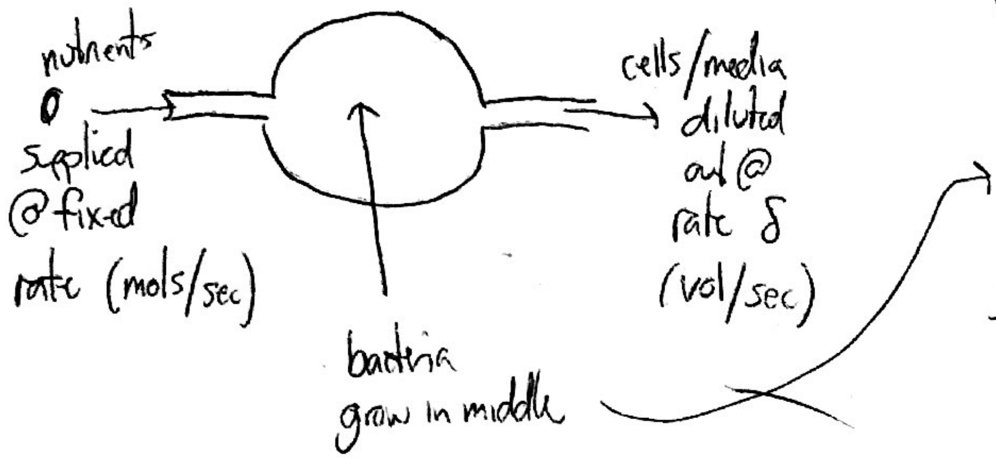
Microscopic models and diffusion limit

we've developed one particular model for dynamics of a mutation frequency based on serial dilution.

$\Longrightarrow$ many other "microscopic" models.

e.g. one experimentally motivated one $\Rightarrow$ a "chemostat" (Homework 1 problem)



reach some steady state size $\bar{N}$ that depends on input nutrients & dilution rate.

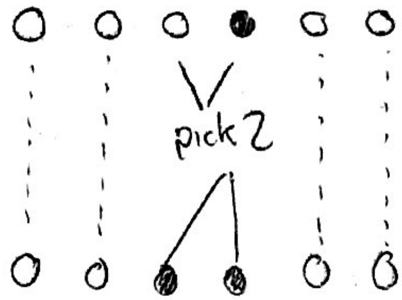
$\Longrightarrow$ people like it (theoretically & physiologically) because no temporal variation during the day.
(in practice, kind of tricky to set up; ~ few replicates)

$\Longrightarrow$ other pure math ones (population genetics)

"ball & urn" style models.

"bean bag genetics"

e.g. "Moran model"



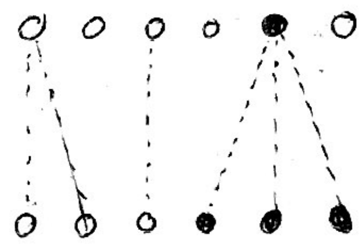
N individuals @ time t $\downarrow$ $t+1/N$

time $t + 1/N$

(need N steps before everyone has reproduced ~1 time: "generation")

one divides & replaces other w/ prob $\dfrac{e^s}{1+e^s}$

e.g. "Wright-Fisher model"



N individuals @ time t

time $t+1$

parent of N new individuals selected randomly w prob $\propto e^{s_i}$

Wright-Fisher model very popular ("canonical") because exact relation for mean & mean squared: (in neutral limit w/o mutations $s=0; \mu=0$)

$$E\left[f(t)\right] = E\left[f(t-1)\right] = f(0)$$

$$E\left[f(t)(1-f(t))\right] = \left(1-\frac{1}{N}\right) E\left[f(t-1)(1-f(t-1))\right] = f(0)(1-f(0))\, e^{-t/N} \to 0$$

(makes sense because @ long times $f=0,1$) $\Rightarrow$ need $t \simeq N$ for $f \to 0,1$

that's about it for exact results, even for such ridiculously simple "bean bag genetics" models.

⇒ how could we hope to make progress for anything remotely resembling real biological organisms (e.g. influenza viruses)?

At same time, you might be surprised to learn that field of pop-gen is basically about applying these simple models (particularly Wright-Fisher) to real data (e.g. Human DNA sequences)... and often does a surprisingly good job* (sometimes)

⇒ why? Humans are definitely <u>not</u> reproducing according to a Wright-Fisher model

today we'll start to get a partial answer to both Qs

⇒ one of my favorite results in classical pop gen

& has some deep connections to concepts of universality, RG, & coarse graining from physics.

to see this, will be useful to step back from evolution context and focus on a basic math problem:

(4)

Discrete random walk:

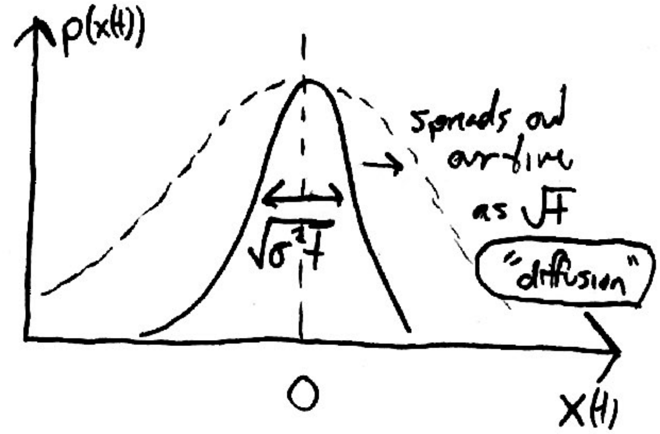
Let $\Delta X_1, \Delta X_2, \ldots \sim \text{Gaussian}(0,\sigma^2)$ and let $X(t) = \sum_{i=1}^{t} \Delta X_i$.

a stochastic process taking $X(t) \to X(t+1)$ (Markov)

Recall that for Gaussian distns, (independent)

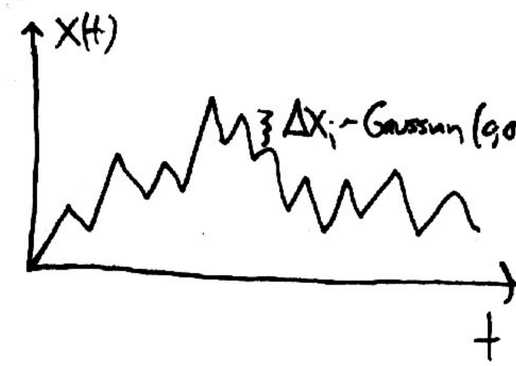
$$\text{Gaussian}(\mu_1, \sigma_1^2) + \text{Gaussian}(\mu_2, \sigma_2^2) = \text{Gaussian}(\mu_1+\mu_2, \sigma_1^2+\sigma_2^2)$$

$$\Rightarrow \quad X(t) \sim \text{Gaussian}(0, \sigma^2 t)$$



Also have probability of arbitrary path: $\underbrace{X(0), X(1)}_{\Delta X_1}, \underbrace{X(2)}_{\Delta X_2}, \ldots, X(t)$ (each Gaussian)

$$p(X(0), X(1), \ldots, X(t)) = \prod_{i=1}^{t} \frac{e^{-\frac{(X(t)-X(t-1))^2}{2\sigma^2}}}{\sqrt{2\pi\sigma^2}}$$



nothing too fancy yet... what if ΔX_i are not Gaussian?

Let $\Delta X_1, \Delta X_2, \ldots \overset{iid}{\sim} p(\Delta X)$ w/ $\langle \Delta X \rangle = \mu$, $\text{Var}(\Delta X) = \sigma^2$

Now when we consider sum: for broad class of dist'ns, $p(\Delta x)$:

$$X(t) = \underbrace{\Delta x_1 + \Delta x_2 + \Delta x_3}_{\text{Gaussian}(\mu\delta t,\, \sigma^2\delta t)} + \ldots + \Delta x_{t-2} + \Delta x_{t-1} + \Delta x_t \xrightarrow{t\to\infty} \text{Gaussian}(\mu t, \sigma^2 t)$$

by <u>central limit theorem</u>

less well appreciated fact:

CLT also applies locally for sub-intervals of length $\delta t \gg 1$.

thus, if <u>coarse-grain</u> over timescale $\delta t \gg 1$ but $\delta t \ll t$, can write sum as:

$$X(t) = \underbrace{\Delta\tilde{x}_1}_{} + \Delta\tilde{x}_2 + \ldots + \Delta\tilde{x}_{(t/\delta t)}$$

$$\Delta x_1 + \ldots \Delta x_{\delta t} \sim \text{Gaussian}(\mu\delta t, \sigma^2\delta t)$$

or in recursive notation,

$$X(t+\delta t) = X(t) + \mu\delta t + \sqrt{\sigma^2\delta t}\; Z_t$$

Z_t: standard Gaussian, $\langle Z_t\rangle = 0$, $\langle Z_t^2\rangle = 1$

people often write this as <u>stochastic differential equation</u> (SDE)

$$\frac{dx}{dt} = \underbrace{\mu}_{\substack{\text{deterministic}\\ \text{part.}}} + \underbrace{\sqrt{\sigma^2}\,\eta(t)}_{\substack{\text{stochastic}\\ \text{part.}}}$$

$\eta(t)$ → "Brownian noise term"

⇒ SDEs have subtle properties, but for our purposes, we can treat them as simply code for series expansion and we'll be fine. will see some egs. soon.

<u>upshot</u>: can now write formula for probability of abitrary path: $x(0), x(\delta t), \ldots x(t)$:

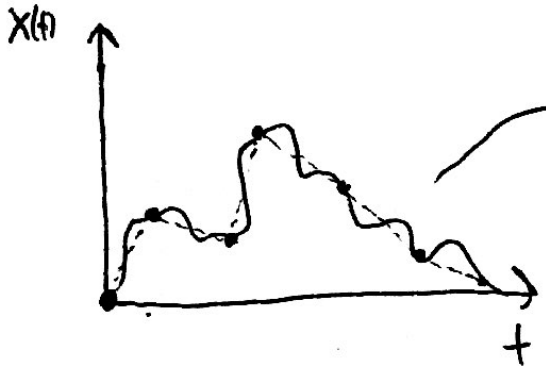
$$p\big(x(0), x(\delta t), \ldots, x(t)\big) = \prod_{i=0}^{(t/\delta t)-1} \frac{1}{\sqrt{2\pi\sigma^2\delta t}} \exp\left[-\frac{\big(x((i+1)\delta t) - x(i\delta t) - \mu\delta t\big)^2}{2\sigma^2\delta t}\right]$$

in differential notation:

"path integral"

$$= \left(2\pi\sigma^2\delta t\right)^{-\frac{t}{2\delta t}} \exp\left[-\sum_{i=0}^{(t/\delta t-1)} \left[\frac{x[(i+1)\delta t] - x(i\delta t)}{\delta t} - \mu\right]^2 \frac{1}{2\sigma^2}\,\delta t\right]$$

$$\exp\left[-\int_0^t \frac{\left(\frac{dx}{dt} - \mu\right)^2}{2\sigma^2}\,dt\right] \mathcal{D}\big(x(t)\big)$$

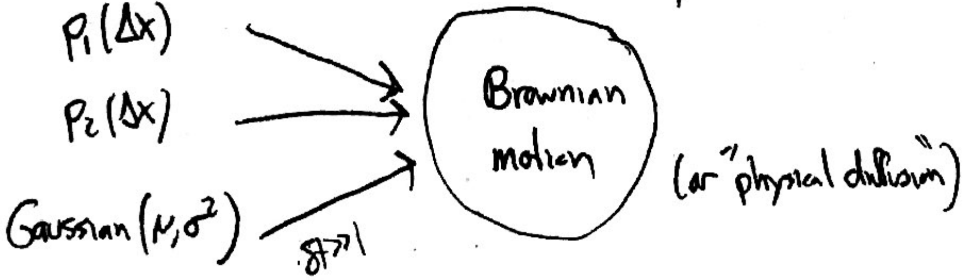


same probabilities as Gaussian model!

$\Rightarrow$ thus, for large class of random walks ($p(\Delta x)$) random paths have similar statistica properties when viewed over sufficiently long length scales ($t \gg 1$, $\delta t \gg 1$)

microscopic models:

"universality class" = microscopic models + limit (e.g. $\delta t \gg 1$)



$\hookrightarrow$ can use any microscopic model to predict behavior of universality class so we choose the one we can solve!

Note: "universal" is slight misnomer here ... it's not that $X(t)$ from $p_1(\Delta x)$ and $p_2(\Delta x)$ are identical. we expect to be able to tell them apart on sufficiently short timescales ($\delta t \sim 1$ or when CLT no longer applies)

$\Rightarrow$ Diffusion/brownian motion comes up a lot in physics (molecular timescales are very fast!)

How does it enter in evolutionary problems? $\rightarrow$ simulated $f(t)$ trajectories suggest that it might.

Basic idea: evolutionary phenomena take place over many generations... maybe diffusion-like behavior emerges when $1 \ll \delta t \ll t$?

Given markov model (e.g. serial dilution, wright fisher) can define

$$\Delta f = f(t+\Delta t) - f(t) = \frac{\text{Poisson}\left(N_0 \frac{f(t)e^{s\Delta t}}{f(t)e^{s\Delta t}+1-f(t)}\right)}{\text{Poisson}\left(N_0 \frac{f(t)e^{s\Delta t}}{f(t)e^{s\Delta t}+1-f(t)}\right) + \text{Poisson}\left(N_0 \frac{1-f(t)}{f(t)e^{s\Delta t}+1-f(t)}\right)} - f(t)$$

Same R.V.

mean and variance depend on $f(t)$!

$$\delta f = f(t+\delta t) - f(t) = \Delta f_t + \Delta f_{t+\Delta t} + \dots \Delta f_{t+\delta t-\Delta t}$$

$\rightarrow$ can't do simple CLT from before.

$\Delta f_t \rightarrow \mu(f(t))$, $\sigma^2(f(t))$

$\Delta f_{t+\Delta t} \rightarrow \mu(f(t+\Delta t))$, $\sigma^2(f(t+\Delta t))$

<u>Key idea</u>: if coarse-graining over many gens ($\delta t \gg 1$) for CLT, but sufficiently few gens that $f(t+\delta t) \approx f(t)$ (or more specifically, $f(t+\delta t) - f(t) \ll f(t)$)

then can still have $\mu(f(t+i\Delta t)) \approx \mu(f(t))$ for all $i\Delta t \leq \delta t$
$\sigma^2(f(t+i\Delta t)) \approx \sigma^2(f(t))$

and can therefore apply CLT just like before:

$$\delta f = \Delta f(t) + \Delta f(t+\Delta t) + \ldots \Delta f(t+\delta t - \Delta t)$$
$$\approx \text{Gaussian}\left(\mu(f(t))\delta t, \sigma^2(f(t))\delta t\right) \equiv \mu(f(t))\delta t + \sqrt{\sigma^2(f(t))\delta t}\, Z_t$$

How can we show that this works (& when it works)?

$\Rightarrow$ use self-consistency argument & series expansions:

$\boxed{\text{Step 1}}$ (dynamics in single timestep)

If $f(t+i\Delta t) - f(t) \ll f(t)$ for all $i \leq \delta t/\Delta t$, must certainly hold for $i=1$. what conditions does this require?

(a) Need $\text{Poisson}\left(N_0 \frac{fe^{s\Delta t}}{fe^{s\Delta t}+1-f}\right) \approx \text{Poisson}(N_0 f) \Rightarrow \boxed{s\Delta t \ll 1}$

(b) saw before that $\frac{\text{Poisson}(N_0 f)}{\text{Poisson}(N_0 f) + \text{Poisson}(N_0(1-f))} \sim f + \frac{g(f)}{\sqrt{N_0}}$

$\Rightarrow \boxed{N_0 \gg 1}$ (strictly speaking, $N_0 f \gg 1$, $N_0(1-f) \gg 1$)

Note: if worried about discrete individuals—good! we will revisit later when we talk about asymptotic matching & low frequency dynamics

(c) In these two limits [$s\Delta t \ll 1$, $N_0 \gg 1$] can calculate leading order contributions to $\mu(f(t)) = \langle \Delta f \rangle$ and $\sigma^2(f(t)) = \mathrm{Var}(\Delta f)$

(i) Argument of Poisson:
$$\frac{f(t)e^{s\Delta t}}{f(t)e^{s\Delta t} + 1 - f} \approx \frac{f(1+s\Delta t)}{fs\Delta t + 1} \approx f(1+s\Delta t)(1-fs\Delta t)$$
$$\approx f + s\Delta t\, f(1-f) + \text{h.o.t.}$$

(ii) Gaussian approx for Poisson:
$$\mathrm{Poisson}\left(N_0\left[f + s\Delta t\, f(1-f)\right]\right) \approx$$
$$N_0\left(f + s\Delta t\, f(1-f)\right) + \sqrt{N_0 f}\, Z_1$$

$$\Rightarrow f(t+\Delta t) = \frac{N_0\left[f + s\Delta t\, f(1-f)\right] + \sqrt{N_0 f}\, Z_1}{N_0\left[f + s\Delta t\, f(1-f)\right] + \sqrt{N_0 f}\, Z_1 + N_0\left[1 - f - s\Delta t\, f(1-f)\right] + \sqrt{N_0(1-f)}\, Z_2}$$

$$= \frac{f + s\Delta t\, f(1-f) + \sqrt{\frac{f}{N_0}}\, Z_1}{1 + \sqrt{\frac{f}{N_0}}\, Z_1 + \sqrt{\frac{1-f}{N_0}}\, Z_2} \approx f + s\Delta t\, f(1-f) + \underbrace{\left(\sqrt{\frac{f}{N_0}} - f\sqrt{\frac{f}{N_0}}\right)Z_1 - f\sqrt{\frac{1-f}{N_0}}\, Z_2}_{\sqrt{\frac{(1-f)^2 f}{N_0} + \frac{f^2(1-f)}{N_0}}\, Z_3}$$

$$\approx \underbrace{f + s\Delta t\, f(1-f)}_{\mu(f)} + \underbrace{\sqrt{\frac{f(1-f)}{N_0}}}_{\sigma^2(f)}\, Z$$

$\boxed{\text{that's one } \cancel{\text{generation}} \text{ cycle}}$

($\Delta t \cdot$ generations)

$\boxed{\text{step 2}}$ add up contribution over δt generations: $\left(\frac{\delta t}{\Delta t} \text{ cycles}\right)$

$$\delta f = s\,\Delta t\, f(1-f)\left(\frac{\delta t}{\Delta t}\right) + \sqrt{\frac{f(1-f)\,\delta t}{N_0\,\Delta t}}\; Z_t$$

$$= s f(1-f)\,\delta t + \sqrt{\frac{f(1-f)\,\delta t}{\underbrace{(N_0 \Delta t)}_{N_e \,=\, \text{effective strength of genetic drift}}}}\; Z_t$$

$\boxed{\text{step 3}}$ check self-consistency: $\delta f \ll f$ [+ $\delta f \ll 1-f$]

$\Rightarrow$ need $\delta t \ll 1/s$, $\delta t \ll N_0 f(1-f)\Delta t$

from $f(t+\delta t) \approx f(t)$ condition.

need $\delta t \gtrsim \Delta t$ from CLT condition. $\Rightarrow$

works when $s \to 0$, $N_0 \to \infty$
$N_0 s \Delta t$ can be anything

when this is true, we have:

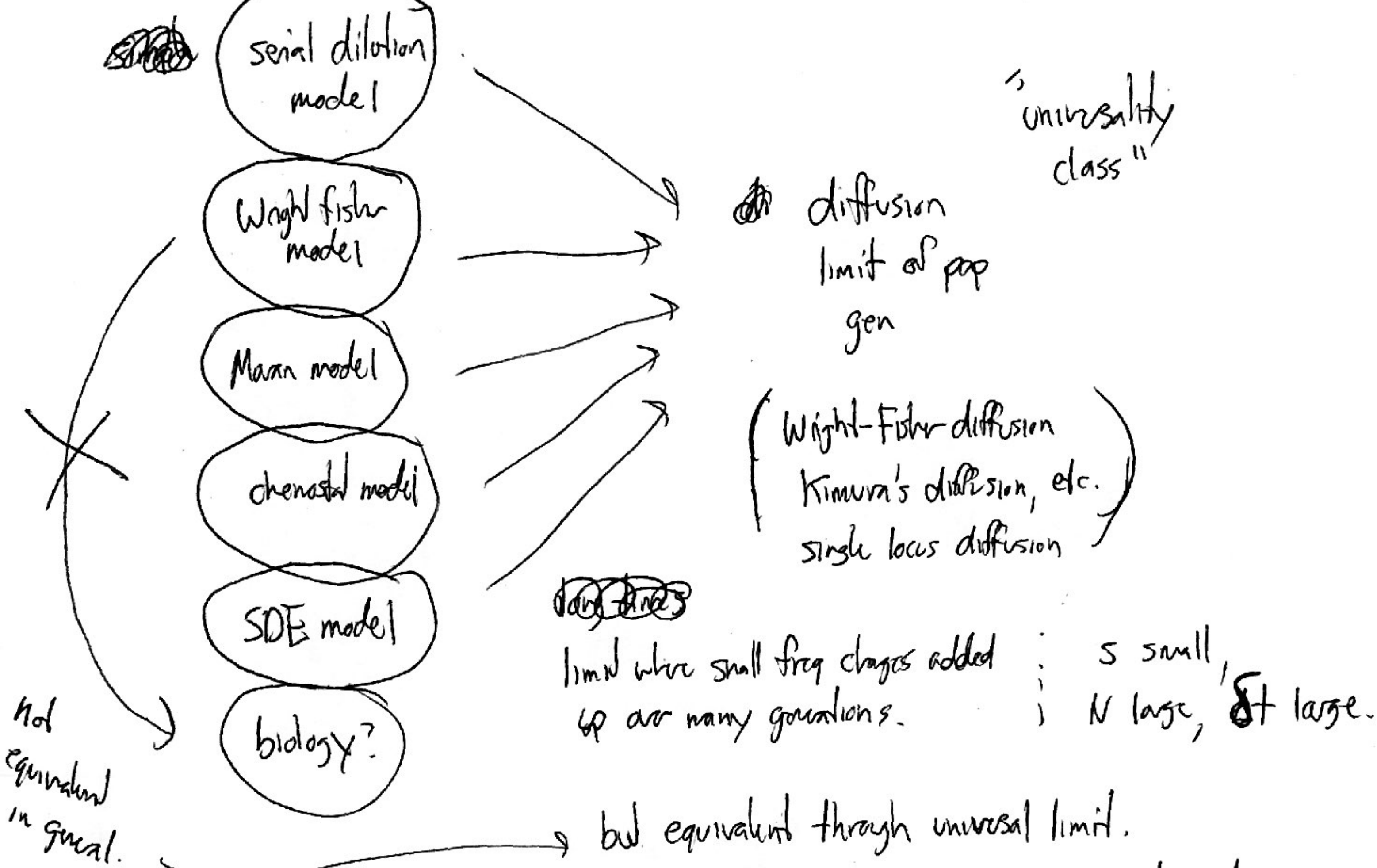
$$f(t+\delta t) \approx f(t) + s f(1-f)\,\delta t + \sqrt{\frac{f(1-f)\,\delta t}{N_e}}\; Z_t$$

$$\Updownarrow$$

SDE: $$\frac{df}{dt} = s f(1-f) + \sqrt{\frac{f(1-f)}{N}}\,\eta(t)$$

$\boxed{\text{the "diffusion limit" of population genetics (Wright, Kimura, Ewens)}}$

was a lot of work to show for simple serial dilution model, but can also show (w/ similar amount of work) that



but equivalent through universal limit.
THAT'S why Wright-Fisher model can sometimes be useful in practice.

⟹ can also see that approach likely to break down on short timescales when details of birth & death process likely important.

<u>IMPORTANT</u>: in pop. gen. literature, can sometimes read that diffusion is approximation to Wright-Fisher model, sometimes want <u>true</u> results. (exact sol'n to WF). can now see that this is not likely to be very useful: In regimes where differences b/w WF & diffusion limit are important, differences between WF & biology ALSO likely to be important ⇒ need detailed model.

Note: this is not the way that the diffusion limit is usually presented in pop. gen. I think the previous derivation gets at the key physical concepts (& is slightly easier to extend to more complex scenarios, as we will see.)

Still, worth presenting the traditional derivation in case you encounter it somewhere else. this derivation works for arbitrary* markov process,

$$X_0 \to X_1 \to X_2 \to \cdots \to X_t \quad w/ \quad Pr(X_i \to X_{i+1} | X_i) = p_1(X_{i+1}|X_i)$$

→ single step transition.

we start by focusing on probability density,

$$p(x,t|x_0) = \text{prob of being at } x \text{ @ time } t \text{ given } x=x_0 \text{ @ } t=0$$

then consider all positions in previous timestep:

$$p(x,t+1|x_0) = \int dx' \, p_1(x|x')\, p(x',t|x_0)$$

→ recursive formula for $p(x,t)$

$$= \int d\Delta x \; p_1(x|x-\Delta x)\, p(x-\Delta x, t|x_0)$$

then expand in time and in Δx:

$$p(x,t+1|x_0) \approx p(x,t) + \partial_t p(x,t|x_0)$$

$$p(x-\Delta x,t|x_0) \approx p(x,t|x_0) - \Delta x\, \partial_x p(x,t|x_0) + \frac{1}{2}\Delta x^2 \partial_x^2 p(x,t|x_0)$$

Expanding & collecting terms, can write as:

$$\frac{dp(x,t|x_0)}{dt} = -\frac{d}{dx}\left[\mu(x)p(x,t|x_0)\right] + \frac{1}{2}\frac{d^2}{dx^2}\left[\sigma^2(x)p(x,t|x_0)\right]$$

where: $\mu(x) = \int \Delta x \, p_1(x+\Delta x|x)\, d\Delta x$

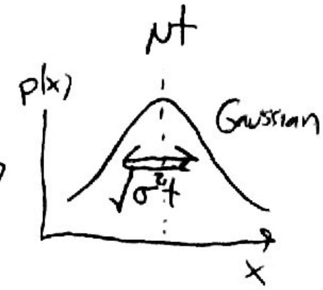
$\sigma^2(x) = \int \Delta x^2 \, p_1(x+\Delta x|x)\, d\Delta x$

this is known as the Fokker-Plank or forward equation.

equivalent to SDE or Langevin equation, $\frac{dx}{dt} = \mu(x) + \sqrt{\sigma^2(x)}\,\eta(t)$

e.g. for Gaussian random walk $(\mu(x)=\mu,\ \sigma^2(x)=\sigma^2)$

$$\Rightarrow \frac{dp}{dt} = -\mu\frac{dp}{dx} + \frac{\sigma^2}{2}\frac{d^2p}{dx^2} \longrightarrow$$ physical diffusion equation $\Rightarrow$



e.g. for Wright-Fisher diffusion,

$$\frac{dp(f,t|x_0)}{dt} = -s\frac{d}{df}\left[f(1-f)p(f,t)\right] + \frac{1}{2N}\frac{d^2}{df^2}\left[f(1-f)\right]$$

this is what you'll normally see in pop. gen. texts. it is equiv to SDE

$$\frac{df}{dt} = sf(1-f) + \sqrt{\frac{f(1-f)}{N}}\,\eta(t)$$

and also to the path integral: $p(f(t)) = \mathcal{D}(f(t))\, e^{-N\int_0^t \frac{\left(\frac{df}{dt} - sf(1-f)\right)^2}{2f(1-f)}dt}$