An Asymptotic Correction to Luce's Choice Axiom for Competing Correlated Cox Processes

Peter Cotton*

August 2026 — working draft

Abstract

Consider N channels firing at random times, with rates driven by a hidden state they all share. Conditionally on that state each channel is Poisson, and an observer records which channels fire and in what order. When the driver mixes fast relative to the firing rates, the probability that channel i fires first approaches the law of Luce's Choice Axiom in the stationary mean rates. We give the first correction in the timescale ratio, governed by the time-integrated covariance K of the rates, and every claim below holds at that order. Recording the first two arrivals recovers all $N^2 - 2N$ observable combinations of K, matching what blocking channels and rerunning the process would give, while the winner alone recovers none of them. The diagonal of K is unobservable, absorbed into the rates, so only correlated modulation leaves a trace. And K is symmetric when the driver satisfies detailed balance, so its observable asymmetry certifies a driver held out of equilibrium, visible with three channels and never with two.

1 Introduction

Consider N channels that fire at random times, whose rates are not constants but are driven by a hidden state all of them share. Conditionally on the path of that state each channel is Poisson, so the whole is a doubly stochastic, or Cox, process in the sense of Cox (1955), with N marks and a common driver. Three fields use this model under three names.

In chemical kinetics a reaction escapes through one of several channels while the surrounding configuration rearranges underneath, which is the picture behind every kinetic Monte Carlo (KMC) selection (Bortz et al., 1975; Gillespie, 1977), and a diffusing particle finding one of several windows is its continuum version. In a distributed system a request is sent to N redundant servers and the first response wins, with response rates driven by shared congestion, and the same race reruns billions of times a day (Dean and Barroso, 2013). In neuroscience, neurons race to threshold under a shared cortical state, and visual categorization completes fast enough that the earliest spikes must be carrying the information (Thorpe et al., 1996), with the race-to-choice reading going back to Bundesen (1990).

The three fields share a predicament and an asset. The rates' common driver is exactly what cannot be observed, and blocking channels, the experiment that would probe it, ranges from costly to impossible. What all three have instead is replication with the order of events logged

*Code and machine verification for every claim in this paper: `https://github.com/microprediction/kinetics` (experiments 7 and 9). Web companion: `https://kinetics.microprediction.org`.

for free: simulation runs, served requests, stimulus trials. Digital commerce sits in the same position, with items competing for a session's first engagement under a shared and unobserved intent, and clickstreams recording the order at scale.[1]

The observer's data is the order of arrivals, and the question is how far down that order the information about the driver reaches. Write ε for the ratio of the driver's mixing time to the timescale on which the channels fire, so that small ε is fast mixing. Our main result is that, to first order in ε, the information saturates at second place. Recording the first two arrivals, from the full system alone, recovers everything the arrival probabilities reveal about the driver at order ε, matching what every possible channel-blocking experiment would yield at the same order, while the winner alone recovers $N-1$ of the N^2-N-1 recoverable quantities.

Making that precise requires knowing how the driver enters the law of the first arrival, and that is a homogenization question. When the driver mixes faster than the channels fire, the winning probabilities converge to softmax in the stationary mean intensities. The first correction is governed by

$$K_{jk} = \int_0^\infty \mathrm{Cov}_\pi(\lambda_j(Y_0), \lambda_k(Y_t))\, dt, \tag{1}$$

the time-integrated covariance of the firing rates. It is a dynamical quantity. Two channels whose rates move together at any instant contribute nothing unless that agreement persists over the time the driver takes to forget itself.

K is the object to be identified, and the mechanism behind the sufficiency of second place is an identity. Deleting a channel changes the outcome only in those realizations the deleted channel arrived first, and in those the runner-up inherits, so second-place frequencies are deletion experiments that have already been run. They identify K modulo an explicit $(N+1)$-dimensional family of directions that, like a rate shift common to every channel, move no branching ratio and so change nothing any observer could see. One pair $(\bar{\lambda}, K)$ serves every availability set. Fluctuations common to all channels cancel exactly, a classical fact (Elandt-Johnson, 1976; Marley and Colonius, 1992). A driver with r modes gives K of rank r.

Section 2 sets up the arrival process and its killed resolvent. Section 3 proves the homogenization expansion and places it against the motional-narrowing literature, where the diagonal of K has been known for seventy years. Section 4 collects the structural properties, Section 5 the second-place identity, and Section 6 the identification and design results. Sections 7 and 8 verify everything, first on finite chains where every quantity is an exact linear solve, then on reflected Brownian motion in a disk with boundary reaction windows.

2 Competing arrivals under a shared driver

Let Y_t be an ergodic Markov process on a state space with generator L/ε and invariant law π. The parameter ε is the ratio of the environment's mixing time to the timescale on which channels fire. Small ε means fast mixing.

Each channel $i \in \{1, \dots, N\}$ fires at intensity $\lambda_i(Y_t) \geq 0$, conditionally Poisson given the path of Y. The first channel to fire wins. Conditionally on the driver the arrivals are Poisson, so the whole is a doubly stochastic Poisson process (Cox, 1955; Grandell, 1976) with N marks, observed until its first event. When Y is a finite chain it is a Markov-modulated Poisson process (Fischer and Meier-Hellstern, 1993) carrying marks in the sense of He and Neuts (1998), and the

[1] Replication is essential. Correlated default in credit risk has the same formal structure but supplies a single realization, one economy along one macroeconomic path.

λ_i are cause-specific hazards in the competing-risks sense. The same marked structure appears in reduced-form credit risk (Lando, 1998), though an economy supplies a single realization, whereas the identification questions of Section 6 presume replication. The arrivals are events of a point process, not the argmin of latent performances, so this is not a Thurstonian race. For an available set $A \subseteq \{1, \dots, N\}$ write $\Lambda_A(y) = \sum_{j \in A} \lambda_j(y)$ for the total hazard, and write $\bar{\lambda}_i = \int \lambda_i \, d\pi$ for the stationary mean intensity.

Let $u_i^A(y)$ be the probability that channel i wins, started from $Y_0 = y$. Conditioning on the first instant and using the conditional Poisson structure gives

$$\left(\frac{L}{\varepsilon} - \Lambda_A\right) u_i^A = -\lambda_i, \qquad u_i^A = \left(\Lambda_A - \frac{L}{\varepsilon}\right)^{-1} \lambda_i. \tag{2}$$

This is a killed resolvent. The generator propagates the environment and the multiplication operator Λ_A kills paths at the total firing rate. On a finite state space it is a linear system, which is what makes the verification in Section 7 exact.

The observable of interest is the win probability from a stationary start,

$$p_i^A = \int u_i^A \, d\pi. \tag{3}$$

Summing (2) over $i \in A$ gives $\sum_{i \in A} u_i^A = 1$, so the p_i^A are a probability vector on A for every ε.

3 Homogenization and the Green–Kubo correction

The expansion below is the averaging principle for a fast ergodic driver (Khas'minskii, 1966), in the operator form of Kurtz (1973). The proof pattern is standard in multiscale analysis (Pavliotis and Stuart, 2008, Ch. 9). An equation $Lg = h$ on the driver's state space has a solution only when h averages to zero under π, and these solvability requirements, imposed order by order, pin down every coefficient of the expansion. Write $\tilde{\lambda}_i = \lambda_i - \bar{\lambda}_i$ for the fluctuation of the ith intensity, and Π for the projection $\Pi f = (\int f d\pi)\mathbf{1}$ onto constants. The deviation operator

$$D = (\Pi - L)^{-1}(I - \Pi) \tag{4}$$

inverts L on the mean-zero subspace, and $Dg = \int_0^\infty e^{tL} g \, dt$ for mean-zero g. Note the sign it carries. Differentiating that integral gives $LDg = -g$, so D is minus the pseudo-inverse of L, and the sign propagates into the proof below. The Green–Kubo matrix is

$$K_{jk} = \int_0^\infty \mathrm{Cov}_\pi(\lambda_j(Y_0), \lambda_k(Y_t)) \, dt = \int \tilde{\lambda}_j \, (D\tilde{\lambda}_k) \, d\pi. \tag{5}$$

The second expression is what one computes. It replaces a time integral by a single linear solve. Note that K need not be symmetric. It is symmetric when the environment is reversible, and the verification in Section 7 deliberately uses a non-reversible chain so that the index placement in the theorem is actually tested.

The diagonal of K is not new. Spectroscopists observed in the 1950s that when a decaying system's rate fluctuates quickly, the decay behaves as if run at the average rate, with a correction given by the time-integrated autocovariance of the rate (the fluctuating quantity there is a transition frequency, an imaginary rate), the phenomenon called motional narrowing (Anderson, 1954; Kubo, 1954) and later organized into a general expansion by van Kampen (1974). For a single channel, the integral $\int_0^\infty \mathrm{Cov}_\pi(\lambda(Y_0), \lambda(Y_t))\, dt$ is exactly that correction. The name Green–Kubo comes from statistical physics, where formulas of Green (1954) and Kubo (1957) express transport coefficients, a diffusion constant or a viscosity, as time integrals of correlation functions computed in equilibrium. K has exactly that form, so we borrow the name. The oldest question about such matrices is Onsager's, whether they are symmetric (Onsager, 1931), and Section 6 shows the answer here carries observable content.

The scalar theory does not see branching ratios. A rate correction common to all channels cancels from a ratio and leaves the choice law untouched, so the choice law is moved only by the off-diagonal and generally non-symmetric entries K_{ji}. Theorem 1 is about those entries.

Each channel also carries an operational clock, its compensator $\Theta_i(t) = \int_0^t \lambda_i(Y_s)\, ds$. The random time-change theorem for point processes (Meyer, 1971; Papangelou, 1972) states that channel i's arrivals, read against the clock Θ_i, form a unit-rate Poisson process. For the stationary driver with generator $\mathcal{L}/\varepsilon$ the clocks concentrate, $\Theta_i(t) = \bar{\lambda}_i t$ plus a fluctuation of order $\sqrt{\varepsilon t}$, and the fluctuations satisfy a joint central limit theorem whose covariance is K up to symmetrization,

$$\frac{1}{t}\,\mathrm{Cov}_\pi\big(\Theta_j(t),\, \Theta_k(t)\big) \;\longrightarrow\; \varepsilon\left(K_{jk} + K_{kj}\right), \qquad t \to \infty.$$

So εK is the diffusion matrix of the vector of clocks, and the first-order correction below is a functional of the clock-fluctuation covariance. In this reading the structural results that follow are statements about clocks. When every channel is modulated by the same factor the N clocks are one shared clock, a time change carries the system to a constant-rate race, and the correction must vanish (Section 4). The shared-clock covariance is $K \propto \bar{\lambda}\bar{\lambda}^\top$, which is the member $d \propto \bar{\lambda}$ of the invisible family $\{d\bar{\lambda}^\top\}$ in Section 6. A shared time change is one direction of the null space. The remaining null direction $\mathrm{diag}(\bar{\lambda})$ is private clock jitter, invisible for the reason the rate gauge makes exact.

Theorem 1 (Homogenized choice law and its first correction)**.** *Let $\bar{\Lambda}_A = \sum_{j\in A} \bar{\lambda}_j$ and suppose $\bar{\Lambda}_A > 0$. Then as $\varepsilon \to 0$,*

$$p_i^A = \frac{\bar{\lambda}_i}{\bar{\Lambda}_A} - \frac{\varepsilon}{\bar{\Lambda}_A}\left[\sum_{j\in A} K_{ji} - \frac{\bar{\lambda}_i}{\bar{\Lambda}_A}\sum_{j,k\in A} K_{jk}\right] + O(\varepsilon^2). \tag{6}$$

In particular $p_i^A \to \bar{\lambda}_i/\bar{\Lambda}_A$, the softmax law in the log mean intensities $\theta_i = \log\bar{\lambda}_i$.

Three literatures arrive at this same limiting law by different routes. It is Luce's choice model (Luce, 1959), the conditional logit of McFadden (1974), and the softmax output layer of Bridle (1990). In the random-utility tradition it follows from Gumbel noise, and Yellott (1977) showed that representation is the only one consistent with the Choice Axiom once three or more alternatives are compared, though not for pairs.[2] Here fast mixing collapses the doubly stochastic race to the classical exponential race run at the averaged rates, which is Cox (1959), so the leading order travels the oldest route of all. The content of Theorem 1 begins at order ε, where the race is no longer exponential. The relabelling $\theta_i = \log\bar{\lambda}_i$ is a change of coordinates, not a derived exponential family.

[2]For pairs other distributions serve equally; Yellott (1977) shows that for triple comparisons "the double exponential representation is unique."

Proof. Multiply (2) by ε,

$$(L - \varepsilon\Lambda_A)\, u_i^A = -\varepsilon\lambda_i,$$

substitute the expansion $u_i^A = u_i^{(0)} + \varepsilon u_i^{(1)} + \varepsilon^2 u_i^{(2)} + O(\varepsilon^3)$, and collect powers of ε:

$$\varepsilon^0 : \quad L u_i^{(0)} = 0, \tag{7}$$

$$\varepsilon^1 : \quad L u_i^{(1)} = \Lambda_A u_i^{(0)} - \lambda_i, \tag{8}$$

$$\varepsilon^2 : \quad L u_i^{(2)} = \Lambda_A u_i^{(1)}. \tag{9}$$

Order ε^0. By ergodicity the kernel of L is the constants, so $u_i^{(0)} \equiv c_i$ for some number c_i.

Order ε^1. An equation $Lg = h$ has a solution only when $\int h\, d\pi = 0$. Applied to (8) with $u_i^{(0)} = c_i$,

$$0 = \int (\Lambda_A c_i - \lambda_i)\, d\pi = \bar\Lambda_A\, c_i - \bar\lambda_i, \qquad\qquad \text{hence} \quad c_i = \frac{\bar\lambda_i}{\bar\Lambda_A},$$

which is the leading softmax term. With this c_i the right side of (8) is mean zero and splits into fluctuations,

$$\Lambda_A c_i - \lambda_i = \left(\bar\Lambda_A + \tilde\Lambda_A\right) c_i - \left(\bar\lambda_i + \tilde\lambda_i\right) = c_i\tilde\Lambda_A - \tilde\lambda_i, \qquad\qquad \tilde\Lambda_A = \textstyle\sum_{j\in A}\tilde\lambda_j.$$

Inverting with $LD = -I$ on mean-zero functions,

$$u_i^{(1)} = -D\left(c_i\tilde\Lambda_A - \tilde\lambda_i\right) + m_i,$$

where the constant $m_i = \int u_i^{(1)}\, d\pi$ is not yet determined, because D maps into mean-zero functions and any constant can be added.

Order ε^2. The same solvability requirement applied to (9) is $\int \Lambda_A u_i^{(1)}\, d\pi = 0$. Substitute $\Lambda_A = \bar\Lambda_A + \tilde\Lambda_A$ and the two parts of $u_i^{(1)}$, and evaluate the four resulting terms:

$$\int \bar\Lambda_A\, m_i\, d\pi = \bar\Lambda_A\, m_i, \qquad\qquad \text{(constant times constant)}$$

$$\int \tilde\Lambda_A\, m_i\, d\pi = 0, \qquad\qquad (\tilde\Lambda_A \text{ is mean zero})$$

$$-\int \bar\Lambda_A\, D\left(c_i\tilde\Lambda_A - \tilde\lambda_i\right) d\pi = 0, \qquad\qquad (D \text{ maps into mean zero})$$

$$-\int \tilde\Lambda_A\, D\left(c_i\tilde\Lambda_A - \tilde\lambda_i\right) d\pi = -c_i \sum_{j,k\in A} K_{jk} + \sum_{j\in A} K_{ji}, \qquad\qquad \text{(by (5))}$$

where the last line uses $\int \tilde\Lambda_A D\tilde\lambda_i\, d\pi = \sum_{j\in A}\int \tilde\lambda_j D\tilde\lambda_i\, d\pi = \sum_{j\in A} K_{ji}$ and its double-sum analogue. Adding the four terms and solving for m_i,

$$m_i = -\frac{1}{\bar\Lambda_A}\left[\sum_{j\in A} K_{ji} - \frac{\bar\lambda_i}{\bar\Lambda_A}\sum_{j,k\in A} K_{jk}\right].$$

Finally, integrating the expansion against π and using that the D-part of $u_i^{(1)}$ is mean zero,

$$p_i^A = \int u_i^A\, d\pi = c_i + \varepsilon\, m_i + O(\varepsilon^2),$$

which is (6). ∎

Lemma 1 (Remainder bound)**.** *Suppose the state space is finite and $\Lambda_A(y) \geq \Lambda_{\min} > 0$ for every y. Let $u^{(1)}, u^{(2)}$ be the correctors constructed in the proof of Theorem 1 and $U = c + \varepsilon u^{(1)} + \varepsilon^2 u^{(2)}$. Then the defect of U is exactly $(L/\varepsilon - \Lambda_A)U + \lambda = -\varepsilon^2 \Lambda_A u^{(2)}$, and*

$$\left\|u^A - U\right\|_\infty \;\leq\; \varepsilon^2 \, \frac{\|\Lambda_A u^{(2)}\|_\infty}{\Lambda_{\min}}.$$

The $O(\varepsilon^2)$ of Theorem 1 is therefore a proven bound with an explicit constant, and repeating the construction one order deeper proves the $O(\varepsilon^3)$ remainder of the corrected law, and so on.

Proof. Apply the operator to the truncation term by term:

$$\left(\tfrac{L}{\varepsilon} - \Lambda_A\right) U + \lambda = \tfrac{1}{\varepsilon}\, Lc + \left(Lu^{(1)} - \Lambda_A c + \lambda\right) + \varepsilon \left(Lu^{(2)} - \Lambda_A u^{(1)}\right) - \varepsilon^2 \, \Lambda_A u^{(2)}.$$

The first bracket vanishes because c is constant. The second vanishes by (8), the equation $u^{(1)}$ was built to solve. The third vanishes by (9). What remains is the displayed defect,

$$\left(\tfrac{L}{\varepsilon} - \Lambda_A\right) U + \lambda = -\varepsilon^2 \, \Lambda_A u^{(2)}.$$

Subtracting this from the exact equation $(L/\varepsilon - \Lambda_A)\, u^A + \lambda = 0$ gives

$$\left(\tfrac{L}{\varepsilon} - \Lambda_A\right)\left(u^A - U\right) = \varepsilon^2 \, \Lambda_A u^{(2)}, \qquad \text{so} \qquad u^A - U = -\left(\Lambda_A - \tfrac{L}{\varepsilon}\right)^{-1} \varepsilon^2 \, \Lambda_A u^{(2)}.$$

The killed resolvent is bounded uniformly in ε. For any bounded f, the Feynman–Kac representation and the bound $\Lambda_A \geq \Lambda_{\min}$ give

$$\begin{aligned} \left|\left(\Lambda_A - \tfrac{L}{\varepsilon}\right)^{-1} f\,(y)\right| &= \left|\mathbb{E}_y \int_0^\infty e^{-\int_0^t \Lambda_A(Y_s^\varepsilon)\, ds} f(Y_t^\varepsilon)\, dt\right| \\ &\leq \|f\|_\infty \int_0^\infty e^{-\Lambda_{\min} t}\, dt = \frac{\|f\|_\infty}{\Lambda_{\min}}, \end{aligned}$$

independently of how fast the generator runs. Applying this with $f = \varepsilon^2 \Lambda_A u^{(2)}$ gives the bound. ☐

The constant is computable. On the four-channel instance committed with experiment 40 it equals 0.33, the defect is ε-independent to 4×10^{-10}, and the bound dominates the measured error at every ε in the sweep with a slack factor of at least 13. The hypothesis $\Lambda_{\min} > 0$ is used only for the resolvent bound. The expansion itself survives a total hazard vanishing on part of the space, as the stress tests of experiment 40 show, but the constant then requires the mean killing time in place of $1/\Lambda_{\min}$.

Proposition 1 (General start)**.** *Let the environment start from an arbitrary law μ_0 rather than from π, and write $p_i^A(\mu_0) = \int u_i^A \, d\mu_0$. Then*

$$p_i^A(\mu_0) = \frac{\bar{\lambda}_i}{\bar{\Lambda}_A} + \varepsilon \left[m_i - \int D\left(\frac{\bar{\lambda}_i}{\bar{\Lambda}_A} \tilde{\Lambda}_A - \tilde{\lambda}_i\right) d\mu_0\right] + O(\varepsilon^2), \tag{10}$$

with m_i the constant appearing in the proof of Theorem 1. The leading softmax term does not depend on the start. The extra term vanishes when $\mu_0 = \pi$, because D maps into mean-zero functions, which recovers (6).

Proof. The expansion in the proof of Theorem 1 is pointwise in y, and integrating $u_i^{(1)} = -D(c_i\tilde{\Lambda}_A - \tilde{\lambda}_i) + m_i$ against μ_0 instead of π leaves the displayed extra term. $\square$

Applied to a system observed away from equilibrium, (6) silently loses the correction. Experiment 40 measures the loss. From a point-mass start the stationary formula (6) is accurate only to order 0.968, while (10) restores order 1.963. Under the invariant law the extra term is 7.3×10^{-17} and the two forms agree.

Remark 1. The two terms in the bracket have distinct roles. The first, $\sum_{j\in A} K_{ji}$, measures how channel i's rate correlates dynamically with the total hazard it competes against. The second subtracts the part of that effect which is common to all channels, and is what keeps $\sum_{i\in A} p_i^A = 1$ at each order.

4 Three structural properties

One computation answers every counterfactual. The pair $(\bar{\lambda}, K)$ does not depend on A. Blocking a channel removes a competitor but does not touch the driver, so the surviving rates keep the mean levels and the pairwise co-movement they always had. All that changes is which pairs compete, and (6) registers exactly that change by restricting its sums to the surviving set. For a system with N channels this reduces 2^N separate resolvent solves to one deviation solve and $O(N^2)$ stored numbers, at the cost of the $O(\varepsilon^2)$ error the expansion carries.

Common-mode fluctuations cancel. Suppose $\lambda_i(y) = a_i c(y)$, so that every channel is modulated by the same environmental factor. The physical mechanism is visible before any algebra. Given that some channel fires at an instant when the driver sits at y, the probability that it is channel i is the ratio

$$\frac{\lambda_i(y)}{\Lambda_A(y)} = \frac{a_i\, c(y)}{c(y)\sum_{j\in A} a_j} = \frac{a_i}{\sum_{j\in A} a_j},$$

and $c(y)$ cancels pointwise, at every state, not merely on average. A common factor is a fluctuating clock speed. It decides when the race ends and has no say in who wins, so no observable built from arrival identities can depend on it. The algebra agrees. Here $\tilde{\lambda}_i = a_i\tilde{c}$ and $K_{jk} = a_j a_k \gamma$ with $\gamma = \int \tilde{c}\,(D\tilde{c})\,d\pi$, and substituting into the bracket in (6),

$$a_i\,\gamma \sum_{j\in A} a_j - \frac{a_i}{\sum_{j\in A} a_j}\,\gamma\Big(\sum_{j\in A} a_j\Big)^2 = 0.$$

The correction vanishes identically.

Remark 2 (Proportional hazards satisfy the Choice Axiom exactly). Suppose $\lambda_i(y) = a_i c(y)$ with $a_i > 0$ and $c \geq 0$ not identically zero. Then for every y, every ε, and every A,

$$u_i^A(y) = \frac{a_i}{\sum_{j\in A} a_j},$$

so the arrival probability satisfies the Choice Axiom from any start, not merely from π and not merely to the order computed above. The verification is one line. The constant $v_i = a_i/\sum_{j\in A} a_j$ satisfies $\Lambda_A v_i = a_i c(y) = \lambda_i$ and is annihilated by L, so it solves (2), and the killed resolvent is invertible.

This is a known result in a new setting. Elandt-Johnson (1976) proved that under proportional hazard rates the failure time and the cause of failure are independent, and did so without assuming the latent failure times are independent. Kochar and Proschan (1991) give an equivalent statement. Marley and Colonius (1992) prove the choice-theoretic version as an equivalence, that proportional cause-specific hazards hold if and only if the chosen option and the time of choice are independent, with the constants of proportionality forced to equal the choice probabilities. They also record the explicit representation $\Pr[t(x) > t] = \exp(-u(x)\Psi(t))$, which yields $u(x)/\sum_y u(y)$ for an arbitrary increasing Ψ.

The present setting adds only that the common factor may be a random functional of the hidden environment rather than a deterministic function of time.

In survival analysis a random multiplier on a hazard is called a frailty (Vaupel et al., 1979), and a common multiplicative modulation is a frailty shared by every channel. The invariance it produces is the proportional-hazards structure of Cox (1972). The proof needs no positivity of c, no time change, and no condition on ε. What it says is that a common modulation factors out of the competition entirely, which is the proportional-hazards normal form. Experiment 40 confirms it where the hypotheses are stressed. With c vanishing on four of six states, with c ranging over nine orders of magnitude, and at $\varepsilon = 100$, the departure from Luce stays below 5.5×10^{-14} at every state.

This is a stronger statement than the vanishing of the correction, and it is practically useful in the opposite direction. It says a measured departure from softmax cannot be produced by any amount of shared environmental fluctuation. Whatever deviation is observed is evidence of differential structure among the channels. The classical negative results run in the opposite direction. The latent joint law of competing risks is not identified from what one observes (Tsiatis, 1975), and identification must be bought with extra structure such as regressor variation (Heckman and Honoré, 1989). The statement here runs the other way. One large class of environmental dependence, the perfectly shared one, leaves no trace whatever in the choice law, so any trace that does appear rules that class out.

Low-rank environments give low-rank corrections. Suppose the intensity deviations are driven by r environmental modes,

$$\lambda_i(y) - \bar{\lambda}_i = b_i^\top z(y) = \sum_{a=1}^{r} b_{ia}\, z_a(y), \qquad \int z\, d\pi = 0,$$

so every channel listens to the same r signals $z_1, \dots, z_r$, each through its own loadings b_i. Substituting into the definition (5) and pulling the loadings outside the integral,

$$\begin{aligned} K_{jk} &= \int \tilde{\lambda}_j \,(D\tilde{\lambda}_k)\, d\pi = \int \Big(\sum_a b_{ja}\, z_a\Big)\, D\Big(\sum_b b_{kb}\, z_b\Big)\, d\pi \\ &= \sum_{a,b} b_{ja}\, b_{kb} \int z_a\,(Dz_b)\, d\pi = \big(B\,\Gamma\, B^\top\big)_{jk}, \qquad \Gamma_{ab} = \int z_a\,(Dz_b)\, d\pi. \end{aligned}$$

Here Γ is the modes' own $r \times r$ integrated-covariance matrix, the same Green–Kubo object one level down, describing the hidden signals rather than the channels. A product of an $N \times r$, an $r \times r$, and an $r \times N$ matrix has rank at most r, no matter how many channels listen.

A second cap belongs to the environment rather than to the loadings. Every deviation $\tilde{\lambda}_i$ is a mean-zero function on the driver's m states, and such functions form a space of dimension

$m-1$. K factors through that space, so

$$\operatorname{rank}(K) \;\leq\; \min(N,\, r,\, m-1)$$

whatever the loadings are. A six-state environment cannot present more than five independent modes, however many channels watch it. Experiment 40 attains both bounds, giving rank exactly r for $r = 1, \ldots, 5$ on a six-state environment and rank $5 = \min(10, 5)$ for generic intensities on ten channels. Rank constraints of this kind, imposed on what a partially observed Markov system can show, go back to Fredkin and Rice (1986).

The rank of the observable part of K counts the driver's effective modes, so the dimensionality of the hidden environment is in principle measurable from arrival data, and experiment 42 develops that count into an exact estimator on the observable class. The count survives the joint estimation of the rates. The invisible family of Section 6 then includes an arbitrary diagonal, but a submatrix whose row and column sets are disjoint contains no diagonal entry, and on such blocks the family shifts K by rank one at most. The largest diagonal-avoiding block of any member of the family has rank $\min(r+1, \lfloor N/2 \rfloor)$, so its rank minus one recovers r whenever $N \geq 2(r+1)$, and otherwise reports that floor. In mixed logit the same low-dimensional factor structure is imposed by hand as a specification choice (McFadden and Train, 2000). Here it is inherited from the environment.

5 The correction predicts the runner-up

The correction has been described so far as a statement about win probabilities, which are what a blocked-set experiment measures. Those experiments are expensive, since each availability set requires its own intervention. There is a cheaper observable that carries the same information.

Let $M_{ij} = \mathbb{P}(\text{runner-up} = j \mid \text{winner} = i)$ be the runner-up kernel, the probability that j finishes second given that i finishes first. Removing channel i changes the outcome only in those realizations where i won, and in those the runner-up inherits the win. This gives an identity that holds for every ε and involves no approximation,

$$q_j^{(-i)} = p_j + p_i M_{ij}, \tag{11}$$

where $q^{(-i)}$ is the deletion counterfactual. The companion work shows M is exactly the object that identifies single deletions, and that winner-only data constrains it not at all. The identity itself is familiar from static discrete choice, where it is the accounting behind diversion ratios, the fractions of a removed product's demand that flow to each substitute (Conlon and Mortimer, 2021). What matters here is that it holds exactly at every ε, which is what lets the expansion be applied to both of its sides.

Applying Theorem 1 to the full set and to $A \setminus \{i\}$ therefore predicts M itself. At leading order the prediction is

$$M_{ij} = \frac{\bar{\lambda}_j}{\bar{\Lambda}_A - \bar{\lambda}_i} + O(\varepsilon), \tag{12}$$

which says the second arrival is drawn from the competition among the remaining channels, the IIA prediction. Multiplying by p_i identifies the leading-order law of the exacta (i, j) as

$$\frac{\bar{\lambda}_i}{\bar{\Lambda}_A} \cdot \frac{\bar{\lambda}_j}{\bar{\Lambda}_A - \bar{\lambda}_i},$$

the sequential ranking model of Plackett (1975), whose paired-comparison ancestor is Bradley and Terry (1952) and whose choice-theoretic basis is the Choice Axiom. The expansion therefore corrects Plackett–Luce for orders of arrival exactly as it corrects the Choice Axiom for winners. Concretely, predicting M_{ij} by the Plackett–Luce kernel above leaves an error of order ε, while substituting the corrected shares of Theorem 1, for the full set and for the set with i deleted, into the identity (11) leaves an error of order ε^2. The correction buys the same extra power of ε for second place that it buys for first.

Experiment 39 verifies this on the same chains. The identity (11) holds to 1.8×10^{-15}, the exactly computed kernel is row-stochastic to 8.0×10^{-15}, which checks the two-stage solve, and the measured convergence orders are 0.967 for the Luce prediction (12) and 1.967 with the Green–Kubo correction. At $\varepsilon = 0.005$ the maximum error over off-diagonal entries falls from 3.0×10^{-4} to 2.1×10^{-6}.

This matters for testing the theory. The second arrival is observed wherever an order of arrivals is recorded, so the correction can be checked against ordinary ranked data rather than against interventions. It also runs the other way. Where the environment is unknown, a measured M constrains the Green–Kubo matrix, which is the beginning of estimating K from data rather than computing it from a known generator.

6 The exacta board suffices

Everything so far computes K from a known generator. A practitioner rarely has the generator. They have observed arrivals. This section asks what K can be recovered from outcomes alone, and the answer has a clean form.

The correction in (6) is linear in K, so stacking it over every availability set gives a linear map from K to observable deviations from softmax. Its null space is exactly what no amount of choice data can reveal.

Proposition 2 (Identifiability of K)**.** *Choice probabilities on all availability sets determine K exactly modulo the $(N+1)$-dimensional family*

$$K \;\longmapsto\; K + d\,\bar{\lambda}^{\top} + t\,\mathrm{diag}(\bar{\lambda}), \qquad d \in \mathbb{R}^N,\; t \in \mathbb{R}, \tag{13}$$

and no smaller family. Exactly $N^2 - N - 1$ combinations of the entries of K are identified.

Proof. Write $B(A,i) = \sum_{j\in A} K_{ji} - c_i \sum_{j,k\in A} K_{jk}$ for the bracket in (6), with $c_i = \bar{\lambda}_i/\bar{\Lambda}_A$. Under the shift $K \mapsto K + d\bar{\lambda}^{\top}$, meaning $K_{jk} \mapsto K_{jk} + d_j\bar{\lambda}_k$, the two sums change by

$$\begin{aligned}
\sum_{j\in A} K_{ji} \;&\longmapsto\; \sum_{j\in A} K_{ji} + \bar{\lambda}_i \sum_{j\in A} d_j,\\
\sum_{j,k\in A} K_{jk} \;&\longmapsto\; \sum_{j,k\in A} K_{jk} + \Big(\sum_{k\in A} \bar{\lambda}_k\Big) \sum_{j\in A} d_j = \sum_{j,k\in A} K_{jk} + \bar{\Lambda}_A \sum_{j\in A} d_j,
\end{aligned}$$

so the bracket changes by

$$\bar{\lambda}_i \sum_{j\in A} d_j - \frac{\bar{\lambda}_i}{\bar{\Lambda}_A}\,\bar{\Lambda}_A \sum_{j\in A} d_j = 0.$$

Under $K \mapsto K + t\,\mathrm{diag}(\bar{\lambda})$, only diagonal entries move, so the first sum gains its $j = i$ term and the second gains its full trace over A:

$$B(A,i) \;\longmapsto\; B(A,i) + t\,\bar{\lambda}_i - \frac{\bar{\lambda}_i}{\bar{\Lambda}_A}\,t\,\bar{\Lambda}_A = B(A,i).$$

Both families are therefore invisible on every availability set. They intersect trivially, since $d\bar{\lambda}^\top$ is diagonal only for $d = 0$, and span $N + 1$ dimensions. That the null space is no larger is a rank computation, carried out in experiment 41 for $N = 3, \ldots, 6$, where the identified dimension equals $N^2 - N - 1$ in every case. □

The two invisible directions have meaning. Adding $d\bar{\lambda}^\top$ changes how each channel's fluctuations correlate with the environment in the direction of the mean intensities, and adding $t\,\mathrm{diag}(\bar{\lambda})$ rescales each channel's own autocorrelation in proportion to its mean rate. Neither moves any branching ratio, for the same reason a common rate shift does not.

Invisible families are the norm when hidden structure is watched through a narrow window. Aggregated Markov models admit whole classes of generators producing identical observations (Kienker, 1989), and in Thurstonian choice the utility covariance is estimable only after normalizations that remove exactly $N + 1$ of the covariance parameters (Dansie, 1985; Bunch, 1991). Whether that count agreeing with the family here is structural or coincidental we leave open. In the other direction, the sufficiency of two places is special to this model. For finite mixtures of Plackett–Luce models, top-two data is known not to suffice (Zhao and Xia, 2019).

Which experiments reach the identified part. The design matters far more than the sample size, and the useful comparison is between three ways of watching the same system. Winner-only data on the full set is nearly useless for this purpose. It supplies $N-1$ independent numbers, so it recovers $N - 1$ of the $N^2 - N - 1$ identified directions and no more. Running every blocked-subset experiment recovers all of them, which is the content of Proposition 2.

Between those sits the observation that makes the theory testable. Recording the runner-up alongside the winner, on the full set alone, recovers everything that the complete set of blocked-subset interventions would, at the first order in ε to which all of this section applies. Under the Choice Axiom the ranking likelihood factorizes into successive applications of the same law, so second choices add efficiency but no new parameters (Beggs et al., 1981), and the ranked specification test built on that factorization detects failures in practice, with estimates drifting as lower ranks are added (Hausman and Ruud, 1987). Here the factorization fails at order ε, and the failure is exactly what carries the new information. Empirical demand estimation reached the same moral by another route. Second-choice survey data is prized because it pins down substitution patterns that market shares alone cannot (Berry et al., 2004), and in this model that practice takes an exact form: at first order, second choices do not merely help, they saturate. Experiment 41 confirms this for $N = 4, 5, 6$, where the ranked design and the exhaustive intervention design have identical rank, $N^2 - N - 1$, while the winner-only design has rank $N - 1$. Interventions are often impossible, and a finishing order is usually free.

When the rates are estimated too. The counts above take $\bar{\lambda}$ as known. In practice the mean rates are estimated from the same observations, and the enlargement of the invisible family then has a closed form.

Proposition 3 (Rate gauge)**.** *For any $\eta \in \mathbb{R}^N$, replacing $(\bar{\lambda}, K)$ by $(\bar{\lambda}+\varepsilon\eta,\ K-\mathrm{diag}(\eta))$ leaves every availability set's share vector unchanged at first order. Under joint estimation of the rates the invisible family is therefore exactly*

$$K \;\longmapsto\; K + d\,\bar{\lambda}^\top + \mathrm{diag}(v), \qquad d, v \in \mathbb{R}^N,$$

of dimension $2N$, and $N^2 - 2N$ combinations of K are identified.

Proof. Two first-order effects must cancel on every availability set A. First, by the diagonal computation in the proof of Proposition 2, shifting K by $-\mathrm{diag}(\eta)$ changes the correction term of channel $i \in A$ by

$$+\frac{1}{\bar{\Lambda}_A}\left[\eta_i - c_i \sum_{l\in A} \eta_l\right].$$

Second, shifting the rates by $\bar{\lambda} \mapsto \bar{\lambda} + \varepsilon\eta$ moves the leading-order share $c_i = \bar{\lambda}_i/\bar{\Lambda}_A$ by ε times

$$\frac{\partial c_i}{\partial \bar{\lambda}_l}\,\eta_l \text{ summed over } l \in A = \frac{\eta_i}{\bar{\Lambda}_A} - \frac{\bar{\lambda}_i}{\bar{\Lambda}_A^2}\sum_{l\in A}\eta_l = -\frac{1}{\bar{\Lambda}_A}\left[-\eta_i + c_i \sum_{l\in A}\eta_l\right],$$

the exact negative of the first effect, for every A and every i. So the substitution changes no first-order observable. The displayed family contains the known-rate family of Proposition 2, and its dimension matches the numerically computed null space of the all-experiments design at every N tested, so it is the whole of it. □

The diagonal K_{ii} is each channel's own integrated autocovariance, which Section 3 identified as the classical motional-narrowing correction to an effective rate. A channel's private fluctuation makes its arrivals burstier without changing how it co-moves with any rival, and at first order burstiness is indistinguishable from a slightly different mean rate. Who beats whom is moved only by co-movement between channels. The proposition says this absorption is exact at first order. When the rates are estimated, the diagonal of K is not merely classical but unobservable, and everything an experiment can see lives in the off-diagonal, the cross-channel dynamical correlations. The practical consequences are unchanged in both readings. Second-place data still matches the complete intervention ensemble, the saturation at second place still holds, and third place adds nothing either way. Experiment 42 verifies the cancellation to 10^{-15} and the family dimension at $N = 5, 6, 7$.

Reciprocity, and what asymmetry means. K is built from equilibrium time correlations, so Onsager's reciprocity question applies to it (Onsager, 1931).

Proposition 4 (Reciprocity)**.** *If the driver satisfies detailed balance then $K = K^\top$.*

Proof. Detailed balance, $\pi(y)L(y,y') = \pi(y')L(y',y)$, is exactly the statement that L is self-adjoint in the inner product $\langle f,g\rangle_\pi = \int fg\,d\pi$, and self-adjointness passes to the semigroup e^{tL}. For a stationary start the covariance is an inner product with the semigroup between the two observables, so

$$\begin{aligned}
\mathrm{Cov}_\pi(\lambda_j(Y_0), \lambda_k(Y_t)) &= \langle \tilde{\lambda}_j,\, e^{tL}\tilde{\lambda}_k\rangle_\pi && \text{(Markov property)}\\
&= \langle e^{tL}\tilde{\lambda}_j,\, \tilde{\lambda}_k\rangle_\pi && \text{(self-adjointness)}\\
&= \mathrm{Cov}_\pi(\lambda_k(Y_0), \lambda_j(Y_t)). && \text{(same formula, } j \text{ and } k \text{ exchanged)}
\end{aligned}$$

Integrating over $t \in [0,\infty)$ gives $K_{jk} = K_{kj}$. □

The asymmetry of K is therefore a signature of a driver held out of equilibrium, and part of it is observable. Under the gauge of Proposition 3 the antisymmetric part of K shifts by $(d\bar{\lambda}^\top - \bar{\lambda}d^\top)/2$, a family of dimension $N-1$, so the asymmetry carries exactly $\binom{N-1}{2}$ gauge-invariant functionals. All of them vanish when the driver is reversible. For $N = 2$ there are none, so two channels can never witness hidden nonequilibrium, and three can, which is the arrival-order analogue of needing a cycle to carry a current. Experiment 43 verifies the counts, the vanishing at 10^{-18}, and class invariance at 10^{-16}, and shows the single $N = 3$ witness growing monotonically with the strength of a cycle drive imposed on an otherwise reversible chain.

For the inference literature that reconstructs hidden nonequilibrium from partial observations, this adds an observable at the opposite extreme of coarseness. No clocks, no trajectories, only orders of arrival across replications, and still enough to certify that the unseen environment breaks detailed balance.

What it costs. Identification is not estimation. Fitting K by least squares on the identified subspace, from shares estimated on pairs and triples and then scored on held-out larger subsets, the estimator beats the uncorrected softmax law only once roughly 10^6 observed arrivals per availability set are observed. The reason is a floor rather than a defect of the estimator. The correction is $O(\varepsilon)$ while multinomial noise on n observations is $O(n^{-1/2})$, so the data requirement scales as ε^{-2}. At the $\varepsilon = 0.05$ of experiment 41 the correction is 1.2×10^{-3} and the crossover duly appears between 10^5 and 10^6 observations.

This bounds the practical claim. Where the generator is known, the correction is cheap and accurate. Where only outcomes are observed, recovering it demands either very many observations or the richer observable of Section 5.

Estimation from the board alone. Experiment 44 turns the saturation result into an estimator. The input is the exacta board of the full set, with the rates unknown. The bridge of Section 5 converts the board into every leave-one-out share without an intervention, the winner shares eliminate $\bar{\lambda}$ by substitution, and least squares on the orthocomplement of the invisible family fits what remains. The design rank is $N^2 - 2N$, the saturation value, so the board determines every identified combination of K in the fast-mixing regime. On exact boards from the chain model the identified part of εK is recovered with relative error $1.5\,\varepsilon$ across $\varepsilon \in [0.01, 0.04]$, first order, which is the order a first-order design can attain. Common-mode rates return zero at machine precision.

The fit is made modulo one more invisible direction beyond the family of Proposition 3. The board carries no clock, so every share is invariant under the global rescaling $(\bar{\lambda}, K) \to (s\bar{\lambda}, sK)$, and the estimator returns K in the time unit where the rates sum to one. In the clock reading of Section 3, choice data cannot see the overall speed of time, only its allocation. Sampled boards obey the ε^{-2} floor of the previous paragraph. At $\varepsilon = 0.02$ the median relative error on the identified part is 1.2 at 10^6 recorded races and 0.34 at 10^7.

7 Verification on finite chains

Everything above is structural. This section and the next test it numerically. On a finite state space every object is an exact linear solve, so the expansion can be checked without simulation error. Experiment 7 runs a six-state chain with ten channels.

Setup. The generator has independent unit-mean exponential off-diagonal rates, giving a nonreversible chain whose spectral gap, the slowest relaxation rate of the environment and so the reciprocal of its mixing time, is 2.63. Channel intensities are lognormal with log-scale 0.8, giving $\bar{\Lambda}_A = 11.4$ on the full set. The seed is 5 and the sweep covers thirteen values of ε from 10^{-3} to

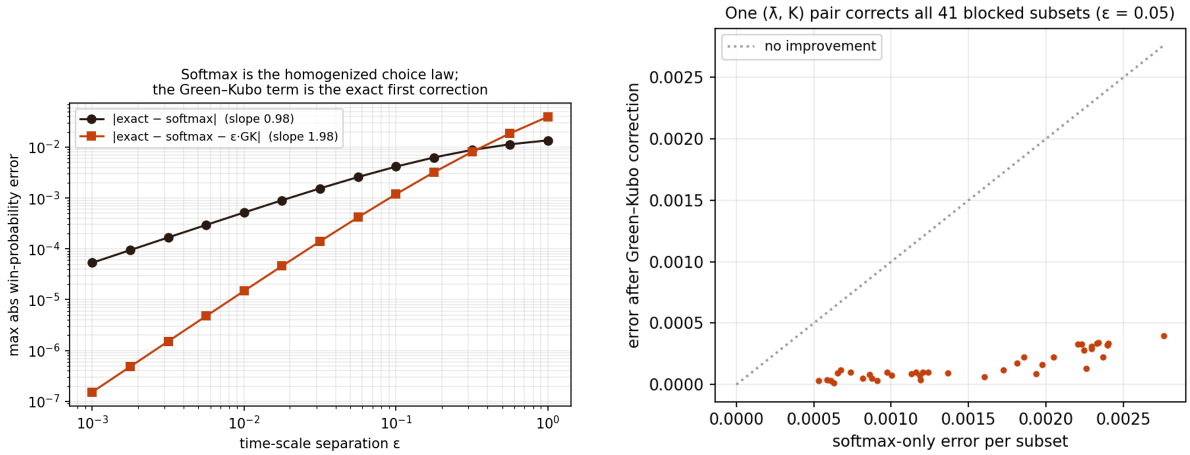

Figure 1: Finite-chain verification (experiment 7). Left, error against ε for the leading softmax law and for the Green–Kubo corrected law, with measured orders 0.985 and 1.985. Right, the same $(\bar{\lambda}, K)$ pair applied to 41 random availability sets at $\varepsilon = 0.05$, with no per-subset computation.

1. The dimensionless small parameter is $\varepsilon\bar{\Lambda}_A/\text{gap} \approx 4.3\,\varepsilon$, so the fitted window ends near 0.08 in those units and $\varepsilon = 0.3$ corresponds to 1.3.

The measured convergence order of the leading term is 0.985, against the theoretical 1. After adding the Green–Kubo correction the measured order is 1.985, against the theoretical 2. Both are fitted over the smallest six of thirteen values of ε, spanning 10^{-3} to 1.8×10^{-2}, since the expansion is asymptotic and the largest ε carry the $O(\varepsilon^2)$ residual. The gain of exactly one order is the signature the expansion predicts.

Experiment 40 repeats the measurement over twenty environments drawn with four to nine hidden states, three to eight channels, and intensity dispersions from 0.3 to 1.2. The leading order has median 0.960 and range $[0.919, 0.982]$, the corrected order has median 1.958 and range $[1.888, 1.979]$, and the gain exceeds 0.948 in every single trial.

Figure 1 shows both. The subset-uniformity claim is tested by fixing one $(\bar{\lambda}, K)$ pair and predicting every availability set. Experiment 40 sweeps all 1013 subsets of size at least two at $\varepsilon = 0.05$, so the maxima here are true maxima rather than samples. The worst error over all subsets falls from 3.1×10^{-3} under softmax to 4.4×10^{-4} with the correction, with no per-subset computation.

The order is uniform in a strong sense. Fitting the convergence order separately on each of the 1013 subsets puts every corrected order in $[1.92, 2.00]$. Small subsets are better rather than worse, because the effective parameter is $\varepsilon\bar{\Lambda}_A/\text{gap}$ and $\bar{\Lambda}_A$ shrinks with the subset. Exactly one subset has the correction slightly worse than softmax, a two-element near-tie where both errors are already two orders of magnitude below the maximum.

Equation (5) is a time integral, and experiment 40 tests whether the integral is necessary or whether equal-time correlation would do. Replacing K by the equal-time covariance $\int \tilde{\lambda}_j \tilde{\lambda}_k \, d\pi$ scaled by a single timescale, with that timescale fitted to give the ablation its best case, leaves the measured order at 1.09 rather than 1.96, and the error at the smallest ε is 8.8 times worse. Static correlation between two channels does not determine the correction. The time integral does.

Common-mode cancellation is confirmed at 2.6×10^{-16} for the Green–Kubo coefficient, and the exact win probabilities agree with Luce to 1.1×10^{-16} at $\varepsilon = 0.3$. The regression suite carries the same check at $\varepsilon = 3.0$, where the environment is ten times slower than the channels rather than faster and the expansion has no claim at all, and Luce still holds to 10^{-12}. That is the content of Remark 2. Rank-two environmental loadings give K with singular values $[0.022, 0.002, 0, 0]$, so the rank is exactly two.

A Gillespie simulation of the same system, event-driven and including the hidden state, agrees with the resolvent solution to 1.9×10^{-3} across 40,000 trajectories at $\varepsilon = 0.3$, against a binomial sampling noise of 2.1×10^{-3}.

The chain is not reversible, so K is asymmetric. On this instance $\max_{jk} |K_{jk} - K_{kj}| = 0.139$ against $\max_{jk} |K_{jk}| = 0.603$, a relative asymmetry of 23% recorded by experiment 40. This is what distinguishes K_{ji} from K_{ij} in (6), and experiment 40 confirms the index order is not a convention. Transposing K in the correction degrades the measured order from 1.94 to 1.58, and K agrees with brute-force time integration of the covariance to 4.1×10^{-6}.

8 Narrow escape

Finite chains verify the algebra. To test the expansion as a statement about physics we take the hidden state to be reflected Brownian motion in the unit disk and the channels to be reaction windows on the boundary, bands in which channel i fires at rate κ. This is the narrow escape geometry (Schuss et al., 2007) with partially reactive patches, and the leading law, proportional to band areas in the slow-reaction regime, is the reaction-limited counterpart of the diffusion-limited patch capture of Berg and Purcell (1977), whose classical law scales with patch size rather than area. The first window to fire wins. Here κ plays the role of ε, since a slow reaction rate means the particle mixes many times before firing.

Here the driver is the physical system itself rather than an external environment. In the language of Section 2, the hidden state Y_t is the particle's position in the disk, and its generator is the reflecting Laplacian. The channels are the windows, and window i fires at rate $\lambda_i(y) = \kappa\, g_i(y)$, where g_i is the fraction of the neighbourhood of y lying inside window i's band, so λ_i is supported near the boundary and vanishes in the interior. Two consequences follow. The mean rates $\bar{\lambda}_i$ are proportional to band areas, which is why the leading law is band-area softmax. And K is built from how long the particle's excursions keep it near one window rather than another, which is a purely geometric quantity.

The small parameter is κ rather than ε. Slow reaction means the particle explores the whole disk many times before anything fires, so small κ is the fast-mixing limit, and blocking a channel means switching a window off.

Setup. Sixteen boundary windows on the unit circle, six clustered within about 50 degrees and the rest scattered, drawn with seed 42. Angular half-widths are lognormal with median 0.030, ranging over $[0.019, 0.055]$ on this draw, and three pairs overlap. The reaction bands extend 0.12 inward in radius. The polar grid is 40×88, giving 3520 cells. The resolvent sweep covers $\kappa \in \{2, 5, 10, 20, 40\}$, and the Brownian simulation runs at $\kappa \in \{5, 20\}$ with time step 2×10^{-4} and 50,000 trajectories.

Experiment 9 separates three layers so that no error source can hide in another. Layer A is a real Brownian simulation with independent per-step firing. Layer B replaces the disk by a polar grid of cells and Brownian motion by the conservative jump process between neighbouring cells

κ	softmax	with Green–Kubo	improvement
2	1.6×10^{-3}	5.0×10^{-5}	31×
5	3.7×10^{-3}	3.0×10^{-4}	12×
10	6.9×10^{-3}	1.1×10^{-3}	6.3×
20	1.2×10^{-2}	3.9×10^{-3}	3.2×
40	2.0×10^{-2}	1.2×10^{-2}	1.6×

Table 1: Maximum absolute error in the win probabilities against the exactly discretized generator, over sixteen boundary windows. The correction gains an order of magnitude at small κ and degrades to nothing by $\kappa = 40$, which is where the neglected $O(\kappa^2)$ term takes over.

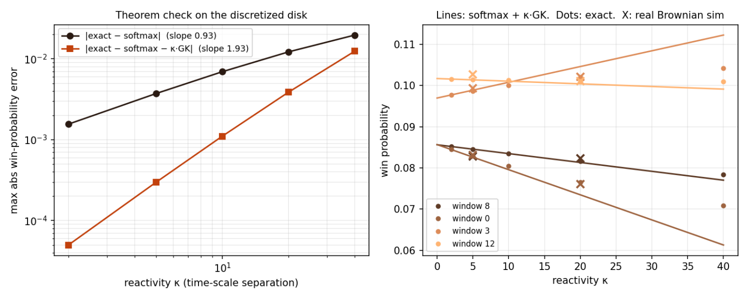

Figure 2: The theorem on continuous physics (experiment 9). Reflected Brownian motion in the unit disk with boundary reaction windows, error against κ for the leading and corrected laws, measured orders 0.93 and 1.93 against an exactly discretized Neumann generator.

that discretizes it, a finite-volume Neumann Laplacian, on which the killed-resolvent solve is exact. The uniform invariant law, reversibility, and conservation are checked numerically rather than assumed. Layer C is the theory, band-area softmax plus the Green–Kubo term from one deviation solve, using the same code that passed the chain tests.

Two distinct homogenizations meet in this geometry. Replacing a patchy boundary by an effective uniform reactivity is boundary homogenization (Berezhkovskii et al., 2004), the same effective-Robin reduction arises for boundaries that switch randomly in time (Lawley and Keener, 2015), and expansions in a small geometric parameter give splitting probabilities among small targets (Bressloff, 2020). The small parameter in this paper is temporal rather than geometric, so the two expansions are orthogonal and compose.

Comparing layers B and C tests the theorem on a continuous-physics generator. The softmax error decays with measured order 0.93 and the corrected error with order 1.93, again a clean gain of one order. Both orders are fitted on the three smallest values of κ, where the expansion applies.

Table 1 gives the full sweep, including where the correction stops paying. The improvement factor falls monotonically from 31× at $\kappa = 2$ to 1.6× at $\kappa = 40$. This is the expected behaviour of a first-order expansion rather than a defect, but it bounds the useful range, and a practitioner should read the correction as buying roughly an order of magnitude only while κ stays in the first half of this table.

Figure 2 shows the three layers together. Comparing layers A and B tests the discretization. The maximum discrepancy is 2.2×10^{-3} at $\kappa = 5$ and 1.6×10^{-3} at $\kappa = 20$, against a sampling noise of about 10^{-3} at 50,000 trajectories.

Two systematic errors. Both would have corrupted a direct comparison of simulation against theory, and neither is visible without the middle layer.

The first is band quantization, worth 5×10^{-2}. Sharp cell indicators mis-size any window narrower than a grid cell, with the mis-sizing ratio running from 2.0 down to zero, since the narrowest window here spans about half a cell and is missed entirely. Fractional area-exact cell coverage fixes it, after which band areas match their analytic values to 2×10^{-17}.

The second is overlap semantics, worth 1.4×10^{-2}. The window geometry contains three overlapping pairs. The original simulation awarded an overlap to the lower-indexed window, while the model stacks intensities as $\lambda = \kappa \sum_i g_i$, which is the physically correct convention for independent receptors. The error's structure identified it, being -0.014 on one window and $+0.006$ on its overlap partner, and its flatness under both time-step and grid refinement confirmed that it was not a discretization artifact.

9 Reproducibility

Every number in this paper comes from a committed script. Experiment 7 gives the chain verification, experiment 9 the narrow-escape study, experiment 39 the runner-up bridge, experiment 40 the replication, the equal-time ablation, the rank bounds, the index-order check, and the remainder-bound verification, experiment 41 the identifiability and estimation results, experiment 42 the mode-count estimator and the rate-gauge verification, experiment 43 the reciprocity and irreversibility-witness results, and experiment 44 the exacta-board estimator. All four run from a clean checkout of `https://github.com/microprediction/kinetics` with NumPy, SciPy, and Matplotlib, and each writes the `results.csv` the corresponding numbers are quoted from. Experiments 7, 39, and 40 take seconds. Experiment 9 takes about three minutes. A regression suite locks every property stated here, including the two convergence orders, the exactness of common-mode cancellation, the rank bounds, and the runner-up identity.

10 Related work

The averaging that produces the leading term is classical (Khas'minskii, 1966; Kurtz, 1973; Pavliotis and Stuart, 2008; Yin and Zhang, 1998), and the closest recent instance is Colantoni (2026), who studies a killing rate modulated by a continuous-time Markov chain, gives Feynman–Kac representations, and proves an averaging principle in the fast-switching limit. That work stops where this one starts. It carries a single killing rate rather than competing channels, so there is no branching ratio to correct, and it establishes the limit without the first-order term.

The ingredient is classical on the other side too. Motional narrowing (Anderson, 1954; Kubo, 1954; van Kampen, 1974) and dynamical disorder (Zwanzig, 1990; Szabo et al., 1982) have used integrated rate autocovariances for seventy years,[3] but for effective rates and survival probabilities, in the scalar setting where no choice law arises. Recent linear-response work on first-passage problems predicts the response of the mean first-passage *time* rather than the identity of the winner.

[3] In the fast limit Zwanzig (1990) notes the fluctuating rate "can be replaced by its time average," which is the leading order here.

For Markov-modulated processes with marked transitions, queueing theory computes mark probabilities exactly by linear algebra on the joint generator (Fischer and Meier-Hellstern, 1993; He and Neuts, 1998), with no expansion and no asymptotic regime. Theorem 1 is the fast-environment asymptotic of those same quantities, and what it adds is not the numbers but the structure the asymptotic exposes. One pair $(\bar{\lambda}, K)$ serves every subset, shared modulation drops out, and low-rank environments give low-rank corrections.

A separate literature identifies hidden structure from timed events. Markov-modulated Poisson processes and aggregated Markov chains are identifiable, up to explicit equivalence, from fully timestamped observation (Rydén, 1996; Fredkin and Rice, 1986), and stochastic thermodynamics infers hidden network properties, entropy production especially, from sequences of visible transitions (Skinner and Dunkel, 2021; Harunari et al., 2022). The reciprocity witness of Section 6 is an observable of that kind at an extreme of coarseness, needing only arrival orders. The observable here is poorer, order of arrivals across replications with no clocks, and the marks-only question appears not to have been asked. In credit risk the direction of inference is the reverse of pricing. Duffie et al. (2009) estimate a hidden frailty from default arrivals by filtering along the single observed path, while the basket-default literature prices order-of-arrival claims without asking what observing them identifies. In psychology, dependent Poisson race models (Ruan et al., 2008) break IIA through shared event streams rather than a mixing driver, with no asymptotic regime.

That a race is at once a competing-risks model and a random-utility model is old. Marley and Colonius (1992) make the identification explicitly, recasting a Thurstonian race in cause-specific hazards, and prove the proportional-hazards equivalence that Remark 2 recovers. Kochar and Proschan (1991) prove the competing-risks equivalence,[4] and Elandt-Johnson (1976) proved the forward direction earlier, with dependent failure times allowed.[5] Their treatment of dependence runs the other way from ours. They record the classical non-identifiability, that any well-behaved structure on a fixed choice set admits an independent representation, so on a fixed set there is nothing to correct. The correction here lives in the comparison across sets, which is exactly where their context-free hazard condition bites. What appears not to have been stated is that softmax is the homogenized choice law of fast hidden kinetics, with a computable first correction.

11 Discussion

Two things come out of this. The expansion explains why softmax is a reasonable default in systems with fast shared dynamics and says how it fails, and the identification analysis says what an observer can hope to recover.

The shape of the failure connects to an empirical literature. In this model the departure from proportional redistribution is a derived consequence of the shared driver, given in closed form by K. In choice data the same departure is an observed regularity. It is documented in measured diversion ratios (Conlon and Mortimer, 2021), in similarity effects among options with shared features (Tversky, 1972), and in the specification test built on the proportional restriction (Hausman and McFadden, 1984). Those observations concern human choice, not arrival processes, and nothing here explains them. The model supplies one mechanism by which such departures arise and a measurement of what they reveal when it operates. The failure is

[4] Their abstract: time and identity of first failure are independent "if and only if the cause specific hazard rates of the components are proportional."

[5] Her abstract: the result "is also valid without the assumption of independence."

governed by integrated dynamical correlations, not by static ones, and it is insensitive to any fluctuation shared across channels.

Four limitations bound the claim. The expansion is in the mixing-to-firing timescale ratio and says nothing when channels fire faster than the environment relaxes. That is the opposite, quenched limit of dynamical disorder (Zwanzig, 1990), in which each realization runs at frozen random rates and outcomes average over the frozen values instead of the rates averaging first. The correction is first order, so systems at moderate separation retain an $O(\varepsilon^2)$ residual, quantified in Table 1, where the improvement decays to nothing by $\kappa = 40$. The theorem is proved for finite chains, and a formal treatment of reversible diffusions with a spectral gap is left open, so the narrow-escape study verifies rather than proves that case. The correction as usually quoted assumes the environment is observed in equilibrium, and Proposition 1 gives the extra term otherwise. Finally the continuum study is one geometry with one window arrangement, and while the chain results replicate over twenty environments, the physics results do not yet replicate over geometries.

One extension looks tractable. The disk's boundary operator diagonalizes in Fourier modes, which may render the leading geometric correction analytic rather than numerical.

Sections 5 and 6 close the gap between the computed correction and the estimated one. The identifiable part of K is characterized exactly, and ranked observations reach all of it. What remains is statistical rather than structural. The ε^{-2} sample requirement is severe at small ε, which is precisely the regime where the expansion is most accurate, so the theory is easiest to trust where it is hardest to fit. Whether the runner-up design also improves the constant in that rate, and not merely the rank, is open.

References

P. W. Anderson. A mathematical model for the narrowing of spectral lines by exchange or motion. *Journal of the Physical Society of Japan* 9 (1954) 316–339.

S. Beggs, S. Cardell, J. Hausman. Assessing the potential demand for electric cars. *Journal of Econometrics* 17 (1981) 1–19.

A. M. Berezhkovskii, Yu. A. Makhnovskii, M. I. Monine, V. Yu. Zitserman, S. Y. Shvartsman. Boundary homogenization for trapping by patchy surfaces. *Journal of Chemical Physics* 121 (2004) 11390–11394.

H. C. Berg, E. M. Purcell. Physics of chemoreception. *Biophysical Journal* 20 (1977) 193–219.

S. Berry, J. Levinsohn, A. Pakes. Differentiated products demand systems from a combination of micro and macro data: the new car market. *Journal of Political Economy* 112 (2004) 68–105.

A. B. Bortz, M. H. Kalos, J. L. Lebowitz. A new algorithm for Monte Carlo simulation of Ising spin systems. *Journal of Computational Physics* 17 (1975) 10–18.

R. A. Bradley, M. E. Terry. Rank analysis of incomplete block designs I. The method of paired comparisons. *Biometrika* 39 (1952) 324–345.

P. C. Bressloff. Asymptotic analysis of extended two-dimensional narrow capture problems. arXiv:2010.11885 (2020).

J. S. Bridle. Probabilistic interpretation of feedforward classification network outputs, with relationships to statistical pattern recognition. In *Neurocomputing*, NATO ASI Series F 68 (1990) 227–236.

D. S. Bunch. Estimability in the multinomial probit model. *Transportation Research Part B* 25 (1991) 1–12.

C. Bundesen. A theory of visual attention. *Psychological Review* 97 (1990) 523–547.

F. Colantoni. Time-dependent Robin heat equation via Markovian switching. arXiv:2604.27901 (2026).

C. Conlon, J. H. Mortimer. Empirical properties of diversion ratios. *RAND Journal of Economics* 52 (2021) 693–726.

D. R. Cox. Some statistical methods connected with series of events. *Journal of the Royal Statistical Society B* 17 (1955) 129–164.

D. R. Cox. The analysis of exponentially distributed life-times with two types of failure. *Journal of the Royal Statistical Society B* 21 (1959) 411–421.

D. R. Cox. Regression models and life-tables. *Journal of the Royal Statistical Society B* 34 (1972) 187–202.

B. R. Dansie. Parameter estimability in the multinomial probit model. *Transportation Research Part B* 19 (1985) 526–528.

J. Dean, L. A. Barroso. The tail at scale. *Communications of the ACM* 56 (2013) 74–80.

D. Duffie, A. Eckner, G. Horel, L. Saita. Frailty correlated default. *Journal of Finance* 64 (2009) 2089–2123.

R. C. Elandt-Johnson. Conditional failure time distributions under competing risk theory with dependent failure times and proportional hazard rates. *Scandinavian Actuarial Journal* 1976 (1976) 37–51.

W. Fischer, K. Meier-Hellstern. The Markov-modulated Poisson process (MMPP) cookbook. *Performance Evaluation* 18 (1993) 149–171.

D. R. Fredkin, J. A. Rice. On aggregated Markov processes. *Journal of Applied Probability* 23 (1986) 208–214.

D. T. Gillespie. Exact stochastic simulation of coupled chemical reactions. *Journal of Physical Chemistry* 81 (1977) 2340–2361.

J. Grandell. *Doubly Stochastic Poisson Processes*. Lecture Notes in Mathematics 529, Springer, 1976.

M. S. Green. Markoff random processes and the statistical mechanics of time-dependent phenomena II. Irreversible processes in fluids. *Journal of Chemical Physics* 22 (1954) 398–413.

P. E. Harunari, A. Dutta, M. Polettini, É. Roldán. What to learn from a few visible transitions' statistics? *Physical Review X* 12 (2022) 041026.

J. Hausman, D. McFadden. Specification tests for the multinomial logit model. *Econometrica* 52 (1984) 1219–1240.

J. Hausman, P. Ruud. Specifying and testing econometric models for rank-ordered data. *Journal of Econometrics* 34 (1987) 83–104.

Q.-M. He, M. F. Neuts. Markov chains with marked transitions. *Stochastic Processes and their Applications* 74 (1998) 37–52.

J. J. Heckman, B. E. Honoré. The identifiability of the competing risks model. *Biometrika* 76 (1989) 325–330.

R. Z. Khas'minskii. A limit theorem for the solutions of differential equations with random right-hand sides. *Theory of Probability and Its Applications* 11 (1966) 390–406.

P. Kienker. Equivalence of aggregated Markov models of ion-channel gating. *Proceedings of the Royal Society of London B* 236 (1989) 269–309.

S. C. Kochar, F. Proschan. Independence of time and cause of failure in the multiple dependent competing risks model. *Statistica Sinica* 1 (1991) 295–299.

R. Kubo. Note on the stochastic theory of resonance absorption. *Journal of the Physical Society of Japan* 9 (1954) 935–944.

R. Kubo. Statistical-mechanical theory of irreversible processes I. *Journal of the Physical Society of Japan* 12 (1957) 570–586.

T. G. Kurtz. A limit theorem for perturbed operator semigroups with applications to random evolutions. *Journal of Functional Analysis* 12 (1973) 55–67.

D. Lando. On Cox processes and credit risky securities. *Review of Derivatives Research* 2 (1998) 99–120.

S. D. Lawley, J. P. Keener. A new derivation of Robin boundary conditions through homogenization of a stochastically switching boundary. *SIAM Journal on Applied Dynamical Systems* 14 (2015) 1845–1867.

R. D. Luce. *Individual Choice Behavior: A Theoretical Analysis*. Wiley, 1959.

A. A. J. Marley, H. Colonius. The 'horse race' random utility model for choice probabilities and reaction times, and its competing risks interpretation. *Journal of Mathematical Psychology* 36 (1992) 1–20.

D. McFadden. Conditional logit analysis of qualitative choice behavior. In P. Zarembka (ed.), *Frontiers in Econometrics*, Academic Press (1974) 105–142.

D. McFadden, K. Train. Mixed MNL models for discrete response. *Journal of Applied Econometrics* 15 (2000) 447–470.

P.-A. Meyer. Démonstration simplifiée d'un théorème de Knight. In *Séminaire de Probabilités V*, Lecture Notes in Mathematics 191, Springer, 1971, 191–195.

L. Onsager. Reciprocal relations in irreversible processes I. *Physical Review* 37 (1931) 405–426.

F. Papangelou. Integrability of expected increments of point processes and a related random change of scale. *Transactions of the American Mathematical Society* 165 (1972) 483–506.

G. A. Pavliotis, A. M. Stuart. *Multiscale Methods: Averaging and Homogenization*, Texts in Applied Mathematics 53, Springer, 2008. Chapter 9, Averaging for Markov chains.

R. L. Plackett. The analysis of permutations. *Journal of the Royal Statistical Society C* 24 (1975) 193–202.

S. Ruan, S. N. MacEachern, T. Otter, A. Dean. The dependent Poisson race model and modeling dependence in conjoint choice experiments. *Psychometrika* 73 (2008) 261–288.

T. Rydén. On identifiability and order of continuous-time aggregated Markov chains, Markov-modulated Poisson processes, and phase-type distributions. *Journal of Applied Probability* 33 (1996) 640–653.

Z. Schuss, A. Singer, D. Holcman. The narrow escape problem for diffusion in cellular microdomains. *Proceedings of the National Academy of Sciences* 104 (2007) 16098–16103.

D. J. Skinner, J. Dunkel. Estimating entropy production from waiting time distributions. *Physical Review Letters* 127 (2021) 198101.

A. Szabo, D. Shoup, S. H. Northrup, J. A. McCammon. Stochastically gated diffusion-influenced reactions. *Journal of Chemical Physics* 77 (1982) 4484–4493.

S. Thorpe, D. Fize, C. Marlot. Speed of processing in the human visual system. *Nature* 381 (1996) 520–522.

A. Tsiatis. A nonidentifiability aspect of the problem of competing risks. *Proceedings of the National Academy of Sciences* 72 (1975) 20–22.

A. Tversky. Elimination by aspects: a theory of choice. *Psychological Review* 79 (1972) 281–299.

N. G. van Kampen. A cumulant expansion for stochastic linear differential equations I. *Physica* 74 (1974) 215–238.

J. W. Vaupel, K. G. Manton, E. Stallard. The impact of heterogeneity in individual frailty on the dynamics of mortality. *Demography* 16 (1979) 439–454.

J. I. Yellott Jr. The relationship between Luce's choice axiom, Thurstone's theory of comparative judgment, and the double exponential distribution. *Journal of Mathematical Psychology* 15 (1977) 109–144.

G. G. Yin, Q. Zhang. *Continuous-Time Markov Chains and Applications: A Singular Perturbation Approach*. Springer, 1998.

Z. Zhao, L. Xia. Learning mixtures of Plackett–Luce models from structured partial orders. In *Advances in Neural Information Processing Systems* 32 (2019).

R. Zwanzig. Rate processes with dynamical disorder. *Accounts of Chemical Research* 23 (1990) 148–152.