Maximum-Likelihood and Domain Constraints

Proof. $\partial_k \ln P(k\,|\,\{\ell\}) = \partial_k \ln P(\{\ell\}\,|\,k)$

By Bayes' rule,

$$P(k\,|\,\{\ell\}) = P(\{\ell\}\,|\,k)P(k)/P(\{\ell\}). \qquad \textbf{[1]}$$

However, we have no prior belief about $P(k)$ so we assume that this is constant, and $P(\{\ell\})$ does not depend on k.

Error in $1/k$.

The calculation of the error bar for $1/k$ is closely related to the error bar in k, as presented in the text. Defining $\theta \equiv 1/k$ (referred to as $\langle \Delta Lk \rangle$ in the text), we have:

$$\frac{d}{d\theta} = \frac{dk}{d\theta}\frac{d}{dk} = -k^{-2}\frac{d}{dk}. \qquad \textbf{[2]}$$

Retaining primes for differentiating with respect to k, then, we find:

$$\begin{aligned} \partial_\theta f &= -k^2 f' \\ \partial_\theta^2 f &= -k^2(-k^2 f')' = k^4 f'' + k^2 f'(k^2)'. \end{aligned} \qquad \textbf{[3]}$$

If we want to evaluate the second derivative at a value k_* (or, equivalently, θ_*) when $f' = 0$, we have:

$$\partial_\theta^2 f\,|_{\theta_*} = k_*^4 f''\,|\,k_* \qquad \textbf{[4]}$$

from which it follows that $\sigma(k)$ and $\sigma(\theta)$, the error bars in k and in $\theta = 1/k$, are related by k^2:

$$\begin{aligned} \sigma(\theta)^{-2} &= \partial_\theta^2 \ln P|_{k_*} = k_*^4 \partial_k^2 \ln P|_{k_*} = k_*^4 \sigma(k)^{-2} \\ \sigma(\theta) &= \sigma(k)/k^2 \end{aligned} \quad \textbf{[5]}$$

In the case $\Lambda \to \infty$, we have :

$$\sigma(\theta) = \sigma(k)/k_*^2 = k_*/\sqrt{N}/k_*^2 = \theta_*/\sqrt{N}. \quad \textbf{[6]}$$

Gaussian Approximation to the Error

If we approximate $P(k\,|\,\{\ell\})$ as a Gaussian, we have:

$$\begin{aligned} P(k\,|\,\{\ell\}) &= \frac{e^{-\frac{1}{2}(k-k_*)^2/\sigma^2}}{\sqrt{2\pi\sigma^2}} \\ \ln P(k\,|\,\{\ell\}) &= -\frac{1}{2}(k-k_*)^2/\sigma^2 - \frac{1}{2}\ln(2\pi\sigma^2) \\ \partial_k^2 \ln P(k\,|\,\{\ell\}) &= -1/\sigma^2. \end{aligned} \quad \textbf{[7]}$$

Calculation of the Functional Form of the Globally Constrained Distribution

The measured distribution in Fig. 2*a* can be analytically calculated and has a simple form in the statistical steady state (*i.e.*, the limit of many simulation rounds, after which the intial condition $s = 0$ has effectively been forgotten). We demonstrate this here.

Consider moving from displacement x to displacement y constrained such that $x, y < z$, where z is the constraint (denoted $\Delta L k_{\max}^0$ in the text). Displacements are incremented in steps s drawn from the true distribution $f(s) = e^{-s}$ (to simplify the calculation, we will measure distances in this section in units such that $1/k = 1$). Given these constraints, we observe a distribution of steps t that obeys the following rules:

- If the addition of the step size s to the initial position x does not exceed the constraint z, or

$$x + s < z \Leftrightarrow s < z - x, \quad \textbf{[8]}$$

then we register a step size $t = s$. The new position *y* then is given by $y = x + s$.

- If the addition of the step size *s* to the initial position *x* exceeds the constraint *z*, or

$$x + s > z \Leftrightarrow s > z - x \,, \qquad \textbf{[9]}$$

then we do not register a step size *t*. The new position *y* is set to 0.

These relationships will be expressed in terms of conditional distributions in the analysis below:

$$\begin{aligned} s < z - x &\Rightarrow p(y \mid x, s) = \delta(y - (x + s)),\, p(t \mid x, s) = \delta(t - s) \\ s > z - x &\Rightarrow p(y \mid x, s) = \delta(y),\, p(t \mid x, s) = \delta(t) \end{aligned} \qquad \textbf{[10]}$$

Definitions

The conditional independence conditions give

$$P \equiv p(y, t, x, s) = p(y \mid x, s)\, p(t \mid x, s)\, p(x)\, p(s) \qquad \textbf{[11]}$$

from which we define

$$\begin{aligned} f(s) &\equiv \int dy\, dt\, dx P \\ p^n(x) &\equiv \int dy\, dt\, ds P \\ p^{n+1}(y) &\equiv \int dt\, dx\, ds P \\ q(t) &\equiv \int dy\, dx\, ds P, \end{aligned} \qquad \textbf{[12]}$$

where n is the number of steps observed. We are interested in calculating the function $q(t)$ in the limit $n \to \infty$. To do this we will first find $p^\infty(x)$ and then evaluate

$$q(t) = \int dx\, ds\; p(t \mid x, s)\, p^\infty(x)\, f(s) \,. \qquad \textbf{[13]}$$

Transition Element $p(y \mid x)$

The conditional distribution $p(y \mid x)$ is the propagator of the distribution of possible lengths from one “roll” to the next. We can calculate this as

$$
\begin{aligned}
p(y\,|\,x) &= \int ds\; p(y\,|\,x,s)\,f(s) \\
&= \int_0^{z-x} ds\; p(y\,|\,x,s)\,f(s) + \int_{z-x}^{\infty} ds\; p(y\,|\,x,s)\,f(s) \\
&= \int_0^{z-x} ds\; \delta(y-(x+s))\,f(s) + \delta(y)\int_{z-x}^{\infty} ds\; f(s) \\
&= \mathrm{e}^{-y+x}\Theta((z-x)-(y-x))\Theta((y-x)-0)) + \delta(y)\mathrm{e}^{-z+x} \\
&= \mathrm{e}^{-y+x}\Theta(z-y)\Theta(y-x) + \delta(y)\mathrm{e}^{-z+x} \\
&= \mathrm{e}^{-y+x}\Theta(y-x) + \delta(y)\mathrm{e}^{-z+x}
\end{aligned} \qquad \textbf{[14]}
$$

(since $z \geq y$) from which

$$
p^{n+1}(y) = \int_0^z dx p(y\,|\,x)\,p^n(x) = \mathrm{e}^{-y}\int_0^y dx p^n(x)\mathrm{e}^x + \delta(y)\mathrm{e}^{-z}\int_0^z dx p^n(x)\mathrm{e}^x. \qquad \textbf{[15]}
$$

The integration over a delta function in Eq. **14**, and the resulting heaviside function, is a special case of the more general caveat

$$
\begin{aligned}
\int_a^b ds f(s)\delta(s-s_0) &= \begin{cases} f(s_0), a < s_0 < b \\ 0, otherwise \end{cases} \\
&\equiv f(s_0)\Theta(s_0-a)\Theta(b-s_0).
\end{aligned} \qquad \textbf{[16]}
$$

Ansatz

Consider the ansatz

$$
p^n(u) \equiv \pi_0^{(n)}\delta(u) + \mathrm{e}^{-u}\sum_{j=0}^{\infty} c_j^{(n)}\frac{u^j}{j!}. \qquad \textbf{[17]}
$$

Clearly when $n = 0$ we have $p^{(0)}(u) = \delta(u)$, so

$$
\begin{aligned}
\pi^{(0)} &= 1 \\
c_j^{(0)} &= 0 \;\forall\; j
\end{aligned} \qquad \textbf{[18]}
$$

and the initial distribution is within this functional form. We need only show that all later distributions have this functional form to have solved for the distribution of lengths for all times.

$$p^{(n+1)}(y) = e^{-y}\int_0^y dx e^x p^{(n)}(x) + \delta(y) e^{-z}\int_0^z dx\, e^x p^{(n)}(x)$$
$$e^x p^{(n)}(x) = \sum_{j=0}^{\infty}\frac{c_j^{(n)}}{j!}x^j + \pi_0^{(n)}\delta(x) \quad \textbf{[19]}$$
$$p^{(n+1)}(y) = e^{-y}\sum_{j=0}^{\infty}\frac{c_j}{(j+1)!}x^{j+1} + e^{-y}\pi_0^n + \delta(y)\left(e^{-z}\sum_{j=0}^{\infty}\frac{c_j}{j+1!}z^{j+1} + e^{-z}\pi_0^n\right).$$

So clearly the functional form is preserved compared to the ansatz (Eq. **17**).

$$c_0^{n+1} = \pi_0^n$$
$$c_{j+1}^{n+1} = c_j^n \quad \textbf{[20]}$$
$$\pi_0^{n+1} = e^{-z}\left(\pi_0^n + \sum_{k=1}^{\infty}\frac{c_{k-1}}{k!}z^k\right).$$

Statistical Steady State

At steady state, we must have $\pi_0^{n+1} = \pi_0^n \equiv \pi_0$, $c_j^{n+1} = c_j^n \equiv c_j$. Consequently,

$$c_0^{n+1} = \pi_0^n \Rightarrow c_0 = \pi_0$$
$$c_{j+1}^{n+1} = c_j^n \Rightarrow c_j = c_0 \quad \textbf{[21]}$$
$$\pi_0^{n+1} = e^{-z}\left(\pi_0^n + \sum_{k=1}^{\infty}\frac{c_{k-1}}{k!}z^k\right) \Rightarrow c_0 = e^{-z}(c_0 + c_0 e^z - c_0)$$

the last of which yields $1 = 1$. Normalization fixes the value for c_0:

$$1 = \int dx c_0\left(\delta(x)+1\right) \Rightarrow c_0 = 1/(1+z)$$
$$p^{\infty}(x) = \frac{\delta(x)+1}{1+z}. \quad \textbf{[22]}$$

Calculation of $q(t)$

We can now finally calculate the distribution of observed steps $q(t)$. Note first the simplifications (recalling the procedure for definite integration over delta functions from Eq. **16**):

$$
\begin{aligned}
\int_0^{z-x} dsf(s)\,p(t\mid x,s) &= f(t)\Theta((z-x)-t)\Theta(t-0) \\
&= f(t)\Theta((z-t)-x) \\
\int_{z-x}^{\infty} dsf(s)\,p(t\mid x,s) &= \delta(t)\int_{z-x}^{\infty} dsf(s) \\
&= \delta(t)\mathrm{e}^{-(z-x)}
\end{aligned}
\qquad \textbf{[23]}
$$

from which

$$
\begin{aligned}
q(t) &= \int_0^z dx\,ds\; p(t\mid x,s)\,p^{\infty}(x)\,f(s) \\
&= \int_0^z dx p^{\infty}(x)\left(\int_0^{z-x} dsf(s)\,p(t\mid x,s) + \int_{z-x}^{\infty} dsf(s)\,p(t\mid x,s)\right) \\
&= \int_0^z dx\left(\frac{\delta(x)+1}{1+z}\right)\left(f(t)\Theta\left((z-t)-x\right)+\delta(t)\mathrm{e}^{-z+x}\right) \\
&= f(t)\frac{1}{1+z}\int_0^{z-t} dx\left(\delta(x)+1\right)+\delta(t)\mathrm{e}^{-z}\frac{1}{1+z}\int_0^z dx\,\mathrm{e}^{x}\left(1+\delta(x)\right) \\
&= f(t)\frac{1+z-t}{1+z}+\delta(t)\mathrm{e}^{-z}\frac{1}{1+z}\left(\mathrm{e}^{z}-1+1\right) \\
&= \frac{(1+z-t)}{1+z}\mathrm{e}^{-t}+\delta(t)\frac{1}{1+z}
\end{aligned}
\qquad \textbf{[24]}
$$

which is properly normalized, since

$$
\begin{aligned}
\int_0^z dt(1+z-t)\mathrm{e}^{-t}+1 &= (1+z)(1-\mathrm{e}^{-z})-(-z\mathrm{e}^{-z}-\mathrm{e}^{-z}+1)+1 \\
&= (1+z)(-\mathrm{e}^{-z}+1)+z\mathrm{e}^{-z}+\mathrm{e}^{-z} \\
&= 1+z.
\end{aligned}
\qquad \textbf{[25]}
$$

Result: Comparison of Cumulative Probability Distributions

$$
\begin{aligned}
P(t' < t \mid t' > 0) &= \frac{P(t > t' > 0)}{P(t > 0)} \\
&= \frac{\int_{0^+}^{t} dt' q(t')}{\int_{0^+}^{z} dt' q(t')} \\
&= \left(\frac{1+z}{z}\right) \frac{\left(z + \left(t - z\right) \mathrm{e}^{-rt}\right)}{1+z} \\
&= 1 - \mathrm{e}^{-rt}(1 - (t/z))
\end{aligned}
\qquad \textbf{[26]}
$$

In Fig. 5, we plot the experimental result (from the simulations described in the text) against the theoretical prediction, for $z = 130$, $r = 60$. Note that for $t << z$, $P(t' < t) \approx 1 - \mathrm{e}^{-rt}$, as expected.

Global Constraints Do Not Permit Estimation of the Distribution Parameter by Simply Counting the Fraction of Discarded Events

Consider the probability distribution for possible step sizes *s*:

$$p(s) = k \exp(-ks) . \qquad \textbf{[27]}$$

A step of size *s* is defined by a transition between two levels, from level *x* to level *y*.
From the probability distribution we can derive P_i^+, the probability that the *i*th step will be terminated by the global constraint:

$$P_i^+ \equiv \int_{s=z-x_i}^{\infty} p(s) ds = e^{-k(z-x_i)} , \qquad \textbf{[28]}$$

where *z* is the highest possible value for *y*, which is given by the constraint. The likelihood Λ of the observed data *D*, summarized in terms of $N_+ / (N_+ + N_-) \equiv N_+ / N$, the fraction of the reaction steps terminated by the global constraint, is (using i_+ and i_- to index steps terminated or not, respectively, by the global constraint)

$$
\begin{aligned}
\Lambda &= P(D) = \prod_{i_+} P_{i_+}^{+} \prod_{i_-} P_{i_-}^{-} \\
\ln \Lambda &= \sum_{i_+} -k(z - x_{i_+}) + \sum_{i_-} \ln(1 - e^{-k(z - x_{i_-})}) \\
&= -N_+ k(z - \langle x_+ \rangle) + N_- \left\langle \ln(1 - e^{-k(z - x_-)}) \right\rangle .
\end{aligned}
\qquad \textbf{[29]}
$$

Eq. **29** makes it clear that we cannot estimate k by only keeping track of N_+, N_-; rather, we must keep track of the initial location of all the steps that did not terminate as well as of $\langle x_+ \rangle$: the average location of the reactions that did terminate.