Probability Approximations via the Poisson Clumping Heuristic: An Update

David Aldous
Department of Statistics
University of California
Berkeley CA 94720
aldous@stat.berkeley.edu

Spring 1992

This is a collection of corrections and additions to the book, and references to subsequent work.

General Remarks

Perhaps the main development since going to press has been continued work by Barbour et al and by Arratia et al, referenced below, on using Stein's method to get explicit bounds for Poisson approximation. However, I have not yet seen any convincing implementation of the Compound Poisson case.

Also of interest is a new book by Berman [14] which collects together his extensive research results on extremes and sojourns: the Gaussian case is emphasized, but also discussed are diffusions, birth-and-death processes and d-dimensional orthogonally invariant processes.

B26.3 Patterns in coin-tossing. The compound Poisson limit theorem, for i.i.d. sequences and geometric clump distributions, is given in Chrysaphinou and Papastavridis [19].

C29 General references. A nice survey of theory and examples of extremes is Lindgren and Rootzen [36]. An early treatment of the regenerative case is in Serfozo [45]. A recent Russian monograph (which I have not seen) which seems related to extremes is [52]. Another recent Conference Proceedings volume is [33].

C31 Asymptotic independence of tails. Smith [48] shows that in this setting Stein's method can be used to get explicit error bounds, even in

the non-stationary case.

C33 Additive processes. An interesting theoretical treatment of tails of stationary distributions occuring as solutions of certain random recurrences (which includes the geometric tails of some of our queueing examples) is given by Goldie [26].

C34 Rice's formula. Formalizations of the "local" version of the heurstic from Rice's formula

$$P(M_t > b) \sim t\rho_b \text{ as } b \to \infty, t \text{ fixed}$$

are given by Albin [2]

C41 Stein's method for extremes of stationary sequences. Barbour et al [13] study the Poisson approximation for the sojourn time of a stationary discrete-time process above a high level in several examples: moving averages with non-negative weights (C5); Gaussian sequnces with non-negative correlations. The Gaussian case in considered further in Holst and Janson [28] and in Arratia et al. [8].

C42 The Compound Poisson Point Process Limit. As remarked in (A12.2), one can formalize a crude version of the heuristic by looking at compound Poisson point process limits. In the context of extremes of stationary processes, the process

$$t \to \text{ sojourn time above } b \text{ during } [0, t]$$

rescales to a compound Poisson counting process, in which the spatial structure of a clump has been collapsed to a point. A theoretical treatment of these issues is in Hsing et al [30].

C43 Extremes for Queues in Heavy Traffic. Serfozo [46, 47] treats the case of maxima of birth-and- death processes and some single-server queues, taking limits as the traffic intensity tends to 1. It is interesting that a non-classical limit distribution

$$P(\xi \leq x) = \exp(-c/(e^x - 1))$$

may occur in this setting.

C44 Extremal Processes. Processes which occur as weak limits of $a_n M_{[nt]} + b_n$, where M_t is the maximum process associated with a stationary process, are called *extremal processes*. Under mixing conditions, such extremal processes can be built out of the Poisson point processes (C3) in a natural way. The case of general stationary processes has recently been of theoretical interest, partly because it requires some unfamiliar technical tools (random upper-semi-continuous functions). See O'Brien et al [41].

C45 Envelopes of Gaussian processes. Extremes of narrow-band Gaussian processes can be studied via the associated *envelope* process. See Vanmarcke [53] for applied discussion and Lindgren [35] for a recent example treated rigorously.

C46 Shot noise. Asyptotics in examples like C13 are treated in Hsing and Teugels [31]

C99 Miscellaneous Husler [32] gives rigorous results for locally stationary Gaussian processes crossing a moving boundary.

D26 Diffusion in random environment. Rick Durrett pointed out that my discussion here is a little casual. Bramson and Durrett [15] construct analogous discrete-space models which in any dimension exhibit the "phase transition" from subdiffusive to diffusive behavior which was claimed in my example.

D33 Brownian motion boundary crossing. Roberts [43] gives formalizations of the heuristic for Brownian motion hitting a square-root-like boundary, via the transformation to the Ornstein-Uhlenbeck process.

D38 Berman's method. Albin [3] gives another formalization of the ergodic-exit method, using discrete approximations, and emphasizing "low-dimensional" technical conditions.

E17 General references. Poisson approximations in "balls in boxes" problems are treated via Stein's method in Barbour et al [12], and via coupling in [29]. Birthday problems are treated via Stein's method in [8].

A sophisticated analytical approach to exact calculations for birthday, coupon collector's and related problems arising in the analysis of algorithms in given in [24].

Stadje [49] contains historical references for the coupon collector's problem.

The promised book of Diaconis-Mosteller has not materialized, but they have a nice paper [23] giving examples and philosophy for coincidences.

E19 The random number test. Marsaglia [37] discusses several powerful tests for random number generators.

F15 Covering i.i.d. blocks. Martin Hildebrand (personal communication) has a brief argument for

$$(\Phi(\mu_m))^{1/m} \to 1/\min_i P(\eta_1 = i) \equiv \alpha, \text{ say},$$

from which it is presumably easy to prove

$$(V_m)^{1/m} \to \alpha \text{ in probability.}$$

F20 Covering problems. Another reference for the uniform i.i.d. covering problem is Mori [40].

G23 Greedy Algorithms. An account of some computer science examples in the spirit of this section is given by Szpankowsky [51]. His emphasis is on determining when a greedy algorithm gives near- optimal results in a maximization problem.

H14 General references. A useful non-technical formula-oriented survey is given by Cowan [20].

H18 One-dimensional problems. In the context of small or large spacings between uniform random points on a circle are studied using Stein's method in Barbour et al [13]; they also discuss the *scan statistic* (essentially our H13).

H19 Clustering problems in 1-dimension. Glaz [25] surveys numerically accurate approximations for the distribution of the "scan statistic" implicit in H19.

H22 Sample random in d dimensions. A natural topic not discussed in the book concerns i.i.d. variables with given distribution in R^d: one could study the maximal inter-point distance $M_n = \max(|X - j - X_i| : 1 \leq i, j \leq n)$. Matthews and Rukhin [38] treat the standard Normal case, showing that clumping does *not* occur, that is for suitable $r_n \to infty$ the number of pairs (i, j) with $|X_j - X_i| > r_n$ has asymptotically Poisson distribution. Presumably with less rapidly-decreasing tails there would be clumping.

I23 Hitting small balls. Gaussian processes hitting small sets (more general than balls) are discussed by Albin [3, 2].

I28 Rough χ^2 processes. Albin [3] gives further results. A statistical application is given in Chan [18].

I29 Smooth stationary processes. Breitung [16] gives further results on estimating the outcrossing rate.

J29 General References. The new book of Adler [1] is a very readable account of extrema of Gaussian processes on infinite-dimensional parameter space.

J38 χ^2 Surfaces. Aronowich and Adler [6] study d-parameter random fields analogous to the 1-parameter χ^2 processes of C28.

J39 Smooth fields in statistics. Some recent statistical work has involved maxima of *smooth* Gaussian fields in high dimensions. As one example, Sun [50] studies how projection pursuit (a method of trying to pick interesting 1-dimensional features out of multidimensional data) works on random data, as a way of assessing significance levels. Here, doing the calculations involves Weyl's formula for volumes of tubes. Moreover, it

appears that "second order" approximations for maxima are required.

K3 The asymptotic distribution of W. Paul Zhang pointed out an error in the argument for (K3a), which asserted

$$P(W(1,T) \le b) \approx \exp(-\kappa b^3 \phi(b) T)$$

for a constant $\kappa = 1$. In the double integral on page 223, we can say that the integral is the same as if the integrand was identically $1/2$, simply because (Z_s^+) and (Z_s^-) have essentially the same distribution. But the argument that we may ignore times of the form $(t_0 + s_1, t_0 + u + s_2)$ is wrong: both processes grow as order $t^{1/2}$ for small t. Thus in the estimation of $EC_{u,v}$ we need to include a term

$$2 \int \int_{s_1, s_2 > 0, s_1 + s_2 \le v - u} P(Z^*_{s_1} < B^*_{s_2}) ds_1 ds_2$$

where (Z^*, B^*) denote BES(3) and standard Brownian motion (Z, B) conditioned on the event

$$Z_t > \max_{s<t} B_s \text{ for all } t.$$

Carrying through this argument suggests that (K3a) holds for $\kappa = (1+4\gamma)^{-1}$, where

$$\gamma = \int_0^1 P(Z^*_s < B^*_{1-s})\, ds.$$

Limit theorems including this example were given by Deheuvels and Revesz [22]. Their results imply a different power (b^1 instead of b^3) in (K3a), and I find that a little hard to believe.

K8 Self-intersections of d-dimensional Brownian motion. Perkins and Taylor [42] study variations on our random set $\mathcal{S}_\delta$ of times $(s,t) : |B_t - B_s| \le \delta$.

K13 Points of increase. A simple and elegant proof of non-existence of points of increase has been found by Burdzy [17].

The bibliography also contains publication data for articles referenced in the book as unpublished.

References

[1] R.J. Adler. *An introduction to Continuity, Extrema and Related Topics for General Gaussian Processes*, volume 12 of *IMS Lecture Notes*. Institute of Mathematical Statistics, Hayward CA, 1990.

[2] J.M.P. Albin. On extremal theory for differentiable stationary stochastic processes. Technical report, University of Lund, 1990.

[3] J.M.P. Albin. On extremal theory for stationary processes. *Annals of Probability*, 18:92–128, 1990.

[4] D.J. Aldous. The harmonic mean formula for probabilities of unions: Applications to sparse random graphs. *Discrete Mathematics*, 76:167–176, 1989.

[5] D.J. Aldous. Stein's method in a two-dimensional coverage problem. *Stat. Probab. Letters*, 8:307–314, 1989.

[6] M. Aronowich and R.J. Adler. Sample path behavior of χ^2 surfaces at extrema. *Adv. Appl. Prob.*, 20:719–738, 1988.

[7] R. Arratia, L. Goldstein, and L. Gordon. Two moments suffice for Poisson approximations: the Chen-Stein method. *Ann. Probab.*, 17:9–25, 1989.

[8] R. Arratia, L. Goldstein, and L. Gordon. Poisson approximations and the Chen-Stein method. Technical report, U.S.C., 1990.

[9] R. Arratia, L. Gordon, and M.S. Waterman. The Erdos-Renyi law in distribution, for coin-tossing and sequence-matching. *Ann. Statist.*, 18:539–570, 1990.

[10] R. Arratia and M.S. Waterman. The Erdos-Renyi strong law for pattern-matching with a given proportion of mismatches. *Ann. Probab.*, 17:1152–1169, 1989.

[11] A.D. Barbour and L. Holst. Some applications of the Stein-Chen method for proving Poisson approximation. *J. Appl. Prob.*, 21:74–90, 1989.

[12] A.D. Barbour, L.Holst, and S. Janson. Poisson approximation in occupancy problems. Technical report, Math. Dept., Uppsala University, 1988.

[13] A.D. Barbour, L.Holst, and S. Janson. Poisson approximation with the Stein-Chen method and coupling. Technical report, Math. Dept., Uppsala University, 1988.

[14] S. M. Berman. *Sojourns and Extremes of Stochastic Processes.* Wadsworth, 1992.

[15] M. Bramson and R. Durrett. Random walk in random environment: A counterexample? *Commun. Math. Phys.*, 119:199–211, 1988.

[16] K. Breitung. Asymptotic crossing rate for stationary Gaussian vector processes. *Stochastic Proc. Appl.*, 29:195–208, 1988.

[17] K. Burdzy. On non-increase of Brownian motion. *Ann. Probability*, 18:978–980, 1990.

[18] K.S. Chan. Percentage points of likelihood ratio tests for threshold autoregression. Technical report, University of Chicago, 1990.

[19] O. Chrysaphinou and S. Papastavridis. A limit theorem on the number of overlapping appearances of a pattern in a sequence of independent trials. *Probability Th. Rel. Fields*, 79:129–143, 1988.

[20] R. Cowan. Objects arranged randomly in space: an accessible theory. *Adv. Appl. Prob*, 21:543–569, 1989.

[21] M.V. Day. Recent progress on the small parameter exit problem. *Stochastics*, 20:121–150, 1987.

[22] P. Deheuvels and P. Revesz. Weak laws for the increments of Wiener processes, brownian bridges, empirical processes and partial sums of i.i.d.r.v's. In *Proceedings 6th Pannonian Symposium on Mathematical Statistics*. Reidel, 1988.

[23] P. Diaconis and F. Mosteller. Methods for studying coincidences. *JASA*, 84:853–861, 1989.

[24] P. Flajolet, D. Gardy, and L. Thimonier. Birthday paradox, coupon collectors, caching algorithms and self-organizing search. Technical report, INRIA, 1987.

[25] J. Glaz. Approximations and bounds for the distribution of the scan statistic. *JASA*, 84:560–566, 1989.

[26] C. M. Goldie. Implicit renewal theory and tails of solutions of random equations. *Ann. Appl. Probab.*, 1:xxx, 1991.

[27] P. Groeneboom. Brownian motion with a parabolic drift and Airy functions. *Probab. Th. Rel. Fields*, 81:79–109, 1989.

[28] L. Holst and S. Janson. Poisson approximation using the Stein-Chen method and coupling: Number of exceedences of gaussian random variables. Technical report, Uppsala University, 1988.

[29] L. Holst, J.E. Kennedy, and M.P. Quine. Rates of Poisson convergence for some coverage and urn problems using coupling. *J. Appl. Prob.*, 25:717–724, 1988.

[30] T. Hsing, J. Husler, and M.R. Leadbetter. On the exceedence point process for a stationary sequence. *Probab. Th. Rel. Fields*, 78:97–112, 1988.

[31] T. Hsing and J.L. Teugels. Extremal properties of shot noise processes. *Adv. Appl. Probab.*, 21:513–525, 1989.

[32] J. Husler. Extreme values and high boundary crossings of locally stationary Gaussian processes. *Ann. Probability*, 18:1141–1158, 1990.

[33] J. Husler and R.-D. Reiss, editors. *Extreme Value Theory*. Springer-Verlag, 1989. Lecture Notes in Statistics 51.

[34] R. Isaac. A limit theorem for probabilities related to the random bombardment of a square. *Acta. Math. Hungar.*, 51:85–97, 1987.

[35] G. Lindgren. Slepian models for χ^2 processes with dependent components with applications to envelope uprocssings. *J. Appl. Prob.*, 26:36–49, 1989.

[36] G. Lindgren and H. Rootzen. Extreme values: Theory and technical examples. *Scand. J. Statist*, 14:241–279, 1987.

[37] G. Marsaglia. A current view of random number generators. In L. Billard, editor, *Computer Science and Statistics: Proc. 16th Symposium of the Interface*, pages 3–10. North-Holland, 1985.

[38] P. C. Matthews and A. L. Rukhin. Asymptotic distribution of the Normal sample range. Technical report, Univ. Maryland, Baltimore, 1991.

[39] P.C. Matthews. Covering problems for Markov chains. *Annals of Probability*, 16:1215–1228, 1988.

[40] T.F. Mori. Maximum waiting time when the size of the alphabet increases. In *Mathematical Statistics and Probability Theory, Vol. B (Bad Tatzmannsdorf, 1986)*, pages 169–178. Reidel, 1987.

[41] G.L. O'Brien, P.J.J.F. Torfs, and W. Vervaat. Stationary self-similar extremal processes. Technical report, York University, Ontario, 1989.

[42] E.A. Perkins and S.J. Taylor. Measuring close approaches on a Brownian path. *Ann. Probab.*, 16:1458–1480, 1988.

[43] G. O. Roberts. Asymptotic approximations for Brownian motion boundary hitting times. *Annals of Probability*, 19:1689–1731, 1991.

[44] H. Rootzen. Maxima and exceedences of stationary Markov chains. *Adv. Appl. Prob.*, 20:371–390, 1988.

[45] R. F. Serfozo. High-level exceedences of regenerative and semi-stationary processes. *J. Appl. Prob*, 17:423–431, 1980.

[46] R.F. Serfozo. Extreme values of birth-and-deathg processes and queues. *Stochastic Proc. Appl.*, 27:291–306, 1988.

[47] R.F. Serfozo. Extreme values of queue lengths in $M/G/1$ and $GI/M/1$ systems. *Math. Operations Research*, 13:349–357, 1988.

[48] R.L. Smith. Extreme-value theory for dependent sequences via the Stein-Chen method of Poisson approximation. *Stochastic Proc. Appl.*, 30:317–328, 1988.

[49] W. Stadje. The collector's problem with group drawings. *Ann. Appl. Probab.*, 22:866–882, 1990.

[50] J. Sun. *Ph. D. thesis*. PhD thesis, Stanford, 1989.

[51] W. Szpankowski. (probably) optimal solution to some problems not only on graphs. Technical report, C.S. Dept, Purdue, 1989.

[52] V.I. Tikhonov and V.I. Khimenko. *Excursions of the Trajectories of Random Processes*. Nauka, Moscow, 1987. In Russian.

[53] E. Vanmarcke. *Random Fields*. MIT Press, Cambridge, MA, 1982.

Minor typos

p. 15. The equation above (A10f) should read $P(M \leq b)$.
p. 46. In (3.), $F(x)$ should be $\overline{F}(x)$.

p. 46. In (7.), ξ should be ξ_3.
p. 47. The top displayed equation should have $\exp(-n\bar{F}(u))$.
p.57 Below (C15a), Y_η should be ηY.
p. 123 In (F8d), C should be B.
p. 141 line -2: delete first log.
p. 189 First line of final paragraph: $X_1(t)$ should be $X_i(t)$.
p. 217 In (J30.4) there should be a 2 in front of the $\sum$.
p. 262 B.F. Logan and L.A. *Shepp* (sorry, Larry!).

[44] [7] [11] [27] [5] [7, 10] [21, 4, 34, 39] [9]