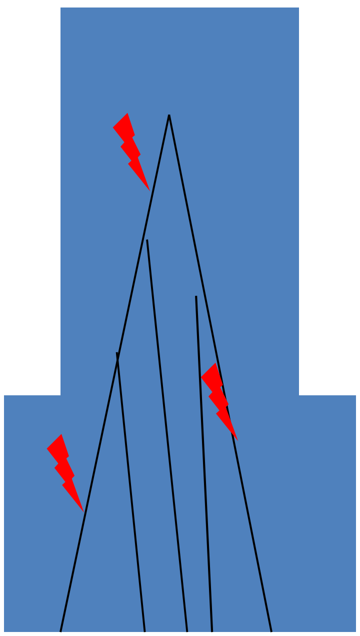
Solution: Summary statistics. Calculate statistics that reflect the properties of the sequence, for example S, π, Tajima's D or measures of LD. Let V designate the vector of summary statistics.

Algorithm I.2

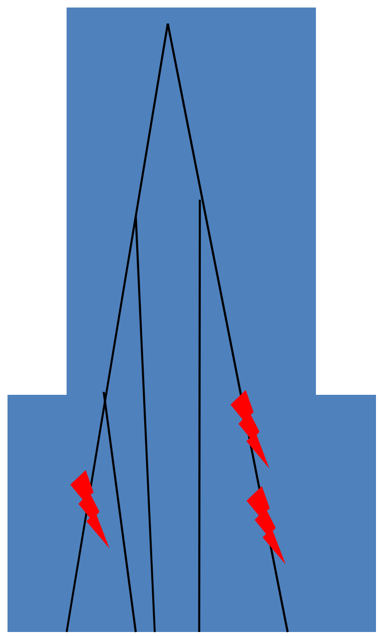
- Repeat n times:
  - Draw T from distribution $P(T|C)$.
  - Draw g from $P(g|T,C)$
  - Calculate $V(g)$
  - If $V(g)=V(G)$: s+=1
- Calculate $P(V(G)|C) \approx s/n$

Example

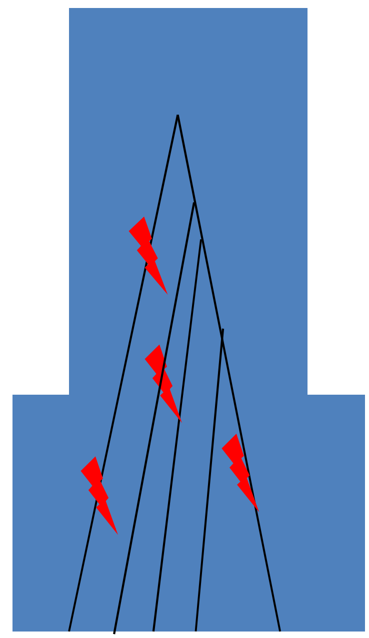
Observed data S=3, π=1.2. Perform 4 simulations for a growth rate λ.



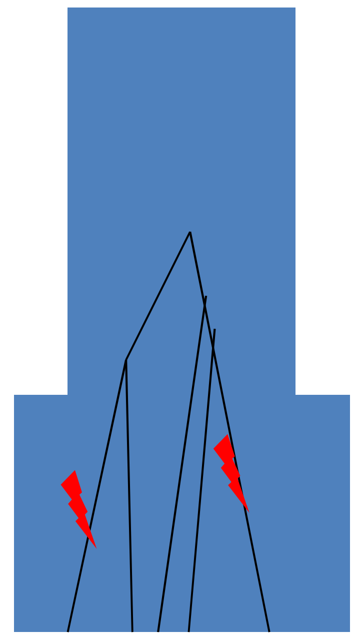
S=3 π=1.4



S=3 π=1.2

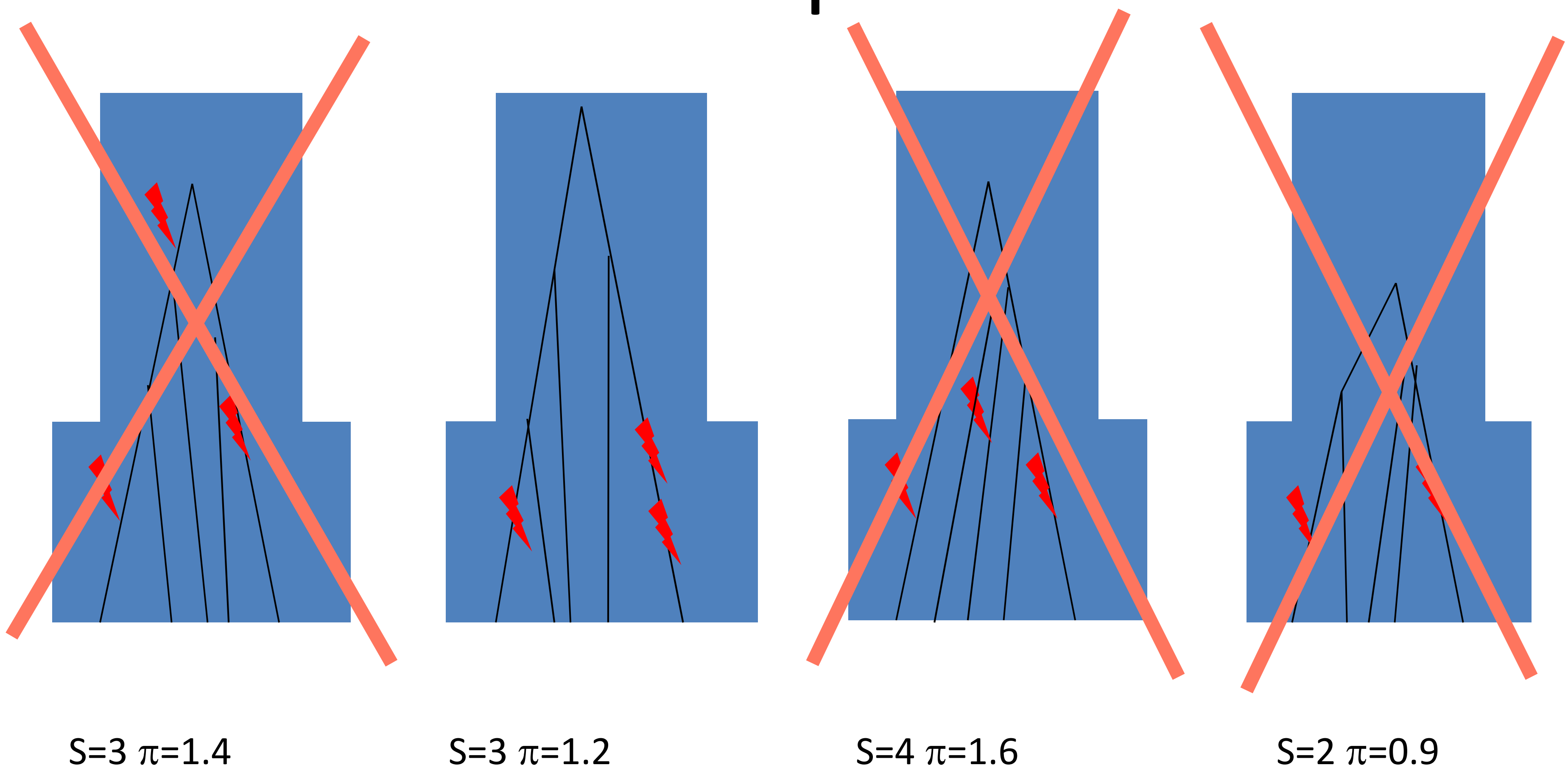


S=4 π=1.6



S=2 π=0.9

Example



$$\boxed{P(S,\pi|\lambda)\approx 1/4}$$

Problem

In reality, replicating exactly the set of summary statistics may still be too improbable.

Solution: Settle for approximate hits. Replace

$$P(G \mid C) = \int_{T,g} 1_{V(g)=V(G)} P(g \mid C,T) P(T \mid C) dT dg$$

with

$$P(G \mid C) \approx \int_{T,g} 1_{|V(g)-V(G)|<\varepsilon} P(g \mid C,T) P(T \mid C) dT dg$$

for an arbitrarily chosen small ε.

Algorithm I.3

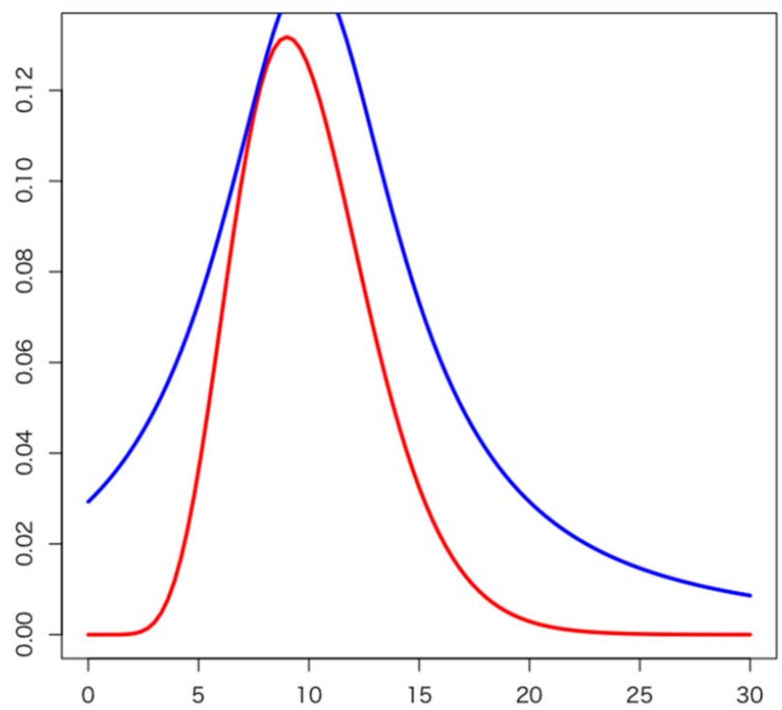
- Repeat n times:
  - Draw T from distribution $P(T|C)$.
  - Draw g from $P(g|T,C)$
  - Calculate $V(g)$
  - If $|V(g)-V(G)|<\varepsilon$: s+=1
- Calculate $P(G|C) \approx s/n$

Potential Challenges

- Sampling may be difficult
  - Rejection sampling
- The most likely trees ($P(T|C)$ is high) may result in configurations where observed data is very unlikely ($P(G|T)$ is low) so that our estimate converges slowly
  - Importance Sampling (not covered today) focuses sampling on most informative trees

Rejection Sampling

Sampling from the density *f* may not be possible, but instead sampling from an envelope function *G* with $G(y) \geq f(y)$ for all *y*.



Rejection Sampling-Algorithm

- Repeat n times:
  - Draw from distribution $Q(T|C)$.
  - Draw from $u[0,1]$
  - If $u<P(T|C)/Q(T|C)$
    - Calculate $P(G|T)$
    - $s+=P(G|T)$
- Calculate $P(G|C)=s/n$

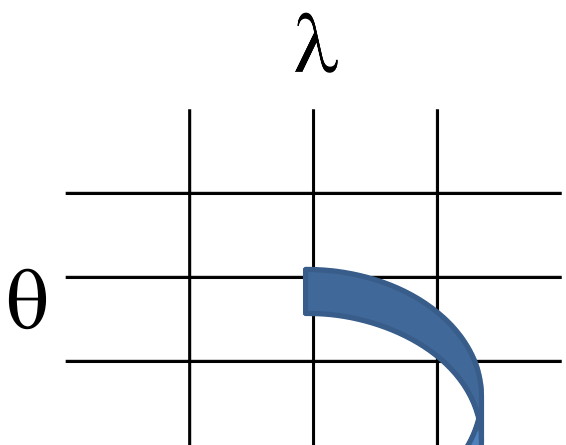
Example for Rejection Sampling

- We can rewrite P(G|C) as ∫P(G|T,S,C)P(T,S|C)dT
- P(T,S|C)=P(S|T,C)P(T|C)
- Hence P(T|C) is an envelope function for P(T,S|C).
- The acceptance probability of a sample from P(T|C) is

$$\frac{P(T,S|C)}{P(T|C)} = P(S|T,C) = \frac{\left(T_{total}\frac{\theta}{2}\right)^{S} e^{-T_{total}\frac{\theta}{2}}}{S!}$$

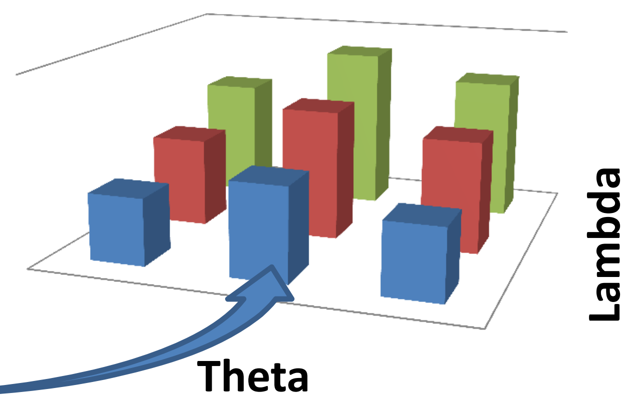
Where T_{total} is the length of tree T.

Exploring a range of parameters



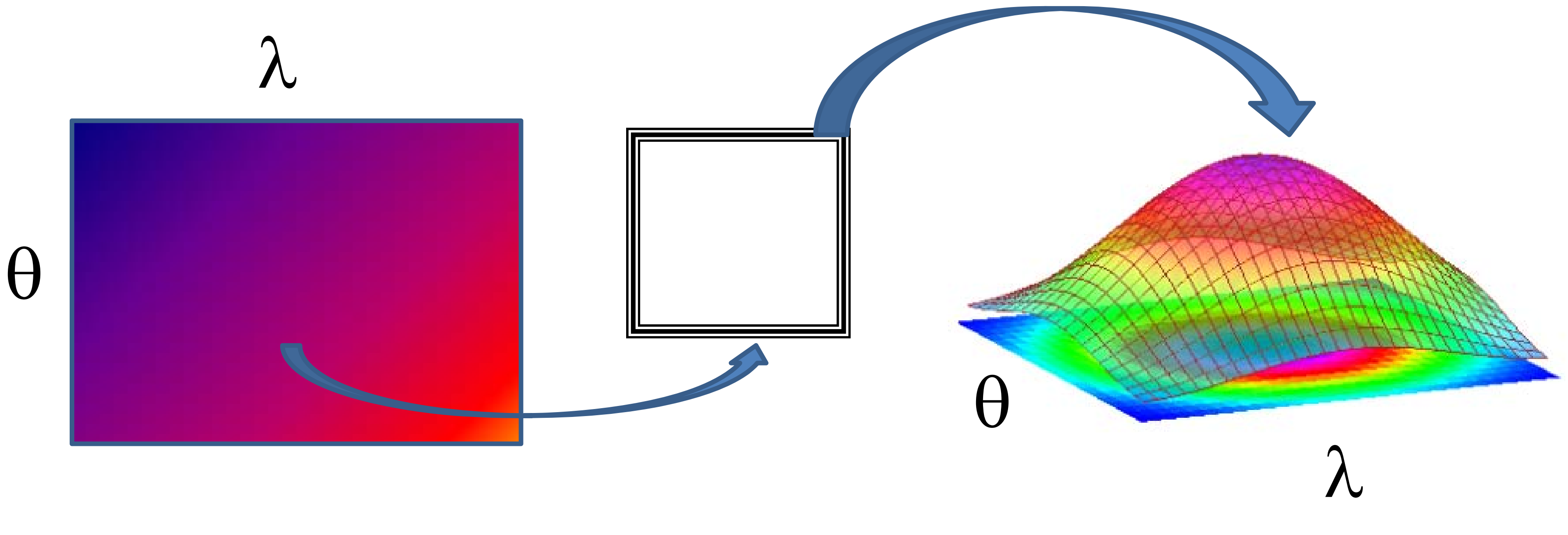
- Repeat n times:
  - Draw T from distribution P(T|C).
  - Draw g from P(g|T,C)
  - Calculate V(g)
  - If V(g)=V(G): s+=1
- Calculate P(G|C)≈c/n

- For each C we can approximate P(G|C)
- P(G|C) is usually calculated under a wide range of parameters C_i, generating a likelihood surface.
- The C_i can be taken from a grid



ABC-Approximate Bayesian Computation

- In a Bayesian framework we want to sample from $\Pr(C|G) \sim \Pr(G|C)\pi(C)$.
- Instead of moving C on a grid, we draw C from its prior.



ABC-Example

- Repeat n times:
  - Draw C from $\pi(C)$.
  - Draw T from distribution $P(T|C)$.
  - Draw g from $P(g|T,C)$
  - Calculate $V(g)$
  - If $|V(g)-V(G)|<\varepsilon$: $s(C)+=1$
- Calculate $P(C|G) \approx s(C)/n$