Genealogy of Wright-Fisher bridges

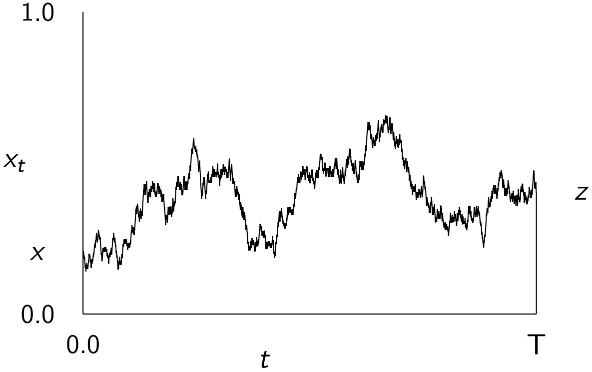
Bob Griffiths

Department of Statistics
University of Oxford

Probabilistic Models in Biology 2015

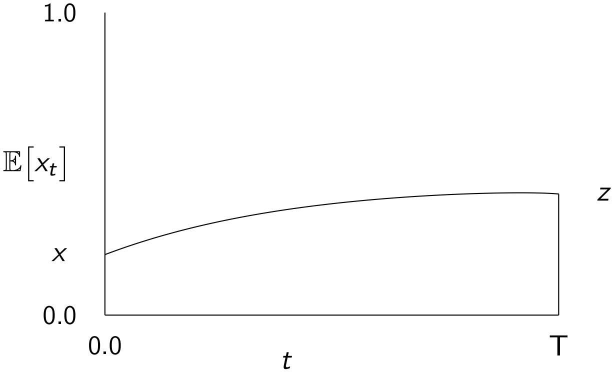
Joshua Schraiber, Robert Griffiths, Steven Evans, (2013), Analysis and rejection sampling of Wright-Fisher diffusion bridges, *Theoretical Population Biology* **89** 64–74

Wright-Fisher diffusion bridge from x to z in time T



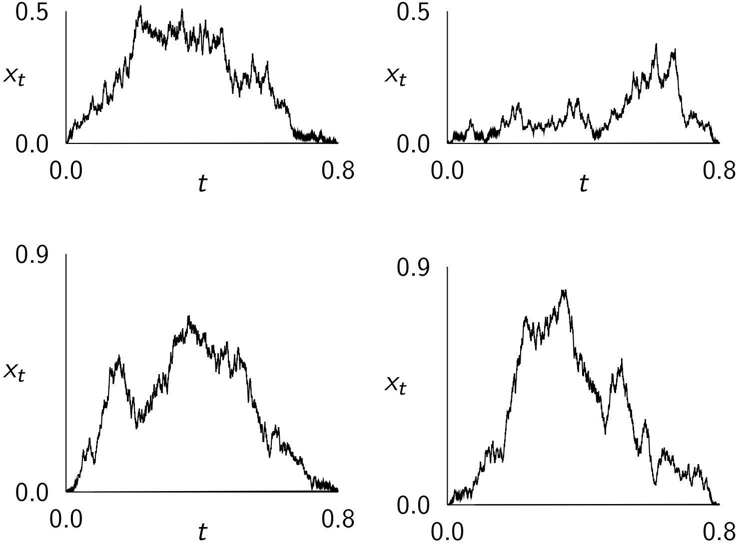
x_t is the frequency of a gene at time t.

Wright-Fisher diffusion bridge from x to z in time T



$\mathbb{E}[x_t]$ is the mean frequency of a gene at time t.

Wright-Fisher diffusion bridges from 0 to 0 in time $T = 0.8$



Wright-Fisher diffusion

A Wright-Fisher diffusion with genic selection is a diffusion process $\{X_t,\ t \geq 0\}$ with state space $[0,1]$ and infinitesimal generator

$$\mathcal{L} = \gamma x(1-x)\frac{\partial}{\partial x} + \frac{1}{2}x(1-x)\frac{\partial^2}{\partial x^2}.$$

When $\gamma = 0$, the diffusion is neutral; otherwise, the drift term indicates the strength and direction of selection.

The transition density $f(x, y; t)$ is the density of $Y = X_t$ given X_0=x.

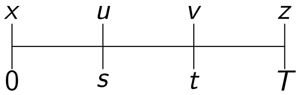
The points $\{0, 1\}$ are absorbing boundaries and much care is needed with them in theory and computation.

Wright-Fisher diffusion bridge

$\{X_t^{x,z,[0,T]}, 0 \leq t \leq T\}$ is the Wright-Fisher diffusion process conditioned to start with value x at time 0 and end with value z at time T.

The bridge is a time-inhomogeneous diffusion and the transition density at t for state v in the bridge conditional on being in state u at time s is

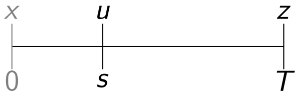
$$f_{x,z,[0,T]}(u, v; s, t) = \frac{f(u, v; t-s) f(v, z; T-t)}{f(u, z; T-s)}.$$



Bridge generator

The time-inhomogeneous generator of the bridge acting on a test function g at time s is

$$\begin{aligned}
\mathcal{L}_{x,z,[0,T];s}\, g(u) &= \lim_{t \downarrow s} \frac{\mathbb{E}[g(X_t) \text{ given } X_0 = x, X_s = u, X_T = z] - g(u)}{t - s} \\
&= u(1-u)\left(\gamma + \frac{\partial}{\partial u} \log f(u, z; T - s)\right) \frac{\partial g}{\partial u}(u) \\
&\quad + \frac{1}{2} u(1-u) \frac{\partial^2 g}{\partial u^2}(u).
\end{aligned}$$



Bridges in a conditioned process

$f_h(x, y; t)$ - transition density of $\{X(t)\}_{t \geq 0}$ conditioned on long term behaviour.
Doob h-transform

$$f_h(x, y; t) = h(x)^{-1} f(x, y; t) h(y).$$

The distribution of bridges is invariant under h-transforms.

Identical bridges

The bridge at times $0 \le t_1 \le \ldots \le t_n \le T$ starting at x at time 0 and ending at y at time T has density

$$\frac{f(x, v_1; t_1) f(v_1, v_2; t_2 - t_1) \cdots f(v_n, y; T - t_n)}{f(x, y; T)}$$

The density for the bridge of the h-transformed process is

$$\frac{f(x, v_1; t_1)\frac{h(v_1)}{h(x)} f(v_1, v_2; t_2 - t_1)\frac{h(v_2)}{h(v_1)} \cdots f(v_n, y; T - t_n)\frac{h(y)}{h(v_n)}}{\frac{h(y)}{h(x)} f(x, y; T)}$$
$$= \frac{f(x, v_1; t_1) f(v_1, v_2; t_2 - t_1) \cdots f(v_n, y; T - t_n)}{f(x, y; T)}.$$

The bridges for the two processes have the same distribution.

Example

p_{xy} is the probability of hitting y from x.

$$p_{xy} = \begin{cases} \frac{1-e^{-2\gamma x}}{1-e^{-2\gamma y}}, & \text{if } y > x, \\ \frac{e^{-2\gamma y}-e^{-2\gamma}}{e^{-2\gamma x}-e^{-2\gamma}}, & \text{if } y < x, \end{cases}$$

when $\gamma \neq 0$ and if $\gamma = 0$,

$$p_{xy} = \begin{cases} \frac{x}{y}, & \text{if } y > x, \\ \frac{1-y}{1-x}, & \text{if } y < x. \end{cases}$$

The transition density conditional on hitting $y > x$ is

$$f_h(x, y; t) = f(x, y; t)/(x/y) = x^{-1} f(x, y; t) y; \;\; h(x) = x.$$

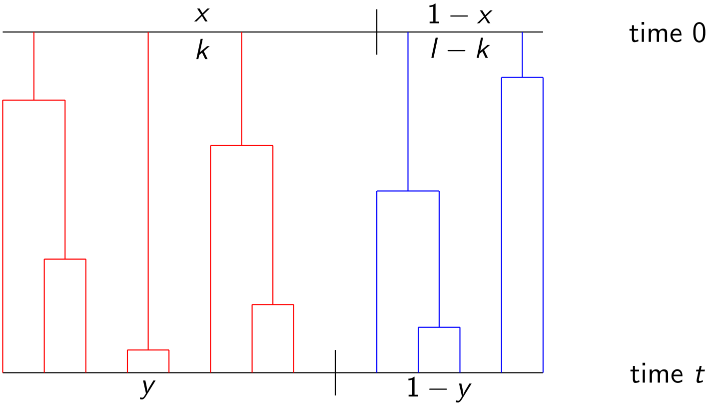
This is the same as conditioning on fixation of the gene type.

The neutral Wright-Fisher diffusion density

The transition densities of the Wright-Fisher diffusion can be expressed

$$f(x, y; t) = \sum_{l=2}^{\infty} q_l(t) \sum_{k=1}^{l-1} \binom{l}{k} x^k (1-x)^{l-k} \mathcal{B}(y; k, l-k).$$

$q_l(t)$ are the transition functions of the number of edges in a Kingman coalescent tree starting with infinitely many leaves at time 0. Edges are lost at rate $\frac{1}{2}n(n-1)$ when n leaves in the tree. $\mathcal{B}(\cdot; \alpha, \beta)$ is the density of the Beta distribution with parameters α and β.



Relative frequencies of the l families have a uniform Dirichlet distribution.

$$f(x, y; t) = \sum_{l=2}^{\infty} q_l(t) \sum_{k=1}^{l-1} \binom{l}{k} x^k (1-x)^{l-k} \mathcal{B}(y; k, l-k).$$

An orthogonal function expansion

The transition density as an expansion in orthogonal polynomials, from Kimura, is

$$f(x, y; t) = x(1-x)\sum_{i=1}^{\infty}(2i+1)i(i+1)P_{i-1}(r)P_{i-1}(s)e^{-\frac{1}{2}i(i+1)t},$$

where $r = 1 - 2x$, $s = 1 - 2y$.

The polynomials are orthogonal on $x(1-x)$, $0 < x < 1$ and scaled so that $P_i(1) = 1$.

The orthogonal function form and the genealogical form are the same density expressed differently.

Asymptotic density as $x \to 0$

The asymptotic form of the density as $x \to 0$ is

$$\begin{aligned} f(x,y;t) &\approx 2x\sum_{l=2}^{\infty}(1-y)^{l-2}\binom{l}{2}q_l(t) \\ &= 2x\sum_{l=2}^{\infty}(1-y)^{l-2}\psi_l(t) \end{aligned}$$

where

$$\psi_l(t) = \frac{1}{2}\sum_{j=l}^{\infty} e^{-\frac{1}{2}j(j-1)t}\frac{(-1)^{j-l}(2j-1)l_{(j-1)}}{l!(j-l)!}$$

is the density of $T_l + T_{l+1} + \cdots$, with independent exponential random variables T_j having rates $\binom{j}{2}$, $j = 2, 3, \ldots$

The density conditional on hitting $y > x$ is

$$x^{-1}f(x,y;t)y \to 2y\sum_{l=2}^{\infty}(1-y)^{l-2}\binom{l}{2}q_l(t) \text{ as } x \downarrow 0$$

Identity

$$2e^t y(1-y)\sum_{l=2}^{\infty}(1-y)^{l-2}\binom{l}{2}q_l(t) = f_{22}(0,y;t)$$

where $f_{22}(0,y;t)$ is the transition density from $0 \to y$ in time t in a Wright-Fisher diffusion with mutation rates $\theta_1 = \theta_2 = 2$ with generator

$$(1-2y)\frac{\partial}{\partial y} + \frac{1}{2}y(1-y)\frac{\partial^2}{\partial y^2}.$$

The identity allows computation via a Jacobi orthogonal polynomial expansion of $f_{22}(0,y;t)$.

Genealogical approach instead of h-transforms

The transition density conditional on at least one family from 0 of the genes type at time t, and neither type being fixed is

$$
\begin{aligned}
f_1(x,y;t) &= \frac{\sum_{l=2}^{\infty} q_l(t) \sum_{k=1}^{l-1} \binom{l}{k} x^k (1-x)^{l-k} \mathcal{B}(y;k,l-k)}{\sum_{l=2}^{\infty} q_l(t)\big(1-(1-x)^l\big)} \\
&\to \frac{\sum_{l=2}^{\infty} q_l(t) l \mathcal{B}(y;1,l-1)}{\sum_{l=2}^{\infty} q_l(t) l} \\
&= \frac{\sum_{l=2}^{\infty} q_l(t) l \mathcal{B}(y;1,l-1)}{\sum_{r=1}^{\infty} (4r-1) e^{-r(2r-1)t}}
\end{aligned}
$$

In the limit there is forced to be exactly one family from 0 of the genes type at time t.

Bridge from 0 to 0 over $[0, T]$

The density of X_t given that $X_0 = x$, $X_T = z$ is

$$\begin{aligned} f_{x,z,[0,T]}(y;t) &= \frac{f(x,y;t)f(y,z;T-t)}{f(x,z;T)} \\ &= \frac{f(x,y;t)f(z,y;T-t)y(1-y)}{f(x,z;T)z(1-z)} \\ &= \frac{x^{-1}f(x,y;t)z^{-1}f(z,y,T-t)y(1-y)}{x^{-1}f(x,z;T)(1-z)}. \end{aligned}$$

In the second line reversibility (before hitting 0 or 1) with respect to the speed measure $y^{-1}(1-y)^{-1}$ is used, that is

$$y^{-1}(1-y)^{-1}f(y,z;T-t) = z^{-1}(1-z)^{-1}f(z,y;T-t).$$

Bridge from 0 to 0 over $[0, T]$ transition density

Take the limit as $x, z \to 0$ in the transition density of a bridge from x to z.
The transition density $f_{0,0,[0,T]}(y;t)$ is

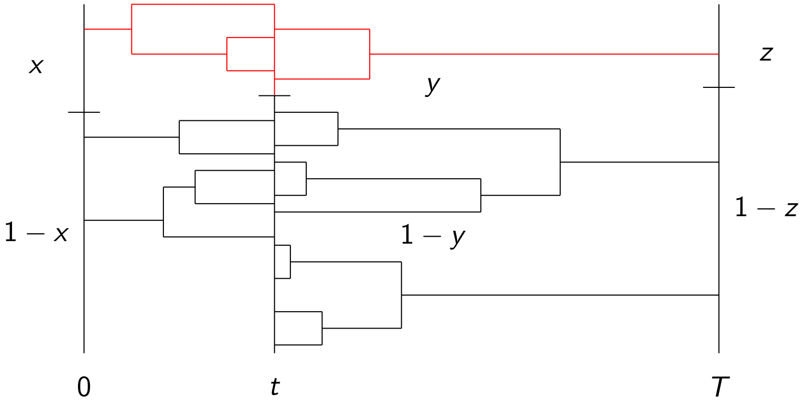
$$\frac{2y(1-y)\sum_{k=2}^{\infty}(1-y)^{k-2}\binom{k}{2}q_k(t) \times \sum_{l=2}^{\infty}(1-y)^{l-2}\binom{l}{2}q_l(T-t)}{\sum_{m=2}^{\infty}\binom{m}{2}q_m(T)}$$

$$= \frac{f_{22}(0,y;t)f_{22}(0,y;T-t)}{2y(1-y)e^{T}\sum_{m=2}^{\infty}\binom{m}{2}q_m(T)}$$

The denominator simplifies because

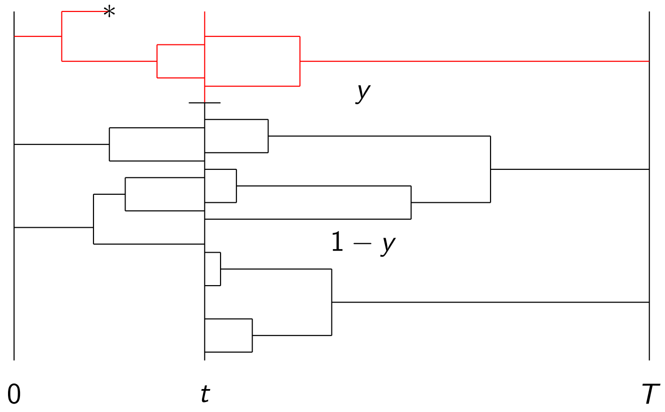
$$\sum_{m=2}^{\infty}\binom{m}{2}q_m(t) = \sum_{n=2}^{\infty}e^{-\frac{1}{2}n(n-1)t}(2n-1)n(n-1).$$

Coalescent bridge picture



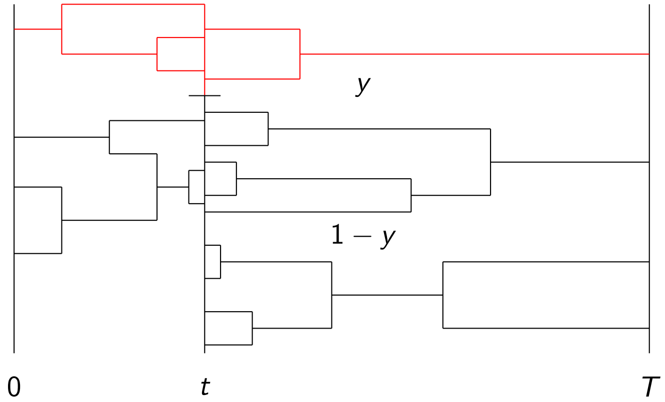
Exactly 1 lineage from 0 to t and from T to t in reverse time with leaves in the frequency y, conditional on y being reached as $x, z \to 0$.

Mutation away from the gene's type at rate $\theta/2$.
This happens in the infinitely-many-alleles model.



$*$ - Lineages can be lost by mutation as well as coalescence.

Coalescent bridge picture with selection



The genes of frequency y coalesce to a single line before 0 and T in the two directions. There is branching and coalescing in the lineages of the genes of frequency $1 - y$.

Genealogy with selection $\gamma > 0$

The transition density is

$$f(x, y; t) = \sum_{\alpha_1, \alpha_2 \geq 1} b_{\boldsymbol{\alpha}}(t, x)\pi_{\boldsymbol{\alpha}}(y)$$

$b_{\boldsymbol{\alpha}}(t, x)$ are transition functions in a two-type birth and death process $\boldsymbol{\alpha}(t) = (\alpha_1(t), \alpha_2(t))$ beginning with an infinite number of individuals of proportions $x, 1 - x$ describing the genealogy.

In this genealogy coalescence occurs within both types of lineages and branching occurs in the second types lineages.

There is eventual absorption into either (1,0) or the face (0,α_2) with $\alpha_2 \geq 1$.

$$\pi_{\boldsymbol{\alpha}}(y) = c(\boldsymbol{\alpha})^{-1} e^{-2\gamma(1-y)} \frac{\Gamma(\alpha_1 + \alpha_2)}{\Gamma(\alpha_1)\Gamma(\alpha_2)} y^{\alpha_1 - 1}(1 - y)^{\alpha_2 - 1},\ 0 < y < 1$$

Transition densities in the neutral 0, 0 bridge

The transition densities of $X_t^{0,0,[0,T]}$, $f_{0,0,[0,T]}(u, v; s, t)$ are

$$f(u, v; t-s)e^{t-s}\frac{f_{22}(0, v; T-t)}{f_{22}(0, u; T-t)}$$

The transition density converges as $T \to \infty$ to

$$e^{t-s}u^{-1}(1-u)^{-1}f(u, v; t-s)v(1-v),$$

the transition density of the neutral Wright-Fisher diffusion conditioned on non-absorption, a process with infinitesimal generator

$$(1-2y)\frac{\partial}{\partial y} + \frac{1}{2}y(1-y)\frac{\partial^2}{\partial y^2}.$$

A limit centering around $\mathbf{T/2}$.

The limiting density of $X_{T/2+t}$ for $-T/2 < t < T/2$ is independent of t and is the same as the quasi-stationary density of $X(t)$

$$\lim_{t\to\infty} \frac{P_x\Big(X(t) \in (y, y+dy)\Big)}{P_x\Big(X(t) \neq 0\Big)} = 6y(1-y)dy$$

for $x \neq 0, 1$.

First passage time distribution

Let $g(\cdot; x, y)$ be the first passage time density from x to y in the neutral Wright-Fisher diffusion.
The density $g(\cdot; x, y)$ is improper because of absorbing states $0, 1$.

$$\int_0^\infty g(t; x, y) dt < 1.$$

Take the Laplace transform of the identity

$$f(x, y; t) = \int_0^t g(\tau; x, y) f(y, y; t - \tau)\, d\tau.$$

The Laplace transform of $g(\cdot; x, y)$ is

$$g^*(\lambda; x, y) = \frac{f^*(x, y; \lambda)}{f^*(y, y; \lambda)}.$$

First passage time from x to y conditioned on hitting y.

The first passage time distribution, conditioned on hitting y has a proper density

$$g(t; x, y)\frac{y}{x}.$$

The first passage time from 0 to y conditioned on hitting y has a proper density

$$g_{\#}(t; y) := \lim_{x \downarrow 0} g(t; x, y)\frac{y}{x}.$$

The mean first passage time from 0 to y conditional on y being hit in $g_{\#}(t; y)$ is

$$2 + 2\frac{1-y}{y}\log(1-y).$$

Maximum in a 0,0 bridge, $\mathbb{P}\{M^{0,0,[0,T]} \geq y\}$

Take a limit as $x, z \downarrow 0$ in the density of the maximum in the x, z bridge. A reversibility argument in the numerator and denominator is needed to take the limit correctly.

Then $\mathbb{P}\{M^{0,0,[0,T]} \geq y)$ is equal to

$$\frac{(1-y)\int_0^T g_{\#}(t;y)\sum_{i=1}^{\infty}(2i+1)i(i+1)P_{i-1}(1-2y)e^{-\frac{1}{2}i(i+1)(T-t)}\,dt}{\sum_{i=1}^{\infty}(2i+1)i(i+1)e^{-\frac{1}{2}i(i+1)T}}$$

The orthogonal polynomial expansion is the best form to do numerical computations.

$g_{\#}(t;y)$ is not explicit, though its Laplace transform is known.

Distribution function of the maximum in a bridge M.

T	0.05	0.10	0.15	0.20	0.25	0.30	0.35	0.40	0.45	0.50
0.5	0.0	0.02	0.17	0.43	0.66	0.83	0.92	0.96	0.99	0.99
1.0			0.0	0.02	0.09	0.21	0.36	0.52	0.66	0.77
1.5				0.0	0.01	0.03	0.08	0.17	0.28	0.40
2.0						0.0	0.02	0.04	0.09	0.17
T	0.55	0.60	0.65	0.70	0.75	0.80	0.85	0.90	0.95	1.0
0.5	1.0									
1.0	0.85	0.91	0.95	0.97	0.99	0.99	1.0			
1.5	0.52	0.63	0.73	0.82	0.88	0.93	0.96	0.99	1.0	
2.0	0.26	0.37	0.48	0.59	0.70	0.97	0.87	0.93	0.97	1.0

T	0.01	0.02	0.03	0.04	0.05	0.06
0.1	0.00	0.01	0.14	0.37	0.59	0.76
T	0.07	0.08	0.09	0.10	0.11	0.12
0.1	0.86	0.93	0.96	0.98	0.99	1.0

References

Barbour, A.D., Ethier, S.N., Griffiths, R.C., 2000. A transition function expansion for a diffusion model with selection. Ann. Appl. Probab. 10, 123-162.

Ethier, S. and Griffiths, R. C. (1993). The transition function of a Fleming-Viot process. *Ann. Probab.* 21 1571–1590.

Griffiths, R. C. and Spanó, D. (2010). Diffusion processes and coalescent trees. Chapter 15 358-375. In: Probability and Mathematical Genetics, Papers in Honour of Sir John Kingman. LMS Lecture Note Series 378. ed Bingham, N. H. and Goldie, C. M., Cambridge University Press.

Jenkins, P. and Spanó, D. Exact simulation of the Wright-Fisher diffusion. arXiv:1506.06998

Schraiber, J. Griffiths, R. C., Evans, S. (2013), Analysis and rejection sampling of Wright-Fisher diffusion bridges,*Theoretical Population Biology* 89 64–74.