Some Monte Carlo methods for phylogenetic problems

Geoff Nicholls

11th of September 2007, Newton Institute, Cambridge

1 Example application

Traits on trees (H&S03, N&G08)

Priors on a tree space

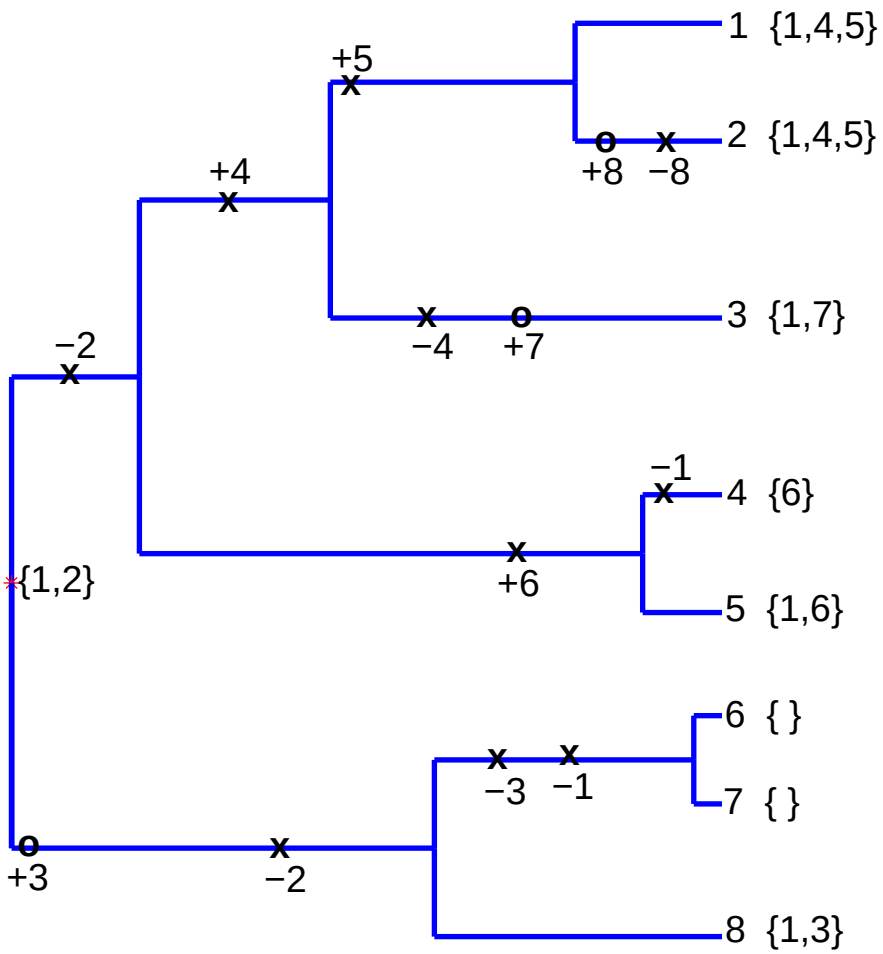
2 Monte Carlo: rejection, ABC, MCMC

3 Parallelizing a single MC run

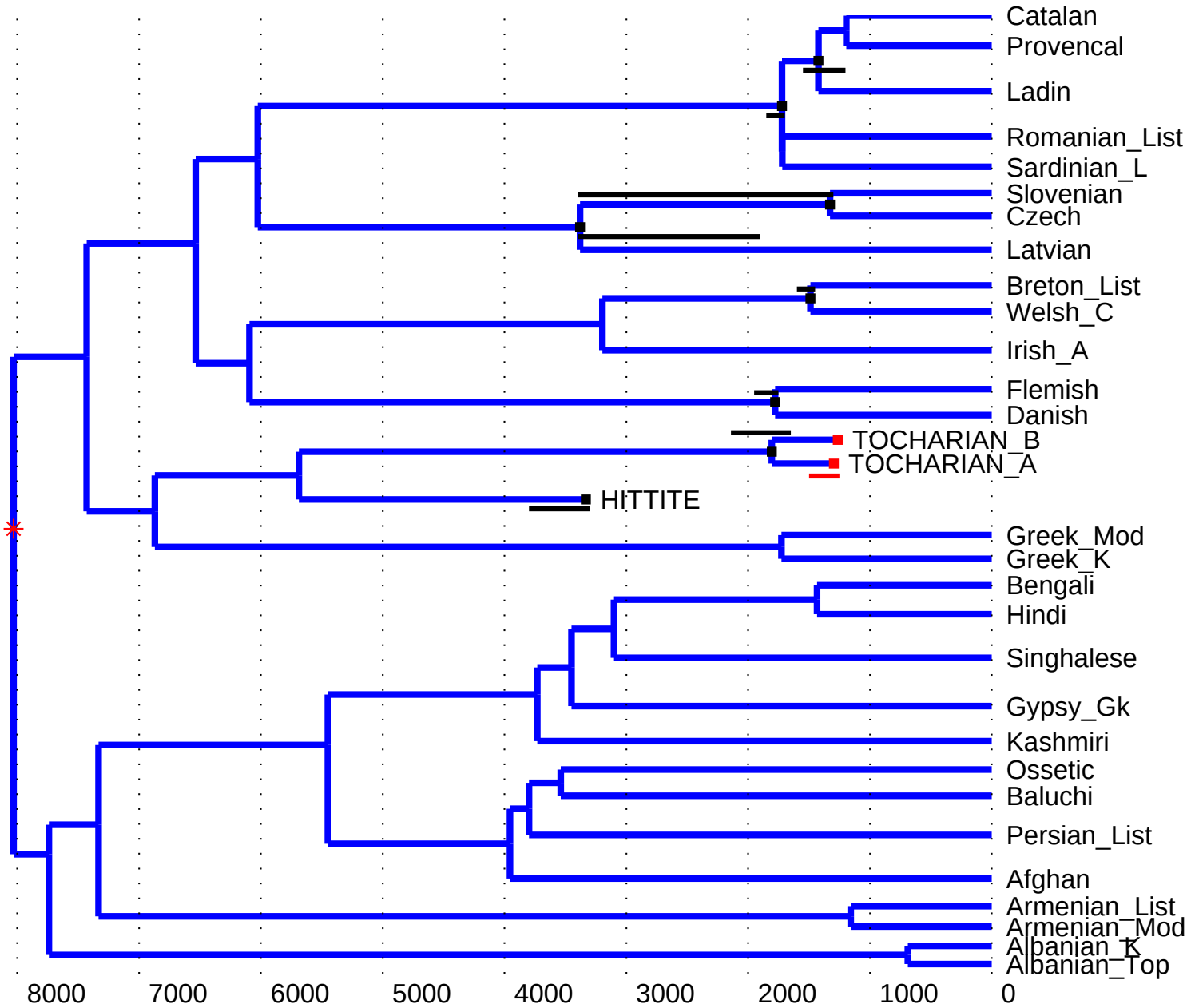
with Tiangang Cui and Allen Rodrigo, UoA

4 Exact MC computation using approximate evaluations

With Colin Fox UoO, Kate Lee QUT, DBR

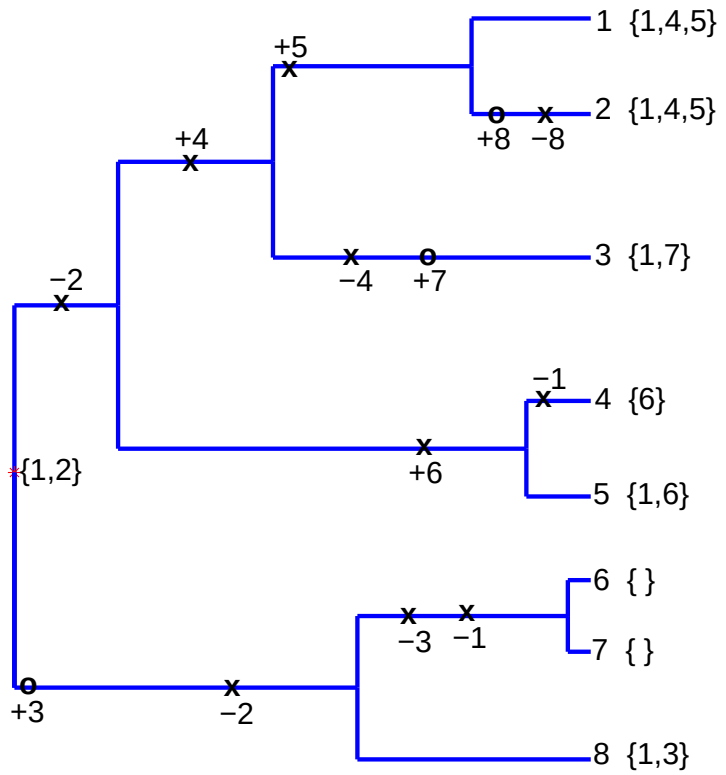


$$D = \begin{bmatrix} (1) & (4) & (5) & (6) \\ 1 & 1 & 1 & 0 \\ 1 & 1 & 1 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \end{bmatrix}$$


Catalan
Provencal
Ladin
Romanian_List
Sardinian_L
Slovenian
Czech
Latvian
Breton_List
Welsh_C
Irish_A
Flemish
Danish
TOCHARIAN_B
TOCHARIAN_A
HITTITE
Greek_Mod
Greek_K
Bengali
Hindi
Singhalese
Gypsy_Gk
Kashmiri
Ossetic
Baluchi
Persian_List
Afghan
Armenian_List
Armenian_Mod
Albanian_K
Albanian_Top
8000
7000
6000
5000
4000
3000
2000
1000
0

$$P(g|D) \propto \left[\int \left(\prod_{s=1}^{L} P(D_{:,s}|x_s, g, \mu)\right) p_X(x|g, \mu, \lambda)dx\right] p(\mu, \lambda)d\mu d\lambda \; f_G(g)$$

$$= p(D|g)f_G(g)$$



A tree prior

$$f_G(g) \propto t_R^{-N+2}, \qquad 0 \leq t_R \leq T$$

Marginal prior for $t_R \sim U(0,T)$ exact if isochronous leaves.

S is a list of free nodes, node $i \in S$ in rooted tree g, $s(g,i)$ is the minimum time node i can achieve in any admissible tree of the same topology as g, then

$$f_G(g) \propto \mathbb{I}_{t_R<T} \prod_{i \in S} (t_R - s(g,i))^{-1}$$

$t_R \not\sim U(0,T)$ but fairly smooth.

Problem: give $f_G(g)$ so that $t_R \sim U(0,T)$ (non-isochronous leaf times and/or interval calibration constraints).

ABC/MCMC

presentation follows Majoram, Molitor, Plagnol, Tavare, PNAS 03

(A version of the) Rejection algorithm

Aim: draw $g \sim P(g|D) \propto p(D|g) f_G(g)$

1. draw $z \sim f_G(z)$ and $u \sim U(0,1)$
2. while $u > p(D|z)$ do
   $z \sim f_G(z)$
   $u \sim U(0,1)$
   end
return $g = z$

(or 1. $z \sim q(z)$ and at 2. the test is $u > p(D|z) f_G(z)/cq(z)$ with $c \geq \max_z p(D|z) f_G(z)/q(z)$).

(Another version of the) Rejection algorithm

Aim: draw $g \sim P(g|D) \propto p(D|g) f_G(g)$

1. draw $z \sim f_G(z)$ and $D' \sim p(D'|z)$
2. while $D' \neq D$ do
   $z \sim f_G(z)$
   $D' \sim p(D'|z)$
   end
return $g = z$

$D' = D$ at 2. occurs with probability $p(D|z)$

(A version of) ABC Rejection algorithm (see Majoram et al. 03)

Aim: draw $g \sim P(g|D) \propto p(D|g)f_G(g)$ (approximately)

Fix $\epsilon > 0$ and some measure of distance $d(D', D)$ between 'data sets'.

1. draw $z \sim f_G(z)$ and $D' \sim p(D'|z)$
2. while $d(D', D) > \epsilon$ do
   $z \sim f_G(z)$
   $D' \sim p(D'|z)$
   end

return $g = z$

No likelihood. Approximate Monte carlo - model mispecification.

Distance definition, ϵ.

MCMC (Metropolis et al. , Hastings, Green)

Aim: $G_t \sim P(\cdot|D)$

Suppose $G_t = g$, and proposal $q(z|g)$ is given.

Then G_{t+1} is determined in the following way

A.

1. draw $z \sim q(z|g)$
2. set $\alpha = 1 \wedge \dfrac{P(z|D)q(g|z)}{P(g|D)q(z|g)}$ and draw $U \sim U(0,1)$

B.

3. if $U \leq \alpha$ set $G_{t+1} = z$ otherwise $G_{t+1} = g$.

Example

if $q(z|g) = f_G(z)$ then $\alpha = 1 \wedge \dfrac{p(D|z)}{p(D|g)}$.

(AVO) MCMC ABC

Aim: $G_t \sim P(\cdot|D)$

Suppose $G_t = g$, and proposal $q(z|g)$ is given.

Then G_{t+1} is determined in the following way

A.

1. draw $z \sim q(z|g)$ and $D' \sim p(D'|z)$
2. set $\alpha = 1 \wedge \dfrac{f_G(z)q(g|z)}{f_G(g)q(z|g)}$ and draw $U \sim U(0,1)$

B.

3. if ($d(D', D) < \epsilon$ and $U \leq \alpha$) set $G_{t+1} = z$ otherwise $G_{t+1} = g$.

Check $P(g|D)k(g,z) = P(z|D)k(z,g)$ when $d = 0$ (discrete case).

Parallel schemes

n-processor MCMC .

Independently (!) by Sohn IEEE TRANSACTIONS ON PARALLEL AND DISTRIBUTED SYSTEMS, VOL. 6, NO. IO, OCTOBER 1995

Suppose $G_t = g$, and proposal $q(z|g)$ is given.

Then $G_{t+1} \dots G_{t+M}, M \leq n$ are determined in the following way

A. For each $s = 1, 2 \dots n$ independently

1. draw $z_s \sim q(z_s|g)$
2. set $\alpha_s = 1 \wedge \dfrac{p(D|z_s) f_G(z_s) q(g|z_s)}{P(D|g) f_G(g) q(z_s|g)}$ and draw $U_s \sim U(0,1)$

B.

3. If ($U_s \leq \alpha_s$) for some $s = s_1, s_2, \dots s_j$,
   then $M = \min\{s_1, s_2, \dots s_j\}$
   and $G_{t+s} = g, s = 1, 2, \dots M-1$ and $G_{t+M} = z_M$,

otherwise $G_{t+s} = g, s = 1, 2, \dots n$.

n-processor MCMC-ABC. (GKN)

Suppose $G_t = g$, and proposal $q(z|g)$ is given.

Then $G_{t+1} \dots G_{t+M}, M \leq n$ are determined in the following way

A. For each $s = 1, 2 \dots n$ independently

1. draw $z_s \sim q(z_s|g)$ and $D'_s \sim p(D'_s|z_s)$
2. set $\alpha_s = 1 \wedge \dfrac{f_G(z_s)q(g|z_s)}{f_G(g)q(z_s|g)}$ and draw $U_s \sim U(0,1)$

B.

3. If $(d(D'_s, D) < \epsilon$ and $U_s \leq \alpha_s)$ for some $s = s_1, s_2, \dots s_j$,
   then $M = \min\{s_1, s_2, \dots s_j\}$
   and $G_{t+s} = g, s = 1, 2, \dots M-1$ and $G_{t+M} = z_M$,
   otherwise $G_{t+s} = g, s = 1, 2, \dots n$.

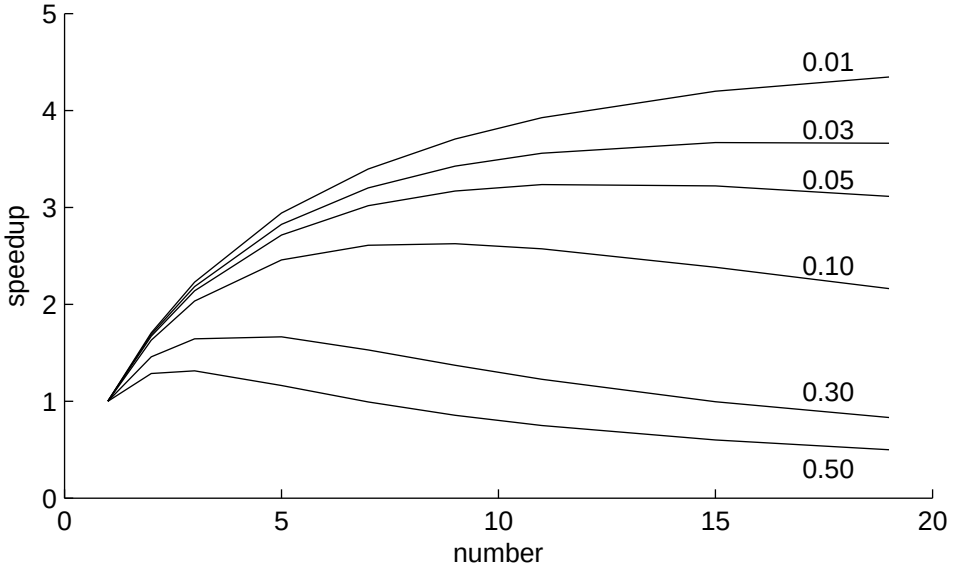
Question: results come back out of s-order (in CPU-time order). Can we pack the MCMC output using the CPU-time sorted results?
Answer: No.

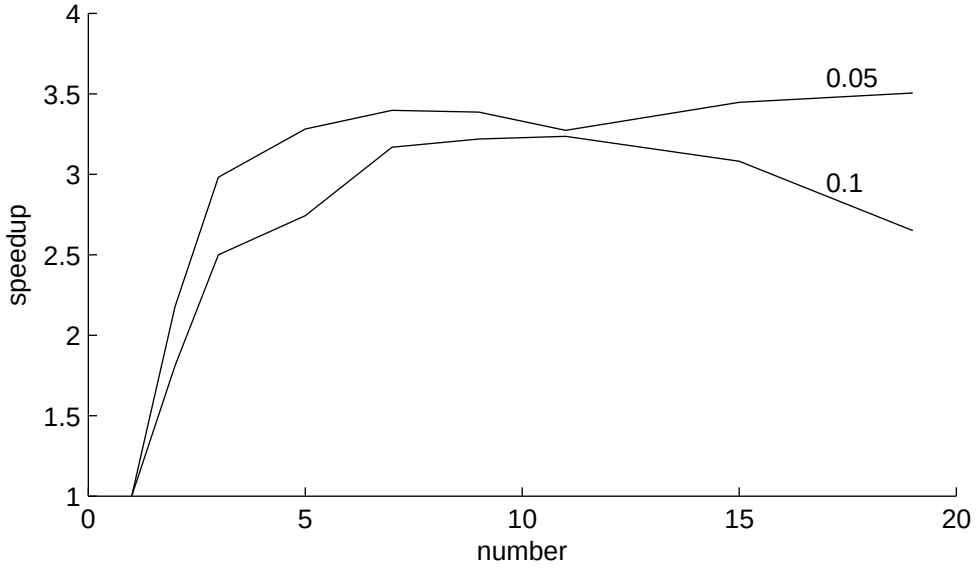
Performance: gross acceptance rate α. $\mathsf{E}\{M\} = (1-(1-\alpha)^n)/\alpha$.

Parallel transaction time nt_T, and $p(D|g)$-time t_D. Let $\rho = t_T/t_D$ ($\simeq 0.2$) serial(parallel)-CPU-secs-per-update t_S (t_P).

$$t_S = t_T + t_D, \qquad t_P = (nt_T + t_D)/\mathsf{E}\{M\}.$$

$$\text{“speedup1”} = t_S/t_P = \alpha^{-1}(1-(1-\alpha)^n)\frac{1+\rho}{1+n\rho}.$$





An idea from the annealing literature (AR).

$M = \log_2(n)$. Jump from G_t to G_{t+M}.

$$\text{“speedup2”} = t_S/t_P = \log_2(n)\frac{1+\rho}{1+n\rho}.$$

compare

$$\text{“speedup1”} = t_S/t_P = \alpha^{-1}(1-(1-\alpha)^n)\frac{1+\rho}{1+n\rho}.$$

Exact computation of MCMC using approximate evaluation

Two MH-MCMC chains $X_t \in \Sigma$, $Y_t \in \Sigma$, $t = 0, 1, 2...$

Common proposal $q(z'|z)$, $z, z' \in \Sigma$

Different acceptance probabilities,

$$\begin{aligned} \alpha_X(z'|z) &= \min\left(1, \frac{\pi_X(z')q(z|z')}{\pi_X(z)q(z'|z)}\right), \\ \alpha_Y(z'|z) &= \min\left(1, \frac{\pi_Y(z')q(z|z')}{\pi_Y(z)q(z'|z)}\right), \end{aligned}$$

π_X and π_Y equilibrium densities of the chains X_t atd Y_t.

At update t, $x' = f_q(x, u_t)$ with f_q chosen so x' has density $q(x'|x)$

$$U_t = u_t,\ u_t = (u_{t,1}, u_{t,2}, ...),\ U_{t,i} \sim U(0,1)\ V_t \sim U(0,1).$$

If $X_t = x$ then $\alpha_{X,t} = \alpha_X(f_q(x, u_t)|x)$ and accept if $v_t < \alpha_{X,t}$.

Couple X_t and Y_t: feed both the same random variates (u_t, v_t)

Start in the same state, $X_0 = Y_0$, separate at first $t > 0$ where

$$\min(\alpha_{X,t}, \alpha_{Y,t}) < v_t < \max(\alpha_{X,t}, \alpha_{Y,t}).$$

Application:
Y_t 'exact' chain ($\pi_Y(y')/\pi_Y(y)$ $y, y' \in \Sigma$ expensive).
X_t approximate ($\pi_X^{(i)}(z) \to \pi_Y(z)$ as $i \to \infty$)

Simulation of X_t gives exactly the same samples we would have had from the simulation of Y_t, from $t = 0$ up to the separation time.

Set i (precision of likelihood evaluation) to smallest value for which the mean separation time exceeds total MCMC run length.

Three kinds of mean separation times.

1. $T(x) = \min(t : X_t \neq Y_t | X_0 = Y_0 = x)$ random separation time from $X_0 = Y_0 = x$.

$$\rho(x) \equiv E_{U,V}\{T(x)\}$$

2. $Y_0 \sim \pi_Y$, $\rho_\pi \equiv E_{U,V,Y_0}\{T(Y_0)\}$.

3. Fix U, V for all t, compute Y_t, $t = ..., -2, -1, 0, 1, 2, ...$ and set $X_t = Y_t$. Define separation marks B_t, $B_t = 1$ if

$$\min(\alpha_{X,t}, \alpha_{Y,t}) < V_t < \max(\alpha_{X,t}, \alpha_{Y,t})$$

so, separation event at step t, and $B_t = 0$ otherwise. First two separation events at N_0, N_1:

$$\begin{aligned} N_0 &= \min\{t; t \geq 0, B_t = 1\} \\ N_1 &= \min\{t; t > N_0, B_t = 1\} \end{aligned}$$

so $T_1 = N_1 - N_0$ is a generic inter-separation interval. Define

$$\rho_{\text{sep}} \equiv E_{U,V}\{T_1\}$$

.

Useful results:
1.

$$\rho_{\text{sep}} = 1/E_{U,V}\{|\alpha_{X,t} - \alpha_{Y,t}|\}$$

2.

$$\rho_{\text{sep}} \geq \rho_{\pi}$$

Use 1. to estimate ρ_{sep} (pilot runs or analytical bound). Increase i so ρ_{sep} exceeds run length. Then by 2. ρ_{π} exceeds run length.

It would be helpful to have a statement in probability.